# Running Time Variability and Resource Allocation: A Data-Driven Analysis of High-Frequency Bus Operations

by

Gabriel Eduardo Sánchez-Martínez

Bachelor of Science in Civil Engineering
University of Puerto Rico, Mayagüez (2010)

Submitted to the Department of Civil and Environmental Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Transportation

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2013

Author . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Department of Civil and Environmental Engineering
October 16, 2012

Certified by . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Harilaos Koutsopoulos
Professor of Transport Science, Royal Institute of Technology
Thesis Supervisor

Certified by . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Nigel H.M. Wilson
Professor of Civil and Environmental Engineering
Thesis Supervisor

Accepted by . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Heidi M. Nepf
Chair, Departmental Committee for Graduate Students

**Running Time Variability and Resource Allocation:**
**A Data-Driven Analysis of High-Frequency Bus Operations**

by
Gabriel Eduardo Sánchez-Martínez

Submitted to the Department of Civil and Environmental Engineering
on October 16, 2012, in partial fulfillment of the
requirements for the degree of
Master of Science in Transportation

## Abstract

Running time variability is one of the most important factors determining service quality and operating cost of high-frequency bus transit. This research aims to improve performance analysis tools currently used in the bus transit industry, particularly for measuring running time variability and understanding its effect on resource allocation using automated data collection systems such as AVL.

Running time variability comes from both systematic changes in ridership and traffic levels at different times of the day, which can be accounted for in service planning, and the inherent stochasticity of homogeneous periods, which must be dealt with through real-time operations control. An aggregation method is developed to measure the non-systematic variability of arbitrary time periods. Visual analysis tools are developed to illustrate running time variability by time of day at the direction and segment levels. The suite of analysis tools makes variability analysis more approachable, potentially leading to more frequent and consistent evaluations.

A discrete event simulation framework is developed to evaluate hypothetical modifications to a route's fleet size using automatically collected data. A simple model based on this framework is built to demonstrate its use. Running times are modeled at the segment level, capturing correlation between adjacent segments. Explicit modeling of ridership, though supported by the framework, is not included. Validation suggests that running times are modeled accurately, but that further work in modeling terminal dispatching, dwell times, and real-time control is required to model headways robustly.

A resource allocation optimization framework is developed to maximize service performance in a group of independent routes, given their headways and a total fleet size constraint. Using a simulation model to evaluate the performance of a route with varying fleet sizes, a greedy optimizer adjusts allocation toward optimality. Due to a number of simplifying assumptions, only minor deviations from the current resource allocation are considered. A potential application is aiding managers to fine-tune resource allocation to improve resource effectiveness.

Thesis Supervisor: Harilaos Koutsopoulos
Title: Professor of Transport Science, Royal Institute of Technology

Thesis Supervisor: Nigel H.M. Wilson
Title: Professor of Civil and Environmental Engineering

*To my family*

## Acknowledgments

I wish to thank the following people for their help and support:

The National Science Foundation, the Lemelson-MIT foundation, and the Frank W. Schoettler Fellowship Fund, for their financial support;

John Barry, Alex Moffat, Rosa McShane, Steve Robinson, Annelies de Koning, Wayne Butler, and Angela Martin from London Buses, as well as Shashi Verma and Lauren Sager Weinstein from TfL, for their help, insight, and support;

My advisors, professors Harilaos Koustopoulos and Nigel Wilson, for their wisdom and insight, for engaging discussions on this and future research, for their continual support and guidance, and for all help editing this thesis;

To John Attanucci and Dr. George Kocur, for the insight they provided throughout the development of this research;

Ginny Siggia, for helping me with administrative matters and hard-to-find references;

Patty Glidden and Kris Kipp, for their help with administrative matters;

The MITES program, for exposing me to the excitement of being at MIT, and Sandra Tenorio, for her warm and welcoming support;

My friends from the transportation program at MIT, Candace Brakewood, Mickaël Schil, Naomi Stein, Laura Viña, Carlos Gutiérrez de Quevedo Aguerrebere, Jay Gordon, Alexandra Malikova, Pierre-Olivier Lepage, Kevin Muhs, Varun Pattabhiraman, Kari Hernández, Janet Choi, and all my other transit lab mates, for their friendship;

Professor González Quevedo, for encouraging me to consider graduate studies in transportation, and for the invaluable knowledge and skills he has taught me;

My cousins in London, for welcoming me and allowing me to stay with them during my summer in London;

My family, especially my parents and my brother, for their unending love and encouragement.

# Contents

# Chapter 1

# Introduction

Bus operations are typically characterized by high uncertainty. Uncertainty manifests itself in various ways that affect operations, one of which is running time variability, or the a-priori uncertainty of a future trip's duration. Running time variability is an important determinant of service quality and the resources required to operate high-frequency transit reliably. Automated data collection systems such as AVL, which provide very large observation samples at low marginal costs, enable the development and use of new data-driven analysis tools that can potentially enhance performance monitoring abilities, and ultimately lead to improved resource allocation and effectiveness.

The objective of this research is to improve running time variability measurement and analysis tools currently used in the bus transit industry, taking full advantage of automatically collected data. This is accomplished by (1) discussing methods for measuring variability at different levels of aggregation and developing visual analysis tools to study running time variability by time of day and segments of a route, (2) exploring how aggregate-level characteristics of a route, such as length, number of stops, and ridership, relate to typical running times and running time variability, (3) presenting a data playback method that can be used to capture and analyze interaction effects at a disaggregate level, (4) developing a framework for discrete event simulation modeling of transit systems, and testing it by implementing a simple simulation model of a real bus route, and (5) developing a simulation-driven budget-constrained resource allocation optimization framework for fine-tuning route fleet sizes in light of their running time variabilities, and illustrating the use of this framework applied to two hypothetical routes.

The principal contributions are an aggregation method for measuring non-systematic variability of arbitrary time periods, visual analysis tools to aid in performance monitoring tasks, a flexible and extensible framework for transit simulation, and a related framework for simulation-driven optimization of resource allocation.

## 1.1 Motivation: Role of Variability in Transit Operations

Running time variability tends to negatively affect the performance of high-frequency transit services. The most direct effect is that riders must estimate their in-vehicle trip time conservatively if they want to arrive at their destination on time. However, there are further indirect effects. Differences in the running times of consecutive vehicles lead to variability in headways. A randomly arriving passenger (one who does not time his arrival at a stop based on either the schedule or real-time arrival information) wishing to arrive at his final destination on time must also conservatively estimate his waiting time at the stop. Passengers, who value their time, find these necessary buffer times burdensome. Researchers have found that reliability is an important factor in mode choice, and that the monetized value of reliability, estimated by discrete choice analysis based on both revealed and stated preference data, is significant (more so in public transport than for driving). (Bates et al., 2001 and Li et al., 2010)

High-frequency services are designed so passengers can "turn up and go". Riders of these services generally do not time their arrivals at stops. This leads to essentially random passenger arrivals (Osuna and Newell, 1972). (This may be changing with the advent of real-time arrival estimates and Internet-enabled portable devices.) It is true that some arrivals are dictated by the arrival times of previous connecting trips in multi-leg journeys, and also that some passengers travel in groups. Nevertheless, these patterns tend to be random as well. When a high-frequency transit service exhibits headway variability, randomly arriving passengers have a higher probability of arriving during a long headway than during a short one. A few lucky riders experience shorter waiting times, but on average passengers wait longer. The additional mean waiting time experienced by passengers due to this effect is called *excess waiting time* (EWT) (Ehrlich, 2010).

Headway variability also leads to variability in crowding levels. Vehicles with longer leading headways find more passengers waiting at stops, so they become more crowded that vehicles with short leading headways. Passengers have a higher probability of encountering a crowded vehicle (after a long wait). (According to Xuan et al., 2011, this was first recognized by Osuna and Newell, 1972, Newell, 1974, and Barnett, 1974.) This effect dissipates (but only in a degenerate, local sense) at the extreme case of headway variability, when consecutive vehicles are *bunched*. When a bunch arrives at a stop, passengers may board the least crowded vehicle, typically the one at the tail of the bunch. In some cases, operator policy might prevent passengers from boarding the leading vehicle.

Headway variability also leads to variability in dwell times. Vehicles with longer leading headways see more people waiting at stops, so they dwell longer to accommodate a higher number of boarding passengers. For these vehicles, it is more probable that there will be at least one person at a stop, so the vehicle also stops more frequently, adding to the trip's total cumulative dwell time. Since vehicles with longer leading headways are more likely to be crowded, it is also more probable that a rider on board will request a stop, even when no one is waiting at that stop to board. Moreover, when a vehicle is very crowded, it takes longer for passengers to make their way to the door to alight. Since dwell time is a significant component of running time, dwell time variability adds to the total running time variability, a detrimental feedback effect (Strathman and Hopper, 1993). Previous research has recognized a generally strong positive correlation between headway variability and running time variability (Abkowitz and Lepofsky, 1990).

The relationships between all these factors are illustrated in the causal diagram shown in Figure 1-1. Arrows indicate causality: those labeled with a + sign indicate a direct relationship (if A increases, B increases), while those labeled with a − sign indicate an inverse relationship (if A increases, B decreases). Two of the arrows leading to the *resources allocated* block are dotted, indicating that the relationship is not automatic; operator re-

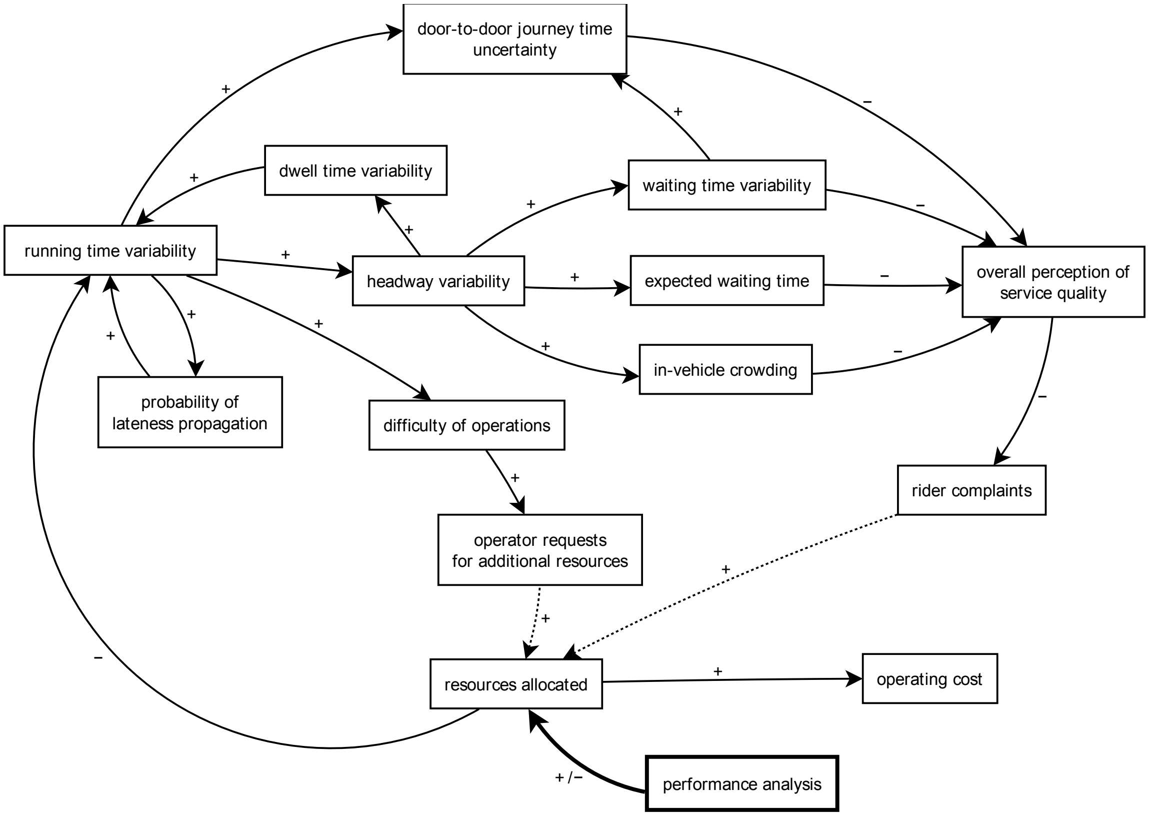


**Figure 1-1:** Role of running time variability in transit operations

quests for additional resources and rider complaints may trigger the allocation of additional resources, but the bus agency must analyze the situation and decide.

Performance analysis, which is the focus of this research, is the only factor shown in the diagram that can change resource allocation in both directions. It is topologically closer to running time variability (with a separation of two arcs) compared to operator requests for additional resources and rider complaints, which shows that its effect on running time variability is more direct. This research aims to improve performance analysis tools so that running time variability can be better understood and managed.

## 1.2 Relationship Between Variability and Resource Level

More resources are required to operate routes which have higher running time variability. In a simple operating environment, the relationship between the number of vehicles $n$, headway $h$, and cycle time $c$ is given by the following equation:

$$c = nh \tag{1.1}$$

Cycle time is a function of the running time distributions and the target dispatch reliability at terminals. It is typically set at a high percentile (typically between the $85^{\text{th}}$ and the $95^{\text{th}}$ percentile) of the two-way running time distribution. The running time distribution arises from characteristics of the route, such as length, number of stops, and ridership. Higher percentiles are used when a higher dispatch reliability is desired. The required fleet size depends on headway and cycle time. Considering a route independently (i.e. no interlining), and having a target reliability, fleet size is given by the following equation:

$$n = \left\lceil \frac{c}{h} \right\rceil \tag{1.2}$$

where $\lceil \cdot \rceil$ denotes rounding up to the next highest integer. Assuming no change in headway, the rounding operation implicitly increases the cycle time so that (1.1) holds.

Headway is determined by a combination of ridership and policy, such that the route has capacity to carry the passengers while maintaining vehicle (and stop) crowding levels under a certain threshold and providing a minimum frequency of service. Ridership is typically considered fixed in the short run. Running time distributions are directly observed from current service. While they depend on a number of exogenous and endogenous factors (for example, driver effects, operator intervention, vehicles, and road geometry, which includes lane widths, stop placement, traffic, signals, junction density, and the availability of bus lanes), they are often also considered fixed in the short run.

Given a particular headway and running time distribution, the only decision variable is the fleet size $n$, which simultaneously determines reliability and cost of service. The fleet size requirement of a route increases with increasing running time variability. Adding vehicles increases operational costs and makes it easier for the operator to run regular service, because the added resources enables greater running times (i.e. farther on the upper tail of the running time distribution) to be covered without propagating delays to the next trip.

Consider a simple hypothetical bus route which operates in a loop with 5 minute headways, and has the running time distribution shown in Figure 1-2, with a mean of 60 minutes and a standard deviation of 10 minutes. The figure indicates the percentiles of running times achieved by different fleet sizes. For example, providing 15 vehicles for service allows 93% of the trips to be completed without delaying the departure of subsequent trips. Providing 12 vehicles allows only half of the trips to be completed without delay propagation, which would lead to disrupted operations.

The buffer time corresponding to a (conservative) cycle time (and thereby, fleet size) can be distributed based on one of several strategies. The simplest approach assigns all buffer time to a terminal. Upon arrival at the terminal after completing a trip, vehicles *stand* (or *lay over*) at the terminal to recover the schedule (or headway). This type of buffer time is commonly called *recovery time*. Following the previous example with a fleet size of 15 vehicles, this would happen for 93% of the trips. In the remaining 7%, vehicles take even longer to complete a trip, so they begin the next trip as soon as possible, with some degree of lateness propagation. A larger fleet size decreases the probability of lateness propagation over several consecutive trips.

