UNIVERSITY RESEARCH

DOT/RSPA/DPB-50/79/5

CONTROL OF SERVICE RELIABILITY IN TRANSIT NETWORKS

MARCH 1979

FINAL REPORT

UNDER CONTRACT: DOT-OS-80018

Prepared for
U.S.DEPARTMENT OF TRANSPORTATION
Research and Special Programs Administration
Office of University Research
Washington, D.C.20590

Technical Report Documentation Page

1. Report No. DOT/RSPA/DPB-50/79/5	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Control of Service Reliability in Urban Bus Networks: Final Report		5. Report Date MARCH 1979
		6. Performing Organization Code
7. Author(s) Mark A. Turnquist, Larry A. Bowman		8. Performing Organization Report No.
9. Performing Organization Name and Address The Transportation Center Northwestern University Evanston, Illinois 60201		10. Work Unit No. (TRAIS)
		11. Contract or Grant No. DOT-OS-80018
12. Sponsoring Agency Name and Address U. S. Department of Transportation Research and Special Programs Admin. Office of University Research Washington, D. C. 20590		13. Type of Report and Period Covered Final Report October 1977 - January 1979
		14. Sponsoring Agency Code DPB-50

15. Supplementary Notes

Technical Monitor: Mark D. Abkowitz, DTS-243
Transp. Systems Center

16. Abstract

Service reliability is important to both the transit user and the transit operator. This report represents the initial phase of a study of various strategies for control of service reliability in transit networks. Specifically, this report focuses on: 1) identification of possible control strategies; 2) development of a bus network simulation model to enable testing of alternative strategies; 3) analysis of the relationship between passenger wait times and service reliability; 4) tests of the effects of network structure on reliability; and 5) preliminary tests of a limited set of control strategies. Major conclusions are: 1) passenger wait time is very sensitive to service reliability; 2) new insights have been gained into the causes of vehicle "bunching" along routes; and 3) a number of potentially useful control strategies have been identified.

17. Key Words Computer simulation, service reliability, urban bus network, control strategy, passenger arrivals, vehicle travel times, transfers	18. Distribution Statement Document is available to the U.S. Public through the National Technical Information Service, Springfield, Virginia 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	111	

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

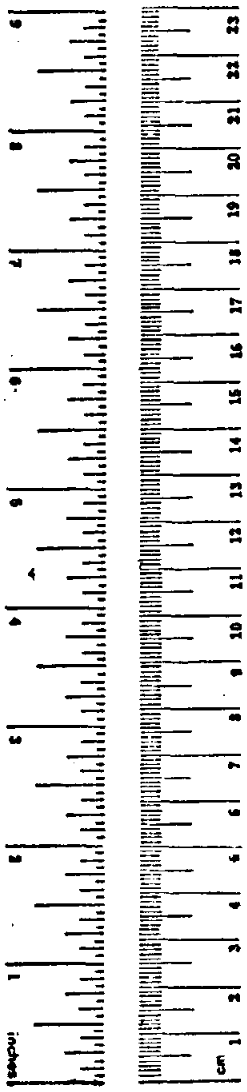
METRIC CONVERSION FACTORS

Approximate Conversions to Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
in	inches	*2.5	centimeters	cm
ft	feet	30	centimeters	cm
yd	yards	0.9	meters	m
mi	miles	1.6	kilometers	km
		AREA		
in^2	square inches	6.5	square centimeters	cm^2
ft^2	square feet	0.09	square meters	m^2
yd^2	square yards	0.8	square meters	m^2
mi^2	square miles	2.6	square kilometers	km^2
	acres	0.4	hectares	ha
		MASS (weight)		
oz	ounces	28	grams	g
lb	pounds	0.45	kilograms	kg
	short tons (2000 lb)	0.9	tonnes	t
		VOLUME		
tsp	teaspoons	5	milliliters	ml
Tbsp	tablespoons	15	milliliters	ml
fl oz	fluid ounces	30	milliliters	ml
c	cups	0.24	liters	l
pt	pints	0.47	liters	l
qt	quarts	0.95	liters	l
gal	gallons	3.8	liters	l
ft^3	cubic feet	0.03	cubic meters	m^3
yd^3	cubic yards	0.76	cubic meters	m^3
		TEMPERATURE (exact)		
	Fahrenheit temperature	5/9 (after subtracting 32)	Celsius temperature	°C

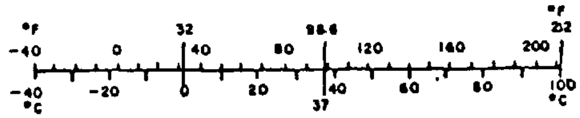
---

*1 in = 2.54 (exactly). For other exact conversions and more detailed tables, see NBS Misc. Publ. 286, Units of Weights and Measures, Price $2.25, SD Catalog No. C13.10:286.



Approximate Conversions from Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
mm	millimeters	0.04	inches	in
cm	centimeters	0.4	inches	in
m	meters	3.3	feet	ft
m	meters	1.1	yards	yd
km	kilometers	0.6	miles	mi
		AREA		
cm^2	square centimeters	0.16	square inches	in^2
m^2	square meters	1.2	square yards	yd^2
km^2	square kilometers	0.4	square miles	mi^2
ha	hectares (10,000 m^2)	2.5	acres	
		MASS (weight)		
g	grams	0.035	ounces	oz
kg	kilograms	2.2	pounds	lb
t	tonnes (1000 kg)	1.1	short tons	
		VOLUME		
ml	milliliters	0.03	fluid ounces	fl oz
l	liters	2.1	pints	pt
l	liters	1.06	quarts	qt
l	liters	0.26	gallons	gal
m^3	cubic meters	35	cubic feet	ft^3
m^3	cubic meters	1.3	cubic yards	yd^3
		TEMPERATURE (exact)		
°C	Celsius temperature	9/5 (then add 32)	Fahrenheit temperature	



Acknowledgment

The work described in this report was performed at Northwestern University under the sponsorship of the Program of University Research, U.S. Department of Transportation, contract number DOT-OS-80018. The advice and assistance of the contract monitor, Mr. Mark Abkowitz have been extremely valuable through the course of the research. Comments on an earlier draft of this report from Professor Nigel Wilson are also appreciated. Several graduate students have provided many hours of effort in collecting and analyzing data. Special thanks are due to Steve Blume, Chester Hashizume and Bill Jordan.

CONTROL OF SERVICE RELIABILITY IN TRANSIT NETWORKS

EXECUTIVE SUMMARY

The concept of "service reliability" has come into increasing prominence in recent years as an important characteristic of the quality of service provided by transportation systems. A basic definition of reliability, as the term is used here, is the variability of a system performance measure over time. The focus is on stochastic variation in performance, rather than on more traditional engineering concepts of probability of component or system failure. The level of service measure most clearly subject to variation is travel time, and this variability is often described in terms of non-adherence to schedule.

Service reliability is important to both the transit user and the transit operator. To the user, non-adherence to schedule results in increased wait time, makes transferring more difficult, and causes uncertain arrival time at the destination. To the operator, unreliability implies that greater "slack" must be built into timetables. This results in less effective utilization of equipment and personnel, and reflects itself in reduced productivity and increased cost in the system's operations. Concern over the effects of unreliability of service has prompted many operators to institute some form of control in an attempt to make service more reliable.

A study of the effectiveness of various strategies for control of unrelaability in transit services is a vital element in the search for ways in which to improve transit productivity and efficiency. Such control strategies have important implications for both planning and management of transit systems.

From the planning perspective, the control strategies of interest include restructuring of routes and schedules, changes in the numbers and location of stops, and, in some cases, the provision of exclusive rights-of-way. Transit management is often concerned with dynamic control strategies, such as holding buses at points along a route, as well as changes of interest to planners. It is thus important to both planners and managers to have tools for analyzing service reliability. The objective of this research is to begin the process of providing those tools.

The first year's work on this project has produced a number of significant results, including:

1. identification of possible planning and real-time strategies for control of service reliability;
2. extension and preliminary validation of a bus network simulation model, to serve as a tool for testing the effectiveness of alternative control strategies;
3. development of a clearer understanding of the relationship between service reliability and passenger wait times; and
4. conclusions regarding the effects of route structure on service reliability.

These results are discussed more fully in the following paragraphs.

Approximately twenty different control actions with potential for improving service reliability have been identified. Some of these actions, such as signal pre-emption, attempt to reduce the number and size of delays a bus encounters. Others, such as skipping stops, attempt to minimize the consequences of those delays.

The control strategies may be divided into two basic groups, planning and real-time. In general, the distinction is that planning strategies involve changes in operations of a persistent nature. For example, changes in route structure have substantial long-term effects on the character of operations, and the decision to make such a change is the result of the service-planning process. On the other hand, real-time control measures are designed to act quickly to remedy specific problems. These actions have immediate effects, but seldom exert any influence on the general nature of operations over a longer time period. Additional discussion of both sets of control strategies is contained in the working paper by Blume and Turnquist (1978).

The basic tool for investigation of the effectiveness of various alternative actions for control of service reliability is a network-simulation model. The basic prototype model was developed under a previous grant from the National Science Foundation. During the first year of the current project, this model has been extended and revised. The model is now capable of analyzing networks of up to 32 routes, with a maximum of 70 zones serving as origins and destinations of trips. Larger networks are possible within the same logical structure; only the amount of central memory available to the program would require changes.

A very important aspect of the simulation is the ability to have different networks, service characteristics and operating policies easily input to the model. This is accomplished through the use of a simulation system, developed at Northwestern University, called Program for Interactive Multiple Process Simulation (PIMP). The PIMP software comprises a set of FORTRAN subroutines for event scheduling, dynamic data management and statistics gathering. The characteristics which make PIMP uniquely appropriate for the current model are the interactive multiple process structure combined with the flexibility of data management through subscripted queues.

A third area in which major results have been achieved in the first year is the understanding of the effects of service reliability on passenger wait times. The simulation model described above can be used to predict changes in the level of reliability, as measured in any of several ways; but in order to evaluate the impacts of service changes, it is vital to understand more clearly the relationship of passenger wait time to reliability.

During the past year, an improved model of the relationship between passenger wait time and the service frequency and reliability of buses has been developed. The model is based on reflecting the choice process involved in passenger decisions as to arrival time at bus stops. It incorporates explicitly the sensitivity of such decisions to both service frequency and reliability.

Another important part of the simulation model refinement is the representation of link travel times for buses. Considerable empirical data were gathered on these travel times for routes in Evanston and Chicago, Illinois, and Cincinnati, Ohio. Statistical analysis of this data confirms the appropriateness of gamma distributions as the underlying probability model for vehicle delays. Overall link travel time can be predicted quite well using a shifted gamma distribution, with the shift accounting for free-flow, non-stop travel time. Furthermore, a procedure has been developed for prediction of the parameters of the distribution, based on the number of boarding and alighting passengers, and the number of signalized intersections through which the bus must pass. This predictive travel time model is useful, with or without the simulation, by an operator interested in more efficiently scheduling bus routes.

Given these refinements to the simulation model , the second major project task was to undertake a preliminary validation of the model against observed data. The bus network of Evanston, Illinois, was selected for this purpose, primarily because substantial data exist for this system, and because the proximity to Northwestern University eased the collection of additional data.

Based on the high correlation of simulated results with observed system performance, this preliminary validation effort was judged successful. It should be noted, however, that this is not viewed as conclusive validation of the model. Because many components of the model have been developed using the Evanston data as a basis, similarity between overall model results and observed system performance is to be expected. Thus, during the second year of this project, a more complete validation effort will be undertaken, using data from a different transit system. This should provide more conclusive evidence with regard to model validity. Nevertheless, results of the preliminary validation using the Evanston data were sufficiently encouraging to allow proceeding with a series of experiments on alternative control strategies.

Simulation experiments to test the effects of various types of potential control strategies have been divided into two sets. The first set of simulation experiments was designed to test the effects of basic network characteristics on several measures of service reliability. Understanding of these basic relationships is an important precursor to actual design and testing of specific strategies. The second set has provided initial investigation of real-time control strategies.

Network effects on measures of reliability were found to depend heavily on the frequency and standard deviation of link travel times. Dynamic instability of vehicles, or "bunching," was found to result from the interaction of frequency and link travel time deviations with the level of user demand. While bunching increases deviations in vehicle arrival times for relatively reliable service, as schedule adherence deteriorates and vehicles group into clusters, a secondary effect is that these clusters tend to be resistant to further departure from schedule.

Characteristics of network shape are most important through their influence on passenger transfers. The density affects the number of transfers made, and the form affects the uncertainty associated with the transfer event.

Finally, a brief investigation of real-time control strategies has indicated a better decision process is needed if headway-based controls are to be of any practical use. Checkpoint control (simply involving enforcement of schedules at time points along the route appears very promising. It is simple and inexpensive to implement, avoids the propagation of delay associated with headway controls, and works to maintain the schedule, which is of critical importance if passengers time their arrivals to match the vehicle timetable. However, an accurate schedule is a necessity. If vehicle arrivals at a checkpoint are consistently either early or late because the schedule is unrealistic, the control will either cause excessive delay or never be enacted.

A strategy of allowing the first vehicle in a bunch to skip several stops in order to move ahead is also a promising form of control. It is well suited for short headway services with a high incidence of vehicle bunching. The major disadvantage of such a strategy is that some passengers will be forced to change buses because the vehicle they are riding will skip the stop at which they wish to alight. In practice, this implies that the strategy may be most useful for routes on which the majority of passengers are destined for one (or a few) stops, such as a rail transit station or the CBD area.

During the second year of this project, the principal work will involve further extension of the simulation model to allow more detailed analysis of control strategies designed to reduce link travel time variability, more complete validation of the extended model, and further testing of a larger number of control strategies.

Table of Contents

List of Figures

List of Tables

CONTROL OF SERVICE RELIABILITY IN TRANSIT NETWORKS

FINAL REPORT

1. Introduction

In recent years, increasing attention has been given to the concept of "service reliability" in transportation systems. In the freight sector particularly, the reliability of service has been widely accepted as an important characteristic of each transport mode, and significant research (Martland, 1972; Sussman, 1972) has been done on reliability in railroad operations. Somewhat less attention has been devoted to research on reliability in the context of urban passenger transportation, but the importance of some measure of reliability to trip-making behavior has been emphasized in several attitudinal studies. For example, Paine *et al* (1966) found that potential users ranked "arriving when planned" as the single most important service characteristic of a transit system. This finding has been substantiated in further studies by Golob (1972) and Wallin and Wright (1974).

Many transit operators also recognize that unreliability in operations is a source of reduced productivity and increased costs. This is due to the need to build substantial "slack time" into timetables in order to absorb deviations from the schedule. This leads to reduced utilization of both equipment and personnel. In light of the current need for more effective public transportation in urban areas and the rather dismal financial picture facing most transit operators, it is important to understand the sources of unreliability, and to investigate the potential of several alternative control strategies to improve both the quality of service provided and the productivity of the equipment and personnel in the system.

Recent work by Turnquist (1978) has taken the first steps in investigating the fundamental relationships between system characteristics (i.e., route length, route density, frequency of service, network connectivity, etc.) and the level of service reliability provided. Understanding of these relationships is vital to the formulation of effective control strategies. The basic mechanism for this investigation was a digital-simulation model of the operation of bus transit networks.

The work undertaken in the first year of the current project has focused on: (1) extension and refinement of the simulation model; (2) testing of the effects of network structure on reliability; and (3) development and initial testing of control strategies to improve service reliability. This report is organized around these three major topics. Section 2 describes the improvements made to the simulation model developed previously. An extensive set of experiments to test the effects of network structure on reliability are described in Section 3. Section 4 includes a discussion of a number of possible strategies for control of service reliability, and describes the results of initial experiments with some of them. Finally, conclusions from the first year's work are presented in Section 5.

2. The Simulation Model

Initial work on the problem of network analysis involved investigation of the potential for closed-form analytic solutions. Models based on queuing theory were attempted, but a number of assumptions necessary in simplifying the problem were unacceptable, given the questions to be investigated in this research. The aspects of the bus network problem which present difficulties in terms of a queuing-theoretic approach include non-random arrivals of passengers at stations, batch service of passengers, and batch arrivals of passengers at transfer points - a result of the interaction of routes.

The rejection of this analytic approach led to simulation as a "method of last resort". Simulation is not usually the preferred way to solve a problem due both to its cumbersome nature, and to the fact that the model does not provide general results. It only provides "point" responses indicating the outcome for a particular set of inputs. However, simulation does allow great flexibility in the representation of individual components of a system. This has provided the impetus for substantial research concerning passenger arrival processes and vehicle travel time distributions. It also has necessitated a test situation for use as a validation exercise, verification of the results provided by the model being vital to promote confidence in the results obtained from experiments later on.

The selection of simulation as the modeling technique also imposes a particular structure on the analysis. By nature it precludes the development of closed form analytical expressions describing relationships between network characteristics and reliability. Instead, an exploratory, experimental approach is required, inferences as to causation being drawn from comparison of experimental outcomes.

Construction of a prototype simulation model was accomplished as part of a previous project (Turnquist, 1978). In the course of the current project, several major refinements and extensions have been made to the model. First, the component of the model dealing with the times of passenger arrivals at bus stops has been improved greatly. Second, the specification of link travel time distributions for buses has been improved, based on substantial empirical observation and statistical testing. Third, additional logic to reflect the enactment of real-time vehicle control policies has been added. Finally, a preliminary validation exercise has been performed on the model as a whole, using data from Evanston, Illinois.

Section 2.1 describes the logical structure of the model, including the major assumptions included and the types of statistics collected for analysis of system performance. A more detailed explanation of the actual mechanics of the simulation program is contained in the Users' Manual (Bowman and Turnquist, 1979). Section 2.2 describes the passenger arrival process in detail, and Section 2.3 discusses link travel time distributions. The validation exercise is described in Section 2.4.

2.1 Simulation Logic

Two required features of the simulation influenced heavily the structure of the resulting model. Testing different network structures requires the network to be an input to the model, not part of the design. Secondly, the desire to collect statistics on the effect of route connections on passenger movements as they travel through the network requires the ability to track individual passengers from origin to destination. Basically, the model structure involves the movement of two kinds of entities through the network: passengers and vehicles. Vehicle movements are simulated by travel times over links subject to service delay from boarding and alighting passengers. Simultaneously, the information in a path matrix is used in making decisions on the assignment of travelers, first to station queues where they wait for service, and then to passenger lists of buses when an appropriate bus for their destination arrives. Transferring routes results in a traveler being reassigned from a vehicle list to the queue of the next stop in his path. The traveler then awaits another bus arrival and the procedure repeats.

The first feature, the inclusion of the network structure as a general input, was accomplished through the use of a simulation language called Program for Interactive Multiple Process Simulation (PIMP), developed by Schaefer (1974). The entire model structure hinges on the use of subscripted queues which can be referenced by variables giving the station numbers in a traveler's trip through the system. By always specifying the next stop in the traveler's path this determines a sequence of queues through which he moves. The interaction of passenger movements (as determined by the next stop in the user's path) with the simulation of buses along routes permits a single model to accommodate any network or service desired.

Kulash (1971) made use of a network simulation model in the analysis of system capacities, but due to his different needs he permitted the steps in a passenger's trip through the network to be "disconnected". The advantage of breaking a traveler's trip into sections is the ability to represent a multi-step trip as several independent no-transfer trips. Zonal trip generation rates are adjusted to account for additional trips created by passenger transfers. Then, instead of deciding whether each individual passenger is to be terminated or transferred, a percentage of the passengers remaining on the vehicle are simply terminated at each stop. In this research, however, analysis of the effects of reliability on transfers required statistics on transfer times and overall origin-destination travel times. Each passenger, when created, was therefore assigned a specific destination and followed through the system until arrival at that point.

To summarize, the ability to trace passenger movements is accomplished through the association of passenger lists with the vehicle entities moving through the network. Vehicles are numbered, and any time a decision is made for a passenger to board or alight a vehicle, the passenger's identification number is added to or removed from the list of passengers currently on that bus. The assignment of a passenger's ID number to bus stop queues and vehicle lists in turn, records the movement of the passenger through the system. The data file retained on passenger movements includes the times at which movements between queues and vehicles occur, as well as the zones where the events take place.

2.1.1 Assumptions

A number of assumptions are present in the experiments conducted, either implicitly due to the model structure, or explicitly as inputs. Tables 2-1 and 2-2 are presented to list the major implicit and explicit assumptions present in the model.

Note that the service delay equations used were taken directly from an earlier study by Boardman and Kraft (1970) assuming an exact fare system with two-door conventional buses. Three equations were obtained, one each for boarding only, alighting only, and simultaneous boarding and alighting of passengers. These equations are presented as equations 2-1 through 2-3.

Table 2-1. Assumptions implicit in structure.

1. Route choice
   Passengers know their alternatives. If expected travel time on the first bus to arrive is less than the expected travel plus wait time for other possible services, then the first to arrive is chosen. Otherwise the passenger will wait for the preferred service.

2. No balking
   Passengers do not give up and leave, even if the first bus is full and they must wait for another.

3. First come first served
   Queues maintain order so if vehicle capacity is exceeded the latest arrivals stay behind.

4. Buses can pass
   If a trailing bus catches up to an earlier bus, the trailing bus is not forced to retain its position, but may pass the leading bus.

5. Order of vehicle dispatches
   If buses pass along a route and arrive at the end point out of order, they will depart on their next cycle of service in the order they arrived.

Table 2-2. Assumptions explicit in inputs.

1. Service delay
   Service delay computed according to the equations from Boardman and Kraft (1970), plus 20 seconds acceleration delay.

2. Scheduled slack time
   10% of the scheduled travel time is provided as driver rest time at the completion of every trip. An additional 10% of the travel time is provided as slack or recovery time. This 10% is available as a buffer to compensate for the possibility of arriving late at the end of the route. The 10% rest time may not be used as recovery time.

Note:
Other assumed inputs are part of the experimental design process. Those mentioned here were fixed throughout the experiments.

Boarding only

$$\text{Delay} = .5863 + 1.9957*\text{On} \qquad (2\text{-}1)$$

Alighting only

$$\text{Delay} = 2.2345 + 1.0792*\text{Off} \qquad (2\text{-}2)$$

Boarding and alighting

$$\text{Delay} = 1.6043+.9588*\text{On}+2.1543*\text{Off}-.0202*\text{On}*\text{Off} \qquad (2\text{-}3)$$

where: Delay = bus dwell time at stop (seconds)
On = number of passengers boarding
Off = number of passengers alighting

2.1.2 Statistics for Measurement

A number of statistics on both passenger and vehicle movements are computed as part of each experimental run. In addition, a detailed file of individual passenger movements is saved. This file can be studied off-line should any specific question be raised not answered by the statistics already provided. A file of bus movements may be printed, if desired; however it is not saved on file. Table 2-3 summarizes the information recorded.

Vehicle statistics are compiled for both deviations about average and scheduled arrival times. This discrepancy between user and operator perceptions of the reliability of the performance of a service can come about when the schedule does not accurately reflect the conditions encountered by the vehicle. A bus consistently early (or late) will seem reliable to the user, but would be labeled unreliable by the criteria employed by most operators. A good review of different statistical measures, including the existence of multiple points of view, is provided by Chapman (1976).

Table 2-3. Statistics incorporated in program.

Users

- Wait time - histogram (15 second intervals) and average.
- Travel time - histogram and average.
- Transfer time - histogram and average.
- Specific route pair transfers - average and deviation for each pair specified. Average also for combined pairs.

Vehicles

- Standard deviation (about average and about schedule) for all stops, each route, and network average.
- Histogram of schedule deviations by route (optional).
- Percentile ranges (about averate and about schedule for each route and network average.
- Load factors - maximum load for each cycle of service on each route.
- Matrix of schedule deviation for each stop, each route, each cycle of service.

2.1.3 Model Size Limits and Computer Requirements

The logical structure of the model is very flexible, and can accommodate bus networks of widely varying size. The major limitations on size of system to be analyzed occur as a result of array dimensions in the program data structure. If desired, these dimensions can be increased to handle larger systems, but the core requirements increase accordingly. The current size limits are as follows:

60	nodes
500	one-way links
24	one-way routes
13	stops per route
24	cycles of service per route
75	passengers per bus
156	vehicles

The simulation model is programmed in FORTRAN. It relies on a subset of the simulation language PIMP (Schaefer, 1974) for memory-management and list processing operations.

The FORTRAN program requires approximately 212,000 (octal) words of core memory, as well as extended core storage (ECS) as provided by CDC 6000-7000 series computers. If necessary, the program can be modified to run without the use of ECS; however, in this case, considerably more central memory would be required.

The simulation language, PIMP, is used for dynamic allocation of core. The current version of PIMP is designed for use with the CDC FORTRAN, Version 3 compiler and includes several routines written in COMPASS, the CDC assembly language. Simple modifications would make it compatible with FORTRAN Version 4. Modification to make the program compatible with machines made by other manufacturers is straightforward, but would require rewriting several routines.

2.2 Passenger Arrivals at Stations

It is proposed that an explicit relationship exists between service reliability and passenger arrival behavior at stations. As a result, it is hypothesized that the true reduction in passenger wait time from improved vehicle reliability will be substantially higher than the benefits estimated via an assumption that passenger arrivals occur independently of vehicle arrivals. Testing this hypothesis is a major goal of this research. If this hypothesis proves true, then there is much stronger justification for taking steps to improve reliability of service.

The times of passenger arrivals at a transit stop are important determinants of passenger wait time, and together with the deviation of bus arrival times, determine the magnitude of increased delay incurred when buses deviate from their scheduled timetable.