The typical bus route structure has two directions, with a terminal at each end of the route. In this case, recovery time can be distributed between the two terminals, providing two opportunities per cycle to recover the schedule (or headway). Equal amounts of recovery time can be assigned to each terminal, or, if running times are more variable in one of the directions, recovery times can be allocated in proportion to running time variability by

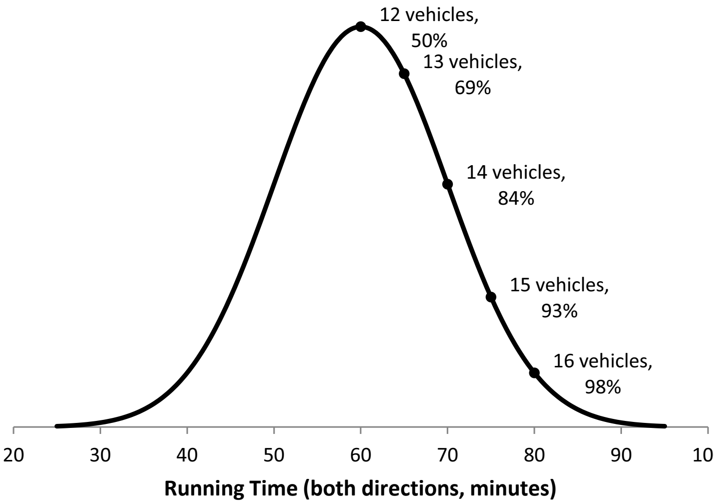


**Figure 1-2:** Running time percentiles achieved by different fleet sizes

direction. The strategy of assigning recovery times to both terminals allows trips of both directions to begin on time (in schedule-based operations) or for headways to be regular at route ends (in headway-based operations). The conservative trip time per direction is referred to as *half-cycle time*, and the sum of the two half-cycle times is the route's cycle time.

In a different strategy, recovery time is distributed both at terminals and en route timing points. When a vehicle arrives at a timing point early, it *holds* to recover the schedule (or headway). In order to distinguish recovery time at terminals from recovery time at timing points, recovery time at a timing point is called *holding time* and recovery time at terminals is called *stand time* or *lay over time*. The percentile threshold that defines when a trip is early need not be the same for holding time and stand time. For example, vehicles may hold at timing points when their cumulative running time is below the median, while stand time can be based on a higher percentile, such as 90$^{\text{th}}$ or 95$^{\text{th}}$ percentile half-cycle times. The threshold choice for holding time is an important decision, because it affects how often and for how long vehicles hold en route (Furth and Muller, 2007). Using higher-percentile thresholds for holding increases the number of locations at which service can be regulated, but at the same time slows trips down. Frequent or long holding times can heighten passengers' perception of a slow trip, which may eventually drive passengers to alternative routes or modes.

The term *resource* is used generally in this research, and is meant to include variable operating costs such as vehicles, crew, fuel, supervision, and management. Capturing all of these factors in detail can be difficult because there may not be accessible data covering each of them. For this reason, fleet size is used throughout this thesis as a proxy for resources. Adding vehicles to a route implies a general increase in the size of operations, which adds crew, fuel, supervision, and management costs.

## 1.3 Sources of Running Time Variability

Almost every component of bus operations is variable, so total running time variability comes from many sources. Some sources, such as traffic, are exogenous factors that introduce variability independently of how service is operated. Other sources, such as operator behavior, are endogenous factors that can add variability by themselves or aggravate the effect of exogenous factors. (For a review of the literature on the sources of running time variability, see Section 1.6.1.)

Vehicle traffic and traffic signals are commonly major sources of variability. They slow down buses in movement, add delays at signals, and make it more difficult for buses to rejoin traffic after serving a stop (in roads without dedicated bus lanes). There is a mix of vehicle types, and each driver has a different response time and level of aggressiveness, which makes running times variable. In roads with dedicated bus lanes, other buses (and in some cases, bicycles) can slow down a bus trip. Bus drivers also introduce running time variability; like drivers of private vehicles, they too have varying response times and levels of aggressiveness. Accidents, incidents, roadwork, and diversions tend to create congestion and make running times more unpredictable.

Demand patterns are variable. Successive vehicles see a varying number of passengers waiting to board at stops, and these passengers have varying destinations. Demand at a major transfer point can be especially variable. Riders coming from a train may arrive in groups. Dwell times are longer when there are more passengers boarding and alighting. Special events and circumstances make running times more variable. Cultural and sporting events can trigger atypical ridership patterns.

Operator actions also contribute. There can be variability in how vehicles are dispatched at terminals, in addition to en route real-time control actions such as holding, short-turning, expressing, dead-heading, and deliberately slowing down or speeding up. These actions can add to or subtract from total variability. The principal motivation for real-time control in high-frequency service is headway regulation, which can lead to decreased running time variability. All the above sources of variability interact in complex and intricate ways, making it difficult to find out what percentage of running time variability is due to each source.

## 1.4 London Buses Environment

The real-world examples used in this thesis pertain to bus service in London. Bus transit in London is privately operated, but planned, procured, managed, and monitored by a government agency called London Buses, a division of Surface Transport (ST) within Transport for London (TfL). London Buses is responsible for all aspects of service planning, including the determination of route alignment, stop placement, span of service, vehicle type, and frequency by time of day. A *minimum performance requirement* is also specified in terms of *excess waiting time* (EWT) (a measure of headway regularity) for high-frequency service and on-time performance for low-frequency service. During the procurement process, London Buses supplies these specifications along with historical information on previous performance and running time. Private operators generate schedules based on these specifications, and must show a good understanding of running times in their submissions. (London Buses, 2009)

Once an operator is selected and begins serving a route, London Buses monitors service performance on a regular basis and calculates performance measures based on automatically collected vehicle location data (AVL). Performance measures are compared against the minimum performance requirements of the route to determine the corresponding incentive payments or penalties according to the contract. Resource levels for a route may be adjusted if changes in running times require it. (For more information on London Buses, see Ehrlich, 2010.)

The business rules at London Buses have led to a relatively advanced analysis-based business practice and a steady reliance on automatically collected data. The management in charge of performance monitoring is aware that the resource requirement of a route depends on running time variability, but unfortunately the current running time reports at their disposal provide limited information about running time variability. This makes it challenging to determine appropriate resource allocation.

Many bus agencies have far less sophisticated analysis practices. Surveys of bus agencies in North America have found that off-line data analysis is seldom included as an objective in projects to equip fleets with AVL systems; emergency response and real-time information provision are the predominant goals. Some agencies do store AVL data electronically, but often there are no systems or business practices to query the data for service planning and performance monitoring purposes (Furth, 2000).

## 1.5 Objectives and Approach

The overarching objective of this research is to improve running time variability measurement and analysis tools currently used in the bus transit industry. This is accomplished by meeting the following specific research objectives:

1. to present running time variability as a key determinant of service quality and resource requirement for high frequency bus service;

2. to develop a framework for descriptive analysis of running time variability, a set of performance measures to quantify variability at different levels of aggregation, and visual analysis tools to study running time variability by time of day and segment of a route;

3. to explore the relationship between typical running time, running time variability, and characteristics of a route at an aggregate level through specification of linear models and empirical estimation of parameters with linear regression;

4. to develop a data playback method for capturing interaction effects at a disaggregate level and analyzing their effect on transit performance;

5. to develop a flexible, extensible, and parsimonious framework for simulation modeling of transit systems, and to test the framework by implementing and validating a simple simulation model of a real bus route;

6. to develop a simulation-driven budget-constrained resource allocation optimization framework for fine-tuning route fleet sizes in light of the running time variabilities of different bus services, and to show how this optimization works through a simple example.

The first step toward better management of running time variability and resource allocation is establishing a consistent measurement practice that allows the members of a service planning or performance evaluation team to communicate ideas about running time variability effectively. Regular monitoring of running times leads to an enhanced understanding of how typical running times and running time variability evolve seasonally, and places management in a better position to adjust resource allocation accordingly. Claims from operators that additional resources are necessary to maintain service quality can be evaluated objectively with analysis focused on running time variability.

Running times vary seasonally, by day of week, and by time of day. The basis of analysis, however, is characterizing variability inherent in operations at a particular time, weekday, and season. Running times observed therein do not exhibit systematic trends and are said to belong to a *homogeneous period* modeled as a *stochastic process*. This purely random, inherent variability quantifies the uncertainty surrounding the duration of future trips. This type of variability is closely related to reliability, cycle time, and fleet size requirement.

Variation by season, weekday, and time of day has a systematic component. For example, morning peak running times may be generally higher than evening running times, and running times in the fall may be generally higher than running times in the summer. These systematic changes respond to changes in the operating environment: higher ridership and more traffic in the morning peak than in the evening, and in the fall than in the summer. There may also be systematic trends in running time variability. In some cases, seasonal variation of typical running times may not be pronounced, but running time variability may change significantly. Systematic trends can be studied and incorporated into the service plan by tuning resources by season, weekday, and time of day.

Running times can be analyzed at the route, direction, and segment level. Route and direction level analyses are useful to make decisions on resource allocation. Segment level analysis can help identify the parts of a route that contribute the most to overall route variability. For such segments, modifications in stop location, addition of bus lanes, or signal priority schemes could lessen the variability and lead to operating cost savings and improved reliability. Having knowledge of segment-level variability is also useful in evaluating route revisions.

In some applications it is desirable to describe the random component of running time variability over a period exhibiting both systematic and random variability. For example, a single-figure route-level variability measure can serve as a screening tool for service planners and performance evaluators to prioritize analysis efforts. On the other hand, aggregate measures do not communicate the level of detail required to support resource allocation and route revision decisions. Disaggregate measures of variability, by time of day and segment, work better for this purpose. Studying large amounts of data in detail can be cumbersome, but visual analysis tools can condense this information and convey meaning in a more natural and appealing fashion, enabling regular assessment of the levels of variability by time of day and segment.

General models explaining typical running time and running time variability can be useful to estimate running times of a proposed route or changes in running times following a route revision. In this manner, they can help service planners predict the resource requirements of different alternatives. The intent of general models is to establish the structure of the relationship between route characteristics and running times. Important route characteristics include distance, number of stops, ridership, and traffic.

The types of analysis discussed above are *descriptive*, in the sense that they characterize the current conditions ex post. With them, it is possible to measure variability in a consistent manner over time and have better management control. In some situations, managers might question whether the current level of resources assigned to a route is appropriate given the running times. To answer this question, it is necessary to predict how service performance responds to changes in fleet size. Since there is no performance history for hypothetical scenarios like this one, it is necessary to rely on predictive models rather than historical observations. Simulation models are well suited for the task because of the flexibility they provide to experiment with how the different elements of bus operations are represented, and also because they are able to capture interaction effects, discontinuities, stochasticity, and nonlinearity.

The simulation model framework developed in this research represents vehicles and passengers as objects. A series of events describing the arrival of vehicles at stops is processed chronologically. The model architecture handles passenger boardings and alightings with vehicle stop visits. In order to model service on a bus route, the route alignment, running time distributions, vehicle types and fleet size, demand, and operating strategies must be specified. Emphasis is placed on obtaining many of these from automatically collected data. Service performance is quantified in terms of running times, waiting times, and loads observed in simulation. Simulation models must be validated before using them for decision support. The most important validation test examines if the simulation model reproduces current operations faithfully (Law, 2007).

Simulation modeling can be used within an optimization framework to find the resource-constrained allocation that maximizes total service quality. Although bus agencies may have set the fleet size of most routes at an appropriate level, the nature of performance monitoring places more emphasis on problem routes. An optimization framework can help detect opportunities to save resources where the fleet size of a route is excessive, and to make more effective resource investments where adding vehicles can have a large benefit in service performance. In effect, an optimization scheme like this one can help management fine-tune resource allocation. The problem can be stated mathematically as an optimization program, with an objective function that captures different aspects of service quality (for instance, excess waiting time and excess load), and constraints to limit the overall fleet size, limit the magnitude of the fleet size change of routes (with respect to the current fleet size), and enforce minimum levels of service quality.

This research focuses on high-frequency bus service, and assumes random passenger arrivals to determine passenger-based measures of waiting time. Nonetheless, many of the concepts presented can be extended to both low-frequency bus service and rail transit. An important simplifying assumption made for the specific implementations of simulation and optimization models is that the operating environment (including running times and ridership) is insensitive to changes in fleet size (and resulting changes in service performance). The simulation and optimization frameworks, however, are extensible and support developing more robust models.

## 1.6 Literature Review

An abundance of scientific literature has been written on the research topics of this thesis, including variability in transit systems, development of performance measures that quantify variability, transit simulation models, and optimization.

### 1.6.1 Variability in Transit

Many researchers have identified the different sources of variability. Cham (2006) listed schedule deviations at terminals, passenger loads, running times, environmental factors (e.g. traffic), and operator behavior as the components of variability of a transit system. Ehrlich (2010) considered operator behavior, inherent route characteristics, ridership, the availability of an automatic vehicle location system for real-time control, the contract structure, and road work as causes affecting service reliability.

Some researchers have classified the different sources of variability by their nature. van Oort (2011) regarded driver behavior, other public transport, infrastructure configuration, service network configuration, schedule quality, driver behavior, vehicle design, and platform design as internal causes of variability, and other traffic (e.g. private vehicles and traffic lights), weather, passenger behavior, and irregular loads as external causes. In classifying types of delays in rail operations, Carey (1999) distinguished between exogenous delays (i.e. equipment failure, delays in passenger boardings and alightings, lateness of operators and crews, which are not caused by the schedule) and knock-on delays, which result from exogenous delays and the trip interdependence in the schedule.

Abkowitz and Lepofsky (1990) analyzed the implementation of a simple holding strategy on two MBTA bus routes in Boston. They found that in some cases dynamic holding control on select timing points can lower running time variability across the entire route and prevent variability from propagating along the route, consistent with the findings of previous research. In addition, they found that because of high correlation between running time variability and headway variability, mean segment running times also decreased, suggesting that resources are used more effectively when headways are regulated in high-frequency transit service.

Strathman and Hopper (1993) carried out an empirical analysis of bus transit on-time performance, using a multinomial logit model to identify the factors determining whether a vehicle is early, on-time, or late. They found that the probability of on-time arrival decreases with increasing number of alighting passengers, at locations further downstream on a route, with greater headways, and for afternoon peak outbound trips. Driver experience was also identified as an important factor. Their approach focuses on schedule adherence rather than running time variability per se.

Pratt et al. (2000) examined traveler response to a diverse range of transportation system changes, based on documented experiences. They found that ridership has responded strongly to changes in reliability. Regular commuters consider arrival at the intended time to be more important than travel time, waiting time, and cost. Balcombe et al. (2004)

conducted similar studies, finding that, although reliability is regarded as one of the most important aspects of service quality, few studies have estimated demand elasticities, largely because changes in reliability are subtle in comparison to a change in fares. An additional difficulty is that reliability is measured differently across agencies. Applications of a theoretical model suggested that excess waiting time is two to three times as onerous as ordinary waiting time (Bly, 1976).

### 1.6.2 Performance Measures and Indicators

Benn (1995) reviewed the state of practice among transit agencies in the United States with respect to bus route evaluation standards, including standards of route design, schedule design, economic and productivity standards, service delivery standards, and passenger comfort and safety. Two types of service delivery standards were identified: on-time performance indicators for schedule-based services, and headway adherence indicators for headway based services. The application of these measures is typically done by classifying timing point departures as early, on-time, or late. For example, a trip might be considered late if it departs a timing point more than five minutes after the scheduled time, and early if it departs any time before the scheduled time. Agencies sometimes use this to report the percentage of trips on time, either at the route or system level. In a survey of performance across bus agencies, Benn found that the majority report performance at or above 90% at the network level. These reports may sometimes differ from the public perception because not all transit vehicles are loaded equally; if most trips during the day are on-time but the most crowded rush hour trips are delayed and crowded, most passengers will perceive relatively lower service quality.

Carey (1999) discussed the use of heuristic measures of reliability to evaluate schedules, distinguishing between exogenous delays (not caused by the schedule) and knock-on delays (initiated by exogenous delays but propagated due to the schedule). The emphasis was on schedule-based rail operations, so the heuristics are less applicable to high-frequency headway-based transit.

Kittelson & Associates et al. (2003) discussed the practice of measuring service quality in the North American transit industry, and identified passenger loads and reliability among other important factors that determine how passengers perceive service quality. Borrowing the concept of Level of Service (LOS) from the 1965 Highway Capacity Manual, they rated different aspects of service quality on a scale from A (best) to F (worst). For example, a service gets an A for on-time performance when 95% or more of the trips are between 0 and 5 minutes late, and an F when the fraction falls below 75%. For headway adherence, a service gets an A when the coefficient of variation of headways is between 0.00 and 0.21, and an F when it is above 0.75. (The coefficient of variation is the standard deviation of observed headways divided by the mean scheduled headway.) Agencies can use tools like these to monitor service quality over time and direct more management attention to routes that are performing poorly.

Cham (2006) developed a framework for understanding bus service reliability using automatically collected data to calculate performance measures. Furth et al. (2006) discussed how automated data collection systems (ADCS), particularly AVL and APC, can be integrated for running time and load analysis, schedule design, and operations planning. Automated data collection systems enable analysis of extreme values, such as the $90^{\text{th}}$ or $95^{\text{th}}$ percentile running times, instead of only mean or median running times. They also enable analysis of any route, at any day of week and time of year, including weekends and holidays.

A recent trend is to quantify reliability from the passenger's perspective, relying primarily in fare card data. Uniman et al. (2010) generated empirical distributions of origin-destination journey times for passengers of the London Underground using fare card data, and developed performance measures to quantify variability of journey times. In particular, they defined *reliability buffer time* (RBT) as the difference between a high percentile (say, $95^{\text{th}}$ percentile) and the median of the journey time distribution, which can be thought of as the extra time a passenger must budget to have a good chance of completing the journey in the planned time. Schil (2012) extended this concept to capture reliability buffer both in waiting and in travel aboard the vehicle.

Ehrlich (2010) presented potential applications of automatic vehicle location data for improving service reliability and operations planning in London Buses. He discussed the performance measures currently used (excess waiting time, percent on-time, and percent lost mileage), which are critical for performance evaluation and to determine incentive payments and penalties. Additionally, he introduced three performance measures more focused on the passenger experience: journey time, excess journey time, and reliability buffer time.

Trompet et al. (2011) evaluated four performance indicators targeting headway regularity in high-frequency urban bus routes: excess waiting time, quadratic mean of deviations from the target headway, and percentage of headways within a tolerance (absolute or relative) of the target headway. They found that, of the three, excess waiting time captures passenger waiting time most effectively. Excess waiting time is already in use by bus agencies such as London Buses.

Passenger perspective measures give important insight, but it is more difficult to relate them to resource allocation levels. It is important to realize that these measures are developed to complement, not replace, direct measurements of running time variability. The research presented in this thesis focuses on running time variability per se.

### 1.6.3 Transit Simulation Modeling

Since simulation provides a convenient way to analyze hypothetical scenarios (in some cases, the only way), many researchers have used the technique to test models and hypotheses. The research presented in this thesis by no means pioneers simulation modeling of transit operations, although development of a generalized and extensible framework is emphasized more than before.

Marguier (1985) developed an analytical stochastic model of bus operations which captures stochasticity in running times and dwell times, and enforces trip chaining. The model ignores the effect that bus bunching may have on the rest of the line, does not enforce capacity constraints, prohibits overtaking, and uses a linear dwell time function. These simplifications, along with distributional assumptions, are necessary to keep the analytical model mathematically tractable. This model remains, to this date, one of the most sophisticated analytical models of cyclical transit operations. Marguier also developed a discrete event simulation model, similar to the one developed in this research, to test the assumptions of his analytical model, and found good agreement between the two in the absence of strong violations to his assumptions. Naturally (given the date of his work), neither of the models take full advantage of automated data collection systems. The simulation model framework presented in Chapter 3 of this thesis provides greater flexibility and can be driven by AVL and AFC data.

Abkowitz et al. (1987) used simulation to evaluate the effectiveness of timed transfers in schedule-based transit operations, in a case study of two hypothetical routes intersecting at a transfer point. Chandrasekar et al. (2002) used a time-stepping microsimulation model to evaluate the effectiveness of combined holding and transit signal priority strategies. Altun and Furth (2009) used Monte Carlo simulation implemented in a spreadsheet as well as traffic microsimulation to analyze transit signal priority. Simulation modeling enabled them to capture random delays at dispatch and at traffic signals, the effect of crowding on dwell time, and to test a variety of signal priority and operational control strategies.

Delgado et al. (2009) evaluated the effectiveness of holding and boarding limits as real time control actions to improve headway regularity in a hypothetical deterministic route. Their approach combines optimization and simulation to improve service. Larrain et al. (2010) presented an optimization framework to select the configuration of express services in a bus corridor with capacity restrictions. They simulated operations in an idealized route to demonstrate the use of the model and identify the factors driving the optimality of express services.

Liao (2011) developed empirical models for dwell time at timing points, vehicle movement time between timing points, as well as a simulation tool in which running times and dwell times are driven by the empirical models. The simulation tool allows a transit planner to estimate the impact of potential changes to a route, such as stop consolidation or offering limited stop service. Although the tool is useful for service planning, the empirical models must be estimated and calibrated beforehand. A generic model (estimated with data of many routes) can be used, but route-specific distributions and correlation structures observable in automatically collected data are lost in the process.

There has also been research with a significant simulation development component, though simulation is always an analysis tool and not an end by itself.

Bly and Jackson (1974) developed a flexible simulation model of a 10 minute headway route. Their model represents passenger arrivals, boarding and alighting times, vehicle capacity constraints, stop-to-stop running times, and service regulation (stand time at terminals to meet the timetable). It is capable of modeling many types of control actions, including terminal stands, holding at timing points and stops, injection of vehicles, skipping stops, limited stop service, and short-turning. Running times, passenger arrivals, and dwell times were based on data collected from manual surveys. The simulation model was used to evaluate control strategies. The researchers found that schedule-adherence at the terminal was one of the most effective control strategies.

Similar work was carried out by Jackson (1977) on a London bus route with 3–6 minute headways. Simulation was used to evaluate adjustments to stand time at the terminals and trip short-turning. Results showed that layover control were more effective at reducing the occurrence of long headways, while short-turns were more effective at reducing lateness of trips (with respect to schedule).

Andersson et al. (1979) developed an interactive simulation model of an urban bus route in peak hour traffic. It is meant to be used as a training tool for route control operators to learn how how their decisions affect service, and potentially during real-time control as a way of testing different ways to respond to a difficult situation. The model accepts control actions, such as short turning a trip or injecting an extra vehicle, from the user, and displays the evolution of operations thereafter. Running times are generated from fitted theoretical distributions, and considerable effort is spent on validating the choice of distributions and testing for goodness of fit.

Moses (2005) developed a simulation model that uses automatically collected data to model operations, demand, and control mechanisms. While he was able to show that control strategies dealing only with bunched vehicles were not effective, the simulation model was not successfully validated with observations of real operations in terms of headways, running times, and load. The largest source of discrepancy was a stronger propagation of headway irregularities in the simulation, which was possibly due to inappropriate modeling of correlations between running times of adjacent segments, either explicitly or implicitly through dwell time modeling.

Milkovits (2008b) also developed a simulation model for use with automatically collected data to model high-frequency bus operations, demand, and control mechanisms. The model was validated in terms of headways, trip time, and schedule adherence, with bus route 63 in Chicago. A tendency to overestimate large gaps in the second half of the route was observed. Milkovits used his simulation model to evaluate the sensitivity of service reliability to passenger demand, terminal departure deviation, minimum recovery time, percentage of missed trips, and holding strategies.

Cats (2011) presented an extensive literature review on adaptations of traffic simulation models to transit. In a review of the development of traffic assignment simulation models, Peeta and Ziliaskopoulos (2001) found that simulation has advantages over analytical approaches for traffic assignment, including better representation of real networks, capturing interactions between different agents in a system, and incorporating stochasticity (as cited in Cats, 2011). Cats argued that the same is true for simulation of transit operations. In his review of previous work on the field, he found a wide range of approaches toward adapting traffic simulation models to incorporate transit, including a microscopic model which simulates operations in the vicinity of stops for analyzing stop designs, the integration of transit into a microscopic simulation model to evaluate transit signal priority strategies, and an application of microscopic simulation to predict travel time between detection and arrival of a bus at an intersection in operations with transit signal priority. Cats also reviewed the literature on applications of simulation to model transit assignment of passengers to the different routes of a network.

Aside from an extensive literature review, Cats (2011) developed *BusMezzo*, a transit extension to the mesoscopic traffic simulator *Mezzo* that includes modeling of rider path choice decisions. Bus movements are modeled at a mesoscopic level, and propagation of delays is captured through trip chaining. Inputs to the model include vehicle schedules. *BusMezzo* is a network level simulation, and emphasis is placed on demand modeling and transit assignment (of riders among the different routes), with insertions of choice generation routines and discrete choice models for path choice. Each traveler is modeled as an adaptive decision maker that responds to path alternatives and their anticipated downstream attributes. The framework is used to evaluate holding strategies and real-time information provision.

### 1.6.4 Resource Allocation Optimization

In this research, optimization is used to allocate a fixed total amount of resources among a group of routes, without changing their frequencies or operating strategies, in a way that maximizes service performance. Optimization has played an important role in planning, frequency determination, and vehicle and crew scheduling for bus transit (Desaulniers and Hickman, 2007). Two relevant examples are highlighted below. The optimization approach followed in this research differs from past approaches in that it optimizes allocation of a fixed total amount of resources over a group of routes rather than determining the fleet size of each route that maximizes total societal cost.

Furth and Wilson (1981) faced a problem similar to the one considered in this research, but focused on frequencies rather than fleet sizes. (Fleet size and frequency are related through (1.1).) Their mathematical model allocates available buses between time periods and between routes in order to maximize net social benefit subject to constraints on total subsidy, fleet size, and levels of vehicle loading. Their user cost is based on waiting times and does not capture reliability.

Furth and Muller (2007) developed a mathematical model that captures user travel time and reliability costs, as well as operator cost, for low-frequency bus service. Their user cost captures waiting time and running time reliability, but is not applicable to high-frequency service. Their model is based on assumptions of how additional resources are distributed between terminal stand time and en route holding. In contrast, the optimization approach of this thesis assumes that operators do not modify their strategy, but that providing additional resources gives controllers greater opportunity to regulate headways following the current strategy.

## 1.7 Thesis Organization

This chapter defined running time variability, discussed sources of variability, and explained why running time variability is an important determinant of service quality and resource allocation. It laid out the research objectives and approach, and presented a review of scientific literature on the topic.

Chapter 2 elaborates on the definition of running time variability, develops variability measures for homogeneous periods, introduces an aggregation method for capturing the random component of variability in heterogeneous periods, presents visual analysis tools for time of day and segment level variability analysis, and explores factors explaining typical running times and running time variability.

Chapter 3 presents a framework for a simulation model evaluating the performance of bus transit lines under various operating conditions and strategies. A simple model based on this framework is implemented and used to illustrate the required procedures for modeling input from automated data collection system databases and for verifying and validating models.

Chapter 4 develops a framework for optimizing fleet allocation, using a simulation model to estimate performance under hypothetical scenarios and an algorithm to adjust fleet size in small steps such that total performance improves while satisfying all relevant constraints. A simple example is used to illustrate the concept.

Chapter 5 summarizes the thesis and identifies topics for future research.

Two appendices can be found at the end. Appendix A elaborates on an aggregate variability measure presented in Chapter 2, which can be used to obtain a route-level variability summary. The procedure for calculating it is presented. Appendix B introduces data playback as a tool for performance analysis and monitoring. Disaggregate datasets containing, for instance, arrival times at each stop of a bus route, can be played back chronologically to recreate what happened on a particular day. This can be visualized as an animation, or custom triggers can be programmed to generate specialized datasets with observations of any aspect of operations (for example, trip curtailments).

# Chapter 2

# Descriptive Analysis of Running Time Variability

## 2.1 Introduction

Chapter 1 introduced the concept of running time variability and discussed the implications it has on resource requirements. Effective management of running time variability requires good measurement and communication practices. Measuring running time variability in a consistent manner allows not only to discern instances of higher or lower variability, but also to establish logical relationships between variability and resource requirements, and to observe patterns of performance over time. It is equally important to use a meaningful and precise language that allows stakeholders to communicate ideas about running time variability, so that effective management actions can follow from variability measurement. Compare, for instance, the following two statements:

1. "We need more vehicles on route X because running times are highly unpredictable."

2. "While typical running times are largely unchanged, we have observed a steady increase in running time variability on route X over the past three months; in that time the running time *spread* in the afternoon peak outbound direction has increased from 10 to 16 minutes, and from 7 to 9 minutes in the inbound direction. We estimate that an addition of one vehicle, combined with improved dispatching discipline at the terminals, should restore performance to where it was three months ago."

Both statements suggest that an additional vehicle should be added to the fleet of route X in order to remedy high running time variability, but otherwise they are very different. It would be difficult to agree or disagree with the first statement, let alone make decisions about resource allocation, without first knowing the meaning of "highly unpredictable". In contrast, the second statement is based on consistent measurements of running time variability and performance over time. The terminology used to make the argument is precise: *spread* has an established mathematical definition, as discussed later in this chapter.

Measurement and communication practices are necessary for understanding running time variability, but not sufficient. Running time variability could be measured incorrectly, or the wrong kind of variability could be measured, or incorrect terminology could lead to a misunderstanding; any of these would lead to poorer understanding of variability, and ultimately to less frequent, sophisticated, and successful efforts to manage it.

The aim of this chapter is to provide a tool set for descriptive analysis of running time variability in high-frequency transit services. Section 2.2 discusses the dimensions and scales on which running times can be analyzed. Section 2.3 presents measures of running time variability for homogeneous time periods (for example, the morning peak period in the inbound direction) and a method to summarize variability occurring over multiple homogeneous periods through an aggregate index. Descriptive analysis tools based on these measures are discussed in Section 2.4. Finally, Section 2.5 presents an exploratory analysis to identify route characteristics that might contribute to longer running times and running time variability, and Section 2.6 presents concluding remarks.

## 2.2 Dimensions and Scales of Running Time Analysis

Running times can be analyzed over two dimensions, time and space, and both can be analyzed at different scales. Temporal scales include seasons, days of the week, and times of the day. Spatial scales include directions and segments. The scale chosen for analysis should be dictated by the decisions the analysis intends to support. In this research, the key decision is how much resource should be allocated to a service. It is feasible to adjust the amount of resource by time of day, day of week, and season, corresponding to the three temporal scales mentioned above.

Analyzing running times by segment can help identify segments that contribute significantly to overall route variability. For such segments, modifications in stop location, addition of bus lanes, or signal priority schemes could decrease variability and lead to cost savings or improvements in reliability. Having knowledge of segment-level running time variability is also useful in evaluating route revisions. For example, if a segment is being added to a route, it may be useful to know not just how much the average running time will increase, but also how overall running time variability will be affected.

Resource allocation depends on both typical running times and running time variability, so measures of both should be considered. Comparisons of running times across seasons, days of the week, times of the day, directions, and segments should be made not only in terms of typical values but also in terms of variabilities. For example, when analyzing running times for a route, even if typical running times might not change much from one season to the next, a change in their variabilities might be reason to adjust resource allocation.

Any meaningful comparison of running time variability for a scale in any dimension requires holding the scale in the other dimension constant. When examining variability by segment, the time range of the data used for analysis must be consistent across segments. For example, observations of each segment could be from April 2011 workdays from 7:30 to 9:30. The results would not have much meaning if data of April were used for some segments and data of October for others. Similarly, when examining variability by time of day, the data used for analysis must be spatially consistent across times of day. For example, observations of end-to-end trips might be used for all times of the day. The comparison would not be useful if observations of end-to-end running times were used for the trips in the morning and observations of only the first half of the route were used in the afternoon.

## 2.3 Measures of Running Time Variability

Underlying the disposition to quantify running time variability is the premise that there are groups of running time observations that, for all practical purposes, emerge from a single operating environment, that is, a *homogeneous period.* For example, one may speak of the variability observed in the afternoon peak during February weekdays. Running time observations in such a period will form a distribution of running times characterizing the period. This distribution, in turn, will be the model of running times for the period. Therefore, before quantifying variability, the level of aggregation of observations must be determined.

In essence, aggregation trades off detail for simplicity. For instance, suppose one month of observations of running times for a route in one direction were examined. The standard deviation of all the observations would be too aggregate a measure, since operations in the morning peak, afternoon peak, and the rest of the day are probably quite different, each with its own factors that contribute to the running time distribution. A single measure would capture both variability within each time period and variability between time periods. Characterizing each time period individually would be more useful for decision support. On the other hand, the standard deviation of each scheduled trip (over multiple days) could be too disaggregate a measure. If all scheduled trips in the morning peak corresponded to a single operating environment, the results could be simplified (and known with greater statistical significance) by aggregating the observations belonging to the morning peak.

The basis of analysis will be measuring variability within homogeneous periods, aggregating observations from different days of the week and weeks of the year only if they are considered to represent the same operating environment. Accordingly, a statement about running time variability could be “the standard deviation of weekday inbound trips on route 1 in the morning peak during October is 7.3 minutes”. Section 2.3.1 discusses measures of running time variability for homogeneous periods.

Running times of different time periods, days of the week, seasons, segments of a route, or routes come from different operating environments. Therefore, changes in running time can be described by comparing both typical running times and running time variabilities. A statement characterizing running time variability in this manner could be “both the mean and standard deviation of running times of weekday inbound trips in the morning peak during October were greater for route 1 than for route 2”. In some applications it is useful to describe the variability of a route in general, or at least across several time periods. Since it cannot be assumed that these running times belong to a single homogeneous period, variability measures described in Section 2.3.1 are inadequate. Aggregate measures of variability for a combination of homogeneous periods, introduced in Section 2.3.2, should be used instead.