Under the assumption of random passenger arrivals (independent of the bus schedule), the expected passenger wait time has been derived by a number of authors including Welding (1963), Holroyd and Scraggs (1966), and Osuna and Newell (1972). The average time users will have to wait before a bus arrives is

$$E(w) = \frac{H}{2} \left[1 + C_v^2\right] \qquad (2\text{-}4)$$

where H = mean headway (service interval)

C_v = coefficient of variation in headways (standard deviation/mean)

The average wait is greater than half the headway because more passengers arrive during the long intervals where the average wait is greater than half the headway, and fewer passengers arrive in the short intervals where the wait is correspondingly shorter. The result is an average passenger wiat time which increases as headway intervals become less uniform.

It is proposed here that under some circumstances passenger arrivals may not be random; rather users will, to some extent, coordinate their activities to coincide with the scheduled service. More reliable service will tend to increase the proportion of users who are aware of the schedule and act accordingly.

2.2.1 Previous Treatments

Most of the research in the field of transit reliability has assumed random passenger arrivals; see for example Barnett (1974),dePirey (1971), Friedman (1976), or Bly and Jackson (1974). Empirical evidence suggests, however, that under certain conditions passengers may coordinate their arrival times with the bus schedule. If this is true, the arrival process is no longer Poisson and the distribution of arrival times should reflect shorter average wait times. Several studies have addressed the question of the manner in which passengers select their arrival times at a station. Three studies which look beyond the assumption of random arrivals are those by Okrent (1974), Jolliffe and Hutchinson (1975), and Turnquist (1978).

Okrent's approach was to estimate a probability distribution which would describe the observed distribution of arrival times directly. He collected data at a number of bus stops in Chicago and Evanston under a variety of conditions, and estimated beta and gamma distributions from the observed data. Okrent concluded that in general 12 to 13 minute headways marked the transition from random to coordinated passenger arrivals. He further went on to comment that a better fit might be obtained by combining market shares of random and coordinated distributions. This would permit a gradual shift to coordinated arrivals as longer intervals prompt more people to learn the schedule.

More recently, Jolliffe and Hutchinson proposed an improved model based upon considering passengers to be of three types: a proportion q, whose arrival is coincident with that of the bus; a proportion (1-q)p who arrive so as to minimize expected wait time; and a proportion (1-q)(1-p), who arrive at random. The proportion q, whose arrivals are coincident with the bus arrival, represents those people who run to the stop because they see the bus coming, and thus wait zero time.

The approach taken by Turnquist (1978) continues in the vein of the work performed by Jolliffe and Hutchinson, but with several changes designed so that, while addressing the question of non-random passenger arrivals, the effect of service reliability might be explicitly incorporated into the study of passenger wait times.

None of these methods describing passenger arrivals includes all the characteristics desired for use in the analysis of schedule reliability and its effects on user level of service. The discrete nature of the models developed by Jolliffe and Hutchinson (1975) and Turnquist (1978) provide useful insight into underlying mechanisms, but offer no guarantee of accurately predicting the magnitude of costs or savings to users of changes in system performance. The continuous distributions fit by Okrent lack a mechanism underlying the passenger's choice of arrival time, rendering it impossible to predict changes in the distribution resulting from improvements in service.

A better model would be a continuous distribution of passenger arrival times at a bus stop, with the distribution being based on understood behavioral mechanisms which could in turn be related to the service levels provided. This implies a positive feedback whereby improving reliability of vehicle arrival times at a stop will result in a distribution of passenger arrivals better coordinated with vehicle arrivals. The potential user savings from improved reliability under such a passenger arrival model are likely to be much greater than those estimated solely on random arrivals.

The next sections describe the derivation of such a model, and the data collection and model calibration needed to verify its reflection of observed behavior. Numerical examples are developed, comparing user savings for scenarios of improved service reliability as predicted by the derived model with its feedback effect, versus the assumption of uniform random arrivals.

2.2.2 Derivation Based on User Utilities and Choice Theory

The approach to modeling the behavior of knowledgeable passenger arrivals uses work done by Cosslett (1976) as a basis. Cosslett's work involved drivers' choice of departure time for auto trips to work, but the structure of his model may be adapted quite readily to consider bus passenger decisions on arrival times at stops. The conceptual basis of the model is as follows. First, divide the service interval (scheduled headway) into a number of smaller sub-intervals. Compute the utility to the passenger of arriving in each interval. Then use these utilities in a logit model to predict individual choice probabilities. Finally, use these probabilities to represent the distribution of arrival times for the population.

The logit model has the following form:

$$P(t) = \frac{e^{U(t)}}{\sum\limits_{j} e^{U(t)}} \qquad (2\text{-}5)$$

where $P(t)$ = probability of choosing interval t

$U(t)$ = utility associated with arrival in interval t

j = number of intervals in one headway.

For the purposes of the passenger arrival time model, this has been generalized to a continuous range of points rather than intervals. In this case, the summation is replaced with an integral, and instead of finite probabilities we obtain a probability density function:

$$f(t) = \frac{e^{U(t)}}{\int_{o}^{H} e^{U(t)}} \qquad (2\text{-}6)$$

where $U(t)$ = value of the utility at time t

H = schedule headway interval.

Implementation of this model requires determination of the passenger's utility as a function of arrival time. We initially assumed utility as a linear function of expected wait time:

$$U(t) = a \cdot E[W] \qquad (2\text{-}7)$$

where a = weighting term (value of time)

$E[W]$ = expected wait time for an arrival at t.

A second form was also tested with a non-linear exponent of wait time, since it was found in calibration that the value of the weighting factor, a, changed for different length headways. The non-linear utility function is shown in equation (2-8).

$$U(t) = E[W]^b \qquad (2\text{-}8)$$

where b = exponent of wait time.

Both versions of the passenger arrival model express user utility as a function of expected wait time. It should be pointed out that expected wait time does not refer to the value obtained by averaging over all arrivals in the service interval. Rather, it reflects the wait time averaged over the distribution of bus arrival times, given the time the user selects to arrive at the station. Also, the exponent in the non-linear model is not reflective of risk-averse behavior; the expected value of the wait time is still used. Inclusion of risk-averse behavior would require, in place of the expected value, the use of a non-linear function integrated over the distribution of possible wait times.

Wait time can be computed as shown in equation (2-9).

$$E[W(t)] = (1-P(t))*W(t) + P(t)*W'(t) \qquad (2\text{-}9)$$

where $E[W(t)]$ = the expected wait time for an arrival at t
$P(t)$ = the probability the intended bus arrives before t
$W(t)$ = the expected wait given the bus arrives after t
$W'(t)$ = the expected wait given the bus arrived before t.

This expression of wait time for an arrival at time t, depends on two factors: the interval length scheduled between successive vehicle arrivals (headway) and the adherence of vehicle arrivals to the schedule (reliability). Longer intervals between buses offer users greater potential to reduce their anticipated wait time through proper selection of their arrival time at the station. Schedule reliability is important due to its effect on the probability of missing an intended bus.

While arrival times of successive vehicles may in fact be related, independence is assumed to allow separate probability density functions for the intended and next bus arrivals.

The distribution of arrival times results from the interaction of the two terms in the wait time expression. As the selected arrival time approaches the scheduled arrival of the bus, the first term in equation (2-9) gets smaller but the second term grows larger. The first term can be associated with the expected wait time until the intended bus arrives, and the second term the penalty or risk of missing it. The overall expected wait time at first decreases, but as the scheduled arrival time of the bus draws near, the risk of missing the bus increases sharply and the total expected wait time goes up. This results in the expression for user utility increasing, reaching a peak at some point before the scheduled arrival of the bus, and then dropping off.

2.2.3 Calibration of Passenger Arrival Model

Testing the performance of the model required collection of empirical data on passenger arrivals at stations. Between spring and fall of 1977, data on vehicle and passenger arrival times were collected at seven locations in Chicago and Evanston. Depending on the headways and level of demand, between five and ten days of observations were collected at each stop. The data were collected for the morning peak period, 6-8 a.m., with the observer recording to the nearest tenth of a minute the arrival times of buses and passengers at the station. Table 2-4 summarizes the data and the stations for which it was collected.

Model calibration was based on estimation of maximum likelihood estimators for both the proportion of total passengers who are aware of the schedule of service, and the parameter (either a or b) of the utility function. Table 2-5 summarizes the proportions in each passenger category, along with the model parameters for linear and non-linear models.

The most important observation is the stability of the estimated exponent in the models. The values obtained for the exponent remain fairly constant at roughly .55. The importance of this finding should be emphasized. If the exponent of wait time can be shown to be a constant, characteristic of all users, the model then becomes more easily transferable. The distribution thus far has only been tested under a limited set of circumstances, restricting any conclusions as to the extend of its transferability. However, the results to date are encouraging.

Table 2-4. Data collected on passenger arrival times.

Location	Route	Days	Buses	Pass.	Headway (min.)	Std. Dev. (min.)
BP	77	6	132	211	5.1	.658
AK	73	6	90	318	7.7	1.395
WR	49B	5	70	161	10.0	1.235
MM	78	5	70	179	10.0	.853
PI	80A	10	100	234	10.0	1.321
CC	201	8	88	188	15.0	.786
MC	215	8	56	151	20.0	1.190

Note: BP - Belmont and Pittsburg (Chicago)
AK - Armitage and Kedzie "
WR - Western and Rosemont "
MM - Montrose and Melvina "
PI - Pioneer and Irving Park (Evanston)
CC - Central and Central Park "
MC - Main and Crawford "

Table 2-5. Calibration results for passenger arrivals.

		Linear			Non Linear		
	Dev	a	r#	L	b	r#	L
BP	.658	*	*	*	*	*	*
AK	1.395	*	*	*	*	*	*
MM	.853	.397	.632	-389.78	.445	1.0	-391.61
WR	1.235	.114	1.0	-367.21	.870	.802	-368.36
PI	1.321	.289	1.0	-513.04	.559	1.0	-513.95
CC	.785	.269	1.0	-440.15	.556	1.0	-445.09
MC	1.190	.227	1.0	-389.00	.542	1.0	-394.47
CM201	1.049	.183	1.0	-565.99	.456	1.0	-567.97
SA201	1.290	.264	1.0	-116.40	.562	1.0	-117.54
SA204	2.210	.218	1.0	-194.25	.544	1.0	-193.78
CR201	1.290	.262	1.0	-133.69	.557	1.0	-134.53
CR203	1.650	.193	1.0	-498.43	.495	1.0	-496.74

a = weighting factor for linear utility
r = proportion of aware passengers
L = likelihood ratio (larger is better)
b = exponent for non-linear utility

* - Essentially random; insensitive to calibration
\# - Market propostions, r, were restricted to lie
in the feasible region between 0 and 1.0

Note:

The last five links identify Evanston bus routes, the first two letters specifying the location.

CM = Chicago at Main
SA = Sherman Arcade (downtown Evanston)
CR = Central at Crawford

A second point of importance is the predominance of near 100% proportions of "aware" passengers. This is of interest because at all locations a spread of arrival times was observed, in some cases yielding only a gentle peaking in the arrival distribution. This can be contrasted with the assumption that knowledgeable passengers arrive at the optimal time (the instant with the minimum expected wait time), the spread being due to the presence of randomly arriving passengers.

In accordance with the concepts put forth by Jolliffe and Hutchinson, one would expect the proportion of users being knowledgeable of the schedule to be a function of several variables including headway, reliability of vehicle arrival time, and trip purpose. All the data collected in this study were from the morning peak period, when most trips are work trips. However, passengers were not interviewed to determine trip purpose.

Variation in service intervals and degrees of schedule adherence lead one to suspect that different market proportions of passengers will make the effort to coordinate their arrival times. The potential gain to the user from making this effort will change, leading to differing levels of motivation. For the locations observed the proportion of aware passengers varied, but the change was in the form of an abrupt transition from 100% to nearly 0%. As seen in Table 2-5, passenger arrivals at the first two locations are essentially random. These locations, Belmont at Pittsburg and Armitage at Kedzie, have 5.1 and 7.7 minute headways respectively. At the remaining locations, with headways of 10 minutes or greater, the proportion of aware passengers jumps in all but one case to a value of 1.0.

It would appear that during the morning peak period even a small time savings will lead to nearly all users making the attempt to coordinate their arrival times. This is reasonable for work trips since little effort is required to learn the schedule of a service which is used every day at the same time. As the service interval drops below 10 minutes, the coordination diminishes. Not only is the potential gain to the user smaller, but randomness in a person's daily routine makes it difficult for individuals to time their arrivals sufficiently accurately to take advantage of the benefits of planning.

The data points contained in the sample range from 20 minute headways with highly reliable service, to headways less than 10 minutes with less reliable arrival times. In the former instance, the coordination results in a prominant peak in the passenger arrival rate about 2.4 minutes before the anticipated vehicle arrival time. The latter demonstrates much less of this peaking effect. Figures 2-1 and 2-2 illustrate this behavior for a 20 and a 5.1 minute interval service.

The reason a single model with a fixed value for the exponent can replicate such diverse behavior is because the headway and deviation are incorporated in the utility expression. Figure 2-3 illustrates the effect the magnitude of the deviation has on the resulting distribution of passenger arrival times. For a given headway, increasing the deviation of vehicle arrival times will increase the probability of missing an intended bus. The expected wait time for an arrival closely coordinated with the schedule will rise more quickly than the average wait for arrival times occurring early enough to provide a larger safety margin. This will reduce the relative advantages of coordinated arrivals, leading to more dispersed patterns. The same holds true for the treatment of different service intervals. Under identical circumstances, the longer interval will exhibit more of a peaking nature, resulting from the lessened impact of the schedule deviation when compared to the longer interval between vehicles.

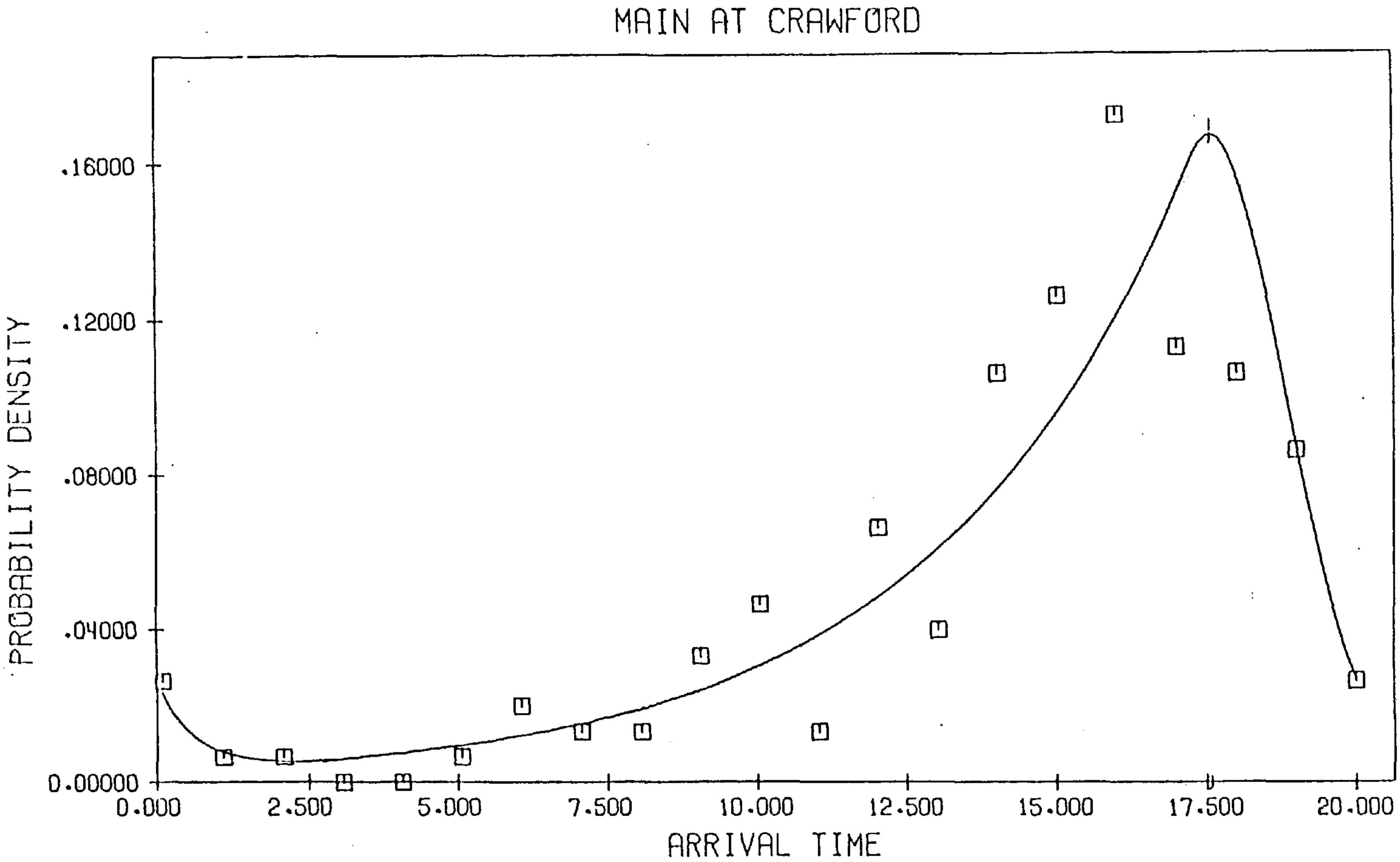


Figure 2-1 Distribution of passenger arrivals for a 20 minute service.

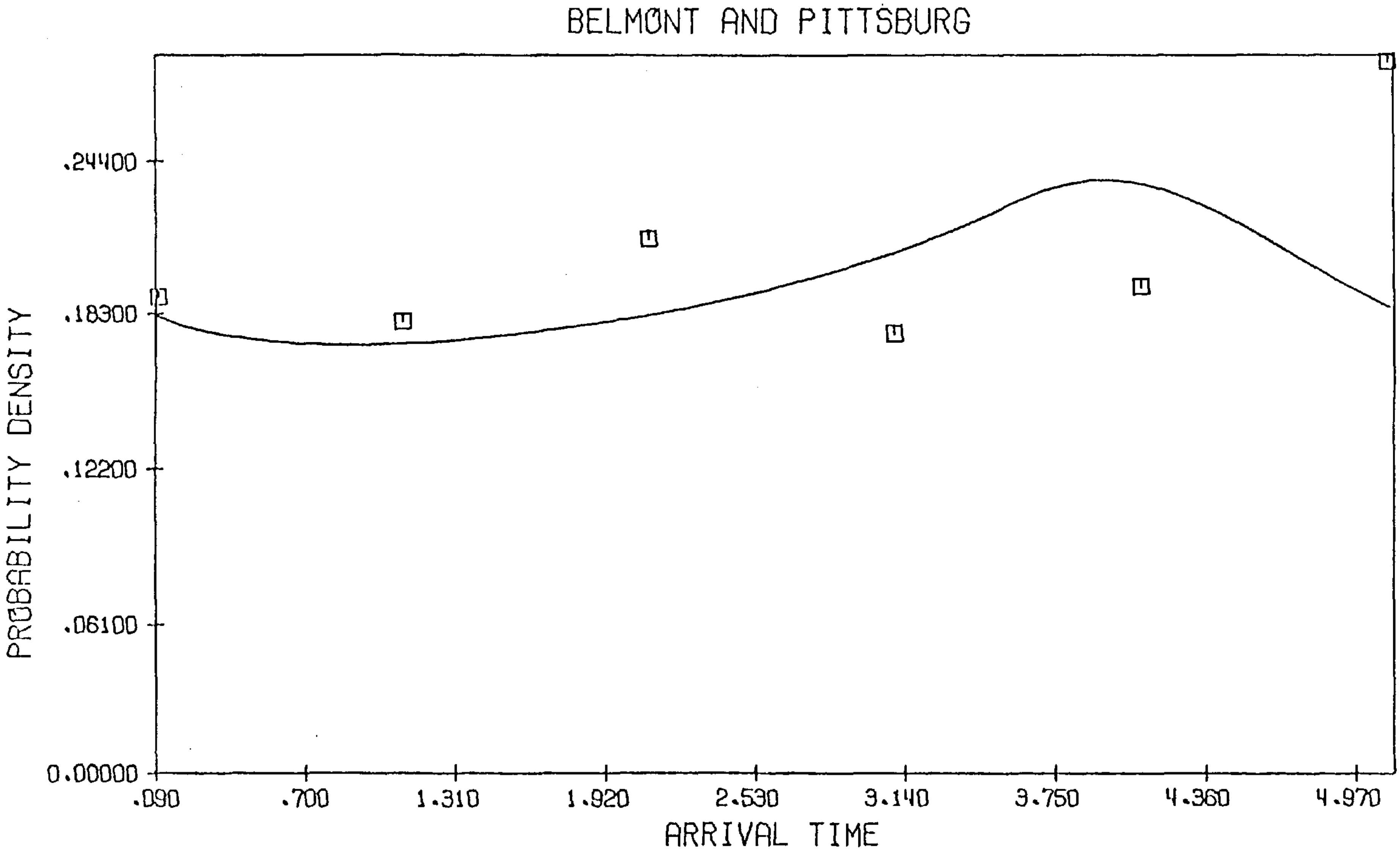


Figure 2-2 Distribution of passenger arrivals for a 5.1 minute service.

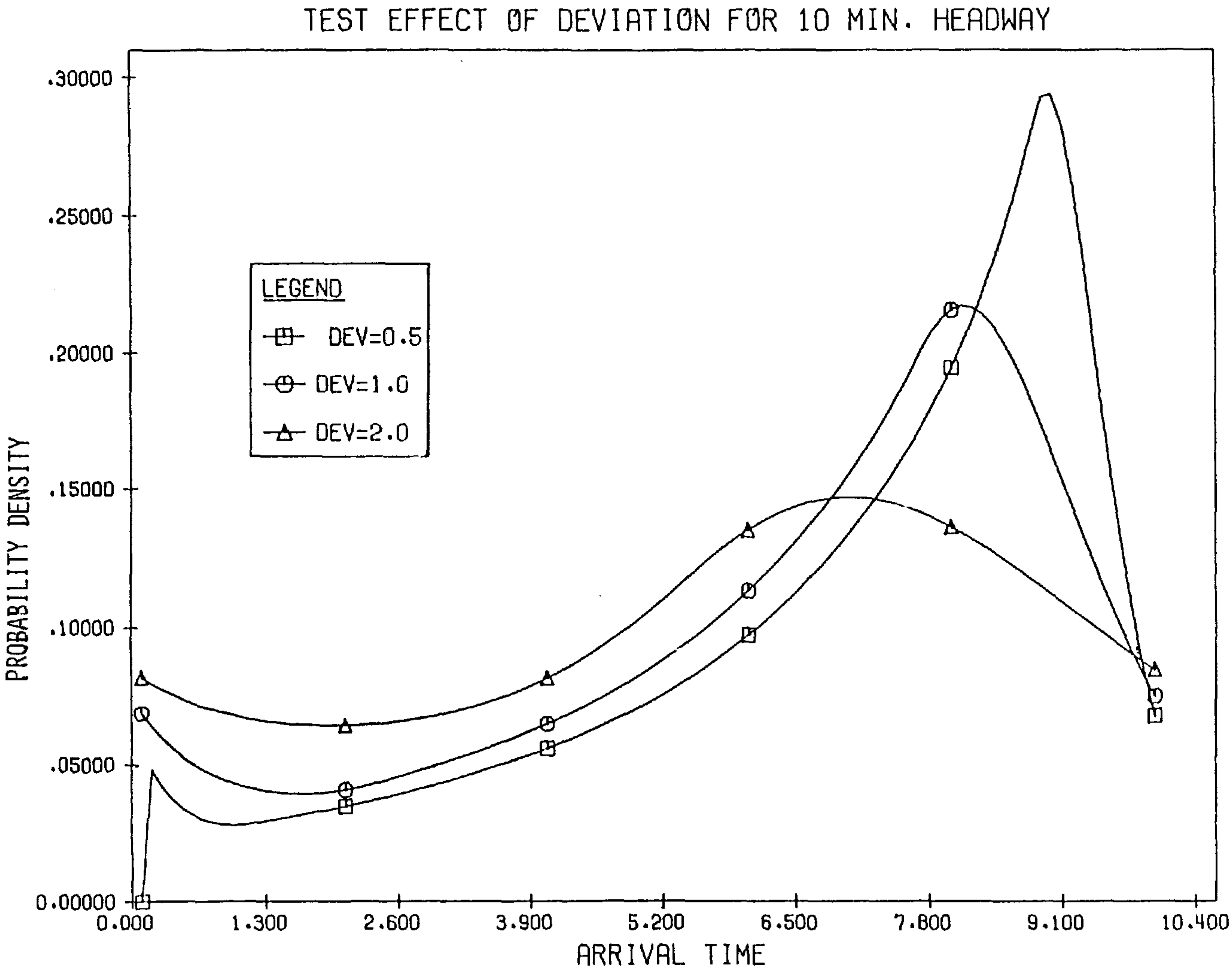


Figure 2-3 Sensitivity of passenger arrival distribution to standard deviation of vehicle arrival times.

One major reason for investigating the arrivals of passengers at bus stops is to develop a better understanding of the benefits that would be realized if service were to be improved. Estimating benefits from the assumption of random arrivals led to equation (2-10),

$$E(W) = (H/2)(1 + C_v^2). \qquad (2\text{-}10)$$

with the only benefits coming about due to more uniform service intervals. To test the hypothesis that much greater benefits in user savings are truly realized (at least for knowledgeable repeat riders), benefits estimated using this equation are compared to savings predicted by the coordinated model, using an exponent of .55. Table 2-6 lists the expected wait times for a variety of headways and deviations, and compares the user savings predicted from a reduction in the standard deviation to a minimum assumed level of .5 minutes. It is easily seen that the benefits predicted when passengers respond to the level of service provided are considerably higher. If benefits are to be used in determining whether a service improvement is worth the cost, this difference may be vital.

As could be expected, the longer the interval the greater the benefits from improved coordination. Improving the 5 minute service results in benefits from the coordinated model of approximately 2.6 times those estimated from the random assumption. Headways of 20 minutes showed benefits from 9 to 26 times those predicted under a random distribution of passengers.

2.2.4 Summary

The behavioral model of user arrival times predicts higher proportions of users as being aware of the scheduled service than do models based on

Table 2-6. User benefits from improved reliability: aware vs. random hypotheses.

	Aware wait	Random wait	Aware gain	Random gain	Ratio: Aware/ Random
Headway - 5 min.					
σ - 0.5	2.171	2.550			
1.0	2.556	2.700	.385	.150	2.566
Headway - 10 min.					
σ - 0.5	3.368	5.025			
1.0	4.020	5.100	.652	.075	8.694
1.5	4.555	5.225	1.187	.200	5.932
2.0	5.004	5.400	1.636	.375	4.362
Headway - 15 min.					
σ - 0.5	4.105	7.517			
1.0	4.956	7.567	.851	.050	17.020
1.5	5.684	7.400	1.579	.133	11.828
2.0	6.323	7.767	2.218	.250	8.872
2.5	6.886	7.917	2.781	.400	6.952
Headway - 20 min.					
σ - 0.5	4.570	10.012			
1.0	5.568	10.050	.998	.037	26.074
1.5	6.439	10.112	1.869	.100	18.690
2.0	7.223	10.200	2.653	.187	14.150
2.5	7.935	10.312	3.365	.300	11.216
3.0	8.583	10.450	4.013	.437	0.162

fitting discrete approximations to the distribution of passenger arrival times. The effect of this difference can be felt in two ways. First, while both models may have the same value for the average passenger wait time, the discrete models will not reflect higher-order moments of the wait time distribution as well. This would present a problem in extending these models to incorporate risk aversion, for example. Second, and more important, the lower estimated proportion of aware passengers obtained via the discrete formulation may result in an over-emphasis on the need for better marketing and promotion of the transit service.

Use of the coordinated arrival model leads to estimated user benefits which are approximately an order of magnitude greater than those estimated using an assumption of random arrivals. The outcome of any decision based on cost-benefit analysis of the effects of implementing reliability strategies will be strongly affected by the approach taken to estimate benefits.

2.2.5 Incorporation of the Passenger Arrival Model in the Simulation

The model described in section 2.2.2 for passenger choice of arrival time gives rise to a probability distribution of time of arrival for any individual passenger. By aggregating over passengers, this distribution becomes an "arrival rate" function. This arrival rate function represents the relative likelihood of a new passenger arrival at any given time, and is the basis for simulating passenger arrivals as a non-stationary Poisson process. By non-stationary, we mean that the arrival rate is not a constant, independent of time.

The usual notion of "random" arrivals is associated with a model of the arrival process as a stationary Poisson process. That is to say, the probability of any individual's time of arrival is a constant over time, and thus the aggregate rate at which people arrive is also constant.

Passenger arrivals to the simulation model are generated by sampling the interarrival times between passengers. For a stationary Poisson process, this simply involves sampling from an exponential distribution. Sampling interarrival times for non-stationary Poisson processes is more difficult, but not substantially so. The technique used in the simulation model is that described by Kaminsky (1977).

2.3 Vehicle Travel Times on Links

The results of any simulation are only as good as the inputs to the model. In the case of transit reliability, the single most important input may be that describing the vehicle movements along transit routes. Schedule reliability at stations derives heavily from the travel times experienced over all preceding links, making a realistic description of the travel time distribution of utmost importance. Also of concern is the fact that evidence exists to suggest that travel times on successive links are not independent. The tendency for deviations over successive links to correct for previous schedule errors has been identified in the study done by IBM for the CTA (IBM, 1974). This phenomenon will be discussed briefly in section 2.3.2. A more detailed discussion of a method used to estimate the magnitude of this effect is contained in Appendix A.

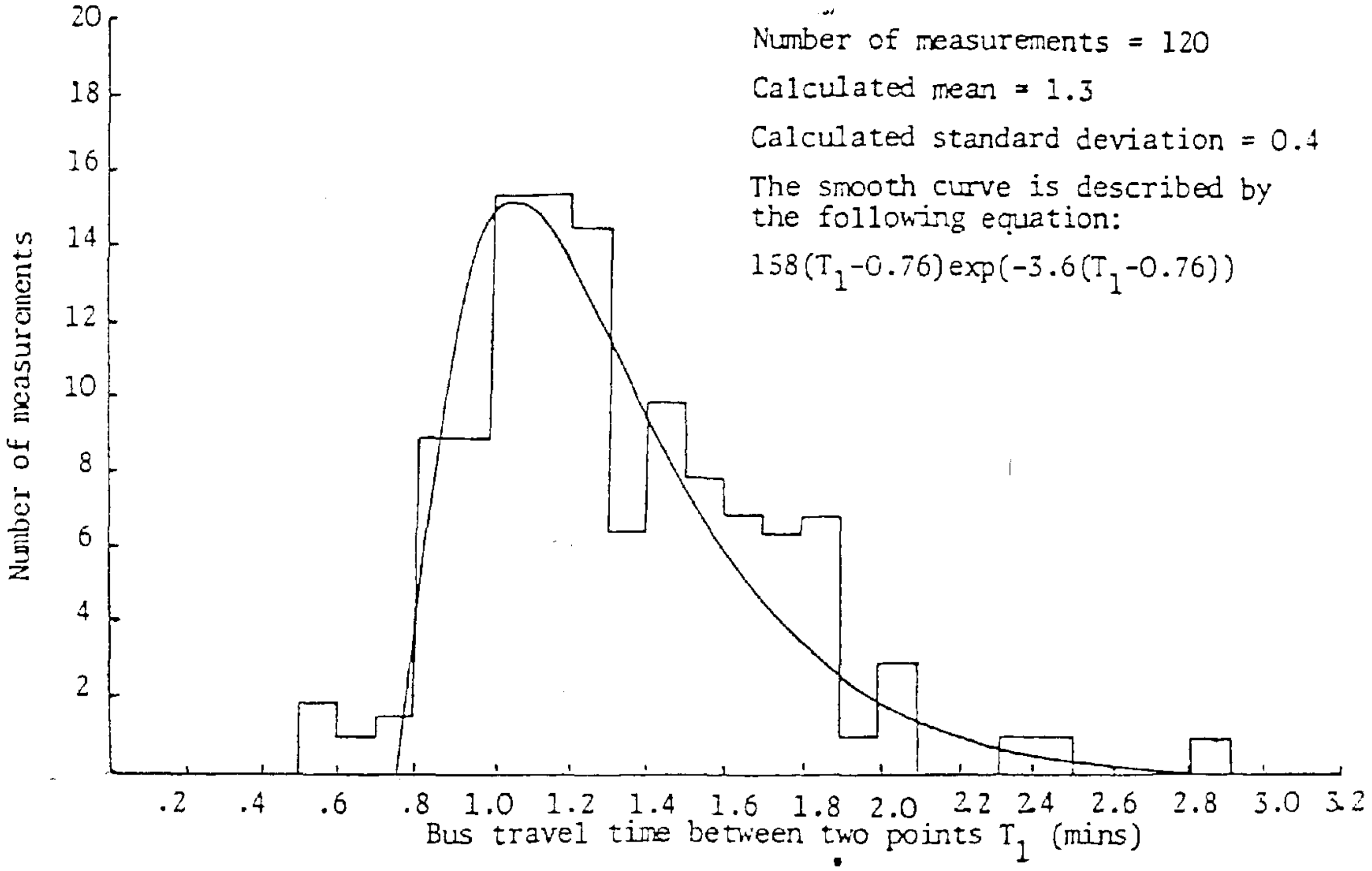
A number of authors have discussed the shape of the distribution of travel times: see for example, Newell (1974), Kulash (1971), dePirey (1971), Polus (1975), and Jenkins (1976). Newell examined dynamic instability for service around a closed loop, travel times being assumed to be uniformly distributed. Kulash, in the analysis of a highly unreliable service, used a simulation model wherein interarrival times of vehicles were allowed to follow an exponential distribution. This is the extreme case of purely random vehicle arrival times. dePirey's work also involved the use of a simulation model. He permitted travel times to be drawn from a Normal distribution, truncating the tails at -.5 and +2.5 deviations to avoid extremely short or long travel times. These three approaches were descriptive, they simply assumed distributions which appeared to have the desired shapes. Polus went a bit further, attempting to determine a distribution with the proper characteristics and then calibrating it. His result was a beta distribution, with parameters T and R, shown in equation 2-11.

$$\text{Travel time} = B\ x \qquad (2\text{-}11a)$$

$$\text{where} \qquad f(x) = \frac{\Gamma(T)}{\Gamma(R)\ \Gamma(T-R)}\ x^{R-1}(1-x)^{T-R-1} \qquad (2\text{-}11b)$$

The shape was determined by assuming a relationship between the two parameters: T=1.5*R. This gives the distribution its overall form,the remaining parameter, B, serving to stretch the distribution over the appropriate time scale. Polus' use of this distribution was to predict reliabilities by predicting the values the beta parameters would take on in specific situations. The resulting first and second moments estimated adequately the corresponding moments of the observed travel times, but the shape of the probability density function did not conform well to the shape of observed histograms. Our needs require the entire distribution to conform to observed data, preventing Polus' work from being utilized in the development of the inputs to our simulation model.

Jenkins discusses the use of a gamma distribution in a particular simulation study. The example cited, shown in Figure 2-4, is shifted away from the origin and is skewed with a long tail to the right (longer travel times). Jenkin's example comes closest to providing a usable descriptor of travel



Number of measurements = 120

Calculated mean = 1.3

Calculated standard deviation = 0.4

The smooth curve is described by the following equation:

$158(T_1-0.76)\exp(-3.6(T_1-0.76))$

Figure 2-4. Gamma distribution from Jenkins (1976).

times; however, it provides only a distribution for a particular link. A more general description is needed, which can be easily applied to any link in a system and requiring only the minimum of data acquisition.

The first step is to identify the desired characteristics of a link travel time distribution, and then to search for a model which could encompass as many of the desired features as possible. The general shape such a distribution should take on has been agreed upon for some time. Studies as early as Welding (1957), and perhaps earlier, have described the travel time distribution as somewhat bell shaped, but skewed with the longer tail extending toward the longer travel times. Ideally, the tail for short times should also be truncated, since there is a minimum reasonable travel time between two points. However, the longer tail could permit exceptionally long times in rare instances.

A set of distributions possessing some or all of the desired features would include the lognormal, gamma and beta distributions. These distributions could be either used directly, or shifted to account for minimum travel time. Based solely on flexibility the shifted beta would seem to be the distribution with the greatest chance of reproducing observed histograms of travel times. The selection of a preferred approach, however, must consider the explanation of why the distribution follows observed behavior, and also the ease in fitting the model to different situations. An equation able to reproduce in perfect detail the most complex circumstances has little value if detailed knowledge of the particular response must be known before calibration can take place. The objective is to be able to estimate the model parameters from readily available descriptors of the proposed service, and use the resulting estimated distribution to predict the travel times for a specific site.

While the shifted beta model offers the most flexibility, the shifted gamma yields the best explanation for observed behavior. The amount shifted can be determined by assuming the maximum legal speed over the length of the link, with the gamma distribution modeling the additional time as the sum of many small delays incurred over the link. These delays are due to signals, traffic conflicts, turning vehicles, congestion, etc. A gamma distribution has two parameters - a shape parameter, a, and a scale parameter, b. The shifted gamma distribution has the following moments:

$$\text{mean} - \text{shift} = ab$$
$$\text{variance} = ab^2$$

The distribution can be viewed as the sum of a exponential distributions each with the same parameter, b. Consider the delay time from each single interference faced by a bus to be exponentially distributed with an average duration of b. Then a bus's total delay as it travels over a link can be viewed as the sum of a independent delays each exponentially distributed with parameter b. This is precisely a gamma distribution with parameters a and b. Note that as the number of delays, a, increases the gamma distribution approaches a normal distribution.

2.3.1 Calibration

Two different data sets have been available for calibration of the link travel time models. The first is a set of travel time data for thirteen links in Chicago and Evanston, Illinois, collected during the summers of 1974 and 1978, while the second is from four routes in Cincinnati, Ohio, collected via automatic vehicle monitoring (AVM) equipment in March and April, 1978.

The travel time data collected in Evanston and Chicago include dwell times to service passengers at stops between the observers. The simulation model does not normally represent each and every bus stop explicitly, since such detail would increase greatly the cost of using the model without concommitant increase it utility of the results. Thus, a link is generally defined as a segment between major stops, often corresponding to "time points" on a route. The data on observed travel times reflect this also, being collected on "links" ranging from 1.6 to 13.5 km. (1.0 to 8.4 miles) in length. It is important to recognize, however, that the data thus collected do not reflect solely in-motion times. While this data set includes dwell time to serve passengers, explicit information on passenger activities (numbers of boardings and alightings) is not included.

The second data set, from four routes in Cincinnati, has been provided by the General Motors Urban Transportation Laboratory. It reflects operations in March and April of 1978, and includes information on numbers of stops, boarding passengers and alighting passengers, as well as overall times. The data are organized by "route segment," corresponding to links between time points. A total of 18 links are represented, ranging in length from 2.7 to 4.3 km. (1.7 to 2.7 miles).

Calibrations of gamma distributions were done separately for the two data sets, allowing comparison of results as an indication of potential transferability. In each case, the first step was to calculate the shift, or free-flow time, for each link. This is computed as the link length divided by the speed limit in effect.

The shift was subtracted from each observed travel time so that, hypothetically, observed delay times remained. Using the maximum likelihood method (see Choi and Wette (1969)) the parameters of a gamma distribution were estimated for delay times on each link. These estimates are denoted as a* and b*. Use of maximum likelihood estimation allowed confidence intervals to be placed on the estimated parameters.

For the Evanston-Chicago data, the hypothesis that travel times on each link are gamma distributed with the indicated shift and maximum likelihood estimators was tested at the 95% significance level with a Kolmogorov-Smirnov (K-S) test. For only 1 of the 13 links could the hypothesis be rejected. Therefore, it was concluded that the shifted gamma distribution is a good model of travel times.

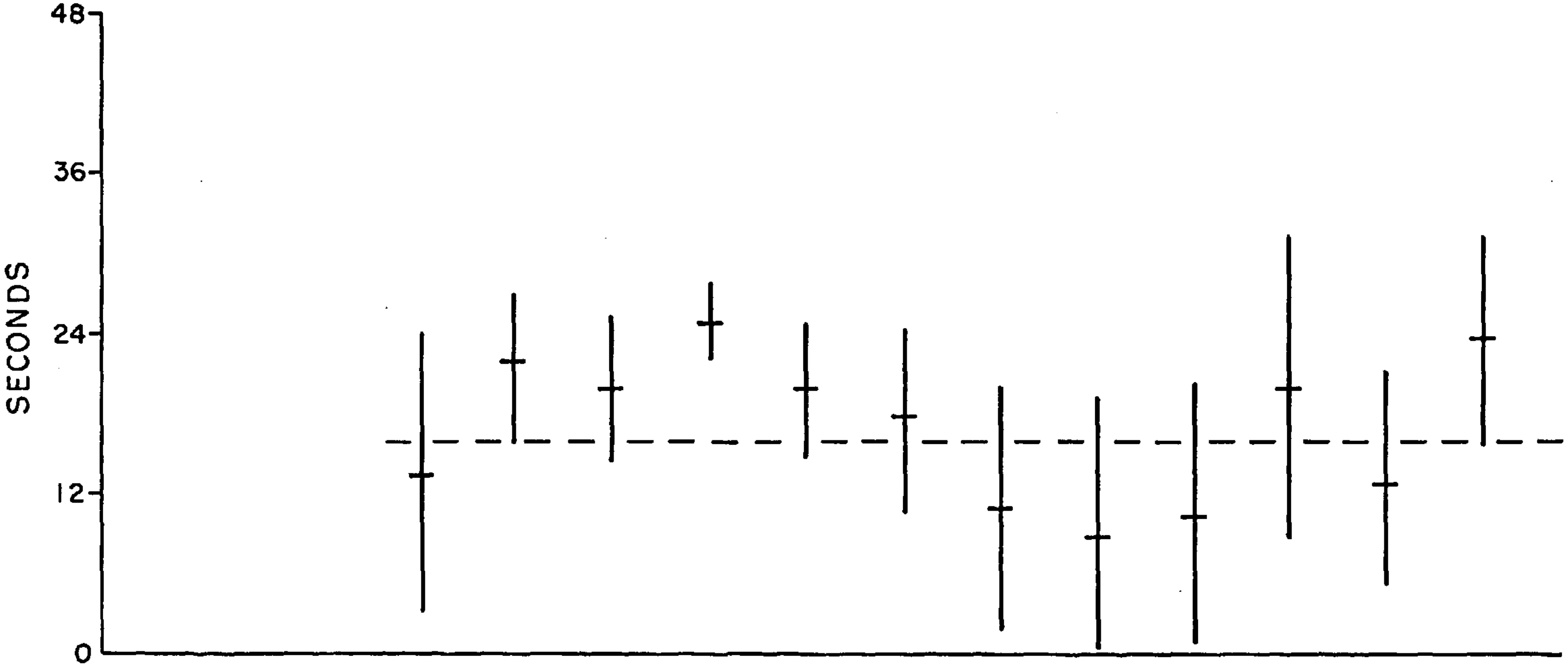
It was now determined whether a single value of b could be used in the gamma travel time distribution for all link types. Figure 2-5 shows the estimator b* and its 95% confidence interval for each of the 12 links for which the gamma distribution was not rejected. The confidence intervals overlap in the range between 16 and 20 seconds. A weighted average of the b* values yields a value of b* = 16.4 seconds. This value falls within the confidence interval of b for all but one of the links. Therefore, it was concluded that this value could be used for both express and local travel time distributions.

Fixing the value of b* at 16.4 seconds, values for the shift and a* were recomputed for each link (excluding link CR to SA for which the gamma was rejected) using the method of moments. That is:

$$a^* = \frac{\text{(observed link travel time variance)}}{b^{*2}} \qquad (2\text{-}12)$$

$$\text{shift} = \text{observed link average travel time)} - a^* b^* \qquad (2\text{-}13)$$

Figure 2-5. 95% confidence intervals for the estimates of b* from links in Evanston and Chicago.



Again a K-S test at the 95% significance level was made testing the hypothesis that the link travel times were gamma distributed with the shift and parameters thus computed. The hypothesis was not rejected for any link.

This analysis was repeated for the data from Cincinnati. Figure 2-6 shows the 95% confidence intervals computed for the value of b* using a total of 18 different links. The weighted average value of b* is 23.4 seconds. As shown in Figure 2-6, this value falls within the 95% confidence limits for 17 of the 18 links. While this value is somewhat higher than that determined for the Evanston-Chicago data, the area of overlap in the confidence intervals for individual links is not markedly different, being generally between 17 and 24 seconds.

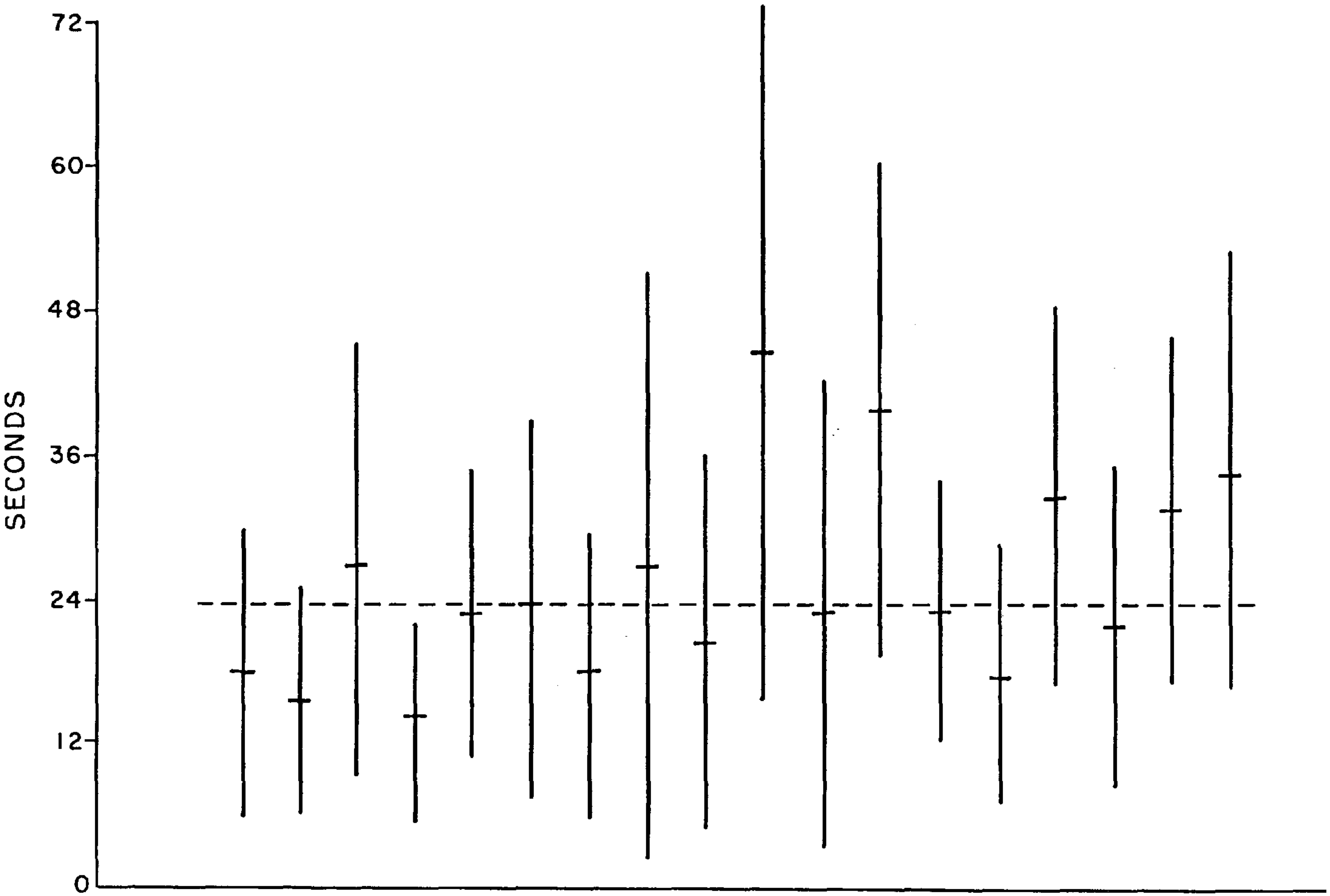
Since the Cincinnati data include information on passenger boardings and alightings at individual stops, as well as general link characteristics, a further attempt was made to see if a simple predictive relationship could be developed for the value of a*. Because boardings and alightings vary substantially by time of day, the data were stratified into morning peak, mid-day and afternoon peak periods. Thus, the 18 links generated 54 data points. Four of these had to be discarded because of insufficient observations to compute reasonable variance estimates, leaving a total of 50.

Linear regression was used to estimate a relationship between a* and link characteristics. The resulting model is shown in equation 2-14.

$$\underset{(4.6)}{a^* = 4.5} + \underset{(.08)}{0.3\,P} - \underset{(1.5)}{0.9\,L} + \underset{(0.2)}{0.5\,S} \qquad (2\text{-}14)$$

$R^2 = .49$

Figure 2-6. 95% confidence intervals for the estimates of b* from links in Cincinnati.



where P = average total boarding and alighting passengers

L = length of link (km)

S = number of signalized intersections along link.

The value in brackets under each estimated coefficient is the standard error of that estimate.

Since the coefficient for length and the constant term are not significantly different from zero, a reasonable predictive model simply involves the variables P and S. Rerunning the regression without the constant term or the L variable results in the following model:

$$a^* = \underset{(.08)}{0.3\ P} + \underset{(.17)}{0.6\ S} \qquad (2\text{-}15)$$

Thus, using a constant value of b*=20 seconds (.33 min.), equation 2-15 to predict a*, and a shift computed as link length divided by the speed limit in effect, the travel time distribution for any link can be determined.

Figures 2-7 and 2-8 illustrate the estimated versus observed means and variances of travel times, respectively, for the Cincinnati data. This provides an overall indication of goodness-of-fit for the entire procedure. As shown in Figure 2-7, the mean value prediction is quite good. The dashed line "estimate = observed" is shown for reference. Figure 2-8 shows that this procedure has a tendency to underestimate variance of travel time. Generally speaking, this is the effect of underestimating the value of b* in some cases as a result of adopting a constant value for all links.