### 2.3.1 Measures of Running Time Variability for Homogeneous Periods

There are many ways of measuring the dispersion of running times belonging to a homogeneous period (which together make a single running time distribution), including mean-based measures like standard deviation and coefficient of variation, and also percentile-based measures such as spread and normalized spread.

Table 2.1 summarizes the four variability measures discussed in this section. Mean-based measures are more traditional than percentile-based measures, and they have distinct mathematical properties that make them the best measures of variability for many applications. The drawback is that there is no straight-forward way of relating these measures to a particular level of reliability without knowing the shape of the running time distribution, which makes their meaning less intuitive than percentile-based measures.

**Table 2.1:** Summary of Variability Measures for Homogeneous Time Periods

| | Absolute<br>(minutes) | Normalized<br>(fraction of typical) |
|---|---|---|
| Mean-Based | Standard Deviation | Coefficient of Variation |
| Percentile-Based | Spread | Normalized Spread |

Since percentile-based measures lack some of the mathematical properties of mean-based measures, they require a large and consistent amount of data to use. If sample sizes are small, running times should be modeled with fitted theoretical distributions and mean-based measures of variability should be used. However, this is increasingly less relevant in transit, and it is certainly not a problem with London Buses's automated data collection systems. Percentile-based measures can be sensitive to sample size, so similar sample sizes should be used when making comparisons. A significant advantage of percentile-based measures is that they can be directly related to a particular level of reliability without knowing the shape of the running time distribution. This makes percentile-based measures more intuitive to use. Percentile-based measures have a visual interpretation not found in mean-based measures.

Variability can be measured in absolute or relative terms. Standard deviation and spread are absolute measures. They quantify variability in units of time (e.g. minutes). Coefficient of variation and normalized spread are relative measures. They quantify variability without units, relative to the typical running time. Absolute variability, and not relative, affects the amount of additional resources required to provide a reliability buffer in transit operations. However, relative measures are useful to compare variability in two sets of running times having different typical running times. For example, a segment with an average running time of 5 minutes may have a standard deviation of 1 minute while another with an average running time of 60 minutes may have a standard deviation of 10 minutes. In absolute terms, the variability is higher on the latter (since $1 < 10$), but relative to the mean, the former is more variable, since $\frac{1}{5} > \frac{10}{60}$.

The four measures of variability presented in Table 2.1 are defined below.

**Standard Deviation** Sample standard deviation is given by the following equation:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{2.1}$$

where $N$ is the sample size, $x_i$ is the $i^{\text{th}}$ observation, and $\mu$, the sample mean, is given by

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.2}$$

Standard deviation uses all observations to describe variability. It is perhaps the most theoretically appealing measure of variability because it is used as a scale parameter for several distributions, including the ubiquitous normal distribution. Being a mean-based measure, however, its meaning is less intuitive than percentile-based spread.

**Coefficient of Variation** The coefficient of variation is the standard deviation normalized by the mean. It is given by the following equation:

$$c_v = \frac{s}{|\mu|} \tag{2.3}$$

The coefficient of variation is used to compare variability of distributions with different means.

**Percentile-Based Spread** Percentile-based spread (or simply *spread*) describes the dispersion of a running time distribution by the difference between a higher and a lower percentile. It may be thought of as the range of a trimmed distribution, in which the top and bottom tails are removed. Percentile-based spread is given by the following equation:

$$\mathrm{S} = p_U - p_L \tag{2.4}$$

where $p_U$ and $p_L$ are upper and lower percentiles of a set of observations belonging to a single homogeneous period.

The choice of percentiles should be wide enough such that most of the spread is captured, but without outliers or observations in the extreme tails of the distribution. In this research, the difference between the $90^{\text{th}}$ and $10^{\text{th}}$ percentile was found to work well. Figure 2-1 shows a sample scatter plot of running times by time of day. A magnified viewing pane depicts various moving percentile lines. (Moving percentile is akin to the better known moving average.) It is seen that the $10^{\text{th}}$ and $90^{\text{th}}$ percentile lines capture most of the spread around the core of running times, without being too sensitive to extremes. Using $90^{\text{th}}$ and $10^{\text{th}}$ percentiles, spread is $\mathrm{S} = p_{90} - p_{10}$. Standard deviation lacks an equivalent visual interpretation. Compared to standard deviation, spread has a more intuitive interpretation of running time variability, even without knowing the shape of the running time distribution.

**Normalized Percentile-Based Spread** Percentile-based spread can be normalized, just like standard deviation is normalized into a coefficient of variation. However, since spread is based on percentiles, it is more consistent to normalize by the median than by the mean.

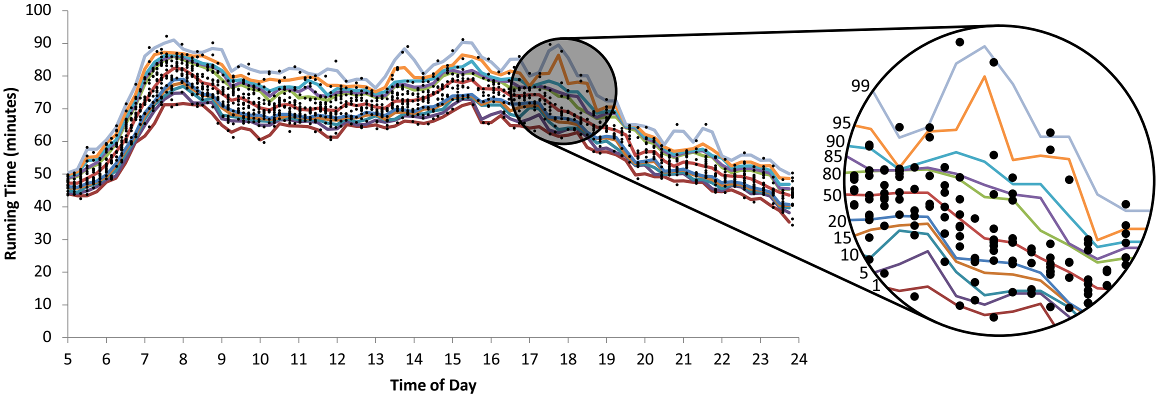


**Figure 2-1:** Running times scatter plot with various percentile lines

Hence, the normalized percentile-based spread is given by the following equation:

$$\tilde{\mathrm{S}} = \frac{p_U - p_L}{p_{50}} \tag{2.5}$$

where $p_U$, $p_L$, and $p_{50}$ are the upper, lower, and 50th percentiles of a set of observations belonging to a single homogeneous period. The tilde on $\tilde{\mathrm{S}}$ denotes normalization. Using 90th and 10th percentiles, normalized spread is $\tilde{\mathrm{S}} = \frac{p_{90} - p_{10}}{p_{50}}$. Again, normalized measures characterize variability relative to typical running times, so they are useful for comparing routes or segments with different typical running times.

### 2.3.2 Aggregate Measures of Running Time Variability

The variability measures discussed above should always be applied to a homogeneous distribution of running times; that is, one that emerges from a particular operating environment, with approximately stationary typical running times and variabilities. Also, they should be applied to a single direction of travel or a single segment within this direction. While this is the basis of variability measures, sometimes it is useful to speak of the variability of a route (in both directions) during one or more time periods. It may also be desirable to characterize running time variability during heterogeneous periods, such as transitions from a morning peak period to the middle of the day.

It is incorrect to pool observations of heterogeneous periods or of multiple homogeneous periods into a single distribution; variability measures calculated from there would not in most cases be representative of the operations during any included period. Figure 2-1, for example, shows running times by time of day on one of the directions of route W15 in London. Observations in the graph were made on weekdays from March 4, 2011 to March 31, 2011. By visual inspection, running times from 7:30 to 8:30 belong to a homogeneous morning peak period, and those from 9:30 to 13:30 belong to a homogeneous mid-day period.

Basic statistics for the two periods are shown in the first two data columns of Table 2.2. The last column shows the same statistics calculated after pooling all observations of both periods. Neither the mean of 73.03 minutes, nor the standard deviation of 5.82 minutes, nor the spread of 15.24 minutes are good indicators of performance in either of the periods. Moreover, the variability measures for the pooled data indicate higher variability than in either of the periods because they are capturing both within-period and between-period variabilities. The two types of variability have very different implications for service planning.

**Table 2.2:** Basic running time statistics for one direction of route W15, London

| | 7:30–8:30 | 9:30–13:30 | pooled |
|---|---:|---:|---:|
| minimum (minutes) | 70.55 | 59.68 | 59.68 |
| maximum (minutes) | 92.20 | 84.60 | 92.20 |
| mean (minutes) | 80.48 | 70.86 | 73.03 |
| standard deviation (minutes) | 4.56 | 4.09 | 5.82 |
| $90^{\text{th}}$ percentile (minutes) | 85.95 | 76.52 | 81.83 |
| $10^{\text{th}}$ percentile (minutes) | 74.37 | 66.30 | 66.59 |
| spread (minutes) | 11.58 | 10.22 | 15.24 |

Between-period variability in running times is usually due to a systematic change in operations. For example, as the morning peak subsides, congestion typically diminishes, which leads to lower running times. In addition, ridership typically also diminishes, which leads to shorter dwell times and less frequent stops. The service plan should reflect these systematic changes and lower the fleet size requirement for the mid-day period. In contrast, within-period variability is of a more random nature. Even though it is sometimes possible to explain why one takes longer than another a posteriori at the microscopic level, it is practically impossible to anticipate these variations in a particular sequence. The service plan recognizes this by providing a reliability buffer time in the form of layover (stand) time at the terminals, which translates to a larger fleet.

Since the two types of variability are managed differently, the aggregate variability measure should have two properties. First, it should be sensitive to within-period stochasticity (the focus of this research), which is related to reliability buffer time. Second, it should not be sensitive to between-period systematic variation, which can be accounted for in scheduling. This introduces a challenge: homogeneous time periods must be identified, but time period analysis is often subjective and difficult to automate, and different routes may have different period definitions.

The method used in this research overcomes this difficulty by defining a regular sequence of short, overlapping periods, or windows, which are assumed homogeneous. Violations to the homogeneity assumption are largely inconsequential because periods are short. One of the variability measures for homogeneous time periods is calculated for each of these windows, and the arithmetic mean of these is calculated. This can be done for one or both directions, for one or more time periods. Using $v(t)$ to denote some measure of variability at time $t$, the aggregate variability measure $V$ can be expressed as follows:

$$V = \frac{\int_{t_i}^{t_f} v(t)dt}{t_f - t_i} \tag{2.6}$$

where $t_i$ and $t_f$ are the beginning and ending times of the analysis period. Figure 2-2 illustrates the concept with hypothetical running times. The aggregate variability from 11:00 to 14:00 is being calculated using a regular sequence of short periods, in this case 30

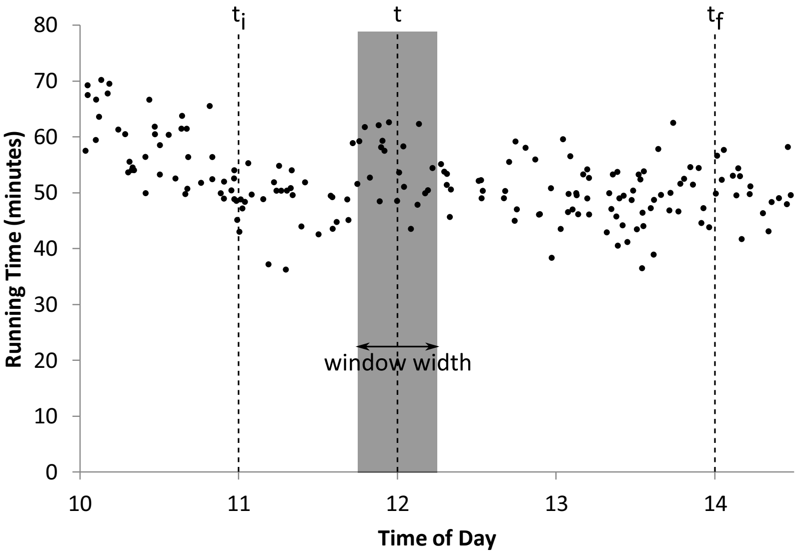


**Figure 2-2:** Method for calculating aggregate variability measures

minutes wide. The first of these is centered at 11:00 (pooling observations from 10:45 to 11:15). The window is then scrolled smoothly until it is centered at 14:00. The aggregate variability measure is the mean of all window variabilities.

If it is desired to obtain an aggregate variability measure for the entire day, the limits of integration $t_i$ and $t_f$ can be chosen to match the times at which service of a particular route begins and ends, respectively. Alternatively, $t_i$ and $t_f$ can be chosen to cover the most representative period of a service day, constant across all routes for more valid comparisons. For example, we could set $t_i$ = 7:30 and $t_f$ = 19:30. Any of the variability measures for homogeneous time periods, absolute or normalized, may be used as $v(t)$.

In the implementation of this measure, the integral of (2.6) is replaced by a summation. Tests showed that using a window size of 30 minutes shifted every 15 minutes works well. The following are defined in order to write an equation for the discretized aggregate variability measure:

- $D = \{d_1, d_2\}$ is the set of directions of a route.
- $W$ is a set of sequential, short, overlapping windows of observations.
- $n_W = |W|$ is the number of windows.
- $T_{d,w}$ is a set of observed end-to-end running times of trips in direction $d$ beginning at times contained in window $w$.
- $v(T_{d,w})$ is a measure of running time variability on a set of observations $T_{d,w}$.

The aggregate variability measure for running times in direction $d$ is given by the following equations:

$$V = \frac{1}{n_W} \sum_{w \in W} v(T_{d,w}) \tag{2.7}$$

To obtain an aggregate variability measure for both directions, the sum of the variabilities in each direction is used:

$$V = \frac{1}{n_W} \sum_{w \in W} \sum_{d \in D} v(T_{d,w}) \tag{2.8}$$

Aggregate variability measures based on spreads for both directions are obtained as follows:

$$V = \frac{1}{n_W} \sum_{w \in W} \sum_{d \in D} \left(p_{90}(T_{d,w}) - p_{10}(T_{d,w})\right) \tag{2.9}$$

where $p_{90}(T_{d,w})$ and $p_{10}(T_{d,w})$ are the $90^{\text{th}}$ and $10^{\text{th}}$ percentile values of a set of observations $T_{d,w}$. This aggregate measure will hereafter be called *mean spread* (MS).

The normalized version is given by the following equation:

$$\tilde{V} = \frac{1}{n_W} \sum_{w \in W} \sum_{d \in D} \frac{p_{90}(T_{d,w}) - p_{10}(T_{d,w})}{p_{50}(T_{d,w})} \tag{2.10}$$

where $p_{50}(T_{d,w})$ is the $50^{\text{th}}$ percentile value of a set of observations $T_{d,w}$. The tilde on $\tilde{V}$ is used to denote normalization. This aggregate measure will hereafter be called *normalized mean spread* (NMS).

Often it is desirable to summarize the running time variability of a route with a single figure. The *diurnal mean spread* (DMS) is defined for this purpose. It is simply the mean spread calculated for both directions of a route during the hours of the day (for instance, starting at 7:30 and ending at 19:30). Although this figure is too aggregate to set resource levels by time period, it is useful to classify routes by their running time variability and to track general variability over time. To ensure that the variability measures are consistent for comparisons, DMS should always be calculated with the same parameters; for example, it may be calculated on observations of 15 successive weekdays from 7:30 to 19:30, using observation windows spaced 15 minutes apart, each 30 minutes wide. An algorithm for computing DMS is presented in Appendix A.

The set of measures presented in this section can be used to quantify variability at many aggregation levels, from a single route-direction-period to the overall DMS that summarizes a route's variability in both directions during the day. The latter is proposed as a screening tool, but it alone does not provide enough information to make resource allocation decisions. Table 2.3 presents mean spreads at the direction and time period level for bus route W15 in London, on weekdays from 2011-03-04 to 2011-03-31. The DMS is shown in the lower right. Similar information could be presented at the segment level.

## 2.4 Visual Analysis Tools

Information about running time variability can be visually organized in order to communicate both aggregate and detailed information quickly. Two visual analysis tools were developed. The first consists of running time scatter plots for each direction of a route as

**Table 2.3:** Summary of Aggregate Running Time Spreads for Route W15

| Direction | morning peak<br>7:30–9:30 | mid-day<br>9:30–15:30 | afternoon peak<br>16:30–18:30 | evening<br>18:30–20:30 | all day<br>7:30–19:30 |
|---|---|---|---|---|---|
| 1 | 11.1 | 10.0 | 10.5 | 9.3 | 10.2 |
| 2 | 9.5 | 11.7 | 25.3 | 14.6 | 14.6 |
| both | 20.7 | 21.7 | 35.8 | 23.9 | 24.8 |

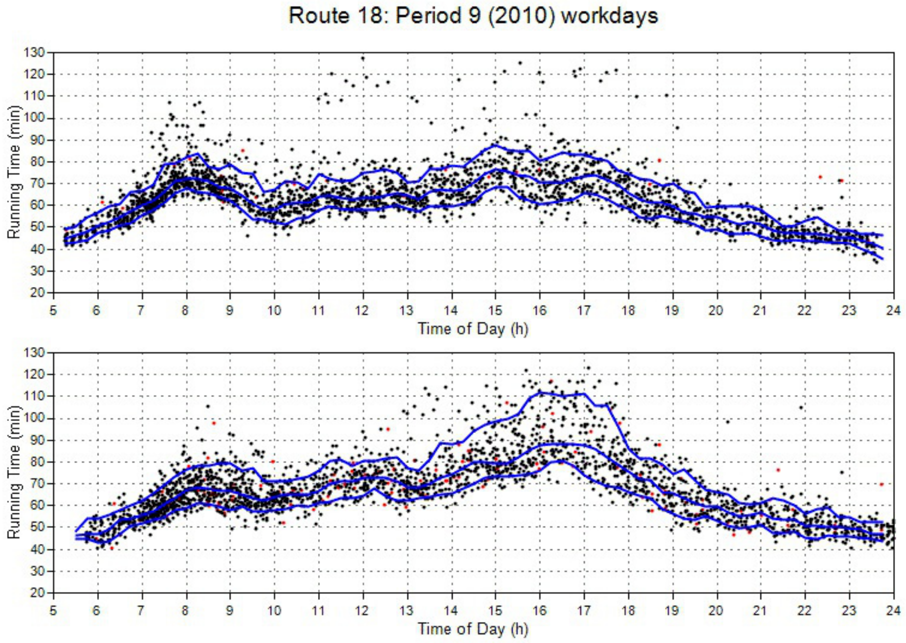


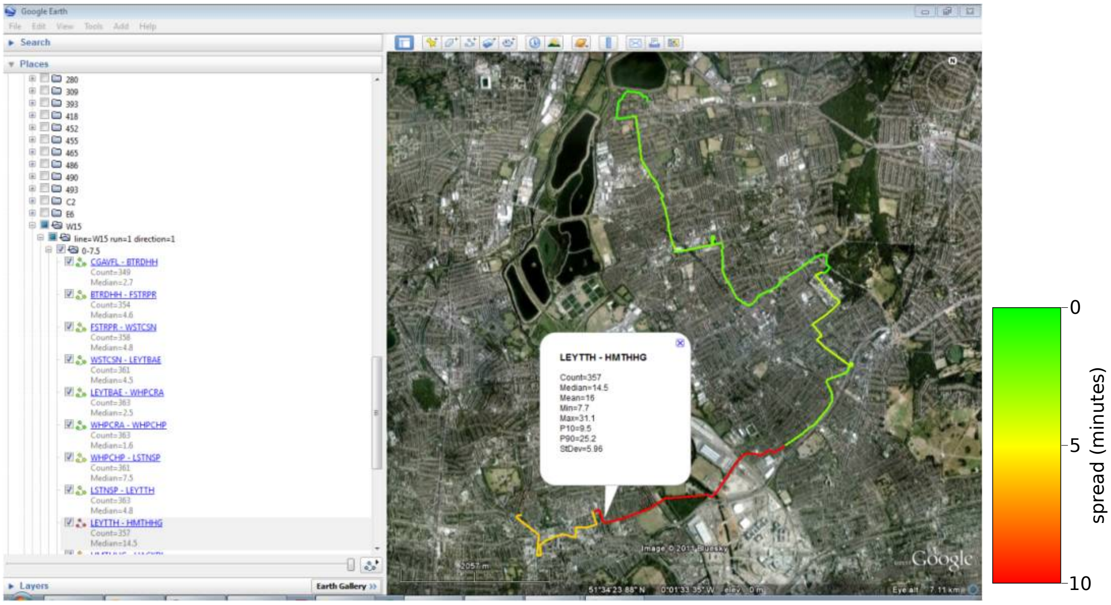
**Figure 2-3:** Example of running time scatter plots with running percentile lines

a function of time of day, showing moving 10th, 50th, and 90th percentiles, an example of which is shown in Figure 2-3.

Comparing several of these graphs of a single route for different seasons or of several routes for the same season enables visualizing how running times and their spreads (represented by the width of the band between the 10th and 90th percentile lines) change seasonally or from route to route. Depending on the motivation driving analysis, additional attention may be given to the extent of peaking, the choice of time periods, or the transitions between periods. Moreover, similar graphs can be made at the segment level.

The second analysis tool consists of plots of a pattern[1] on a map or aerial picture, coloring segments according to the performance measure of interest. Coloring according to the aggregate *mean spread* has the advantage that only the random component (and not the systematic component) of running time variability is captured, irrespective of how time periods are defined. This concept was tested using Google Earth KML layers. An example

[1] A pattern is a longitudinal alignment of a planned service with a sequence of segments, turns, and stops which vehicles follow from the beginning terminal to the ending terminal. Typically routes have two patterns, one for each direction, but some routes may have additional patterns in order to accommodate variations such as school trips and scheduled short-turns.



280
309
393
418
452
455
465
486
490
493
C2
E6
W15
line=W15 run=1 direction=1
0-7.5

CGAVFL - BTRDHH
Count=349
Median=2.7

BTRDHH - FSTRPR
Count=354
Median=4.6

FSTRPR - WSTCSN
Count=358
Median=4.8

WSTCSN - LEYTBAE
Count=361
Median=4.5

LEYTBAE - WHPCRA
Count=363
Median=2.5

WHPCRA - WHPCHP
Count=363
Median=1.6

WHPCHP - LSTNSP
Count=361
Median=7.5

LSTNSP - LEYTTH
Count=363
Median=4.8

LEYTTH - HMTHHG
Count=357
Median=14.5

**LEYTTH - HMTHHG**

Count=357
Median=14.5
Mean=16
Min=7.7
Max=31.1
P10=9.5
P90=25.2
StDev=5.96

Source: Google Earth

**Figure 2-4:** Example of route path segments colored according to percentile-based spread

is shown in Figure 2-4. Clicking on a segment opens a balloon with relevant statistics.

This visual illustrates running time variability per segment for a particular time period and direction of a route. The color gradient is green on one end, for segments with no variability, and red on the other, for segments with high variability. The transition from green to yellow to red is linear on *mean spread*. Tests showed that setting the saturation at the red end of the scale to ten minutes of spread yields a good balance of segment colors; five minutes tends to produce too many red segments, and twenty tends to produce too many green segments. Nevertheless, it is a parameter that can be changed according to the particularities of the task at hand.

It may be useful to color segments according to other performance measures, such as the percentage of scheduled trips observed, with green indicating that the number of trips observed equals or exceeds the number of planned trips, and red indicating a lower number was observed, which could happen at route ends if an operator often curtails trips before reaching the terminal. Segments could be also colored according to expected or excess waiting time.

## 2.5 General Patterns of Running Time Variability

An exploratory analysis of running times was carried out to identify factors contributing to higher or more variable running times. An understanding of running time patterns could be useful for service planning. For example, when evaluating a potential extension of a route, knowing typical running times and what variability to expect in the added segments (according to their characteristics) would help obtain good estimates of resource requirements, especially when the agency has no experience operating on the new segment.

Various linear models of either median running time or running time variability as a function of route characteristics were specified, and their parameters estimated on running time observations of a representative sample of 41 bus routes in London, which included radial and circumferential routes, short and long routes, as well as routes that operate smoothly and others that are more problematic. AVL records containing departure and arrival timestamps for each stop of each trip during three four-week periods of the year were obtained from the iBus database:

- Summer, 2010-07-24 to 2010-08-20
- Fall, 2010-11-13 to 2010-12-10
- Spring, 2011-03-05 to 2011-04-01

Weekends and holidays were excluded. Analysis procedures for parsing the files, filtering data, identifying trips, fixing errors in data, calculating statistics, and generating output were implemented using the .NET Framework. A single set of time periods was chosen in order to compare route performance consistently:

- Start of service to 7:30
- Morning peak, from 7:30 to 9:30
- Mid-day, from 9:30 to 15:30
- Transition to afternoon peak, from 15:30 to 16:30
- Afternoon peak, from 16:30 to 18:30
- Evening, from 18:30 to 20:30
- Night, from 20:30 to 23:00
- 23:00 to end of service

Patterns were broken into segments defined by timing points, since these are available in the stop lists of the database, making it possible to associate running times to their segments without geospatial calculations.

Several approaches exploring general patterns of typical running times and running time variability were followed, including (1) a visual inspection of route maps with segments colored according to running time variability, (2) a comparison of mean spread statistics by season, (3) a linear model of typical peak running time as a function of typical mid-day running time, (4) a linear model of mean spread as a function of route characteristics, season, and operator, (5) a linear model of median running time at the period level as a function of route characteristics, and (6) linear models of spread at the period level as a function of route characteristics. Each of these is discussed below.

### 2.5.1 Running Time Variability by Segment and Time of Day (Visual)

Running time variability in the London Buses network varies greatly from route to route. Figures 2-5 through 2-10 show segment variabilities of several routes for the morning peak, mid-day and afternoon peak during the spring, with colors according to percentile-based spread. The color scale used is the one shown in Figure 2-4.

While morning peak variability is more prominent in the inbound direction, mid-day and afternoon peak variability is similar in both directions. No strong geographical pattern is immediately apparent. Segments with high running time spreads are dispersed throughout London. Further analysis would be required to understand the factors driving variability at the segment level. For example, the effect of roadway geometry, availability of bus lanes, stop density, passenger flow through stops, intersections, and traffic could be studied.

### 2.5.2 Aggregate Mean Spreads by Season

Table 2.4 shows descriptive statistics of route aggregate mean spreads for the three seasons studied. The variability measures were calculated using (2.9) with 30 minute windows spaced at 15 minute intervals, from the beginning to the end of service[2]. Consistent with a-priori beliefs of London Buses staff, mean spreads are lowest in the summer and highest in the fall. An ANOVA f-test for the three periods yielded a p-value of $2.8 \times 10^{-4}$, confirming that the scores of the three periods are statistically different. A t-test for the fall and spring periods yielded a two-tailed p-value of $7.4 \times 10^{-5}$, confirming that these two are also statistically different.

**Table 2.4:** Aggregate Mean Spread Statistics by Season

| Statistic | Summer | Fall | Spring |
|---|---:|---:|---:|
| minimum | 11 | 14 | 12 |
| median | 17.5 | 20 | 19 |
| maximum | 25 | 32 | 31 |
| mean | 17.58 | 21.66 | 19.24 |
| standard deviation | 3.32 | 5.08 | 4.68 |

All statistics are reported in minutes.

### 2.5.3 Peak to Mid-day Running time Relationship

The following linear model was specified to study the relationship between typical running times in the middle of the day and those in the morning and afternoon peaks:

$$\text{MEDIAN\_PEAK} = \beta_1 + \beta_2 \text{MEDIAN\_MD} \tag{2.11}$$

Parameters $\beta_1$ and $\beta_2$ were estimated separately for the morning and afternoon peaks. Regression results are shown in Table 2.5. $\bar{R}^2$ are approximately 0.90 or greater, confirming strong linear relationships. All $\beta_2$ coefficients are close to 1, and t-tests confirm that they are not statistically different from 1 at the 10% significance level, which suggests that, *on average*, the two-way morning peak and afternoon peak median running times are a constant amount $\beta_1$ above the mid-day running times.

The signs and magnitudes of the $\beta_1$ coefficients suggest the following:

[2] This aggregate variability measure could lead to inconsistent route-to-route comparisons because routes that operate more hours of service during the night might have lower mean spreads. DMS can be used for consistent comparisons. An underlying assumption in this analysis is that spans of service did not change from season to season, so season-to-season comparisons of aggregate variability are consistent.

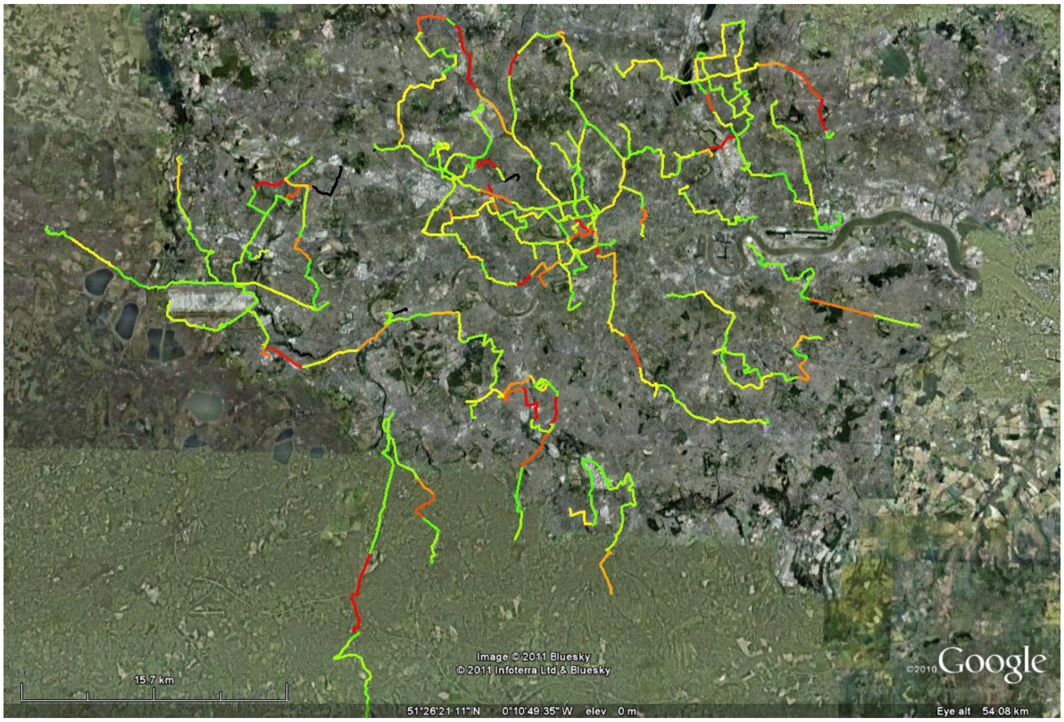


Image © 2011 Bluesky
© 2011 Infoterra Ltd & Bluesky
15.7 km
51°26'21.11" N 0°10'49.35" W elev 0 m
Eye alt 54.08 km
©2010 Google

Source: Google Earth

**Figure 2-5:** Running time variability per segment, morning peak inbound

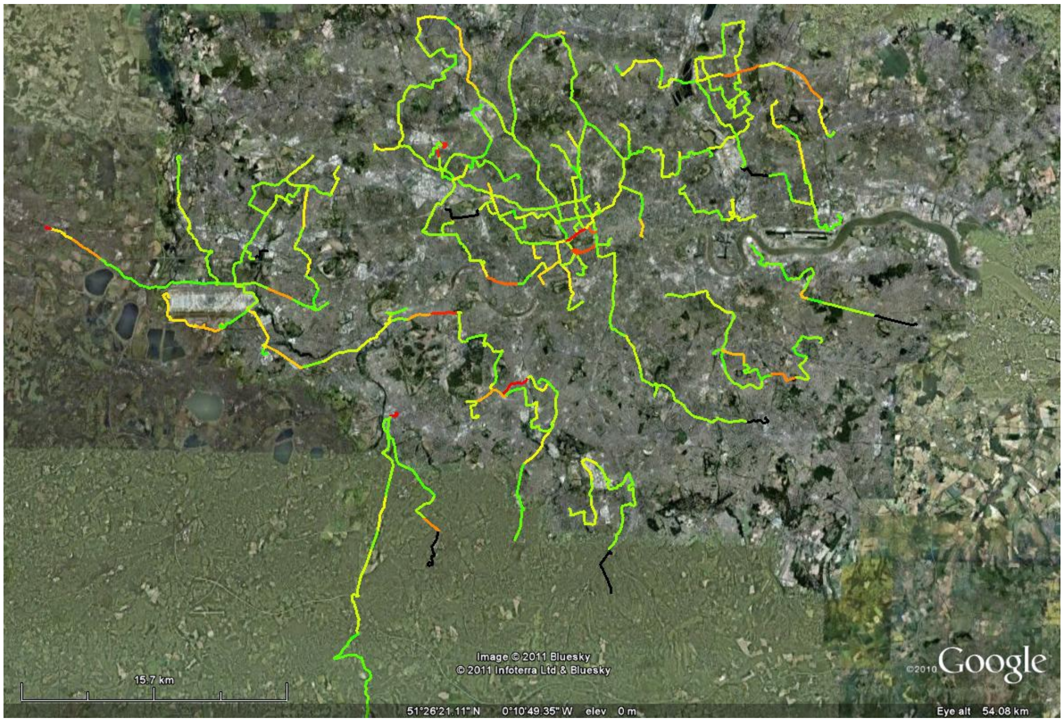


Image © 2011 Bluesky
© 2011 Infoterra Ltd & Bluesky
15.7 km
51°26'21.11" N 0°10'49.35" W elev 0 m
Eye alt 54.08 km
©2010 Google