Quite clearly, this model for parameter estimation of the link travel time distributions has some shortcomings. However, it is a very simple model which explains a substantial portion of total variation in link travel time characteristics. As such, it appears to be a useful procedure. It

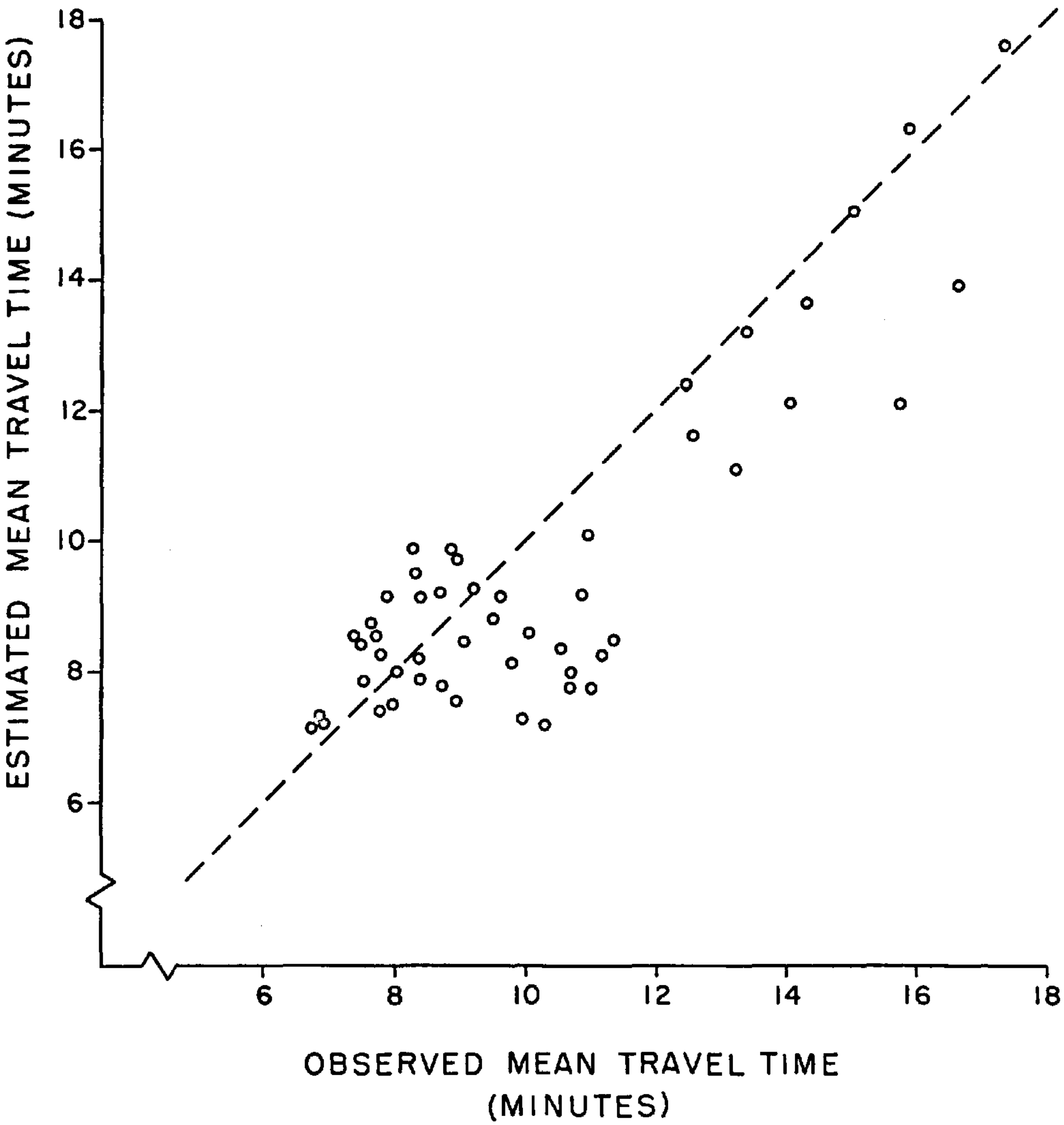


Figure 2-7. Estimated versus observed mean travel time on links.

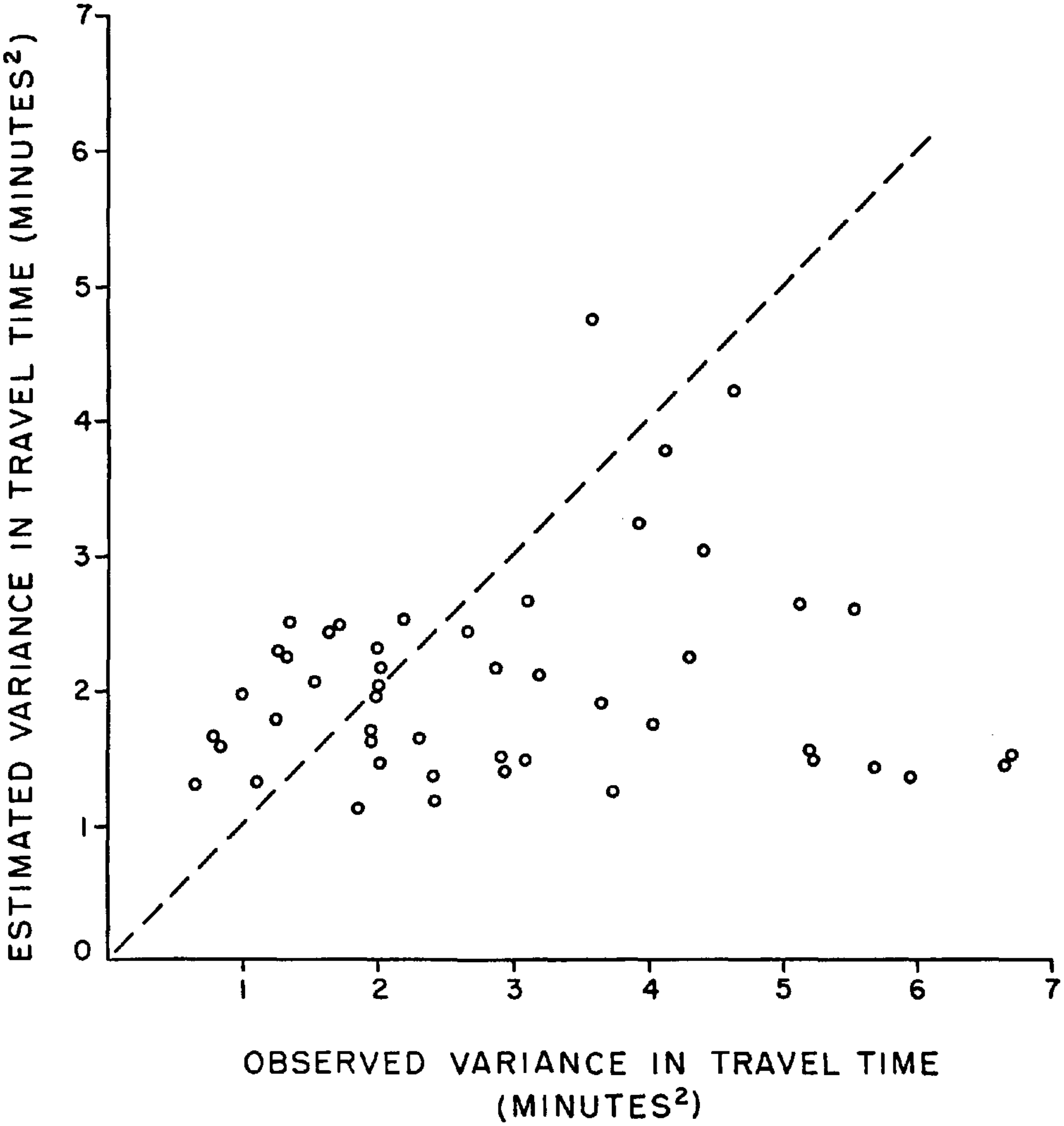


Figure 2-8. Estimated versus observed variance in link travel time.

should be noted that this estimation procedure could be used by a transit operator or analyst, quite independently of the simulation model, to characterize travel times on route segments. This could provide useful information for the scheduling and run cutting process.

2.3.2 Recovery Tendency

Although most studies of reliability along routes assume independence between travel times on successive links, there is evidence to suggest that this is not necessarily the case. Welding (1957) made reference to a tendency for drivers to regain their schedule. More recently, an IBM report prepared for the Chicago Transit Authority (IBM, 1974) made use of what were referred to as "coupling curves." The name, it would appear, comes from the dependency of one link travel time on the travel time for the preceding link. The basic idea expressed by these coupling curves is that while on average a late (or early) arrival at one station will be followed by a late (or early) arrival at the next, the total amount of schedule error tends to decrease. For example, a vehicle which was previously 3 minutes late, on average was found to be only 2 minutes behind schedule when next observed.

The magnitude of impact from such a self-correcting mechanism will be dependent on a number of factors: traffic conditions, tightness of schedule, driver knowledge of schedule, and motivational factors. The results in the IBM report, while illustrative of the phenomenon in question, do not provide sufficient information for general application. As a result, data from Evanston and Chicago were employed in an attempt to describe the effects more precisely.

Two routes were selected, one in Chicago and one in Evanston. The data from the route in Evanston confirmed the tendency for deviations from schedule to be at least partially offset during travel over links further along the route. However, data from the route in Chicago did not. Details of the analysis are presented in Appendix A. The different results from the two routes can be attributed to the different environments in which the routes operate. First of all, the ability of the driver to compensate for schedule error will depend on the presence of slack (excess) time in the schedule, as well as on traffic conditions. Travel in Evanston is less congested and therefore offers the driver more opportunity to speed up when necessary. The second influence is the presence of temporal variation. Positive correlation between link travel times caused by changes in congestion over time will serve to obscure the presence of any negative correlation induced by a driver's efforts at regaining the schedule. The recovery factor observed in Evanston was used in validating the simulation model (to be discussed in section 2.4) but it was not included in the experiments run for analysis, since it may not be a generally applicable phenomenon.

2.4 Validation

Validation of the simulation model requires both validation of the individual components, and validation of the system model as a whole. Statistical analyses of empirical data described in sections 2.2 and 2.3 have served to verify the behavior of component models reflecting passenger and vehicle arrivals. However, validation of the system as a whole requires an existing bus system to be simulated. The comparison of simulated and observed performance provides the basis for evaluating the validity of the model. Since data collected in Evanston is by far the most complete set available to us, the Evanston bus system is the logical choice for such a comparison. The Evanston route network is illustrated in Figure 2-9.

In general, it is preferable to validate the model against a system not

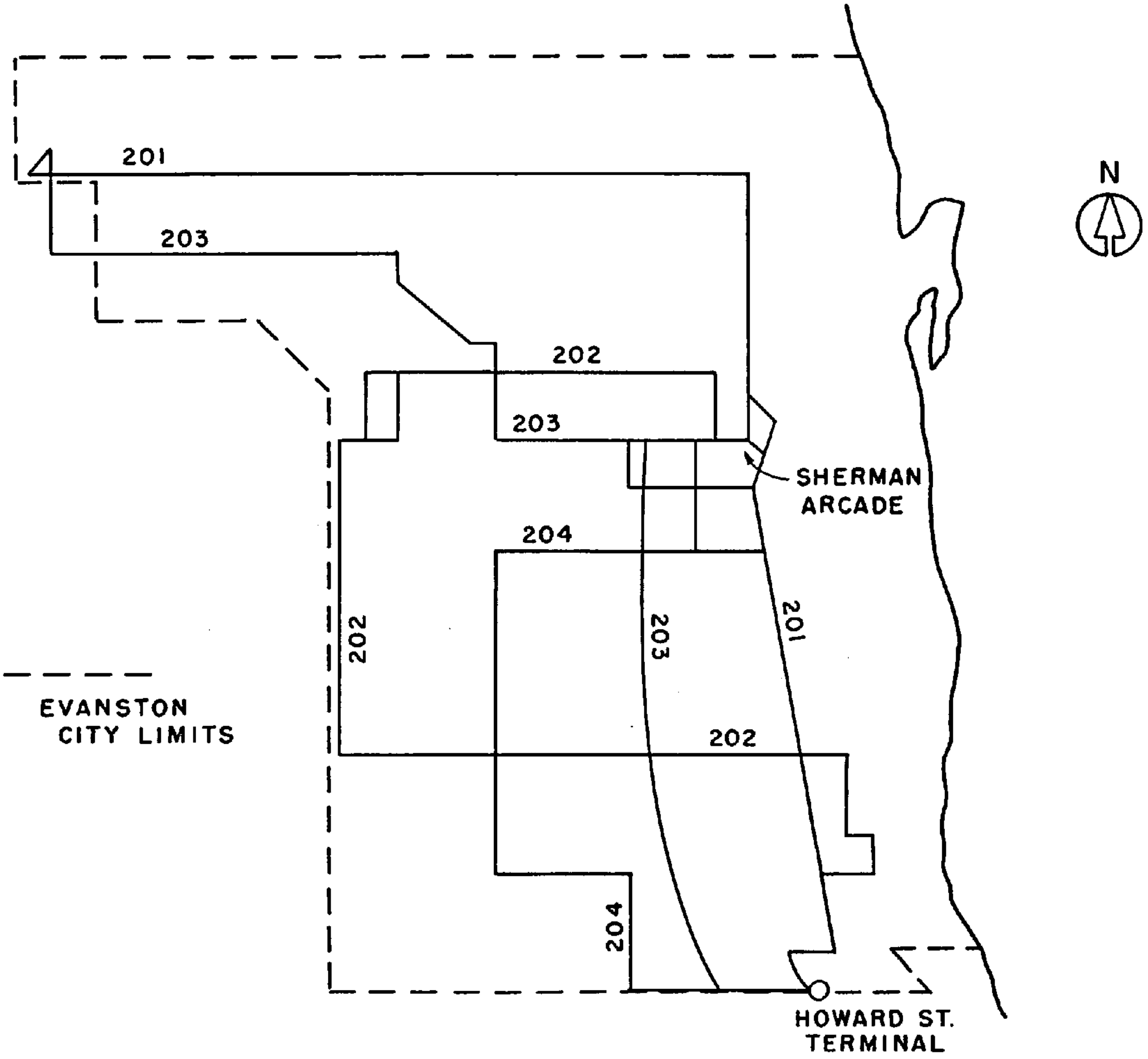


Figure 2-9. Evanston Bus Routes.

involved in the calibration of parts of the model. However, since the purpose of this part of the validation is to verify the interaction of the individual components, the structure of which was not involved in the component calibrations, comparing model performance to Evanston data is still useful. During the second year of this project, more complete validation against another system will be undertaken. In the simulation of Evanston, measurements of the actual schedules and deviations were used, so validation did not have to be concerned with the soundness of the assumed values. Validation concentrated instead on the overall performance of the model.

Seven simulated morning rush hour periods were compared to seven days of observations collected in Evanston in 1974. Analysis of the performance of model interactions will address questions in three categories:

1. Schedule adherence
   Do vehicle schedule deviations resemble those observed in the real system?

2. Transfer volumes
   Do zonal generation-attraction rates result in the desired loadings? Do origin-destination flows correspond to those observed? And do passengers make correct decisions as to route choice?

3. Transfer delay
   How do the transfer times in the model correspond to those observed in the field?

2.4.1 Vehicle Schedule Adherence

Measurements were taken on two major attributes of vehicle departure times from stops:

1. The average difference between the actual and scheduled departure times for given stops, denoted "lateness".

2. Standard deviations of actual departure times from given stops.

These data were available for a total of 21 stops in the Evanston system, including a route terminal and a major transfer point. These data can be compared with the simulated performance as a measure of validity of the model.

Figures 2-10 and 2-11 illustrate plots of simulated versus actual data for these two attributes of schedule adherence. For the lateness measure, the correlation coefficient between simulated and observed values is .893. The line indicating "simulated - actual" is included in Figure 2-10 for comparison to the results obtained. While the observed results are not explained entirely by the simulation, there is close agreement. Major departures from schedule are reflected quite well.

The standard deviations of departure times from stops are not modeled as accurately, but the principal source of the inaccuracy has been identified and corrected. The data points shown in Figure 2-11 are separated into two groups. Those indicated by an "X" represent the initial stop on a given route, while all other stops are indicated by a circle (o). For most stops, the deviations are reflected reasonably well by the model. However, in the observed data, there exists substantial variability in the departure times from initial stops along routes, which was not originally reflected in the simulation. This occurred because dispatches from terminals were made in the model exactly on schedule, so that deviations from schedule at the first few stops along the route were generally quite small. However, observed data indicates that dispatches from the terminal are distributed about the schedule with a standard deviation of approximately one minute. Incorporating this into the simulation has improved the accuracy of the model.

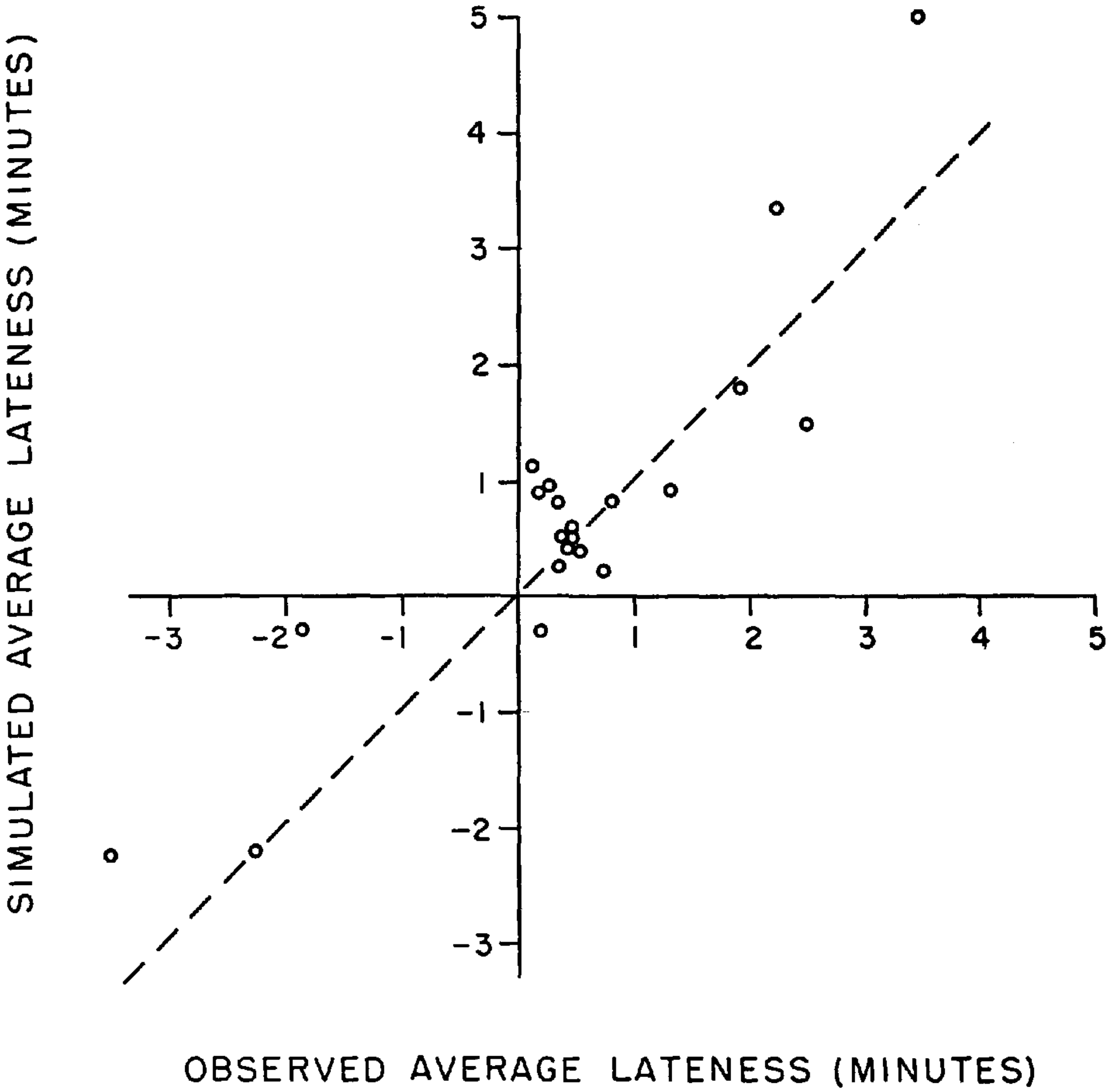


Figure 2-10. Comparison of vehicle lateness at stops

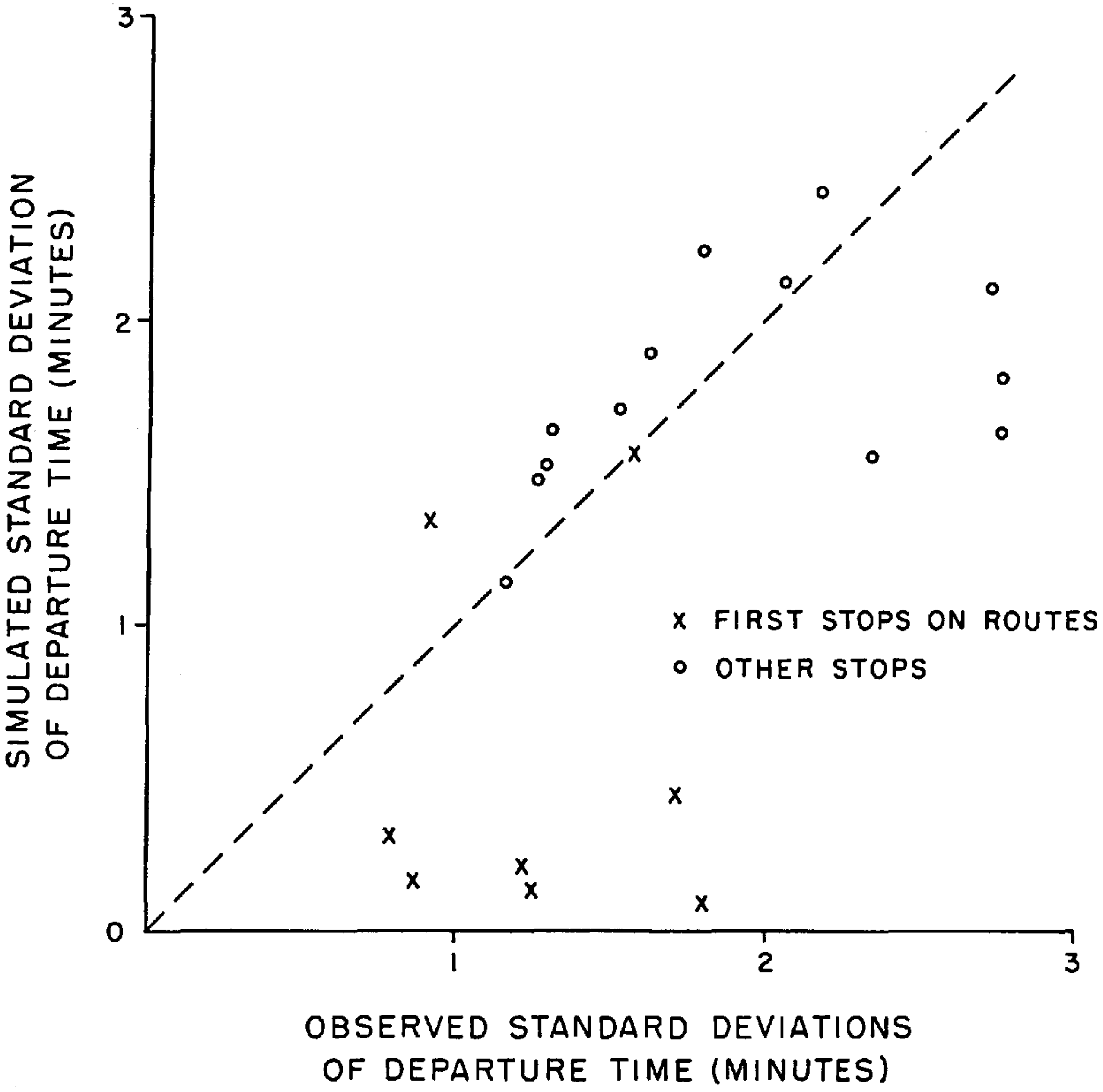


Figure 2-11. Comparison of standard deviations of departure times from stops.

2.4.2 Number of Transfers

Numbers of transfers at various stops are an indication of whether the passenger route selection logic is working properly. Figure 2-12 illustrates simulated and observed numbers of transfers at all major transfer points in the system. Quite clearly, the model is reflecting paths through the system very accurately. This is particularly important for observations at the single stop (Sherman Arcade in downtown Evanston) where approximately one-helf of all transfers take place.

2.4.3 Transfer Delay for Passengers

Because transfers are a major contributor to uncertainty in overall passenger trip time, it is important that the model reflect the characteristics of transfer time accurately. Accurate predictions of transfer times can also be taken as a measure of model validity, since it represents a principal result of network interactions among individual bus routes. Observed and simulated results for two different attributes of transfer time have been compared for various stops in the system - the mean transfer delay for all passengers transferring, and the standard deviation of those delays.

Figure 2-13 illustrates the comparison for average transfer delay. The correlation coefficient between simulated results and actual data is .94, and Figure 2-13 shows quite clearly that average transfer delays at individual stations are reproduced very well by the model.

The comparison of the standard deviations of transfer times is shown in Figure 2-14. In general, the results are quite good, with a correlation coefficient between observed and simulated values of .85. The model

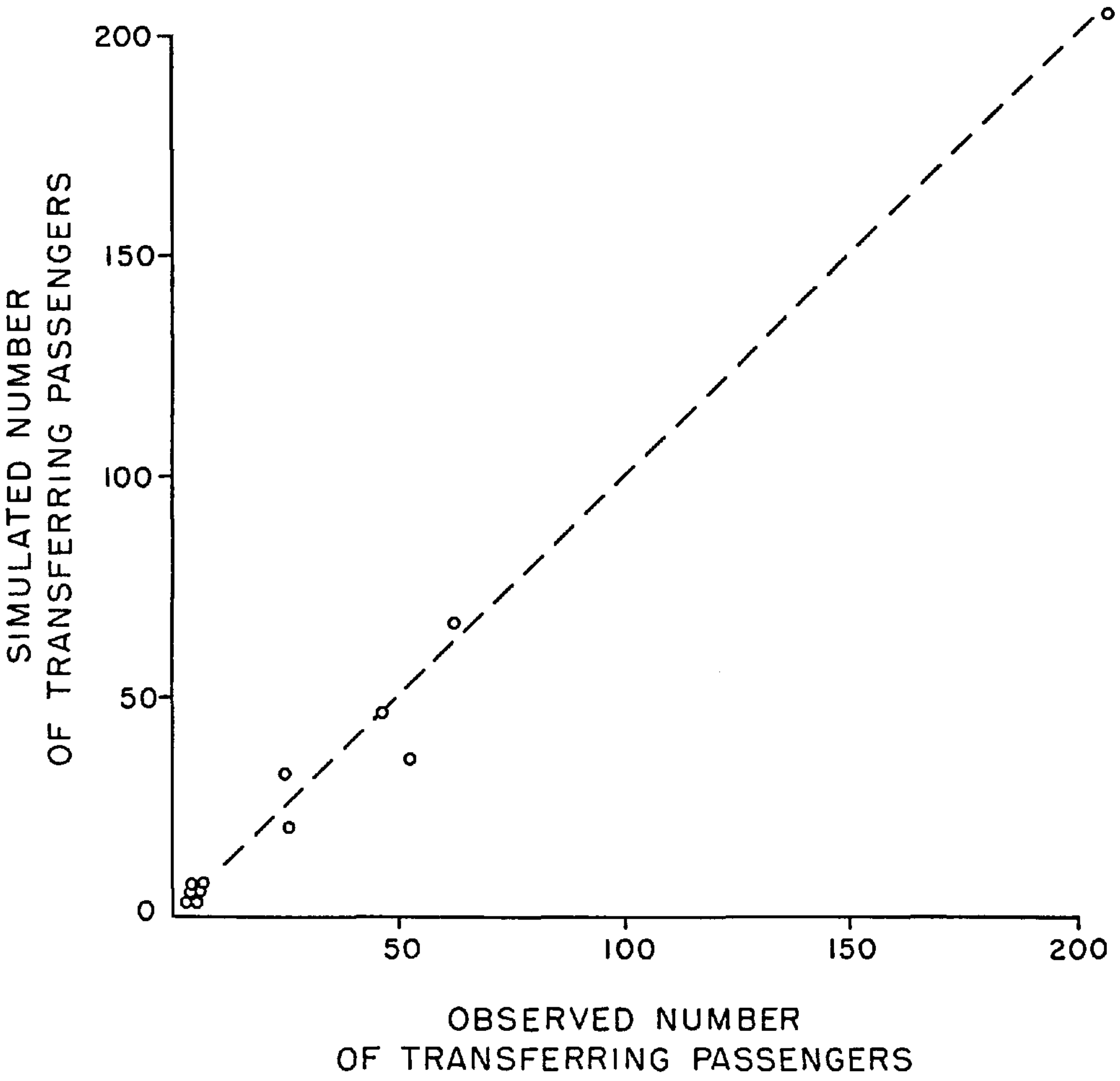


Figure 2-12. Simulated versus observed numbers of transfers at various stops.

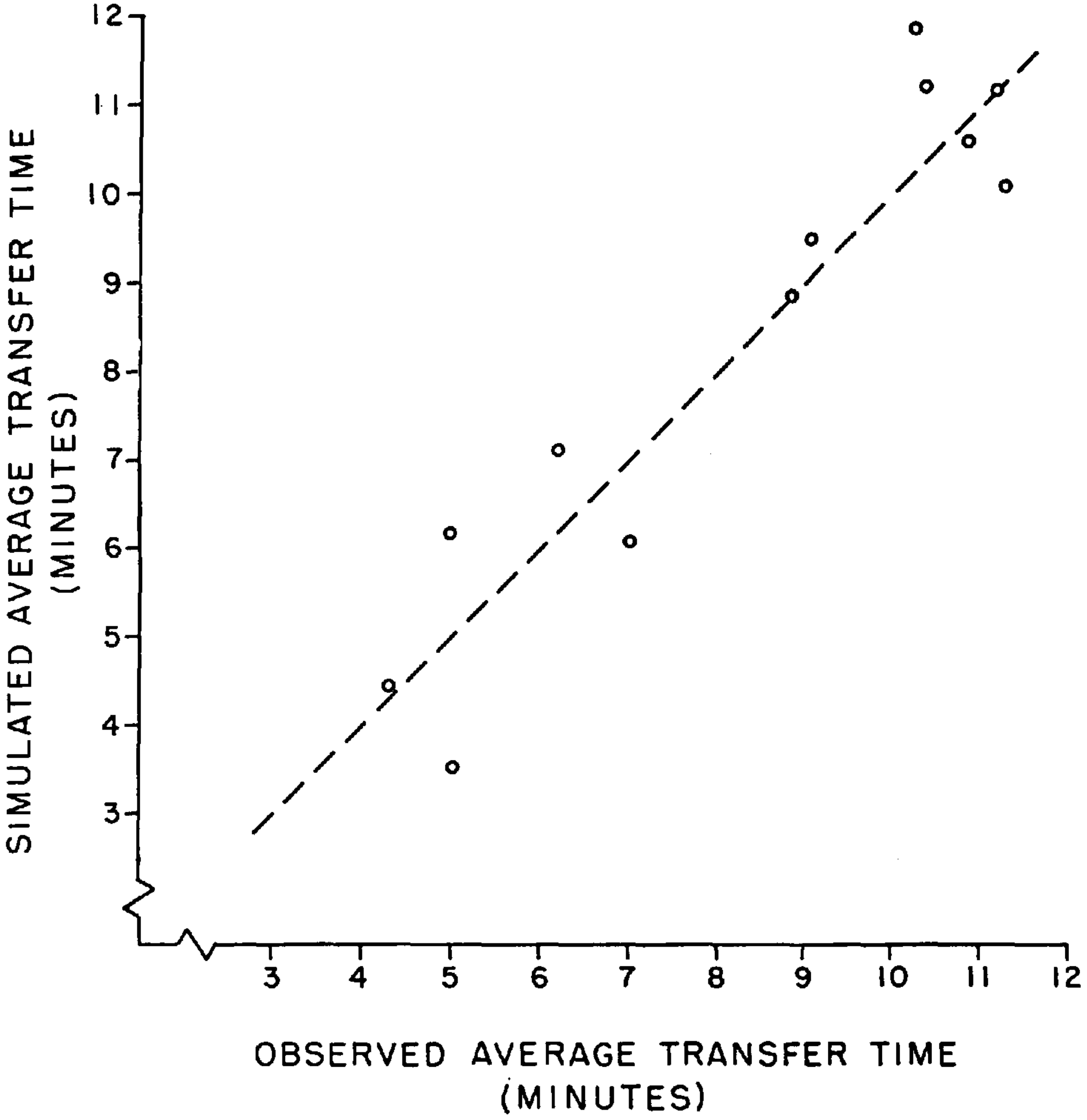


Figure 2-13. Comparison of average transfer delay.

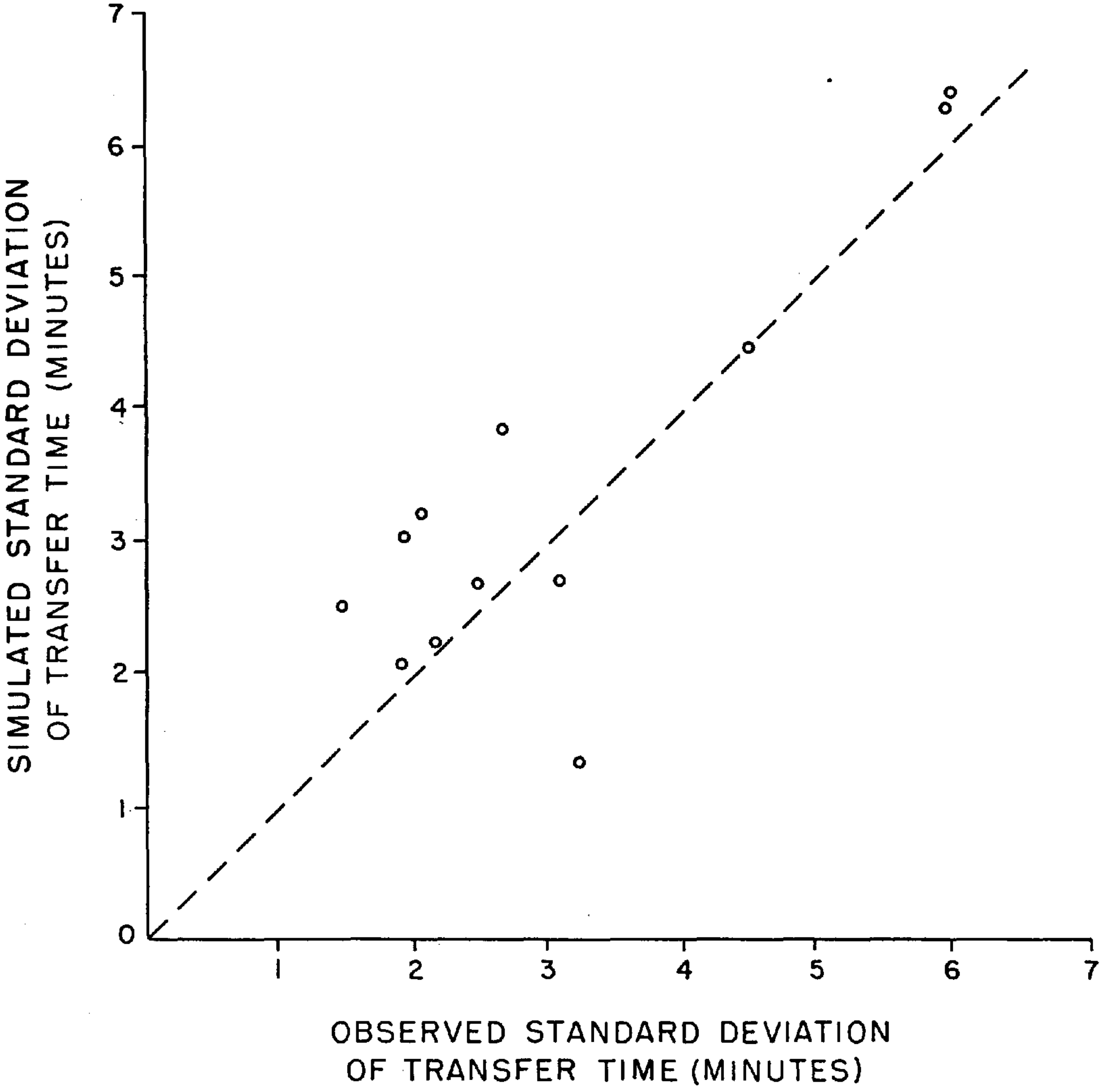


Figure 2-14. Comparison of standard deviations of transfer delay.

results are somewhat less accurate for these deviations than for the mean values, particularly in the 1-3 minute range. However, it is encouraging to note that the model reflects well the differences between stops in the 1-3 minute range and those in the 6-7 minute range.

2.4.4 Summary of Validation Tests

In general, the simulated results conform quite well to the observed data from the Evanston system. The comparisons illustrate that it is easier to predict means than to predict deviations from the mean, but even when dealing with measures of deviation the simulation results are reasonably accurate.

The validation exercises indicated one aspect in which the model could be improved significantly. This involved the degree of schedule adherence on departure from the dispatching point at a route terminal. This change was incorporated into the model before the final experiments were run.

On the basis of these validation tests, it was decided to proceed to a phase of experimentation with alternative route structures, densities, traffic conditions, volumes, etc. These experiments form the basis for conclusions regarding strategies for the control of reliability.

3. The Effects of Network Characteristics on Reliability

Strategies for the control of service reliability can be classified in several ways. One of the most obvious distinctions is between what might be termed "planning" strategies and "real-time" strategies. In general, the distinction is that planning strategies involve changes of a persistent nature, while real-time strategies do not. For example, changes in route structure can have substantial effects on the basic characteristics of service provided, and the decision to make such changes is the result of the operations planning process. On the other hand, real-time control measures such as delaying some vehicles at schedule checkpoints, are designed to act quickly to remedy specific problems. These actions have immediate effects, but seldom exert any influence on the general nature of operations over a longer time period. A comprehensive review of both types of strategies is contained in the recent paper by Blume and Turnquist (1978).

This section describes simulation experiments conducted to analyze the effects of several network characteristics on service reliability. The results of these experiments form the basis for conclusions regarding the potential effectiveness of a number of different planning strategies for control of service reliability. Experiments conducted to assess the potential of several real-time control strategies are discussed in section 4.

3.1 Experimental Design

In order to insure that the experiments conducted will produce the information desired at reasonable cost, attention must be paid to the question of experimental design. A properly designed set of experiments will yield information on all the important terms while minimizing the number of runs necessary. Four steps are involved in setting up an effective set of experiments: (1) the important variables must be identified; (2) the interaction terms which may be of significance should be recognized; (3) the levels of the variables to be tested must be determined; and (4) an appropriate experimental layout must be selected.

Five variables were identified with the most significant potential effects on the reliability of transit networks. The first four variables, frequency of service - FREQ, standard deviation of link travel time - STDDEV, demand/capacity ratio - DEM, and route density - DENS, are quantitative. The fifth variable, network form - FORM, is descriptive. Frequency of service and the standard deviation of link travel times are self explanatory. Form refers to the shape or structure of the network, route density is defined as the miles of two-way transit routes divided by the area in square miles, and the demand/capacity ratio refers to the passenger miles carried per hour divided by the available "space" miles (both seated and standee spaces) per hour on all vehicles in the network.

The number of levels at which the variables are to be tested depends on the type of response being investigated. Linear responses can be estimated from measurements at two levels, but nonlinear responses require a minimum of three levels for estimation. Two of the quantitative variables, frequency and link deviation, were investigated for non-linearities in their effect on the reliability of service, requiring that they be tested at three levels. The other two quantitative responses, demand and route density, were tested only for linear responses, so two levels were sufficient. The descriptive variable, form, takes on two levels, corresponding to grid and radial patterns of routes. Table 3 1 lists the five chosen variables along with their important interaction terms

Once the number of levels to be tested was determined for each variable, it remained to identify exact values corresponding to each level. Existing services were surveyed to aid in the selection of values for the frequency, density and link variation variables. Values for demand were chosen to

Table 3-1. Main effects and important interactions.

MAIN EFFECTS	INTERACTIONS
FREQ	FREQ-STDDEV
STDDEV	FREQ-DEM
DEM	DENS-STDDEV
DENS	DEM-STDDEV
FORM	FORM-DENS
	FORM-STDDEV

cover as broad a range as possible, with the highest level pressing the practical limits of the system.

The densities selected represent a compromise between the desire to model large areas with routes sufficiently long for deviations to build, and consideration of the cost of model execution for larger networks. Five areas in Chicago and the surrounding region were defined. The areas bounded and the route miles contained in each were measured, yielding five values of route density. Table 3-2 summarizes the densities obtained. While the densities observed ranged from 1.3 to 3.0 route miles per square mile of area, the levels used in the experiments were 1.0 and 1.8. The range selected for the experiments is lower than those observed because of the need to compromise between the number of zones and the length of bus routes. Desired route lengths were taken from a review of the lengths observed in the Chicago bus system. The average Chicago bus route is about 7.9 miles long. Most transit systems can be expected to have shorter routes than Chicago, so a minimum length of 4 to 5 miles seem reasonable, with a preference for something a little longer. Figure 3-1 illustrates the four network configurations of grid and radial networks at the two density levels chosen.

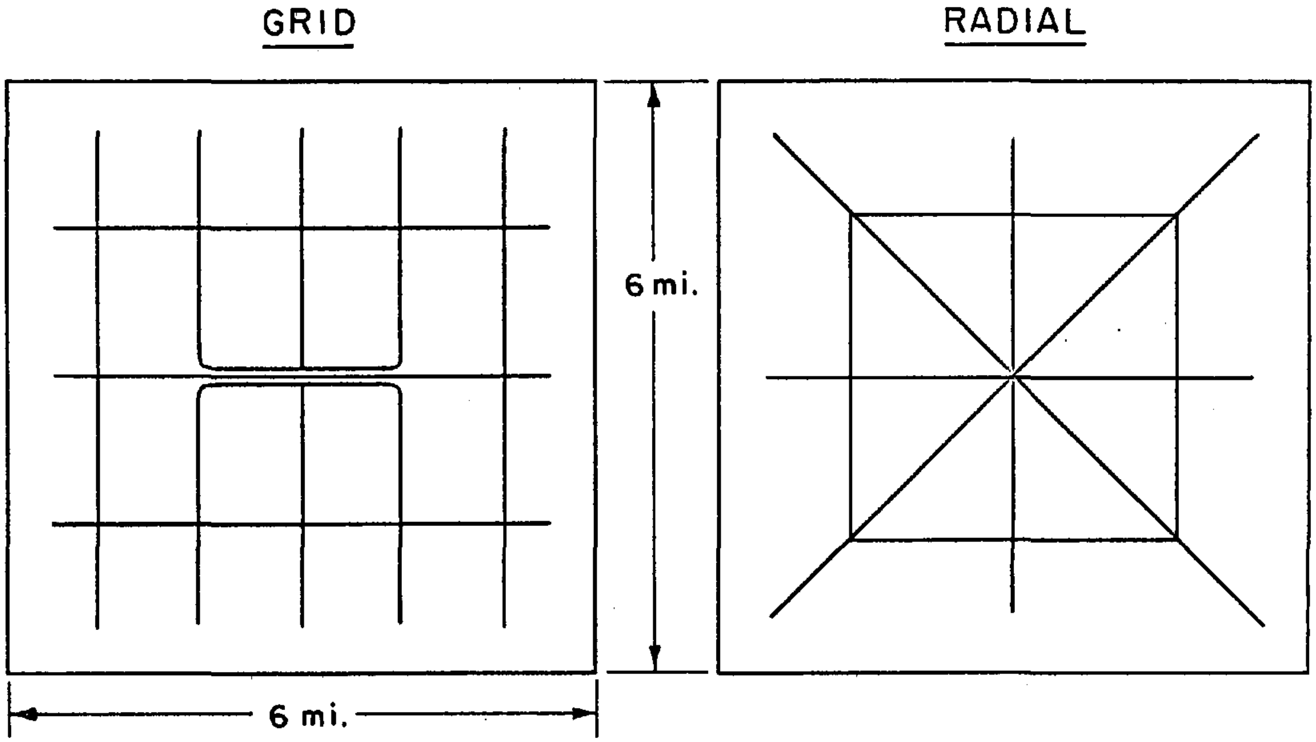
Table 3-2. Sample route densities.

Region	Area	Rte-miles	Density
NortranA	18.2	24.4	1.34
NortranB	25.7	49.6	1.93
Wilmette	5.80	10.6	1.87
Evanston	7.92	21.4	2.70
Chicago			2-3 @

@ Common spacings for transit routes in large parts of Chicago are a route every one or one-half mile. Densities of two or three route-miles per square mile of area will result when routes in a grid are spaced one mile apart in both directions, or one mile in one direction and one-half mile in the other.

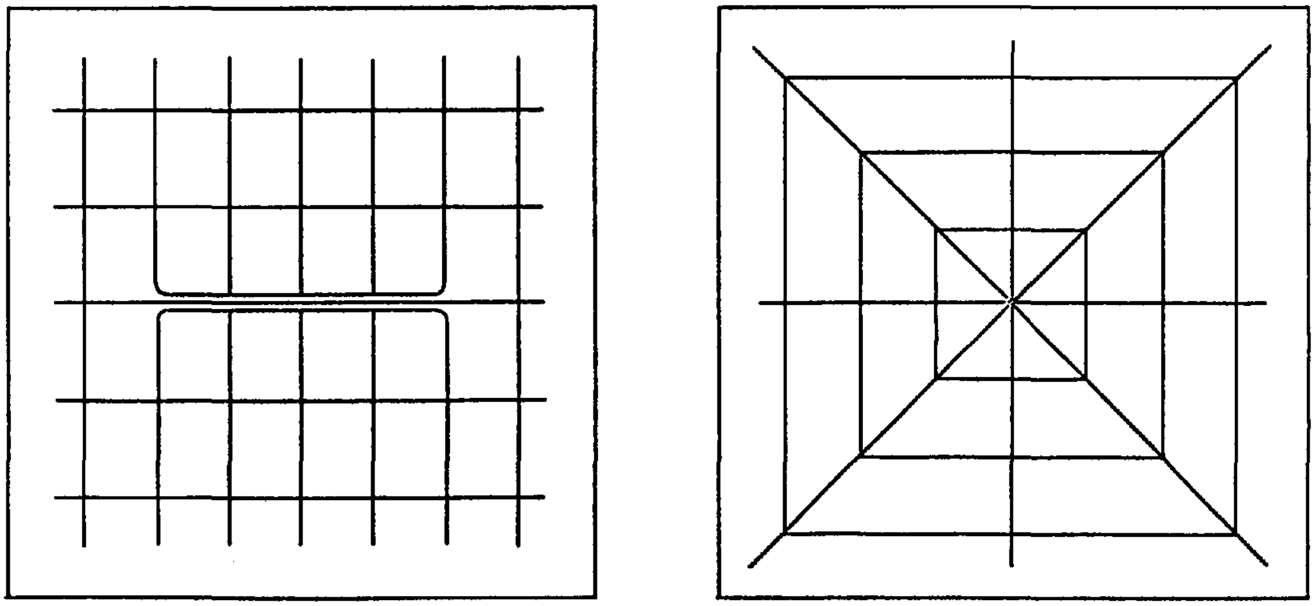
In determining a range of values for the standard deviation of link travel times,data from Evanston and Chicago were again employed. The average and deviation of travel times on 48 links were used to compute the standard deviation as a fraction of the mean travel time. The ratios of the deviations to the mean travel times ranged from .02 to .22. The three levels of STDDEV selected for use in the experiments are 2, 12, and 22% of the mean link travel times.

The frequency variable was set at 3, 7.5, and 12 departures per hour, corresponding to headways of 20, 8, and 5 minutes respectively. The final quantitative variable, demand was set at 10% and 40% of the network capacity. The average passenger trip length was computed for each network under the assumed pattern of demand, so the capacity of the system reflects the network structure. Demand for transit was assumed to be essentially uniformly distributed across the network, but with the attractors of the central zones set at twice the normal level to reflect a general CBD orientation in trip making. An upper bound of 40% was chosen because due to high demand links and peaking in demand, even at 40%, local demand



Route miles - 40

Route miles - 36



Route miles - 60

Route miles - 60

Figure 3-1. Network configurations.

sometimes exceeds vehicle capacities. Table 3-3 summarizes the levels assumed for all five variables.

Table 3-3. Variable levels tested.

Frequency	3.0 - 7.5 - 12.0 buses/hour
Deviation	2 - 12 - 22% of link travel time
Demand	10 - 40% of capacity (in passenger miles)
Density	1.0 - 1.8 route miles/square mile of area
Form	grid-radial

The levels selected for the four quantitative variables and the two forms leads to a set of experiments with two variables at three levels and three variables at two levels. Conner and Young (1961) present the necessary experiments to carry out a one-half replication of a factorial design with this combination of variable levels. It requires 36 experiments, preserves independence between the estimates of the five main effects, and retains estimates of all the first order interaction terms. Table 3-4 lists the 36 experimental runs in the analysis of network interactions.

3.2 Measures of Network Reliability

While reliability has generally been agreed upon as one of the most important attributes of transit service, there are not widely accepted measures of reliability. The analysis in this study is based on four measures, reflecting different aspects of system performance. The forms of measurement chosen, listed in Table 3-5, include the standard deviation in vehicle arrival times at stops, the coefficient of variation of arrival times (standard deviation divided by the mean service interval), the coefficient of variation in transfer times, and an aggregate measure of deviations in arrival and transfer times. The aggregate measure represents uncertainty in

Table 3-4. Primary set of experiments for analysis of network effects.

*********Grid*********				******Radial******			
FREQ	STDDEV	DEM	DENS	FREQ	STDDEV	DEM	DENS
0	0	0	0	0	0	1	0
1	2	0	0	1	2	1	0
2	1	0	0	2	1	1	0
0	0	1	1	0	0	0	1
1	2	1	1	1	2	0	1
2	1	1	1	2	1	0	1
1	0	1	0	1	0	1	1
0	1	1	0	0	1	1	1
2	2	1	0	2	2	1	1
1	0	0	1	1	0	0	0
0	1	0	1	0	1	0	0
2	2	0	1	2	2	0	0
2	0	1	0	2	0	1	1
0	2	1	0	0	2	1	1
1	1	1	0	1	1	1	1
2	0	0	1	2	0	0	0
0	2	0	1	0	2	0	0
1	1	0	1	1	1	0	0

Table 3-5. Selected reliability measures

Name	Form	Purpose
DEV	Standard deviation in vehicle arrival time.	Uncertainty in arrival time.
DEV/H	Coefficient of variation of arrival time.	Important in studies of average passenger wait time.
TFDEV/H	Coefficient of variation of transfer delays.	Indicates effectiveness of transfer connections.
AGDEV	Twice the deviation of arrival time plus the deviation in transfer time times the expected number of transfers.	Measures uncertainty in total origin-destination trip time.

total origin-destination trip time, reflecting the variability in arrival time of the bus at the origin stop and the destination stop, plus the expected impact of variability in transfer time for each passenger.

Recent work by Hashizume (1978) tested forms of reliability measures appropriate for different levels of analysis; single stops, single routes, or multiple route systems. His discussion focuses on two aspects of service relevant to a consideration of reliability - the overall degree of variability and measures of extreme performance.

Hashizume concluded that either the standard deviation or a measure of percentile range is most appropriate for measuring variability, with a slight preference expressed for the percentile range form (a percentile range is the smallest range in arrival time values encompassing a specified percentage of observations). A similar measure is suggested for measuring extreme performance. A percentage level is chosen, with extreme performance indicating the value above/below which the desired percentage of observations fall.

Although Hashizume preferred the percentile range measure, standard deviation was chosen to represent central tencency in the analysis. For distributions of arrival times which are approximately normally distributed, the information provided by both of these measures will be the same. A related approach, outlined in a paper by Berry and Belmont (1950), uses percentile measures from observed data to estimate the parameters of the underlying distribution. It has been shown that adequate estimates of the deviation can be obtained even when the underlying distribution is highly non-normal. Given their similarity, the ease of computation favored use of the standard deviation. This choice also avoided the need to arbitrarily select a value for use in computing percentile ranges.

The second aspect of reliability discussed by Hashizume (1978) was the measure of extreme performance. One effect of large departures from schedule is the failure to make connections as planned. Since we are dealing with networks, extreme performance is represented by the uncertainty in delay due to transfers. This check on reliability offers the advantage of reflecting the importance of extreme deviations within the context of network interactions. The value of transfer deviations turns out to predominantly reflect frequency of service. Therefore the value of the standard deviation was scaled by the length of the mean service interval, improving the sensitivity to other factors.