Source: Google Earth

**Figure 2-6:** Running time variability per segment, morning peak outbound

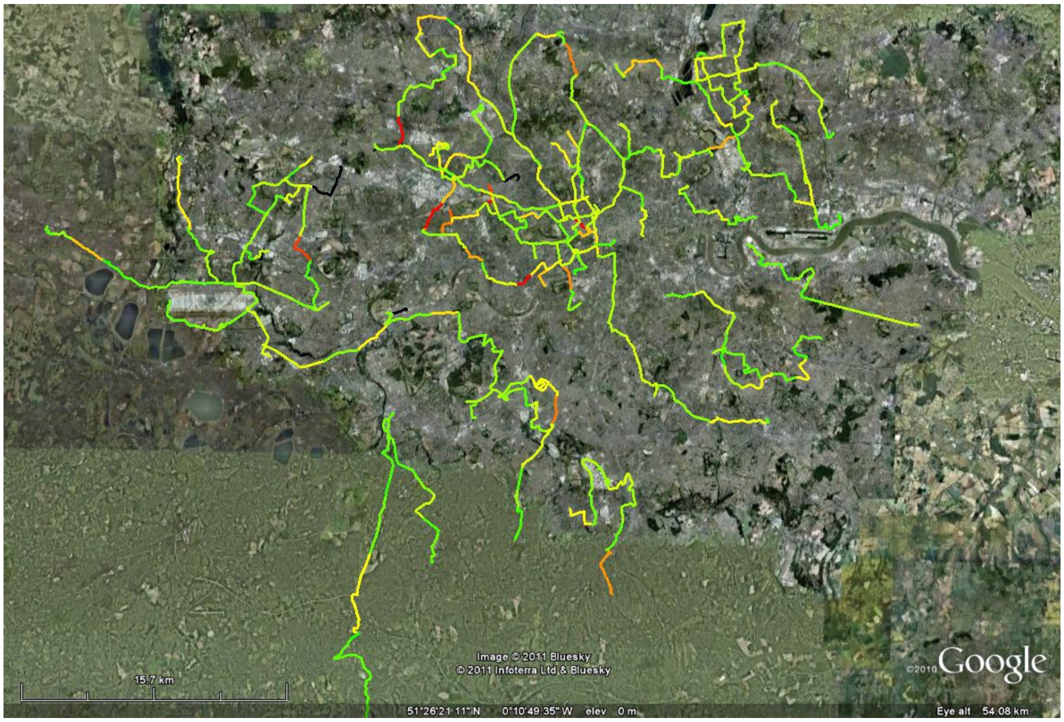


Image © 2011 Bluesky
© 2011 Infoterra Ltd & Bluesky
©2010 Google
15.7 km
51°26'21.11" N 0°10'49.35" W elev 0 m
Eye alt 54.08 km

Source: Google Earth

**Figure 2-7:** Running time variability per segment, mid-day inbound

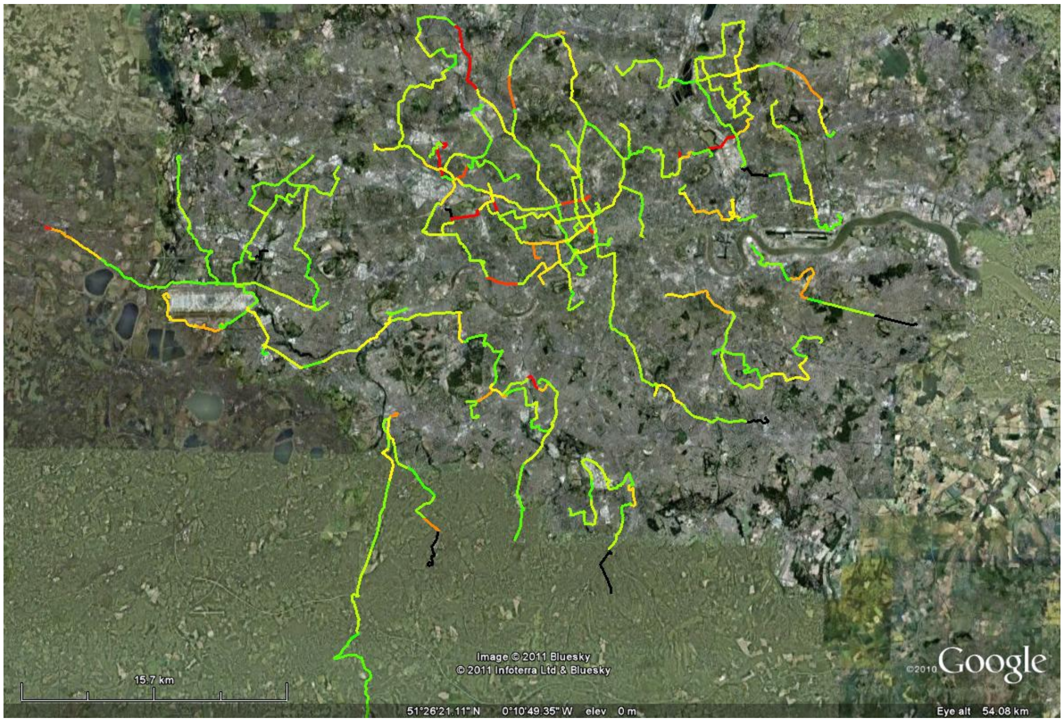


Image © 2011 Bluesky
© 2011 Infoterra Ltd & Bluesky
©2010 Google
15.7 km
51°26'21.11" N 0°10'49.35" W elev 0 m
Eye alt 54.08 km

Source: Google Earth

**Figure 2-8:** Running time variability per segment, mid-day outbound

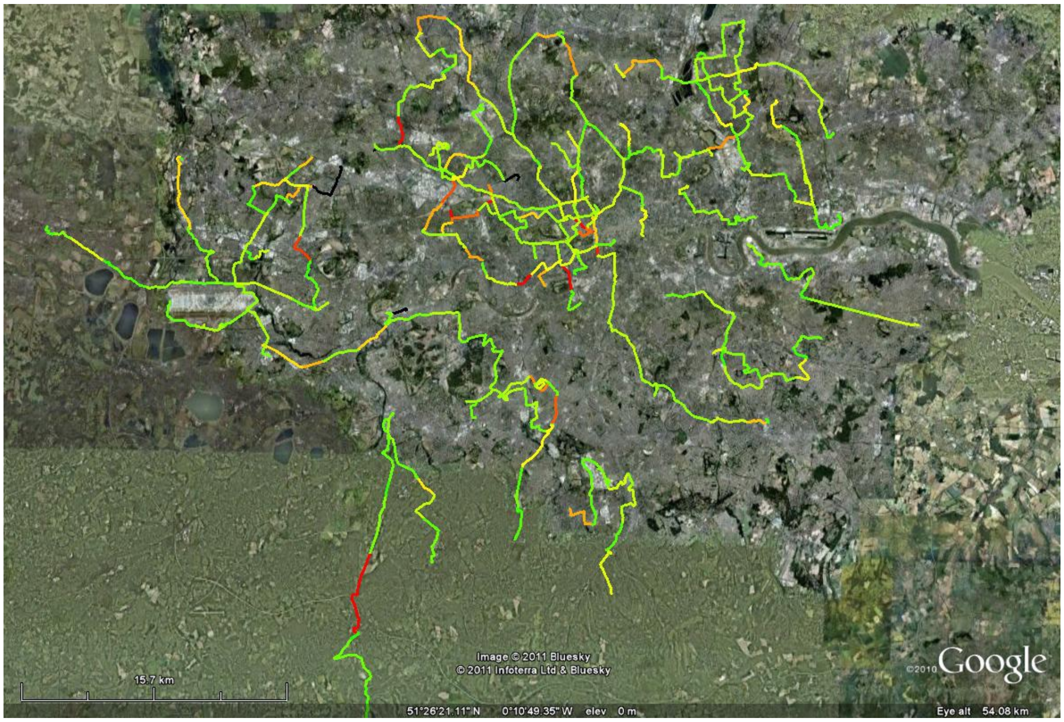


Image © 2011 Bluesky
© 2011 Infoterra Ltd & Bluesky
15.7 km
©2010
Google
51°26'21.11" N 0°10'49.35" W elev 0 m
Eye alt 54.08 km

Source: Google Earth

**Figure 2-9:** Running time variability per segment, afternoon peak inbound

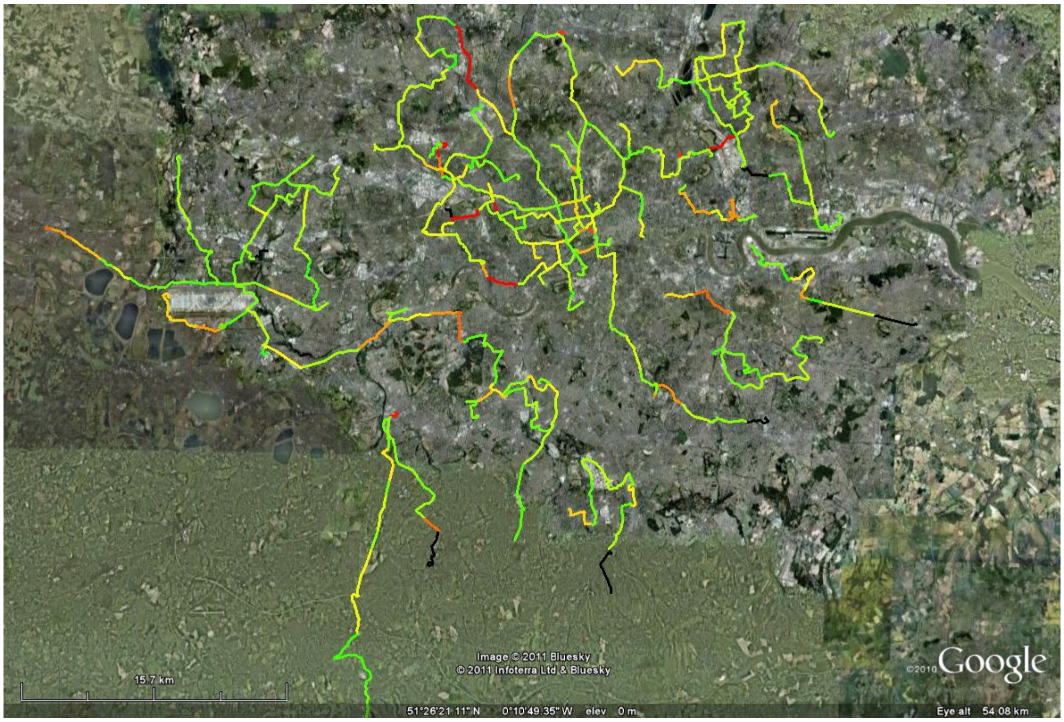


Image © 2011 Bluesky
© 2011 Infoterra Ltd & Bluesky
15.7 km
©2010
Google
51°26'21.11" N 0°10'49.35" W elev 0 m
Eye alt 54.08 km

Source: Google Earth

**Figure 2-10:** Running time variability per segment, afternoon peak outbound

**Table 2.5:** Summary of Regression Results for Peak to Mid-day Running Times

| | AM to mid-day | | | | PM to mid-day | | | |
|---|---|---|---|---|---|---|---|---|
| | Summer | Fall | Spring | all | Summer | Fall | Spring | all |
| $\beta_1$ | -1.856 | 9.755 | 6.991 | 3.414 | 3.378 | 4.663 | 5.407 | 4.252 |
| $\beta_2$ | 0.993 | 0.960 | 0.981 | 0.991 | 1.029 | 1.034 | 1.013 | 1.027 |
| $\bar{R}^2$ | 0.949 | 0.899 | 0.931 | 0.909 | 0.962 | 0.953 | 0.962 | 0.959 |

- The three seasons have different peaking characteristics.
- $\beta_1$ is positive except for the morning peak of the summer, meaning that morning peak two-way running times are higher than mid-day two-way running times in these periods.
- $\beta_1$ is negative for the morning peak of the summer, meaning that morning peak two-way running times are lower than mid-day two-way running times in the summer, possibly an effect of school holidays.
- $\beta_1$ is higher for the fall than the spring in the peaks, consistent with TfL's experience that the fall season is a more challenging operating environment.

### 2.5.4 Linear Model of Mean Spread

A second linear model was specified relating aggregate mean spread to two-way route distance and dummy variables for seasons, radial routes, routes with at least a part in a central business district (CBD), and operators. Spread was computed using (2.9) with 30 minute windows at 15 minute intervals, covering the whole span of service of each route. A group of dummy variables was used to identify operators. In this model, all operators but one have a corresponding dummy variable; one must be left out in order to estimate parameter values. Estimated values of dummy variable coefficients are relative to the operator left out, which is used as the basis of comparison. The arbitrary choice was to use Selkant as the base operator.

Regression statistics can be found in Table 2.6. The signs of parameter estimates for period dummy variables, route distance, the radial route dummy variable, and the CBD dummy variable suggest that longer routes, radial routes, and CBD routes are more variable on average. Most of the parameter estimates are statistically significant; the parameter estimate of distance is not very significant, but it was left in the model because experience shows it is a relevant factor. Moreover, the parameter estimates of the operator dummy variables are significant as a group. Nonetheless, $\bar{R}^2 \approx 0.51$ suggests that a large random component has not been captured by the variables.

The interpretation of the parameter estimates for the operator dummy variables is that, for the routes considered by this model, London Central and Quality Line operate with the least variability, while Arriva, Abellio, and London United operate with the greatest variability. Great care must be exercised in taking this interpretation further and arriving at conclusions concerning the skill with which services are run, as the model is based on a small sample of routes. In addition, even if the dataset contained all routes, it is possible that some operators exhibit greater average running time variability because they operate

**Table 2.6:** Linear Regression on Aggregate Mean Spread

| Coefficient | Estimate | Std. Error | $t$ value | Pr(> $\|t\|$) | |
|---|---|---|---|---|---|
| (Intercept) | 10.49 | 1.81 | 5.79 | 7.16E-8 | *** |
| SUMMER | -1.88 | 0.74 | -2.53 | 1.28E-2 | * |
| FALL | 2.40 | 0.73 | 3.27 | 1.43E-3 | ** |
| DISTANCE | 6.71E-5 | 5.21E-5 | 1.29 | 2.01E-1 | |
| RADIAL | 2.57 | 0.92 | 2.80 | 6.07E-3 | ** |
| CBD | 2.49 | 0.99 | 2.52 | 1.32E-2 | * |
| OP_Abellio | 6.03 | 1.33 | 4.53 | 1.56E-5 | *** |
| OP_Arriva | 7.77 | 1.28 | 6.09 | 1.84E-8 | *** |
| OP_EastLondon | 3.80 | 1.53 | 2.48 | 1.49E-2 | * |
| OP_First | 4.61 | 1.53 | 3.01 | 3.22E-3 | ** |
| OP_LondonCentral | -2.08 | 2.19 | -0.95 | 3.45E-1 | |
| OP_LondonGeneral | 4.72 | 1.67 | 2.83 | 5.58E-3 | ** |
| OP_LondonUnited | 6.22 | 1.42 | 4.40 | 2.64E-5 | *** |
| OP_Metrobus | 0.11 | 2.88 | 0.04 | 9.71E-1 | |
| OP_Metroline | 4.52 | 1.12 | 4.02 | 1.10E-4 | *** |
| OP_QualityLine | -1.43 | 2.70 | -0.53 | 5.97E-1 | |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 3.318 on 106 degrees of freedom
$R^2$: 0.5677, $\bar{R}^2$: 0.5066
F-statistic: 9.282 on 15 and 106 DF, p-value: 1.839E-13

more challenging routes. For example, there might be correlation between the operator dummy variables and proxies for difficulty of operations, such as the CBD dummy. Since the dependent variable is mean spread over the whole span of service of each route, operators having routes with more night hours might have lower mean spreads. (See footnote 2.) Using DMS as the dependent variable would yield more consistent results.