Measures of the coefficient of variation of vehicle arrival times are important to service as it relates to passenger wait time. Turnquist (1978) developed simple analytic models to investigate the influence of reliability on user behavior. The expression for the optimal user arrival time, and resulting average passenger wait time, suggest that the mean service interval, as well as uncertainty in vehicle arrival times, should be represented in the measure chosen. The coefficient of variation is a convenient measure reflecting this aspect of service reliability.

Hashizume also touched on the question of reliability in the context of multiple route systems (networks). Service at a network level is best represented by the uncertainty in door-to-door trip time. Ideally this would be computed by determining the variation in individual O-D travel times, and weighting zone pair deviations by their respective passenger volumes to arrive at a single measure. Zone pairs have extremely small sample sizes, however, presenting significant problems in data collection. Attempts to stratify O-D pairs at a less specific level proved unsuccessful.

The purpose of an aggregate measure is to represent the total uncertainty in O-D travel time in a form which can be easily measured. This is accomplished by combining the deviations in arrival and transfer times. The aggregate measure contains twice the deviation in arrival times (uncertainty in getting on the bus at the origin and off at the destination), plus the deviation in transfer delay multiplied by the expected number of transfers in a passenger trip.

3.3 Analysis of Experimental Results

Linear regression was used to analyze the data from the experiments described in the previous section. The variables included in the final regression equations are summarized in Table 3-6 and discussed in the remainder of this chapter. Separate regressions were run with and without the presence of interaction terms, and only those variables included at the 10% significance level were allowed in the equations. To aid in comparison of the relative impacts of variables in the equations, the coefficients have been converted to normal standardized regression coefficients.

In analyzing the network impacts on the reliability of service, the influences of the various network parameters will be contrasted along each of the four dimensions of reliability. Table 3-7 provides a graphical representation of the effects which are found to exert significant influence on the four measures.

Tables 3-6 and 3-7 can be referred to directly for an interpretation of the influence of a particular variable and measure of interest. The discussion of the results in the text will concentrate only on the major underlying processes.

Two underlying processes are present which are worth discussing individually. One is the effect on reliability of vehicle bunching; bunching is affected by frequency (FREQ), the standard deviation of link travel time

Table 3-6. Regression equations of reliability measures.

DEV		AGDEV		DEV/H		TFDEV/H	
VAR	COEF	VAR	COEF	VAR	COEF	VAR	COEF
* * * * * * * * * * * * MAIN EFFECTS ONLY * * * * * * * * * * * *							
STDDEV	136.83	STDDEV	344.52	FREQ	.4138	FREQ	.1717
FREQ	-40.67	FREQ	-287.40	STDDEV	.2309	STDDEV	.1334
R^2 = .706		R^2 = .770		FORM	.0337	FORM	.0359
				Constant	-.1581	DENS	-.0299
				R^2 = .937		Constant	.1890
						R^2 = .781	
* * * * * * * * * * * * MAIN AND INTERACTION TERMS * * * * * * * * * * * *							
STDDEV	306.27	STDDEV	811.39	FREQ-STDDEV	.1715	FREQ	.1717
FREQ-STDDEV	-215.27	FREQ-STDDEV	-610.58	FREQ	.3254	STDDEV	.1334
DEM-STDDEV	-143.44	DEM-STDDEV	-324.28	STDDEV	.3565	FORM	.0359
FREQ-DEM	133.60	FREQ-DEM	278.55	FREQ-DEM	.2446	DENS	-.0299
FORM-STDDEV	26.44	FORM-DENS	96.64	DEM-STDDEV	-.2126	Constant	.1890
Constant	28.53	FREQ	-575.11	FORM-DENS	.0425	R^2 = .781	
R^2 = .942		FREQSQ	452.84	FREQSQ	-.1197		
		Constant	418.21	STDDEVSQ	-.1051		
		R^2 = .960		Constant	-.1575		
				R^2 = .989			

Note:

FREQ = Frequency
STDDEV = Standard deviation of link travel time
DEM = Demand
DENS = Density
FORM = Form
FREQSQ = Frequency*Frequency
STDDEVSQ = STDDEV*STDDEV

Table 3-7. Significant variables for each measure

	DEV	AGDEV	DEV/H	TFDEV/H
FREQ		6(-)	2	1
STDDEV	1	1	3	2
FORM				3
DENS				4(-)
FREQSQ		7	7(-)	
STDDEVSQ			8(-)	
FREQ-STDDEV	2(-)	2(-)	1	
FREQ-DEM	4	4	4	
DEM-STDDEV	3(-)	3(-)	5(-)	
FORM-DENS		5	6	
FORM-STDDEV	5			

Note: the numbers indicate the order in which variables enter the regression equations. The sign in parentheses indicates a negative sign on the coefficient.

(STDDEV), and the level of user demand (DEM). The other is the effect of the network shape parameters - form and density - on the level of reliability experienced. The shape parameters contribute to reliability primarily through their influence on the transfer process between vehicles on connecting routes.

3.3.1 Vehicle Bunching

The tendency of vehicles to bunch when traveling along a bus route, and the effects of the factors which interact to bring about bunching, are by far the stronger of the two processes affecting service reliability. Research on the dynamic instability of spacing between successive vehicles suggests that higher frequency services should be more susceptible to the tendency for vehicles to group together in their travel along routes; see Osuna and Newell (1972). In bunching, the uneven spacing of vehicles is aggravated by the uneven service delay resulting from more passenger arrivals occurring during longer service intervals, and relatively few passengers arriving during the shorter intervals. The instability refers to the tendency for long intervals to get longer and short intervals shorter.

The influence of frequency on bunching suggests the coefficient of FREQ should be positive in the regression estimates of the reliability measures; a positive coefficient means high frequency service is more unreliable than low frequency service. However, it can be seen in Table 3-6 that as an independent main effect, the coefficient of frequency is positive only for the two scaled reliability measures; it takes on negative values for the reliability measures DEV and AGDEV.

A possible explanation for this result involves the interaction with the standard deviation of link travel times. As seen in Figure 3-2, when STDDEV is low, higher frequency does in fact lead to higher values of DEV, the deviation in vehicle arrival times. However, as link deviations become larger, higher frequency leads to smaller deviations.

This behavior is probably traceable to the effects of vehicle bunching. As the deviations in link travel time increase, the higher frequency services will be the first to undergo bunching. When vehicles form clusters the collective platoons of buses will be less easily influenced by individual deviations in travel times than the buses were singly. The reason bunching under high link deviations helps to constrain the magnitude of the deviations from schedule is because clustering will occur sooner on high frequency routes. Vehicles on low frequency routes have to get further off schedule before clustering begins to limit the effects of the standard deviation of link travel times. However, as Figure 3-3 shows for the scaled measures,

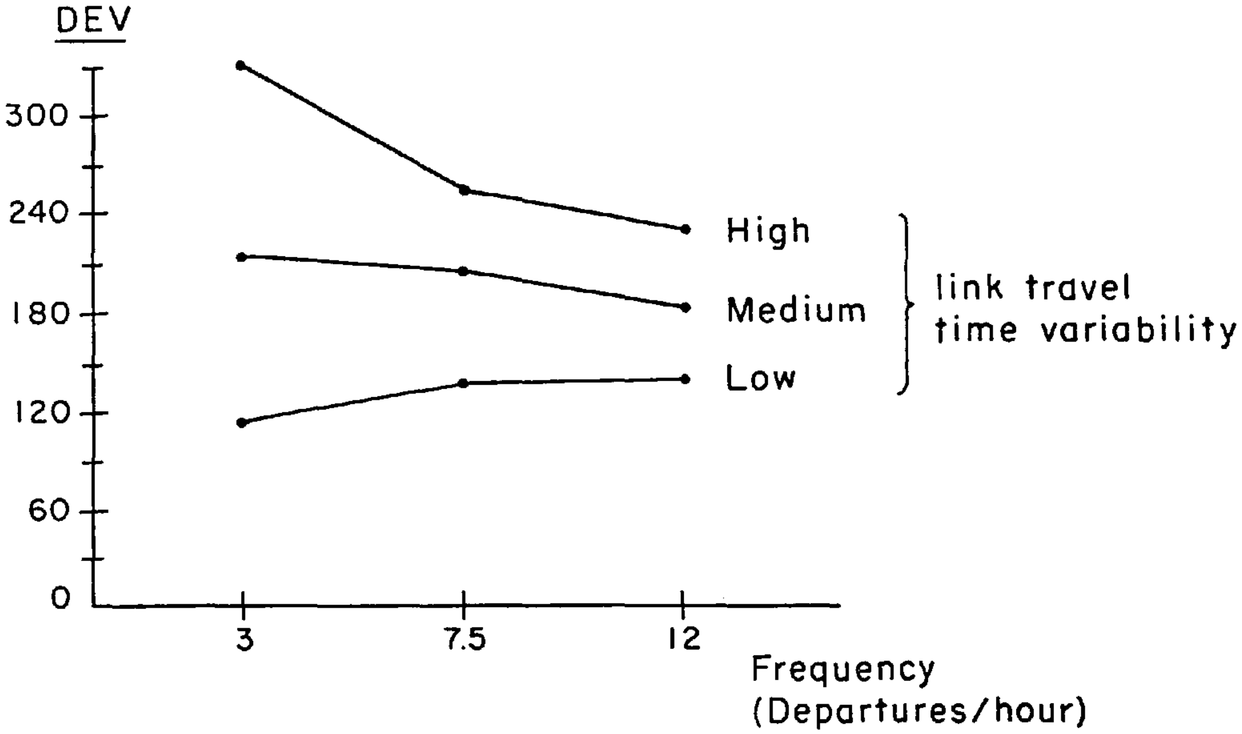


Figure 3-2. Response of arrival time deviation to frequency.

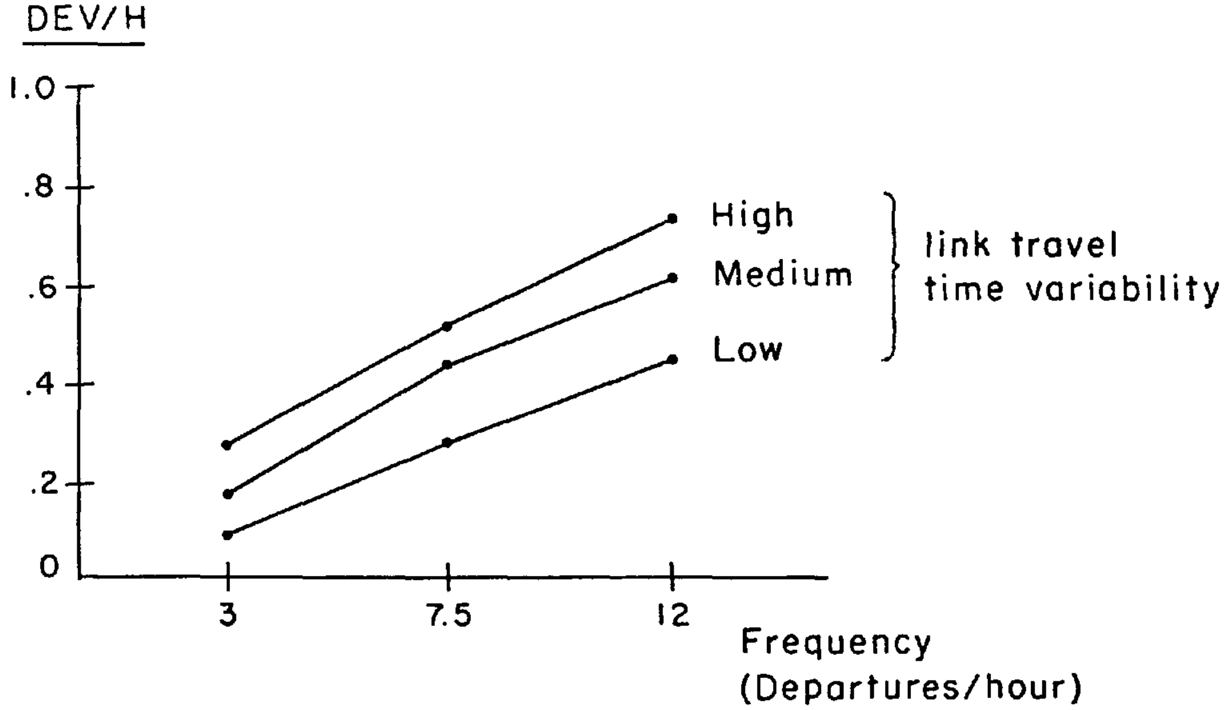


Figure 3-3. Response of scaled deviation of arrival time to frequency.

DEV/H and TFDEV/H, the extent of schedule disruption still increases relative to the spacing between vehicles.

The level of user demand is the third component in the process affecting the bunching of vehicles in their service along the transit routes. As was mentioned before, the instability which promotes bunching is due to the greater number of waiting passengers when a slow bus arrives at a station. The greater delay to board these passengers results in slow buses getting further behind schedule, and in early buses getting ahead. The relative impact of passenger service delay will be the greatest when the spacing between vehicles is small. The interaction between demand and frequency should, therefore, be positive. As is shown in Table 3-7, the interaction, FREQ-DEM, is significant for three of the network measures of reliability.

Only TFDEV/H, the scaled measure of the deviation in transfer times, is relatively unaffected by the level of user demand. This may be due to the fact that TFDEV/H depends on the order in which vehicles from connecting routes arrive at a station, and not just the deviation in arrival times.

3.3.2 Effect of Form and Density on Transfers

Network shape parameters contribute to uncertainty of travel primarily through their influence on transfers. Two points which must be considered are the number of transfers made, and the uncertainty associated with each transfer.

A recent paper by Turnquist and Bowman (1977) expressed the average number of transfers as the product of two terms: the probability of having to transfer, and the expected number of transfers given that a transfer must be made.

$$E(T) = P(T \geq 1) * E(T \mid T \geq 1) \qquad (3\text{-}1)$$

where T = number of transfers per passenger trip.

These parameters relate to the density of the route structure. Generally speaking, higher densities tend to increase the probability that a transfer is necessary, but decrease the conditional expectation of the number of transfers required. Table 3-8 summarizes the number of transfers required in each of the four density-form combinations tested. Due to the coarseness of zone spacing in the experimental networks, the conditional expected number of transfers in the radial networks increases with density. This may be due to the addition of a circumferential route which creates a number of zones not having direct access to the central transfer point.

The uncertainty associated with an individual transfer event depends on the amount of overlap provided in the scheduled arrival times of the two routes. To minimize planned transfer delays, arrival times of vehicles on different routes can be coordinated to some extent. The network form becomes important at this point because of the greater coordination of route connections possible in radial networks, as compared to grid networks. Tabe- 3-9 summarizes the uncertainty in transfer times for the four network shapes tested.

Table 3-9 reveals that radials require 5.3% fewer transfers than grid networks, but that the uncertainty in the length of a transfer delay is 12.7% higher. The combined effect is an estimated 6.7% greater uncertainty in the amount of delay to users from transferring in radial networks. Similarly, high density networks have an estimated 13.4% less uncertainty due to transfers than do low density networks.

While TFDEV/H, the scaled deviation in transfer delay, is the system measure least affected by the vehicle bunching process, it has the strongest response to the effects of form and density. The more direct relationship

Table 3-8. Transfer probabilities

	$P(T \geq 1)$	$E(T \mid T \geq 1)$	$E(T)$
Grid - avg.	.7259	1.174	.8512
low dens.	.7114	1.229	.8744
high dens.	.7403	1.118	.8280
Radial - avg.	.7027	1.147	.8062
low dens.	.7001	1.060	.7424
high dens.	.7052	1.234	.8700

Table 3-9. Estimated deviation in transfer times (seconds)

	E(TFDEV)	E(TFDEV/H)	E(T)
Grid - avg.	223	.3974	.8512
low dens.	240	.4189	.8744
high dens.	225	.3758	.8280
Radial - avg.	262	.4344	.8062
low dens.	263	.4408	.7424
high dens.	261	.4280	.8700

of TFDEV/H to these parameters is evident in Table 3-7, by the presence of both the FORM and DENS main effects as significant variables in the regression equation which estimates the value of this reliability measure. The interaction, FORM-DENS, is not present, but due to the small sample size on which the regression is based this point is not significant.

The remaining three system measures - DEV, DEV/H, and AGDEV - are less directly influenced by the transfer process and hence are less sensitive to the form and density of the transit network. The regression equations estimating these measures each contain one term related to the network shape. AGDEV and DEV/H are sensitive to the interaction of FORM-DENS. The final measure, DEV, is affected by the interaction term, FORM-STDDEV. This difference in interaction terms is not particularly important, because due to the small sample size, a correlation of .7621 exists between the FORM-DENS and FORM-STDDEV terms.

3.4 Comments

Understanding of the network relationships which influence the reliability of service provides insight into the types of approaches available in efforts to favorably influence service. The user level of service, as affected by vehicle bunching, may be improved either by preventing bunches from forming, or by somehow breaking up clusters of buses after they form. The more effective technique may depend on the sensitivity of service to bunching. Short headway services in particular, may provide a practical application for controls which depend on the use of intervention to maintain uniform vehicle spacing.

The importance of transferring to the users' level of service, and the sensitivity it displays to the scheduled overlap in vehicle arrivel times, reveals the problem of the tradeoff between the length of the scheduled wait time and the risk of missing the intended connecting bus. Where arrivals can be scheduled to coincide, the application of controls to the operation of transit service has the potential to permit more closely scheduled arrivals on connecting services while maintaining a reasonable assurance that the intended connection will be successful. Such closely timed services are most common in radially oriented networks. It is possible to schedule grid networks so as to operate in this way, but successful implementation is much more difficult.

The importance of reliability to the users' level of service, and the presence of network relationships offering the potential for improvement through active intervention in transit operations, argues for the need to investigate the effectiveness of different types of control strategies and establish the types of circumstances under which different controls might effectively be employed.

4. The Effects of Real-Time Control Strategies

Several real-time approaches to correcting service disruptions have been discussed in the literature. A recent paper by Blume and Turnquist (1978) provides a review and catalog of these approaches to control departures from scheduled service. Several examples of real-time controls are listed in Table 4-1. The first three strategies can be used routinely in making adjustments to service. Strategies four and five are appropriate when service has been seriously disrupted, and drastic measures are required.

The study of the effectiveness of real-time controls in this section will be limited to the first two possibilities, holding buses or skipping stations. The third strategy refers to letting buses pass one another, a policy already in effect for all runs of the simulation model. The last two controls were omitted from the analysis because they are less practical as regular control measures. Labor agreements often provide for driver rest time at route endpoints. Therefore vehicles cannot always be turned back early. The use of stand-by vehicles is a difficult question requiring determination of the number of vehicles to withhold from normal service, and careful location of these extra vehicles. Its usefulness lies more in substituting for buses suffering mechanical failure, a possibility not treated by the simulation model.

Table 4-1. Real time control strategies.

1. Skip stops (express service).
2. Hold vehicles at control points.
3. Place the following bus ahead.
4. Turn vehicles back short of the route end.
5. Send in a stand-by vehicle.

4.1 Development of Holding Parameters

The decision to hold a bus involves finding the answers to two questions;

1. Under what conditions should a bus be held?
2. How long should the hold be applied?

The answers to these questions will, of course, depend on one another. There are two ways in which an early service can be held; as a fraction of the amount the vehicle is early, or as a predetermined fixed delay.

Barnett (1974) formulated a mathematical optimization approach to analyze the question of vehicle holding. His approach assumed random arrival of passengers independent of vehicle arrival times and was based on the use of a discrete approximation to the distribution of vehicle arrivals. The probabilities representing the discrete arrival times were chosen to retain the mean and variance of the observed arrivals, as well as the covariance between successive vehicle arrivals.

The hold decision in Barnett's model was derived so as to minimize the total passenger delay to passengers arriving at all stops on the route. Total delay was expressed as:

$$D = \gamma E(d) + (1-\gamma)*E(w) \tag{4-1}$$

where D = total delay plus wait time

$E(d)$ = expected delay for passengers already on board

$E(w)$ = expected waiting time for passengers arriving at the control stop

γ = a weighting constant to reflect the relative volumes of arriving to delayed passengers.

The expected delay to passengers already on the bus is simply the average amount a bus will be held under a given strategy. The assumption that passengers arrive independently of the schedule was retained in expressing the expected wait time of passengers arriving at the control point. The expected wait time is then represented by equation (4-2).

$$E(W) = \frac{E(h^2)}{2E(h)} \qquad (4\text{-}2)$$

At this point our derivation departs from Barnett's approach. Instead of using a discrete probability model to minimize the expression, the logic is preserved but applied to a continuous distribution of vehicle arrivel times.

Assuming the average headway will remain unchanged, E(h) will equal the scheduled service interval or headway, H. Selection of a holding time is intended to reduce the second moment of the headway, $E(h^2)$, subject to the restriction that the expected delay term, E(d), not grow so large as to negate the benefits from more regular intervals. For simplicity in notation the discussion will focus temporarily on the $E(h^2)$ term.

If a vehicle is held some unknown time, x, the length of the held service interval can be written as:

$$h = H + L_1 - L_2 + x \qquad (4\text{-}3)$$

where h = interval between vehicles one and two

H = the scheduled interval

L_i = the lateness (earliness) of vehicle i $(i=1,2)$

x = delay due to holding of vehicle two.

The second moment of the headway is then:

$$E(h^2) = E[H^2+L_1{}^2+L_2{}^2+x^2+2x(H+L_1-L_2)+2H(L_1-L_2)-2L_1L_2] \qquad (4\text{-}4)$$

If, on average, arrivals are on schedule, $E(L_i) = 0$. This implies that $E(L_i^2)$ is the variance of the i′th vehicle arrival time. If we further assume that successive vehicle arrivals are independent of one another, equation (4-4) can be written as:

$$E(h^2) = H^2+2Var(A_i)-2HE(L_1-L_2)+E(x^2)+2E\,[x(H+L_1-L_2)] \qquad (4\text{-}5)$$

where A_i = arrival time of i′th vehicle.

The two approaches to holding vehicles can be referred to as "percentage" and "all-or-nothing" strategies. A cursory investigation was given to the mathematics involved in arriving at the optimal values for implementation of these approaches. The intent was not to derive a rigorous mathematical optimization of the hold lengths, but to make necessary simplifying assumptions to permit the derivation of an expression to provide an approximate magnitude of hold delay appropriate for a given set of circumstances.

The percentage hold, x, is represented by $-P(L_2-L_1)$. (L_2-L_1) is the amount by which the preceding interval differs from the scheduled headway, and P is the fraction of that difference the bus is to be held. The fraction of the earliness which the bus is to be held is determined by the interval length, the standard deviation of vehicle arrivals at a given stop, and the weighting factor, γ, which reflects the delay to passengers already on the bus. The mathematics are presented in appendix B. The result is as follows:

$$P = 1.0 - \left(\frac{\gamma}{1-\gamma}\right)\sqrt{\frac{2}{\pi}}\,\frac{H}{\sigma} \qquad (4\text{-}6)$$

Determination of the value for an all-or-nothing approach is more complex mathematically. The value of x, the length of time to be held, cannot be solved directly, but equation 4-7 can be solved iteratively for x.

$$0 = \Phi(\frac{-x}{\sigma}) [\gamma + \frac{x(1-\gamma)}{H}]$$

$$+ \; N (\frac{-x}{\sigma})[\gamma x + \frac{1-\gamma}{H} \; (\frac{x}{2} + 2x^2 - 4\sigma^2)] \qquad (4-7)$$

Φ = the cumulative normal distribution

N = the value of the normal probability density function

H = scheduled headway

σ = standard deviation in vehicle arrivel times

γ = the weighting factor for delayed passengers.

Sample values of these two approaches were computed for illustravive test situations. As can be seen in Table 4-2, the all-or-nothing hold is applied over a wider range of values of γ, the fraction of users being delayed by the holding strategy. The length of time a bus is to be held decreases as γ, the fraction of delayed riders, increases because the total user delay from holding outweighs the potential user benefits. The all-or-nothing hold is only applied when the vehicle departs from schedule by an amount greater than the length of the fixed hold. The greater deviation required for it to be implemented results in less frequent holding and a lower expected delay to riders.

4.2 Controls Tested

Four variations of real-time control strategies were tested; three based on holding vehicles at stations, and the fourth representing express or skip-stop service. All holding strategies were of the percentage form, and delayed the vehicle 50% of the amount the bus initially was early. Table 4-3 summarizes the four strategies.

Table 4-2. Sample values for percentage and all-or-nothing holding strategies.

Headway = 10 min. σ = 1.0 min.		Percent	All-or-nothing
	$\gamma = 0.0$	100%	1.16 min.
	$\gamma = 0.1$	11.4%	.81 min.
	$\gamma = 0.11$	0.0%	.39 min.
Headway = 10 min. σ = 2.0			
	$\gamma = 0.0$	100%	2.58 min.
	$\gamma = 0.1$	56%	2.28 min.
	$\gamma = 0.2$	0%	1.91 min.

Table 4-3. Real time controls tested.

Name	Description
EXPRESS	An express or skip-stop strategy. When buses begin to cluster together the leading bus skips stations to try to pull ahead and prevent bunching.
CHEKPT	A checkpoint holding strategy where the hold decision is based on whether or not the vehicle is ahead of schedule.
HEADWY	A hold decision based solely on the length of the service interval to the preceding bus.
PREFOL	A hold decision based on the lengths of the service intervals to both the preceding and following buses.

The EXPRESS (skip-stop) decision is assumed to be made on essentially line-of-sight information. Not until two buses begin to close in on one another is one of them sent ahead. Holding strategies make use of different types of information, including both schedule and interval data. The least information is used by the CHEKPT strategy, but the type of information and the way in which it is used is different from the other two holding strategies. Since the PREFOL strategy uses more information than the HEADWY strategy, it may be expected to be more effective.