### 2.5.5 Linear Model of Median Running Time at the Period Level

A linear model was specified with median running time as a function of distance, number of stops, average number of boardings per trip, and dummy variables for routes entering the congestion charging zone in central London. The model was estimated on data aggregated at the direction and time period level (in contrast to the previous model, which was aggregated at the route level). Rather than using a single CBD dummy variable (as in the previous model), dummy variables for to-, from-, and through- the Congestion Charging Zone were used. Electronic Ticketing Machine (ETM) data was used to obtain the average number of boardings per trip. At the time this data was retrieved, it was available only for the spring season, so the model was estimated using only spring data and so seasonal factors were excluded from the model.

Table 2.7 shows the regression results for this model. The signs of all the parameter estimates agree with general experience: longer running times are related to routes going to, coming from, and passing through the congestion charging zone, greater run distances, greater number of stops, and greater ridership. The statistical significance of the estimates is very high, and the overall fit, with $\bar{R}^2 \approx 0.74$, is moderately high. The average number of boardings per trip is a better characterization of how busy the operating environment is in comparison to morning peak and afternoon peak dummy variables (not included in the model) because it captures pattern-specific characteristics. For example, ridership is related to dwell time and is perhaps a proxy for traffic along the corridor.

**Table 2.7:** Linear Regression on Median Running Time

| Coefficient | Estimate | Std. Error | $t$ value | Pr($>\|t\|$) | |
|---|---|---|---|---|---|
| (Intercept) | 6.34 | 2.33 | 2.72 | 7.05E-03 | ** |
| FromCBD | 5.33 | 1.34 | 3.97 | 9.42E-05 | *** |
| ToCBD | 9.59 | 1.33 | 7.20 | 8.19E-12 | *** |
| ThruCBD | 11.95 | 2.71 | 4.40 | 1.63E-05 | *** |
| Distance | 1.32E-3 | 1.87E-4 | 7.07 | 1.75E-11 | *** |
| Nstops | 0.53 | 0.09 | 6.01 | 6.86E-09 | *** |
| Boardings | 0.13 | 0.02 | 7.58 | 7.78E-13 | *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 6.508 on 236 degrees of freedom
$R^2$: 0.7482, $\bar{R}^2$: 0.7418
F-statistic: 116.9 on 6 and 236 DF, p-value: < 2.2E-16

### 2.5.6 Linear Models of Spread at the Period Level

The effect of the variables in the previous model on running time variability was also studied. A linear model was specified relating percentile-based spread at the time period level to distance, number of stops, average number of boardings per trip, and dummy variables for routes entering the congestion charging zone in central London. Spread was computed using (2.4)[3].

Table 2.8 shows regression results for this linear model. Once again, the signs of the parameter estimates agree with general experience: running time variability is more predominant in runs that enter the congestion charging zone, in runs of greater distance and greater number of stops, and when ridership is higher. Parameter estimates related to the congestion charging zone are statistically significant, while the others are not. An $\bar{R}^2 \approx 0.28$ is indicative of a poor linear fit.

Regression results for a second model of spread at the period level are shown in Table 2.9. This linear model explores the effect of average number of boardings per trip, median running time, median running speed, and dummy variables for the congestion charging zone on spread. All parameter estimates are statistically significant, and the linear fit is better than in the previous model ($\bar{R}^2 \approx 0.43$), but is still not high.

The signs of the parameter estimates for congestion charge dummy variables, median running time, and median running speed agree with general experience, while the sign of the parameter estimate for average number of boardings is counterintuitively negative. However, it is low in magnitude with respect to median running time. A possible interpretation is that spread is more strongly influenced by median running time, but for a given running time (which is influenced by distance and number of boardings), variability decreases

[3] Aggregate mean spread (MS) could have been used instead to isolate the within-periods variability.

**Table 2.8:** Linear Regression on Spread

| Coefficient | Estimate | Std. Error | $t$ value | Pr($>|t|$) | |
|---|---:|---:|---:|---|---|
| (Intercept) | 7.28 | 1.03 | 7.06 | 1.90E-11 | *** |
| FromCBD | 2.67 | 0.59 | 4.49 | 1.12E-05 | *** |
| ToCBD | 4.21 | 0.59 | 7.13 | 1.19E-11 | *** |
| ThruCBD | 6.16 | 1.20 | 5.13 | 6.09E-07 | *** |
| Distance | 8.90E-5 | 8.28E-5 | 1.08 | 2.83E-01 | |
| Nstops | 0.01 | 0.04 | 0.17 | 8.69E-01 | |
| Boardings | 0.01 | 0.01 | 1.49 | 1.36E-01 | |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.881 on 236 degrees of freedom
$R^2$: 0.2955, $\bar{R}^2$: 0.2776
F-statistic: 16.5 on 6 and 236 DF, p-value: 7.294E-16

**Table 2.9:** Linear Regression on Spread

| Coefficient | Estimate | Std. Error | $t$ value | Pr($>|t|$) | |
|---|---:|---:|---:|---|---|
| (Intercept) | 9.76 | 1.45 | 6.74 | 1.23E-10 | *** |
| FromCBD | 1.50 | 0.54 | 2.77 | 5.99E-03 | ** |
| ToCBD | 2.36 | 0.57 | 4.14 | 4.87E-05 | *** |
| ThruCBD | 3.87 | 1.10 | 3.51 | 5.30E-04 | *** |
| Boardings | -0.02 | 0.01 | -2.56 | 1.10E-02 | * |
| RTmedian | 0.11 | 0.01 | 7.57 | 8.63E-13 | *** |
| Distance/RTmedian | -0.02 | 4.37E-3 | -4.78 | 3.09E-06 | *** |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.552 on 236 degrees of freedom
$R^2$: 0.4474, $\bar{R}^2$: 0.4333
F-statistic: 31.84 on 6 and 236 DF, p-value: $<$ 2.2E-16

slightly both with median running speed and number of boardings; this is possible if routes with higher ridership tend to have corridors with better road geometry (for example, more bus lanes), leading to slightly lower variabilities.

### 2.5.7 Other Models and Future Work

Other model specifications were tested, which suggest that seasonal running time variability is significant and that distance, number of stops, ridership, and servicing the congestion charging zone tend to increase both median running time and running time variability. Unfortunately, it was not possible to obtain very good linear fits that relate running time variability to general route characteristics. The available data do not include information on important factors such as weather conditions, road work, and traffic. The omission of these may introduce bias in the estimates. Therefore, the linear models presented in this section should not be used to forecast running time variability.

In chapter 4 of his master's thesis, Ehrlich (2010) explored the relationship between the ratio of Actual Waiting Time (AWT) to Scheduled Waiting Time (SWT) and the following variables: iBus (having the system or not), precipitation, ridership, route length, percent lost mileage due to traffic[4], location in London, operators, and periods. The AWT:SWT ratio is a measure of headway variability rather than running time variability. Nonetheless, Ehrlich's results are comparable to those presented in this report: ridership, route length, location in London, operators and periods were all estimated with statistical significance, but overall linear fits (i.e. values of $\bar{R}^2$) were poor.

Improvement of the linear models for running time and its variability will require a larger dataset with variables such as traffic and characteristics of each corridor, in addition to a larger sample size. Observations of service affected by roadwork or atypical weather can be excluded.

The phenomena that give rise to running time variability are multiple and complex in the way they interact. For this reason, analysis driven by AVL data specific to the question being asked (rather than a general model of running time variability) remains a better way of assessing, for instance, the adequacy of resource assignments. Moreover, approaches for exploring running time variability that capture interactions between the different factors at a less aggregate level should be explored. The data playback tools presented in Appendix B and the simulation model introduced in Chapter 3 are steps in this direction.

## 2.6 Conclusion

Good management of running time variability requires good measurement and communication practices. The measures introduced in this chapter facilitate precise, consistent, and regular assessment of running time variability. The measures work at different levels of aggregation in two dimensions: time and space. The basis of these measures is the quantification of running time variability of homogeneous periods, either in absolute terms or relative to typical running times. Percentile-based measures, referred to as *spread*, were introduced for applications in which large quantities of running time data are available from an AVL database, primarily because they can be related to the amount of reliability buffer time that needs to be provided, even when the shape of running time distributions is unknown.

Running times vary randomly within periods and also transition systematically from one period to another. Systematic trends are known before hand and are addressed in the service planning stage by adjusting fleet size throughout the day. Within-period random variation follows no pattern (practically speaking) and is dealt with by providing a reliability buffer time, which translates to additional resources (i.e. vehicles and crew). Since most practical cases deal with heterogeneous time periods or a mix of time periods, an aggregation method was designed to capture the random within-period component of variability while being less sensitive to systematic between-period variability. The method calculates variability on successive, short, overlapping time periods, which are assumed homogeneous. When *spread* is used as the basis (to measure the variability of each period), the aggregate *mean spread*

[4] Percent lost mileage is the percentage of revenue vehicle miles in schedule not operated. Operators enter reason codes for lost mileage. Percent lost mileage due to traffic includes only the portion which the operator claimed was due to traffic. In the absence of direct observations, Ehrlich used this as a proxy for traffic. Unfortunately, since this variable is operator-reported, it may have introduced endogeneity.

(MS) variability measure is obtained. This measure can be used on one or both directions of a route, and on any range of times of day.

The *diurnal mean spread* (DMS) was developed as a standard measure of overall random variability for both directions of a route during daytime operations. Even though the general variability measures and the aggregation technique are flexible, DMS was defined more specifically, so that consistent measurements of variability can be made on different routes and over time. An algorithm used to compute DMS is presented in Appendix A. Detailed measures, at the direction, time period, and segment level, can be combined with DMS to provide a complete profile of running time variability for a route.

In addition to defining measures, two visual analysis tools were introduced. The first consists of running time scatter plots for each direction of a route, showing running times throughout the day and moving 10th, 50th and 90th percentile lines to indicate typical running times and running time spread. The second illustrates variability on a map, with segments color-coded according to aggregate spreads or any other performance measure.

Various linear models were developed with the goal of finding general patterns of route characteristics leading to higher or lower typical running times and running time variability. Their parameters were estimated on running time observations of a sample of bus routes in London. It was found that running time variability varies greatly from route to route, that typical running times are lower in the summer and higher in (late) fall relative to what they are in spring, and that routes with higher typical running times also tend to have more variable running times. Routes entering central London tend to have higher and more variable running times. Distance, number of stops, and ridership all tend to contribute to higher and more variable running times as well. Routes exhibiting greater operational speeds have slightly lower running time variabilities.

Linear models of median running times were estimated with better goodness of fit than those of variability. This suggests that there are other factors not explored here driving running time variability. Variables such as traffic, weather conditions, corridor characteristics, road work, ridership patterns (at a disaggregate level), operator behavior at the terminal and mid-route, and even fleet size itself could have significant impact on variability.

# Chapter 3

# Simulation Modeling of High-Frequency Bus Transit Lines

## 3.1 Introduction

Performance analysis concepts discussed so far are based on statistical analysis of observations of actual operations made through automated data collection systems. They allow us to quantify running time variability and identify when and where it is greatest. The results of this analysis, together with other evaluation criteria, may lead management to reconsider the resources assigned to a particular service. For example, they may decide to add vehicles in the morning peak in light of high running time variability. The question that arises is how many vehicles to add. Typically, this is a difficult question to answer, and unfortunately the analysis tools discussed in the previous chapters do not shed much light on the matter.

The complexity of deciding vehicle allocation stems from the variety of effects that adding or removing vehicles has on level of service and operating cost. Adding a vehicle allows the operator to serve a route with better headway regularity, potentially decreasing expected waiting time for passengers and balancing in-vehicle loads. The cost of operating the service increases. On the other hand, the operator sees an increase in revenue, usually accompanied by easier operations (in terms of meeting the headways specified in the contract) and marginally more difficult supervision and control requirements. Finally, adding vehicles increases congestion at stops and terminals, where there may be a space limit on the number of vehicles that can stand simultaneously.

The cost of adding vehicles, or the savings associated with removing vehicles, are usually known to the service provider with sufficient precision to analyze options. This knowledge may come in the form of past experience, cost models of varying complexity, or simply direct negotiation with the private operator. It is more challenging to estimate the benefits gained from an addition, or the penalties incurred with a reduction, in terms of service performance. Nevertheless, the latter estimate is required for cost-benefit analysis.

In order to estimate the effects of changes in resource allocation on service performance, we must predict how operations will be with a different number of vehicles than currently used. In other words, we must rely on performance prediction models rather than observations. Simulation is well suited for the task because it is flexible in four ways: it is adaptable, it is extensible, it can capture interaction effects naturally, and it lends itself well to data-driven analysis.

Adaptability refers to the relative ease with which we can make changes to the model. For example, we may wish to test different dispatching strategies at terminals. Doing so is easier with a simulation model than an analytical model because the behavior of the simulation can be defined in modules. For instance, there may be a module for schedule-based dispatching and another for headway-based dispatching. Switching between the two merely requires specifying which module to use when a vehicle reaches a terminal. Although this will also affect what happens beyond the terminal (as intended), we do not have to make changes elsewhere. In particular, there is no dependent analytical expression that loses validity in response to the switch.

Closely related to adaptability is extensibility. An extensible model is one that allows the addition of new components without major changes to the original model structure. For example, we can add the representation of individual riders to a simulation model that initially only represents vehicles. If the model in question were event-based, as is the one developed in this research, then in the initial case all events might correspond to vehicles arriving at stops, while in the adapted model there might be a mix of events, some representing vehicle movement and others passengers arriving at stops. Since simulation models are extensible, we can build upon existing models to explore new questions.

Another benefit of simulation is its natural support for modeling interactions between different elements of the system. Rather than relying on simple analytical formulations, simulation models explicitly represent objects in different states, their actions, and their reactions to other objects' actions, depending on the circumstances. It is possible to capture some degree of interactions with complex multi-regime analytical forms, but expressions quickly become intractable. For this reason, analytical models of transit lines seldom feature vehicle overtaking or deal with capacity constraints.

A fourth attribute of simulation models that makes them relatively flexible is that they naturally accommodate data-driven analysis. This allows us to use large sets of data directly to answer research questions with less concern that the data might not meet some distributional assumptions. In the case of modern transit operations, AVL, AFC, and APC systems provide large sets of disaggregate observations. Since analytical models are based on theoretical distributions, they cannot use data directly. Furthermore, it is unlikely that fitted distributions of the type assumed in a particular analytical model will pass goodness-of-fit tests on the data in every case. With simulation, however, we can build empirical distributions of running times based directly on observations, and they can take on any shape. At the multivariate level, this means that correlation structures can also be arbitrary.

The advantage of using empirical distributions comes at the expense of greater difficulty in interpretation of the results and the need for larger sample sizes. It is generally easier to identify the cause of changes in results with analytical models; this is more difficult with simulation, especially when distributions are arbitrary. Moreover, a large sample size should be used to build empirical distributions if the model is to capture the effects of low-probability events, which often drive our interpretation of the results and any decisions that depend on it. The sample size should depend on the variability and overall distribution of the variable: if events that occur only once a month matter, then empirical distributions should be built on data that spans at least a month, and perhaps more.

Generally speaking, we should strive to specify parsimonious models; that is, models that represent the system being studied with sufficient detail to answer the questions at hand, but without excess complexity. For instance, simulating weather patterns as part of a transit model would be inappropriate unless we were interested in studying the effect of weather on transit operations and have data reflecting differences in rider behavior and vehicle speeds under different weather conditions. Both the question being answered and data availability should drive model specification.

The remainder of this chapter presents a basic simulation model of bus operations that can be used to forecast performance as the resources invested in a route are altered. Section 3.2 introduces a general framework on which the specific simulation model discussed in Section 3.3 is based. Section 3.4 presents algorithms to obtain the parameters necessary for simulation from automated data collection systems. Following this, Section 3.5 covers post-processing of data obtained from the simulation, and Section 3.6 discusses some of the limitations of the model. Section 3.7 briefly describes the implementation of the simulation model. Verification and validation of the model are covered in Sections 3.8 and 3.9, before Section 3.10 ends the chapter with a brief summary and concluding remarks.