The analysis of the effectiveness of control strategies was performed at an exploratory level. The purpose was to get an indication of the potential these controls offer, and the conditions under which they may be applicable.

Due to the number of parameters involved, a full set of experiments would have been much too expensive. Instead, six scenarios were selected, with four controls and a base run being performed on each. The six scenarios were chosen to represent as broad a spectrum of network configurations as possible. Table 4-4 indicates the scenarios tested.

Table 4-4. Scenarios for testing control strategies.

	FREQ	STDDEV	DEM	DENS
Grid	0	2	0	0
	1	2	2	0
	2	1	0	0
Radial	0	1	2	0
	1	1	0	0
	2	2	2	0

Note: 0 = low
1 = medium
2 = high

4.3 Analysis of Results

Any intervention with the normal operation of a transit service will involve a redistribution of user costs. Improving the reliability of service benefits those passengers yet to board a bus, but costs, in terms of delay, accrue to riders held at stops or passed up by express buses. It is this tradeoff that is reflected in the value of γ, the weighting factor in Barnett's total delay equation.

Analysis of controls involves comparison of their effects in three areas: measures of network reliability, average passenger wait and travel times, and delays to passengers due to the controls. Tables 4-5, 4-6, and 4-7 summarize the results of the control strategies with respect to these three types of impacts.

4.3.1 Network Measures of Reliability

The more information available on which to base a decision, the better the results are expected to be. This suggests that the best performances of the holding strategies should be PREFOL, HEADWAY, and CHEKPT in descending order. However, the results of the PREFOL and CHEKPT strategies are extremely close, each performing better in two of the four network measures. This may be related to the different way in which the information is used by the CHEKPT strategy. The HEADWAY strategy is not only weaker than the other holding controls, but in three of the four measures its results are worse than the uncontrolled network. However, it does provide the best results of all the controls with respect to reduction in TFDEV/H, the standard deviation of the scaled measure of transfer times. The successful connection of services in a passenger's trip depends on the order in which vehicles from connecting routes arrive at the transfer point. Since the HEADWAY control maintains time intervals between arrivals, it may provide the steadiest

Table 4-5. Effect of controls on network measures of reliability

	Scenario					
	G0020	G0121	G0210	R0011	R0110	R0221
DEV (in seconds)						
BASE	372	250	161	301	224	208
CHEKPT	416	235	128	331	181	199
HEADWAY	468	236	153	308	219	235
PREFOL	314	244	119	280	246	213
EXPRESS	360	244	140	296	195	204
AGDEV (in seconds)						
BASE	1184	716	476	875	613	541
CHEKPT	1172	665	365	957	522	530
HEADWAY	1299	669	414	909	575	603
PREFOL	1033	737	358	840	648	555
EXPRESS	1110	686	392	861	549	517
TFDEV/H (in fraction of a headway)						
BASE	.4083	.5313	.5733	.3392	.4583	.5876
CHEKPT	.3158	.4833	.4100	.4442	.4208	.6267
HEADWAY	.3358	.4854	.4067	.2708	.3833	.6233
PREFOL	.3742	.6354	.4467	.2975	.4333	.6100
EXPRESS	.3625	.4938	.4267	.3317	.4417	.5200
DEV/H (in fraction of a headway)						
BASE	.3100	.5208	.5367	.2508	.4667	.6933
CHEKPT	.3467	.4896	.4267	.2758	.3854	.6633
HEADWAY	.3900	.4917	.5100	.2567	.4563	.7833
PREFOL	.2617	.5083	.3967	.2333	.5125	.7100
EXPRESS	.3000	.5083	.4667	.2367	.4063	.6800

Note:

The letter, G or R, refers to either grid or radial form. The four numbers indicate the levels taken by the four quantitative variables. They indicate in order:

Density
Frequency
Travel time deviation
Demand

Table 4-6. Effect of controls on user level of service (seconds)

	G0020	G0121	G0210	R0011	R0110	R0221
<u>Average Wait time</u>						
BASE	666	254	156	594	253	163
CHEKPT	671	246	144	605	239	154
HEADWAY	628	236	152	587	246	165
PREFOL	628	257	145	580	266	161
EXPRESS	661	254	153	592	275	165
<u>In-Vehicle Travel Time</u>						
BASE	1204	1223	1051	1154	1031	1093
CHEKPT	1241	1198	1074	1135	1051	1095
HEADWAY	1283	1196	1099	1152	1107	1126
PREFOL	1202	1196	1090	1128	1099	1094
EXPRESS	1235	1203	1218	1183	993	1132
<u>Transfer Time</u>						
BASE	708	288	184	764	293	190
CHEKPT	718	301	174	694	300	214
HEADWAY	675	274	165	758	249	183
PREFOL	665	293	158	734	310	223
EXPRESS	664	344	257	709	361	246

Table 4-7. Delay per passenger due to controls. (in seconds)

	G0020	G0121	G0210	R0011	R0110	R0221
CHEKPT	4.5	1.9	12.0	1.9	2.3	1.8
HEADWAY	53.0	45.7	40.9	27.2	86.3	49.2
PREFOL	4.0	41.7	33.5	6.7	43.5	26.6
EXPRESS	8.5	37.9	14.2	1.7	8.0	57.7

Note:
The passenger delays listed here are already represented in the wait, travel, and transfer times in Table 4-6.

performance in terms of the sequence of vehicle arrivals on connecting routes. The CHEKPT strategy was not effective at controlling the deviation in transfer times, but vehicles were so frequently late that the hold was seldom applied.

The poor performance of the headway control, HEADWAY, can be attributed to the propagation of lateness to following buses. If the extreme case is considered where intervals are always required to be at least 100% of their scheduled length, a single late bus arrival will cause every bus thereafter to leave control point late as well. Enforcing a minimum interval length truncates the short side of the distribution of headways. Without short headways to balance out the longer ones, average interval length must increase and vehicles must slow down.

The headway control which also takes into consideration the following bus, PREFOL, contributes less to the propagation of slow travel. An early trailing bus will partially offset the effect of a late preceding bus, cancelling the tendency to hold the control bus.

The skip-stop strategy, EXPRESS, does not always give the greatest improvement in service, but it appears to be the most consistent. A skip-stop strategy is most suitable for implementation on high frequency services (see Table 4-8). Bunching must occur before running vehicles express can be of help, which limits its usefulness on long interval services. The average number of times express service was enacted in the simulation experiments is 5, 26 and 51 times for low, medium, and high frequency service respectively.

Table 4-8. Percentage improvement in reliability measures due to use of skip-stop strategy at different service frequencies.

	20 min.	8 min.	5 min.
DEV	2.5%	7.4%	6.8%
AGDEV	4.3%	7.3%	10.6%
DEV/H	7.1%	5.5%	18.5%
TFDEV/H	2.5%	7.4%	6.8%

4.3.2 Passenger Wait Times and Control Delays

The passenger level of service measures presented in Table 4-6 show that the average total trip time changed by only a few seconds out of an average trip time of over one-half hour when the control strategies were enacted. The wait and transfer times are slightly lower for the holding strategies, but in-vehicle travel time is higher. The HEADWAY strategy has the greatest tendency to pass delay on to later vehicle trips.

Another potential benefit from regular service intervals, in addition to the effect on wait time, is more uniform vehicle loadings. If a bus route is operating near its capacity, uneven loadings can cause vehicles to overfill, causing passengers to be turned away. At lower volumes, uniform vehicle loadings have the benefit of minimizing the number of passengers who must stand. Standing or being passed up by a full bus are very annoying to users, making it valuable to consider these aspects of service as well.

The number of passengers turned away from a full bus is higher for all control runs than for the base case. The total number of passengers turned away from the scenarios with high demand are:

RUN TYPE	PASS. TURNED AWAY
BASE	602
CHEKPT	650
HEADWAY	873
PREFOL	889
EXPRESS	1237

Holding strategies tend to slow down the average vehicle speed. Lower speeds mean fewer round trips per day and less productivity, increasing the occurrences where vehicles fill to capacity and turn passengers away.

The express strategy is the worst offender in this respect, turning away 637 more passengers than the base case. Sending a vehicle ahead non-stop causes the local passengers from the express bus to have to transfer to the trailing bus, and also leaves to the trailing bus the passenger loads at the next several stops. Under high demand, when both buses are heavily loaded, the following bus will frequently be unable to absorb the passengers transferred to it from the express bus and still have sufficient capacity to pick up the passengers along the route. In operation of a real system a bus may not be sent express under such conditions, and a check of this type should be included in the model.

4.4 Conclusions

No one strategy will work best under all circumstances. Different situations are more easily handled by different controls; some controls, however, may be a poor choice in any situation. The strategies based on headways have problems in that they transmit vehicle delay to later arrivals. One must be careful not to hold too often or too long, or vehicle delay propagates to later arrivals, and level of service deteriorates. These results tend to confirm earlier findings reported by Koffman (1978). His experiments also indicated that an interval-based holding strategy can cause vehicle delays to be passed along to affect the service of later arriving buses. The fact that the checkpoint-based holding strategy does not share this behavior leaves hope that the "propagation" phenomenon is not a general characteristic of holding strategies; rather, it is more specifically related to the particular type of interval length enforcing decision tested. Future usefulness of headway interval controls depends on formulating a headway decision which can correct for irregular interval lengths in a sufficiently selective manner that excessive delay is not transmitted to the following buses.

The checkpoint strategy is the simplest control to apply and may also have valuable applications, especially in coordinating transfers at major route connections. A major advantage which ties in with the non-random distribution of passenger arrivals is the fact that it is the only strategy which strives to maintain the schedule. The purpose of the others is to maintain regular vehicle spacing along the route. If passenger arrivals coincide well with the scheduled vehicle arrival times, hold and skip-stop strategies will not be desirable. The benefits of improved regularity in headway spacing will probably be greatly outweighed by higher wait times from controls which alter the arrival times of vehicles.

For short headways, or other circumstances where passenger arrivals are not expected to be coordinated with the scheduled vehicle arrivals, the EXPRESS and PREFOL strategies have potential. The problem with the PREFOL control is cost. It requires information not easily obtained without some sort of central monitoring system. The necessary capital expenditure makes this more suitable for major cities which can support the equipment necessary.

Running buses express to break up vehicle bunching has two desirable features. First, it is extremely consistent. The benefits are not always as great as with other control strategies, but in the experiments run to date, it has never caused service to deteriorate. Second, the cost of implementing it can be negligible. The information required can be as simple as line-of-sight; drivers taking action only when they see the trailing bus begin to catch up. Radios could also be used to let drivers check the relative vehicle loads before passing groups of waiting passengers. The need for buses to cluster before express service can be of use means the express service is best suited to short headways. The shortest tested was five minutes, but in many major cities there are routes with scheduled headways of two minutes or less. The benefits of a skip-stop control should therefore be tested on shorter services to see whether the benefits continue to increase. Finally, skipping stations should only be done to break up clusters of buses, not to maintain vehicle spacing as tested by Koffman (1978). In Koffman's experiments, the increased passenger wait times more than offset any improvement in line-haul time. Skipping passengers without a bus following close behind is not only unlikely to be acceptable to users or operators, but the user delay will be so large that the actual level of service will deteriorate.

The EXPRESS strategy also has the drawback of forcing some local passengers to leave the bus they are on, and board a trailing bus, because their destinations are stops to be skipped. This may be a significant problem in implementing such a strategy in many cases. Situations in which the majority of passengers are bound for a single stop (e.g., a rapid transit station or the CBD) might be most amenable to such a strategy.

5. Summary and Conclusions

The major tasks undertaken in this project have been: (1) the extension and refinement of a digital simulation model of operations in bus transit networks; (2) preliminary validation of that model against observed data; (3) identification of possible strategies for control of service reliability; (4) tests of the basic effects of network structure on service reliability; and (5) preliminary tests of selected real-time control strategies.

The basic tool for investigation of the effectiveness of various alternative actions for control of service reliability is a network-simulation model. A prototype model was developed under a previous grant from the National Science Foundation. During the first year of the current project, this model has been extended and revised. The model is now capable of analyzing networks of up to 32 routes, with a maximum of 70 zones serving as origins and destinations of trips. Larger networks are possible within the same logical structure; only the amount of central memory available to the program would require changes.

A very important aspect of the simulation is the ability to have different networks, service characteristics and operating policies easily input to the model. This is accomplished through the use of a simulation system, developed at Northwestern University, called Program for Interactive Multiple Process Simulation (PIMP). The PIMP software comprises a set of FORTRAN subroutines for event scheduling, dynamic data management and statistics gathering. The characteristics which make PIMP uniquely appropriate for the current model are the interactive multiple process structure combined with the flexibility of data management through subscripted queues.

The basic structure of the bus simulation model hinges on the use of subscripted queues which share common storage. A path matrix contains the zone numbers through which a passenger moves in completing his trip, and determines a sequence of queues in which the passenger is placed. The interaction of this path matrix with the movement of buses along routes permits great flexibility in the specification of the route network and service characteristics. As passengers move through the system, a data file of their movements, travel times and delays is constructed. This file provides the source of information for all statistics, which are computed by a post-processor. More complete documentation of the simulation model is provided in the User's Manual by Bowman and Turnquist (1979).

An integral part of the model extension and refinement has been the development of an improved understanding of the effects of service reliability on passenger wait times. The simulation model described in this report can be used to predict changes in the level of reliability, as measured in any of several ways; but in order to evaluate the impacts of service changes, it is vital to understand more clearly the relationship of passenger wait time to reliability.

The commonly used model which asserts that average wait is one-half the scheduled headway is true only under very special circumstances. This model requires that all passengers arrive at random, and that headways be perfectly regular. Several improvements have been suggested to the simplistic model, but generally have not treated non-random arrivals of passengers.

During the past year, an improved model of the relationship between passenger wait times and the service frequency and reliability of buses has been developed. This model is described in detail in section 2.2. The model is based on the choice process underlying passengers' decisions as to the time of arrival at bus stops. This decision process is dependent on both service frequency and the degree of reliability in vehicle arrival times.

Another important part of the simulation model refinement is the representation of link travel times for buses. Considerable empirical data were gathered on these travel times for routes in Evanston and Chicago, Illinois, and Cincinnati, Ohio. Statistical analysis of this data confirms the appropriateness of gamma distributions as the underlying probability model for vehicle delays. Overall link travel time can be predicted quite well using a shifted gamma distribution, with the shift accounting for free-flow, non-stop travel time. Furthermore, a procedure has been developed for prediction of the parameters of the distribution for any link, based on the number of boarding and alighting passengers, and the number of signalized intersections through which the bus must pass. This predictive travel time model is useful, with or without the simulation, by an operator interested in more efficiently scheduling bus routes.

Given these refinements to the simulation model, the second major project task was to begin to validate the model against observed data. The bus network of Evanston, Illinois, was selected for this purpose, primarily because substantial data exist for this system, and because the proximity of the system to Northwestern eased the collection of additional data. This preliminary validation effort indicated high correlation of simulated results with observed system performance. As a result, a decision was made to proceed with initial experiments on alternative control strategies.

Approximately twenty different control actions with potential for improving service reliability have been identified. Some of these actions, such as signal pre-emption, attempt to reduce the number and size of delays a bus encounters. Others, such as skipping stops, attempt to minimize the consequences of those delays.

The control strategies may be divided into two basic groups, planning and real-time. In general, the distinction is that planning strategies involve changes in operations of a persistent nature. For example, changes in route structure have substantial long-term effects on the character of operations, and the decision to make such a change is the result of the service-planning process. On the other hand, real-time control measures are designed to act quickly to remedy specific problems. These actions have immediate effects, but seldom exert any influence on the general nature of operations over a longer time period. Additional discussion of both sets of control strategies in contained in the working paper by Blume and Turnquist (1978).

Simulation experiments to test the effects of various types of potential control strategies have been divided into two sets, corresponding to this same classification. The first set of simulation experiments was designed to test the effects of basic network characteristics on several measures of service reliability. Understanding of these basic relationships is an important precursor to actual design and testing of specific strategies.

Network effects on measures of reliability were found to depend heavily on the frequency and standard deviation of link travel times. Dynamic instability of vehicles was found to be a function of the interaction of frequency and link travel time deviations with the level of user demand. While bunching adds to deviations in vehicle arrival times for relatively reliable service, as schedule adherence deteriorates vehicles group into clusters and inhibit further departure from schedule.

Characteristics of network shape are most important through their influence on passenger transfers. The density affects the number of transfers made, and the form affects the uncertainty associated with each transfer.

Finally, a brief investigation of real-time control strategies has indicated a better decision process is needed if headway-based controls are to be of any practical use. The checkpoint control appears promising. It is simple and inexpensive to implement, avoids the propagation of delay associated with headway controls, and works to maintain the schedule, which is of critical importance if passengers time their arrivals to match the vehicle timetable. The only shortcoming is the need for an accurate schedule. If bus arrivals at a checkpoint are consistently either early or late, the control can either cause excessive delay or never be enacted.

The express strategy is also a promising form of control. It is well suited for short headway services with a high incidence of vehicle bunching, particularly in situations where most passengers alight at one (or a few) stop(s) on the route.

5.1 Practical Implications of the Research

The first result which is of importance to a transit operator is the significance of the improved model of passenger arrivals. The substantial effect of reliability on the average passenger wait time shows reliability to be an aspect of system performance which should be of concern to the operator.

Given the importance of reliability in the operation of a transit system, this research has several practical implications for providing a higher level of service to transit users. First of all, as a result of the influence of vehicle bunching, when relatively high deviations exist in link travel time not much is to be gained by increasing the frequency of service (assuming it is not required to meet the level of demand). Rather than ending up running groups of buses, the resources may be better used to operate a denser route structure thus reducing the access portion of the users' total travel time.

The influence of transfers on the level of service points out the need to pay special attention to the on-time arrival of vehicles at major transfer stations. This is especially true for radially oriented network structures. As a rule, providing excess slack time in the route schedule is to be avoided, due to its effect on slowing down travel time and vehicle productivity. However, where a large number of passenger transfers can be aided by creating enough slack time to assure successful connections, allowing a short delay may be highly beneficial.

The operator also has the option of exercising active or real-time control over the operation of service. The two controls thus far showing potential for practical applications are the EXPRESS and CHEKPT strategies. The EXPRESS control strategy is used to counter the effects of vehicle bunching. As a result, the EXPRESS strategy is best suited for application to high frequency routes.

The CHEKPT control strategy is important as being the only control tested which works to restore vehicles to scheduled operation. Considering the fact that the greatest portion of reduced passenger wait time comes about due to the users' attempts to coordinate their arrivals with those of the buses, the important contribution of a reliability control strategy would seem to be in maintaining vehicle arrival times as scheduled, and not in altering service to even out the lengths of service intervals. The headway-based holding strategies, in fact, will cause one vehicle's deviation from schedule to result in later vehicles also arriving off schedule, potentially worsening the level of service as perceived by the user.

5.2 Plans for the Second Year

One way to classify the sources of unreliability in transit service is to distinguish those elements which give rise to variability in link travel time from those which give rise to variability in dwell time at stops. The results of the first year's experimentation indicate the importance of a thorough investigation of strategies designed to reduce both types of variability.

Strategies to reduce variability in dwell time at stops include changes in the number and location of stops, reductions in passenger boarding and alighting time, etc. The effects of such strategies can be assessed quite adequately with the existing model. Investigation of strategies for reducing link travel time variability, however, requires further enhancement and extension of the present model.

There are several strategies for reducing travel time variability, including bus priority at traffic signals, bus lanes and busways, and zone scheduling. Evaluation of strategies such as priority signalization, which involve the interaction of buses with the traffic stream in which they are moving, requires a rather detailed representation of bus and traffic movement. This point has been made very effectively by Waksman and Schmieder (1978) in their review of bus route simulation models.

The simulation model developed in this project does not currently have such a detailed representation of traffic flow. It views link travel time macroscopically - i.e., as a random variable with specified distribution from which samples are drawn. While this point of view was valuable for purposes of initial experimentation with the model, it is not adequate for in-depth examination of a number of promising strategies for control of service reliability.

During the second year of this project, the simulation model will be extended to incorporate greater detail on traffic flow, and the interactions between buses and general traffic. It is not the intent, however, to model traffic at the level of each individual vehicle.

Such a level of detail is unnecessary for the purposes of this project, and is very expensive computationally. A more appropriate level of detail is that represented by the SUB model developed by Radelat (1973). Traffic is modeled as platoons, to avoid the cost of simulating individual vehicles while retaining the ability to represent important interactions among buses and the general traffic stream.

In its current form, the SUB model has four significant weaknesses for use in evaluating strategies to improve service reliability. First, its representation of passenger arrivals and movements is very simplistic. However, the simulation model developed in this project is quite detailed in this regard, so incorporation of parts of the SUB model into the current model will make use of the complementarity that exists.

Second, SUB does not represent operations in a network of routes. Again, however, that aspect of the problem will be handled by the current model. It is likely that detailed traffic representation will not be necessary on all routes in a given exercise. Only those routes on which changes are being made need to be examined in detail; the others can be represented macroscopically, as in the present form of the model. The user will have the flexibility to specify the level of detail required in each portion of the network.

The third weakness of the SUB model is in the logic for representing interactions between buses and the general traffic stream. Because the general traffic is always moved first at each period on each link, the "interaction" is always in one direction only. Traffic affects bus performance, but not vice-versa. This representation can be enriched in a straightforward manner, without abandoning the basic philosophy of modeling traffic in platoons. A basic model of this type has been developed by Unger (1977) to evaluate bus preemption of traffic signals on Archer Avenue in Chicago. The experience gained with this prototype model will be quite valuable in extending the network simulation model.

The fourth aspect of the SUB model requiring improvement is the logic for buses passing one another. This is a relatively minor modification, but one which would allow much greater flexibility in representing a variety of circumstances.

To summarize, the current network simulation model and the SUB model have complementary strengths, and extension of the current model to incorporate the more detailed traffic simulation capabilities of the SUB model is a logical and highly useful step. The resulting model will have greater capabilities than any other existing model for analysis of a variety of strategies for improving the quality of bus transit service. Once this extension is completed, a number of these strategies can be tested much more carefully than was possible during the first year of this project.

REFERENCES

Barnett, A., "On Controlling Randomness in Transit Operations," Transportation Science, Vol. 8, No. 2, May, 1974, pp. 102-116.

Berry, D.S. and D.M. Belmont, "Distribution of Vehicle Speeds and Travel Times," Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, August, 1950.

Blume, S. and M.A. Turnquist, Improving the Reliability of Bus Service: A Catalog of Possible Strategies, working paper for USDOT contract DOT-OS-80018, April, 1978.

Bly, P.H. and R.L. Jackson, "Evaluation of Gus Control Strategies by Simulation," TRRL 637, 1974.

Boardman, T.J. and W.H. Kraft, "Predicting Bus Passenger Service Time - Part II," Traffic Engineering Vol. 40, February, 1970, pp. 36-40.

Bowman, L.A., and M.A. Turnquist, Users' Manual for the Bus Network Simulation Model, report prepared for the Office of University Research, U.S. Dept. of Transportation, January, 1979.

Chapman, R.A., "Bus Reliability - Definition and Measurement," Transport Operations Research Group, University of Newcastle-upon-Tyne, working paper no. 18, March, 1976.

Choi, S.C. and R. Wette, "Maximum Likelihood Estimation of the Parameters of the Gamma Distribution and their Bias," Technometrics, Vol. 11, No. 4, November, 1969.

Connor, W.S. and S. Young, "Fractional Factorial Designs for Experiments With Factors at Two and Three Levels," National Bureau of Standards Applied Mathematics Series, 1961.

Cosslett, S., "Trip Timing Analysis: The Effect of Congestion on Scheduling Trips to Work by Auto," working paper, Dept. of Economics, University of California, Berkeley, March, 1976.

dePirey, Y.A., Simulation of a Bus Line as a Means to Analyze and Improve Bus Transit Reliability, unpublished M.S. thesis, Dept. of Civil Engineering, Northwestern University, August, 1971.

Friedman, R.R., Statistical Models of the Mean and Standard Deviation of Passenger Wait Time in Urban Bus Transit, unpublished M.S. thesis, Transportation Center, Northwestern University, June, 1976.

Golob, T.F. et al., "An Analysis of Consumer Preferences for a Public Transportation System," Transportation Research, Vol. 6, March, 1972, pp. 81-102.

Hashizume, C., Measures of Network Reliability, unpublished M.S. thesis, Northwestern University, Dept. of Civil Engineering, June, 1978.

Holroyd, E.M. and D.A. Scraggs, "Waiting Times for Buses in Central London," Traffic Engineering and Control, Vol. 8, No. 3, July, 1966, pp. 158-160.

IBM, Schedule Control and Management Information System Study, prepared for Chicago Transit Authority, November, 1974.

Jenkins, I.A., "A Comparison of Several Techniques for Simulating Bus Routes," TORG, working paper no. 14, January, 1976.

Jolliffe, J.K. and T.P. Hutchinson, "A Behavioral Explanation of the Association Between Bus and Passenger Arrivals at a Bus Stop," Transportation Science, Vol. 9, No. 4, November, 1975, pp. 248-282.

Kaminsky, F.C. and D.L. Rumpf, "Simulating Non-Stationary Poisson Processes: A comparison of alternatives including the correct approach," Simulation, Vol. 29, No. 1, July, 1977, pp. 17-20.

Koffman, D. "A Simulation Study of Alternative Real-time Bus Headway Control Strategies," Transportation Research Record 663, 1978, pp. 41-46.

Kulash, D.J., Routing and Scheduling in Public Transportation Systems, Ph.D. dissertation, Dept. of Civil Engineering, MIT Urban Systems Laboratory, January, 1971.

Martland, C.D., "Rail Trip Time Reliability, Evaluation of Performance Measures and Analysis of Trip Time Data," Studies in Railroad Operations and Economics, Vol. 2, MIT, 1972.

Newell, G.F., "Control of Pairing of Vehicles on a Public Transportation Route, Two Vehicles, One Control Point," Transportation Science, Vol. 8, No. 3, August, 1974, pp. 248-264.

Okrent, M.M., Effects of Transit Service Characteristics on Passenger Waiting Time, unpublished M.S. thesis, Dept, of Civil Engineering, Northwestern University, September, 1974.

Osuna, E.E., Control Strategies for Idealized Bus Systems, Ph.D. dissertation, University of California at Berkeley, November, 1969.

Osuna, E.E. and G.F. Newell, "Control Strategies for an Idealized Public Transportation System," Transportation Science, Vol. 6, No. 1, 1972, pp. 57-72.

Paine, F.T. et al., Consumer Conceived Attributes of Transportation: An Attitude Study, University of Maryland, Dept. of Business Administration, June, 1967.

Polus, A., Measures of Transportation System Reliability: Concepts and Applications, unpublished Ph.D. dissertation, Dept. of Civil Engineering, Northwestern University, June, 1975.

Radelat, G., Simulation of the Operation of a Mass Transportation System: Urban Buses on an Arterial Street, Ph.D. dissertation, Catholic University of America, 1973.

Schaefer, B.M., Program for Interactive Multiple Process Simulation (PIMP), Dept, of Industrial Engineering, Northwestern University, 1974.

Sussman, J.M. et al,. Reliability in Railroad Operations: Executive Summary, MIT, Dept. of Civil Engineering, Research Report R73-4, December, 1972.

Turnquist, M.A. Network Reliability in Urban Transit Systems, final report to the National Science Foundation, November, 1978.

Turnquist, M.A. "A Model for Investigating the Effects of Service Frequency and Reliability on Bus Passenger Waiting Time," TRR 663, pp. 70-73, 1978.

Turnquist, M.A. and L.A. Bowman, Network Effects on Transit Reliability, prepared for presentation at International Symposium on Travel Supply Models, Montreal, November, 1977.

Unger, J.C. Evaluation of Bus Priority Signal Alternatives on Archer Avenue, M.S. thesis, The Transportation Center, Northwestern University, June, 1977.

Waksman, R., and D. Schmieder, Bus Route Simulation Models for Studying Service Improvement Strategies, Staff Study, Transportation Systems Center, U.S. Dept. of Transportation, Cambridge, Mass., November, 1977.

Wallin, R.J. and P.H. Wright, "Factors Which Influence Modal Choice," Traffic Quarterly, Vol. 28, No. 2, April, 1974, pp. 271-290.

Welding, P.I., "The Instability of Close Interval Service," Operational Research Quarterly Vol. 8, No. 3, September, 1957, pp. 133-148.

Appendix A

Estimation of Correlation in Travel Times on Successive Links

Quantitative estimates of link travel time correlations were constructed using data from two routes one in Evanston and one in Chicago. The two routes were each divided into two sections. Comparison of the total route variances with the variances added over both links in each route provided the information on the dependence or correlation between link travel times.

In developing an expression relating travel times over successive links, the correct analytic method would be to derive a bivariate joint probability density function. This was rejected both for reasons of insufficient data and enormous complexity. Since this approach offered little hope, a simpler method was adopted. The correcting tendency was assumed on the average to be a simple fraction of the travel time error on the previous link in the route. This results in the following expression:

$$T_i = \gamma_i + \theta e_{i-1} \tag{A-1}$$

where

$$e_{i-1} = T_{i-1} - U_{i-1} + e_{i-2} \tag{A-2}$$

U_i = scheduled travel time over link i.

The recovery tendency affects the variation in vehicle travel time; therefore the variance expression will be used to determine the value of θ . The variance in travel time over a number of links can be expressed as:

$$V(W_n) = \sum_1^n V(T_i) + \sum_2^n 2\mathrm{Cov}(T_i,T_{i-1}) + \sum_3^n 2\mathrm{Cov}(T_i,T_{i-2})$$

$$+ \ldots + \quad 2\mathrm{Cov}(T_i,T_{i-n+1}) \tag{A-3}$$

where W_n = cumulative travel over n links

$$\mathrm{Cov}(T_i,T_{i-j}) = E[(T_i-U_i)(T_{i-j}-U_{i-j})] \qquad \text{(A-4)}$$

If the link travel times are independent all covariance terms will be zero and the total variance will simply be the sum of the individual link variances; the $V(T_i)$. Replacing T_i in equation (A-4) by equation (A-1) the covariance for travel times one link removed can be expressed as:

$$\mathrm{Cov}(T_i,T_{i-1}) = E[(\gamma_i - U_i + \theta(T_{i-1}-U_{i-1}+e_{i-2}))(T_{i-1}-U_{i-1})]$$

$$= E[(\gamma_i-U_i)(T_{i-1}-U_{i-1})] + E[\theta(T_{i-1}-U_{i-1})^2]$$

$$+ E[\theta e_{i-2}(T_{i-1}-U_{i-1})] \qquad \text{(A-5)}$$

But since γ_i is independent of travel times on previous links, equation (A-5) can be written as:

$$\mathrm{Cov}(T_i,T_{i-1}) = \theta E[(T_{i-1}-U_{i-1})^2] + \theta E[e_{i-2}(T_{i-1}-U_{i-1})] \qquad \text{(A-6)}$$

The first term is the variance of the travel time on link i-1. Substituting equation (A-1) for T_{i-1} we get:

$$\mathrm{Cov}(T_i,T_{i-1}) = \theta V(T_{i-1}) + \theta E[e_{i-2}(\gamma_{i-1}-U_{i-1}+\theta e_{i-2})] \qquad \text{(A-7)}$$

$$= \theta V(T_{i-1}) + \theta E[e_{i-2}(\gamma_{i-1}-U_{i-1})] + \theta^2 E[e_{i-2}{}^2]$$

Again, the γ_i are independent of schedule errors so the expected value of the second term is zero. The last term, $E[e_{i-2}^2]$, is the variance of the cumulative error up through link i-2. The error is simply the observed travel time minus its average, so the variance of the error is the same as the variance in the total travel time through link i-2.

Equation (A-7) can now be written as:

$$\mathrm{Cov}(T_i,T_{i-1}) = \theta V(T_{i-1}) + \theta^2 V(W_{i-2}) \qquad \text{(A-8)}$$

Extending this approach, the covariance can be found between travel times on links two steps removed:

$$\mathrm{Cov}(T_i,T_{i-2}) = E[(T_i-U_i(T_{i-2}-U_{i-2})] \qquad \text{(A-9)}$$

where $T_i = \gamma_i + \theta\ e_{i-1} = \gamma_i + \theta\ (T_{i-1}-U_{i-1}+e_{i-2})$

$$= \gamma_i + \theta[(\gamma_{i-1}+\theta e_{i-2})-U_{i-1}+e_{i-2}]$$

$$= \gamma_i + \theta\ (\gamma_{i-1}-U_{i-1}) + \theta\ (1+\theta)\ e_{i-2} \qquad \text{(A-10)}$$

Substituting (A-10) into equation (A-9) and simplifying we get:

$$\begin{aligned}\mathrm{Cov}(T_i,T_{i-2}) &= E[(\gamma_i-U_i)(T_{i-2}-U_{i-2})]\\ &\quad + \theta\ E[(\gamma_{i-1}-U_{i-1})(T_{i-2}-U_{i-2})]\\ &\quad + \theta(1+\theta)\ E[e_{i-2}(T_{i-2}-U_{i-2})] \qquad \text{(A-11)}\end{aligned}$$

Because of the independence of the γ_i the first two terms equal zero. Substituting equation (A-2) for e_{i-2} we get:

$$\mathrm{Cov}(T_i,T_{i-2}) = \theta(1+\theta)\ E[(T_{i-2}-U_{i-2}+e_{i-3})(T_{i-2}-U_{i-2})] \qquad \text{(A-12)}$$

which as before can be shown to be:

$$\mathrm{Cov}(T_i,T_{i-2}) = \theta(1+\theta)\ [V(T_{i-2})+\theta V(W_{i-3})] \qquad \text{(A-13)}$$

Assuming the initial error is zero, $e_0 = 0$, this procedure can be extended for all covariance terms until an expression is obtained solely in terms of observed variances and the unknown θ.

$$\mathrm{Cov}(T_i,T_{i-j}) = (1+\theta)^{j-1}[\theta V(T_{i-j})+\theta^2 V(W_{i-j-1})] \qquad \text{(A-14)}$$

$$\begin{aligned} V(W_n) = {} & \sum_1^n V(T_i)+2\sum_2^n [\theta V(T_{i-1})+\theta^2 V(W_{i-2})] \\ & +2\sum_3^n (1+\theta)[\theta V(T_{i-2})+\theta^2 V(W_{i-3})] \qquad \text{(A-15)} \\ & \vdots \\ & +2\sum_{n-1}^n (1+\theta)^{i-2}[\theta V(T_{i-n+1})+\theta^2 V(W_{i-n})] \end{aligned}$$

Collecting terms, equation (A-15) can be rewritten as:

$$\begin{aligned} V(W_n) = {} & V(T_n) + \sum_1^{n-1} V(T_{n-i})(1+2\theta(1+\theta)^{i-1}) \\ & + \sum_2^{n-1} V(W_{n-i})\theta^2(1+\theta)^{i-2} \qquad \text{(A-16)} \end{aligned}$$

For the simple case of two links, equation (A-16) reduces to:

$$V(W_2) = V(T_2) + V(T_1)(1+2\theta) \qquad \text{(A-17)}$$

which can be solved for θ as follows:

$$\theta = \frac{V(W_2) - V(T_1) - V(T_2)}{2V(T_1)} \qquad \text{(A-18)}$$

The estimated correction factor, θ, for peak hour Evanston service is -.21. This implies that 21% of the previous deviation is corrected over the next link, on the average. The same route segments for the non-peak period in Evanston yielded a value of -.615. This is a rather extreme case where highly reliable service combines with relatively uncongested streets and generous amounts of built in schedule slack time. The other route tested, route 84 in Chicago between Central and Bryn Mawr on Peterson, showed no correcting tendency at all. During the peak period, a value of +.001 was obtained. This could be due to the steady congestion during the peak period. Travel is always slow, permitting the driver no latitude to speed up or slow down. During the off-peak period the results are even less encouraging, with a value of +.299. This is not to suggest that travel is worse during the off-peak hours. The deviation of travel times may increase in the off-peak because at these times the congestion along Peterson may not be as consistent. Sometimes travel may be unimpeded, while on other days congestion may slow travel significantly. The positive value of the correcting factor could easily come about due to the correlation of congestion on consecutive links.

Appendix B

Length of Hold Delay for Control Strategies

Two basic forms of holding strategies are addressed: holding a fraction of the time a vehicle is early, or fixing a threshold beyond which the vehicle is always delayed a predetermined amount. In addition two ways of measuring the earliness of a vehicle are tested. Early may refer to a service interval which is shorter than intended, or it may be measured with respect to the scheduled arrival time.

A number of assumptions were required in order to make these estimates of the magnitude of holding delay. Bus arrivals are assumed to be independent and normally distributed about the scheduled arrival time. If vehicle arrival times are negatively correlated, the benefits from holding will be greater. The lengths of the holds obtained are also based on the assumption of random passenger arrivals.

B.1 Methodology

These explorations are based on the approach taken by Barnett (1974). The objective is to minimize a function combining the average wait time for randomly arriving passengers and the delay to passengers already on the bus due to holding.

$$D = \gamma E(d) + (1-\gamma)\, E(w) \qquad (1)$$

where D = total user delay

γ = weighting factor to account for delay to passengers already on the bus

$E(d)$ = average delay to a vehicle due to holding

$E(w)$ = the average wait time at this stop for randomly arriving passengers.

The average passenger wait time has been shown to be related to the variance in the length of the service interval (Osuna and Newell, 1972).

$$E(w) = \frac{H}{2}(1 + C_v^2) = \frac{E(h^2)}{2H} \tag{2}$$

where C_v = the coefficient of variation of headway intervals

H = the average headway interval

$E(h^2)$ = the second moment of headway intervals.

The aim of holding an early vehicle is to reduce the value of the second moment of the headway, $E(h^2)$.

B.2 Percentage Based Holding

This approach to a holding strategy can be expressed as:

$$d = \max\{0, Pe\} \tag{3}$$

where d = the hold time

e = the amount the vehicle is early (shortness of headway)

P = the fraction of the early value the bus is to be held ($0 \le P \le 1$).

Applying this holding policy changes the expression for the now delayed interval lengths.

$$h = H - e + Pe \tag{4}$$

and

$$\begin{aligned} E(h^2) &= E(H^2+e^2+P^2e^2-2Pe^2-2He+2PHe) \\ &= H^2 + E(e^2) - 2HE(e) + E[(P^2-2P)e^2] \\ &\qquad + 2HE(Pe) \end{aligned} \tag{5}$$

Vehicles are expected to be on time on the average so $E(e) = 0$ and $E(e^2)$ is the variance of interval lengths.

The shortcoming of this approach is that the use of a continuous distribution of vehicle arrival times precludes following Barnett's approach to account for the effect of a holding decision on the following interval. The last term in equation(5)represents the effect of holding which lengthens the average interval between buses, but since this also shortens the next interval the average should not be affected. Hence, the effect of this term is deleted from future consideration.

The average delay due to holding, E(d), is also expressed in terms of holding.

$$E(d) = E(Pe) \tag{6}$$

Assuming headway lengths are distributed independently and normally, the expected values of terms containing the variable e are as follows;

$$\begin{aligned} E[(P^2-2P)e^2] &= \int_{-\infty}^{0} 0\, e^2\, N(0,\sigma^2)de + \int_{0}^{\infty} (P^2-2P)e^2\, N(0,\sigma^2)de \\ &= 0 + (P^2-2P)\,\frac{\sigma^2}{2} \end{aligned} \tag{7}$$

$$\begin{aligned} E(Pe) &= \int_{-\infty}^{0} 0e\, N(0,\sigma^2)de + \int_{0}^{\infty} Pe\, N(0,\sigma^2)de \\ &= P\,\frac{\sigma}{2\pi} \end{aligned} \tag{8}$$

Substituting into equation (1) gives the expression of total delay which is to be minimized.

$$D = \frac{\gamma P \sigma}{2\pi} + \frac{(1-\gamma)}{2H}\left[H^2 + \sigma^2 + (P^2-2P)\,\frac{\sigma^2}{2}\right] \tag{9}$$

Taking the derivative and setting it equal to zero yields:

$$0 = \frac{\gamma\sigma}{2\pi} + \frac{(1-\gamma)}{2H}\,(2P-2)\,\frac{\sigma^2}{2} \tag{10}$$

which can be solved for the value of P.

$$P = 1.0 - \left(\frac{\gamma}{1-\gamma}\right)\sqrt{\frac{2}{\pi}}\ \frac{H}{\sigma} \qquad (11)$$

An identical approach can be used where earliness is defined with respect to the schedule and not the interval length. The minimization, however, still refers to the second moment of the headway. The assumption that successive vehicle arrivals are independent permits the headway variance to be represented by twice the variance of vehicle arrival times.

$$\mathrm{Var}(h) = E(h^2) - E(h)^2 = 2\ \mathrm{Var}\ (\mathrm{veh.}) \qquad (12)$$

Converting this to the form required by the previous procedure;

$$E(h^2) = E(h)^2 + 2\mathrm{Var}(\mathrm{veh.}) = H^2 + 2\ E(e^2) - E(e)^2 \qquad (13)$$

The variable, e, now refers to the earliness in terms of the schedule. Incorporating the effect of holding on the second moment we get;

$$E(h^2) = H^2 + 2\{E[(1-P)^2e^2] - E[(1-P)e]^2\} \qquad (14)$$

where $$E[(1-P)^2e^2] = \int_{-\infty}^{0} (1-0)^2e^2\ N(0,\sigma^2)de + \int_{0}^{\infty} (1-P)^2e^2\ N(0,\sigma^2)de$$

$$= \frac{\sigma^2}{2} + (1-P)^2\ \frac{\sigma^2}{2} \qquad (15)$$

and $$E(1-P)e = \int_{-\infty}^{0} (1-0)e\ N(0,\sigma^2)de + \int_{0}^{\infty} (1-P)e\ N(0,\sigma^2)de$$

$$= \frac{-\sigma}{\sqrt{2\pi}} + (1-P)\ \frac{\sigma}{\sqrt{2\pi}} = \frac{P\sigma}{\sqrt{2\pi}} \qquad (16)$$

Substituting into equation (1) yields a new expression for total user delay.

$$D = \frac{\gamma P \sigma}{\sqrt{2\pi}} + \frac{(1-\gamma)}{2H} \left[H^2 + 2\sigma^2 + (P^2-2P)\sigma^2 - \frac{P^2\sigma^2}{\pi}\right] \qquad (17)$$

Setting the derivative to zero and solving for P yields;

$$P = \frac{\pi}{\pi-1} \left[1.0 - \left(\frac{\gamma}{1-\gamma}\right) \frac{H}{\sigma\sqrt{2\pi}}\right] \qquad (18)$$

B.3 All-Or-Nothing

In the all-or-nothing approach, a threshold is set regarding the interval length or schedule deviation. Then every time the service exceeds the threshold, the vehicle is held a fixed length of time. For the all-or-nothing strategy tested, the threshold was assumed to be the same as the magnitude of the time a bus is to be held.

The method for determining an appropriate value for the fixed holding time follows the same approach as was used previously. However substituting the expression for the fixed delay results in a much different expression for the second moment of the headway interval. The delayed interval length is now;

$$\begin{array}{llll} h = H - e + x & \text{if} & e>x & \text{(Hold)} \\ h = H - e & \text{if} & e<x & \text{(Do not hold)} \end{array} \qquad (19)$$

which results in a second moment of the headway interval of;

$$E(h^2) = E[H^2 + e^2 + x^2 - 2He + 2Hx - 2xe] \qquad (20)$$

The effect of holding on the average length of the following interval is still not recognized, hence the presence of the fifth term - 2Hx - which causes the average interval length to increase as holding is implemented. This cannot yet be treated in connection with continuous vehicle arrivals, and the approximate solution sought here will assume that these effects cancel out in successive service intervals.

Due to the special form of the percentage based holding strategy, the integrals in the expected value expressions turned out to be integrable. This is no longer always the case for the all-or-nothing hold. The 2Hx term in equation (20) has been deleted, the remaining expressions of x are $E(x^2)$ and $E(2xe)$. These are evaluated as follows:

$$E(2xe) = 2 \int_{-\infty}^{x} 0\ e\ N(0,\sigma^2)de + 2 \int_{x}^{\infty} xe\ N(0,\sigma^2)de$$

$$= 0 + x\ \sigma \sqrt{\frac{2}{\pi}}\ (\tfrac{1}{2} \exp-(\frac{x}{\sqrt{2}\ \sigma})^2\ ) \qquad (21)$$

$$E(x^2) = \int_{-\infty}^{x} 0^2\ N(0,\sigma^2)de + \int_{x}^{\infty} x^2\ N(0,\sigma^2)de$$

$$= 0 + x^2\ \Phi\ (\frac{-x}{\sigma}) \qquad (22)$$

where $\Phi\ (\frac{-x}{\sigma})$ is the cumulative normal distribution.

The expected delay to passengers already on board is:

$$E(d) = \int_{-\infty}^{x} 0\ N(0,\sigma^2)de + \int_{x}^{\infty} x\ N(0,\sigma^2)de$$

$$= x\ \Phi\ (\frac{-x}{\sigma}\ ) \qquad (23)$$

Substitution into equation (1) again yields an expression for total delay.

$$D = \gamma x\ \Phi\ (\frac{-x}{\sigma}) + \frac{(1-\gamma)}{2H}\ [H^2+\sigma^2+x^2\ \Phi\ (\frac{-x}{\sigma})-2x\sigma\sqrt{\frac{2}{\pi}}\ (\tfrac{1}{2}\exp-(\frac{x}{\sqrt{2}\ \sigma}\ )^2] \qquad (24)$$

The derivative of total delay can be taken but the resulting equation cannot be solved directly for x. Finding the value of x requires an iterative solution.

$$0 = \Phi\left(\frac{-x}{\sigma}\right)\left(\gamma + x\frac{(1-\gamma)}{H}\right) + N\left(\frac{-x}{\sigma}\right)\left[\gamma x + \frac{(1-\gamma)}{H}\left(\frac{x}{2} + 2x^2 - 4\sigma^2\right)\right] \qquad (25)$$

where $\Phi\left(\frac{-x}{\sigma}\right)$ = the cumulative normal distribution

$N\left(\frac{-x}{\sigma}\right)$ = the value of the standardized normal distribution.

The all-or-nothing strategy can also be applied to schedule, rather than headway deviations Recalling equation (12)

$$\begin{aligned} E(h^2) &= H^2 + 2\ \mathrm{Var}\ (\mathrm{arr.}) \\ &= H^2 + 2\ \{E[(e-x)^2] - E(e-x)^2\} \\ &= H^2 + 2\ \{\mathrm{Var}\ (\mathrm{arr.}) - 2E(ex) + E(x^2) - [E(e)-E(x)]^2\} \qquad (26) \end{aligned}$$

Substituting for the expressions of the expected values of x terms, and inserting the second moment back into equation (1), we get;

$$\begin{aligned} D = \gamma x\ \Phi\left(\frac{-x}{\sigma}\right) + \gamma N\left(\frac{-x}{\sigma}\right) + \frac{(1-\gamma)}{2H}\Big\{ & H^2 + 2\ \mathrm{Var}(\mathrm{arr.}) + 2x^2\ \Phi\left(\frac{-x}{\sigma}\right) \\ & -2\left[xN\left(\frac{x}{\sigma}\right)\right]^2 \qquad (27) \\ & -4x\sigma\sqrt{\frac{2}{\pi}}\left[\tfrac{1}{2}\exp-\left(\frac{x}{\sqrt{2}\,\sigma}\right)^2\right]\Big\} \end{aligned}$$

Taking the derivative and solving for zero yields:

$$\begin{aligned} 0 = {} & \Phi\left(\frac{-x}{\sigma}\right)\left[\gamma + (1-\gamma)\frac{2x}{H}\right] + N\left(\frac{-x}{\sigma}\right)\left[\gamma x + \frac{x^2(1-\gamma)}{H}\right] \\ & + N\left(\frac{-x}{\sigma}\right)^2\ \frac{-2x(1-\gamma)}{H} + N\left(\frac{-x}{\gamma}\right)^3\ \frac{2x^3(1-\gamma)}{H\sigma^2} \\ & + \frac{(1-\gamma)}{H\sigma}\ (x^2 - \sigma^2)\sqrt{\frac{2}{\pi}}\ \left[-\tfrac{1}{2}\left(\frac{x}{\sigma}\right)^2\right] \qquad (28) \end{aligned}$$

To solve equation (28) iteratively for the value of x, even for a small number of points, a computer is desireable.

Considerable further work needs to be done investigating the optimal forms of holding strategies. Two extensions which further research may be able to incorporate are the correlation between successive vehicle arrival times, and the average wait time if passengers coordinate their arrivals with the scheduled vehicle arrival times.

REQUEST FOR FEEDBACK TO
The DOT Program Of University Research

DOT/RSPA/DPB-50/79/5 "Control of Service Reliability in Urban Bus Networks: Final Report"

YES	NO	
☐	☐	Did you find the report useful for your particular needs? If so, how?
☐	☐	Did you find the research to be of high quality?
☐	☐	Were the results of the research communicated effectively by this report?
☐	☐	Do you think this report will be valuable to workers in the field of transportation represented by the subject area of the research?
☐	☐	Are there one or more areas of the report which need strengthening? Which areas?
☐	☐	Would you be interested in receiving further reports in this area of research? If so, fill out form on other side.

Please furnish in the space below any comments you may have concerning the report. We are particularly interested in further elaboration of the above questions.

Cut Out Along This Line

COMMENTS

Thank you for your cooperation. No postage necessary if mailed in the U.S.A.
FOLD ON TWO LINES, STAPLE AND MAIL.

RESEARCH FEEDBACK

Your comments, please . . .

This booklet was published by the DOT Program of University Research and is intended to serve as a reference source for transportation analysts, planners, and operators. Your comments on the other side of this form will be reviewed by the persons responsible for writing and publishing this material. Feedback is extremely important in improving the quality of research results, the transfer of research information, and the communication link between the researcher and the user.

FOLD ON TWO LINES, STAPLE AND MAIL.

Fold Fold

DEPARTMENT OF TRANSPORTATION
RESEARCH AND SPECIAL PROGRAMS Administration
WASHINGTON D.C. 20590

Official Business

PENALTY FOR PRIVATE USE, $300

POSTAGE AND FEES PAID
DEPARTMENT OF
TRANSPORTATION
DOT 513

U.S.MAIL

OFFICE OF UNIVERSITY RESEARCH (DPB-50)
Research and Special Programs Administration
U.S. Department of Transportation
400 Seventh Street, S.W.
Washington, D.C. 20590

Fold Fold

REQUEST FOR INFORMATION FROM THE UNIVERSITY RESEARCH PROGRAM

☐Check here if you would like to be placed on the mail list for the University Research Program Solicitation Booklet (DT-63C)

IF YOU WISH TO BE ADDED TO THE MAIL LIST FOR FUTURE REPORTS, PLEASE FILL OUT THIS FORM.

Name ______________________________ Title ________________
Use Block Letters or Type

Department/Office/Room ______________________________

Organization ______________________________

Street Address ______________________________

City ____________________ State ____________ Zip ____________

U.S. DEPARTMENT OF TRANSPORTATION
RESEARCH AND SPECIAL PROGRAMS ADMINISTRATION
WASHINGTON, D.C. 20590

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE, $300

POSTAGE AND FEES PAID
RESEARCH AND SPECIAL PROGRAMS
ADMINISTRATION
DOT 513

U.S. MAIL