## 3.2 Simulation Model Architecture

### 3.2.1 General Framework

This section introduces a flexible framework for transit simulation models. Although it is developed with simulation of a single bus route in mind, it is extensible to simulation of multiple interacting routes or rail systems. The simulation model of a single bus route presented in Section 3.3 is based on this framework.

A class diagram of the simulation model is shown in Figure 3-1. At the top left is the simulation object itself, which is associated with a route and a series of events. (There could be multiple routes, but only single-route examples are discussed in this thesis.) The route has a set of locations, each of which can have a set of distributions and a *location controller*. Distributions are used to model, for instance, segment running times and extra time at the ends of the route. Distributions can be multivariate to capture correlation with running times in previous segments. Locations may have multiple distributions, each for a different time period.

Each location has a *location controller* that governs the behavior of a vehicle upon its arrival at the location, including duration of stay, the next location the vehicle will visit, and the distribution used to determine travel time to the next location. There are two principal types of locations in bus services: terminals and bus stops. Terminals are locations where vehicles can be brought into service or pulled out service, as well as lay over (i.e. stand) to recover the schedule or headway. Bus stops are locations where passengers board and alight vehicles. Whether a location represents a terminal or a stop is determined by the location controller. In stops, controllers govern dwell time and real-time control decisions such as holding or short-turning. In terminals, controllers govern the removal of vehicles from service and dispatching strategies. The notion of directions is deliberately left out of the architecture framework in order to provide the flexibility to model different kinds of transit services. For example, shuttle-type services operating in a loop do not have different directions. Complex bus routes and rail transit lines often have different branches, and the destination of vehicles may change en route.

The simulation model is driven by a set of *discrete events*, processed chronologically, that represent changes in the state of the system. Each event is associated with a time, a location, and a dynamic object as shown in Figure 3-1. Vehicles, drivers, and passengers are dynamic objects because their state changes and they may enter and leave the system. Hence, events can mark the arrival of a vehicle at a terminal or bus stop, the arrival of a passenger at a stop, and so on. Algorithmically, a heap data structure is used to store events as they are generated (in arbitrary order) and retrieve them chronologically. Not everything that happens in the model has to be an event, however. For example, the boarding process, instead of being represented by an event, can be part of what happens when a vehicle arrives at a stop. In this case, the action is associated with the vehicle, rather than the passenger.

An entire day of operations is simulated, from its start (typically in the early hours of the morning) to its end (typically past midnight), so there are no boundary conditions to specify (or mis-specify). It is possible to simulate operations in a specific time period with given boundary conditions, but it turns out that a convenient way of obtaining them is

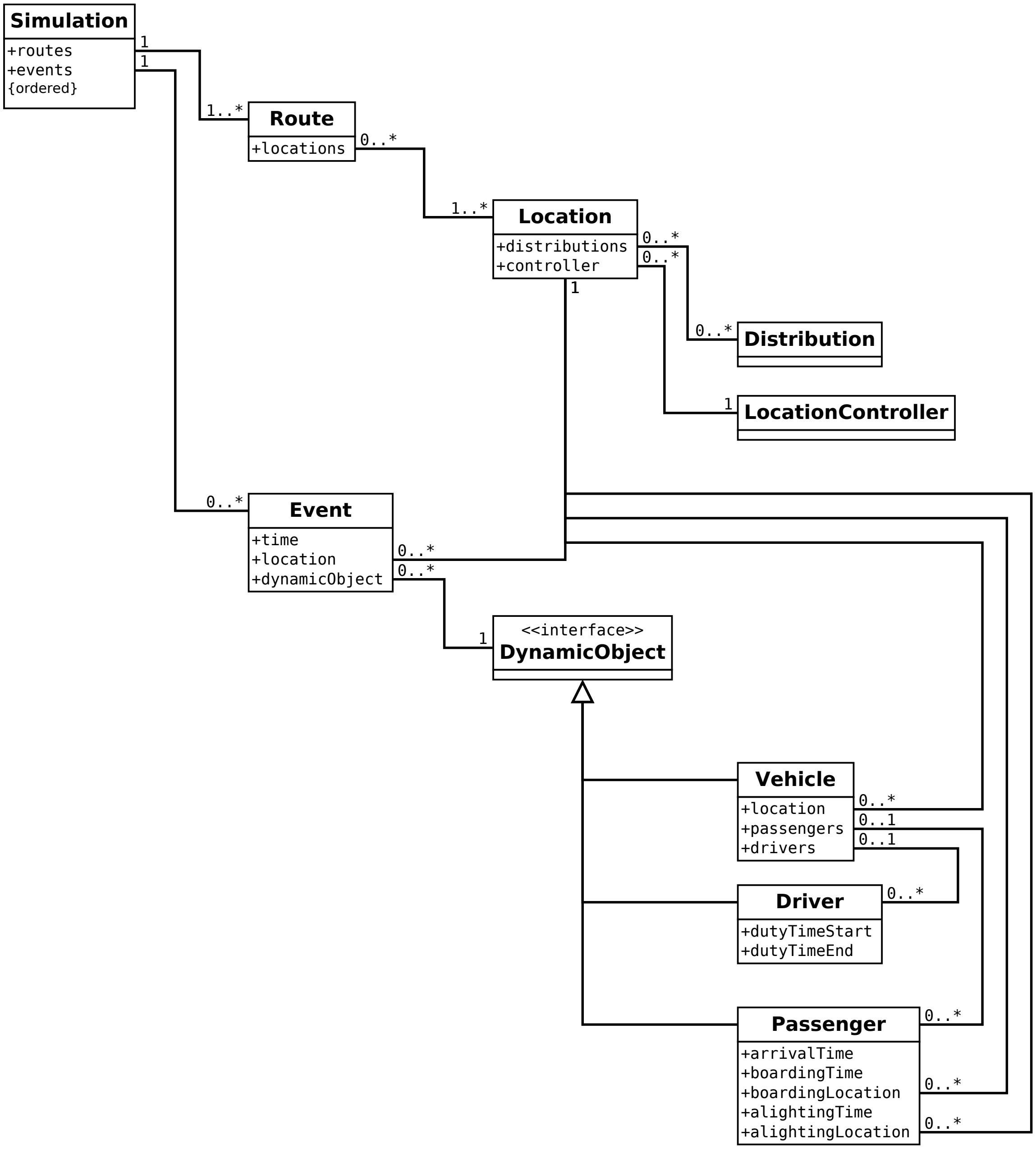


**Figure 3-1:** Class diagram of simulation model

*warming up* the model using simulation from the beginning of the day to the beginning of the selected period. In this sense, the simulation is *terminating*.

Steady-state simulation (as opposed to terminating) may not be appropriate for the targeted applications. For instance, it may be difficult to identify homogeneous time periods, since a system may never operate in steady-state. It is possible, and even likely, that in the course of operations of a transit service on any given day, passenger arrival rates and traffic conditions change at intervals shorter than the cycle time. The performance of such a service at any given moment is a result of a transient operating environment. In this case, even setting initial conditions for a steady-state simulation may be a challenge.

The outcome after simulating a single day of operations is a set of observations from which corresponding performance measures can be calculated. For example, from the distribution of headways it is possible to obtain the means and standard deviations of headway by stop and time period, as well as expected waiting time at the stop, direction, and route level. From observations of trip duration it is possible to obtain mean, median, and 90th percentile running times, both at the segment level and end-to-end. From observations of vehicle loads it is possible to obtain measures indicating levels of in-vehicle comfort experienced by passengers.

If we were to simulate a single day two independent times—or in simulation terminology, run two *replications*—the resulting performance measures would almost certainly be different. Performance measures are a result of the particular realizations of running times and other random variables in each replication. For example, the headway between the first two trips departing a terminal after 8:00 might be short one day and long another, leading to different estimates of expected waiting time. As such, it is best to run many replications to increase the sample size of observations on which performance measures are based, and to establish both central tendencies and confidence intervals around them. This matter is discussed further in Section 3.5.

A high-level activity diagram of the simulation algorithm is shown in Figure 3-2. No source and sink nodes are present because the model is designed to run as a component of a larger process, such as the optimization algorithm developed in Chapter 4. In order to run the simulation, route specification, running time distributions, vehicles, demand representation, location controllers, and any additional controller-dependent parameters must be specified. Replication runs follow, each with an initialization and a data collection stage. Finally, observations of all replications are pooled to obtain performance measures. The different inputs to the simulation model are discussed below.

### 3.2.2 Input

#### Route Specification

Route structure is modeled as a set of locations. Ordered lists of stops by direction, typically available in transit agency databases and also published for the convenience of the general public, can be used to define location objects. Each location's controller determines if the location is a terminal or a bus stop, and the next location in the sequence. When modeling a bus route at the segment level, locations are created for terminals and stops at segment ends, but not for stops within a segment.

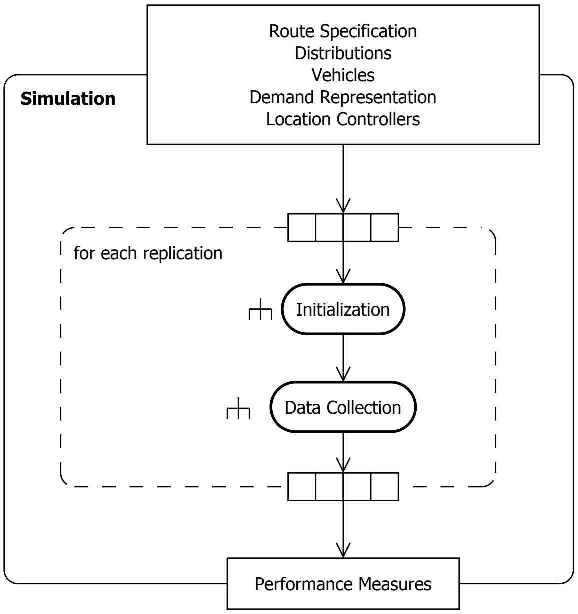


**Figure 3-2:** Simulation model activity diagram showing inputs, outputs, and high-level tasks

### Running Time Distributions

Distributions associated with each location are used to model running time to the next location. Different distributions may be used for different times of the day and for the different locations a vehicle may visit next. Distributions may be theoretical or empirical, univariate or multivariate. It is now common for transit agencies to have databases with large amounts of AVL data at the stop or segment level. This makes it possible not only to obtain good empirical distributions of segment running times, but also to capture correlations between them.

Distributions associated with a terminal and a stop within it model extra time at the end of the route. In many practical situations, arrival at the last bus stop of a route is equivalent to arrival at the terminal, and no extra time is needed. In such cases, the distributions may be defined as a constant zero so that vehicles jump instantaneously between terminals and the bus stops within them. On the other hand, if the route being modeled requires extra time at route ends (for instance, reflecting travel time from the last stop to a parking space designated for layover), a fitted theoretical or empirical univariate distribution may be used.

### Vehicles

Vehicles in the simulation are dynamic objects with identity, so a set of vehicles must be specified as input to the model. At the very least, an insertion time and location must be used to initialize the event heap, but additional properties may also be specified if needed. For example, if the removal of vehicles from service depends on a target time and location, these can be specified. In a similar manner, vehicle seating and standing capacities can be specified to capture capacity constraints and measure passenger comfort. In simulations that model dwell times explicitly and have different types of vehicles, dwell time controllers can be added to the framework in order to determine dwell times for different vehicle types. This is useful for simulations of rail transit, where different trains may have a different number of cars, and therefore different capacities and dwell time functions.

### Demand Representation

Passengers are also modeled as dynamic objects with identity, so a set of passengers must be specified as input. Each passenger has an origin stop with an arrival time and a destination stop. Passengers are typically picked up by the first vehicle to visit their origin stop and subsequently dropped off at their destination stop. (They do not board and alight themselves, since boardings and alightings are handled by vehicles.) This process gives the passenger a boarding time, an alighting time, and possibly some measure of in-vehicle comfort.

Although demand representation is set a-priori, it is not restricted to be constant. A different set of passengers can be given for each replication in order to introduce demand variability. The process that generates sets of passengers may be based on empirical data from an AFC database, especially if it has been processed to infer destinations and expanded to capture ridership not registered in the AFC database. (See Gordon, 2012 for details.) In this case, inferred historical boarding and alighting events of different days can be the demand input for different replications. Alternatively, the demand representation can be based on a random arrival process with assumed rates $\lambda_{o,d}(t)$ for passengers going from $o$ to $d$ at time of day $t$. The simulation model can also be used without demand representation.

Passenger origins and destinations must be defined locations in the route specification. Therefore, if the route specification is based on segments, each covering multiple stops, it may be necessary to pool the demand for all stops within each segment. This complicates the interpretation of vehicle load capacity and in-vehicle comfort, especially if there is significant turnover of passengers within the segment. In light of this, an all-stop route specification allows greater accuracy of load-based performance measures.

### Location Controllers

Location controllers govern the behavior of vehicles at each location. At terminals, they are responsible for removing vehicles from service and for setting layover times. Vehicle removals can be based on target removal times for each vehicle. For example, there may be a rule to remove vehicle 235 from service when it completes a trip scheduled to end at 18:00. Alternatively, they can be based on a vehicle profile, in which case the controller removes vehicles when the active fleet size at a particular time exceeds that specified by a vehicle profile. An example is a rule to remove the first vehicle that arrives at the terminal if there are more than 12 active vehicles at 18:00. Controllers also determine layover time for each vehicle upon arrival at the terminal. Layover times can be based on a schedule or on headway, as discussed in Section 3.2.3.

At bus stops, location controllers govern dwell times, the next location the vehicle will visit, and the time it will take the vehicle to arrive at that location. In the simplest case, dwell times are included in the running time distributions, so they are not explicitly modeled. If running times do not include dwell times, then a dwell time model can be used, with the benefit of correlating boardings, alightings, and load with dwell time. Controllers can extend dwell times (beyond what is required for boardings and alightings) to model holding. Typically, the next stop a vehicle visits is the next downstream in the direction of travel, but controllers may send the vehicle to other locations, which allows modeling of expressing, off-service dead-heading, and short-turning. In any case, the time of arrival at the next location is based on the distributions associated with the stop.

### 3.2.3 Terminal Dispatch Strategy

Several terminal dispatch strategies can be used to determine the stand time of a vehicle arriving at a terminal. These include dispatching based on schedule, preset half-cycle times, target headway, and even headway.

#### Schedule-Based Dispatching

In schedule-based dispatching, the objective is for vehicles to begin trips according to a specified timetable. If a vehicle arrives at the terminal before its next scheduled departure is due, the vehicle holds until the scheduled time. If the vehicle arrives after the scheduled departure time, it departs immediately upon arriving at the terminal. Using $\eta$ to denote stand time,

$$\eta = \max\left(0, t_{\text{scheduled departure}} - t_{\text{actual arrival}}\right) \tag{3.1}$$

Schedule-based dispatching is most appropriate for schedule-based operations, which are typical in low-frequency services but also used by some operators in high-frequency service.

#### Preset Half-Cycle Time Dispatching

Recovery times can also be used to determine when a vehicle should lay over (i.e. stand) at a terminal. Total recovery time $r(t)$ at time of day $t$ is calculated as follows:

$$r(t) = n(t)h_s(t) - p_{50,1}(t) - p_{50,2}(t) \tag{3.2}$$

where $n(t)$ is the number of vehicles, $h_s(t)$ is the scheduled headway, and $p_{k,d}(t)$ denotes the $k^{\text{th}}$ percentile end-to-end running time for direction $d$. In the absence of constraints, total recovery time is allocated to each direction according to running time variability:

$$r_1(t) = \frac{p_{U,1}(t) - p_{50,1}(t)}{p_{U,1}(t) - p_{50,1}(t) + p_{U,2}(t) - p_{50,2}(t)} \cdot r(t) \tag{3.3a}$$

$$r_2(t) = \frac{p_{U,2}(t) - p_{50,2}(t)}{p_{U,1}(t) - p_{50,1}(t) + p_{U,2}(t) - p_{50,2}(t)} \cdot r(t) \tag{3.3b}$$

where $r_d(t)$ is the recovery time allocated to direction $d$ and $p_{U,d}$ is an upper percentile (such as 90th) of direction $d$ running times.

When half-cycle dispatching is used, vehicles stand at the terminal until they complete the half-cycle time from the previous trip, and depart immediately if they arrive late. Half-cycle time is the sum of median running time $\tilde{\tau}$ and recovery time $r$ for a given direction of a route. Mathematically, stand time is given by

$$\eta = \max\left(0, \tilde{\tau} + r - \tau'\right) \tag{3.4}$$

where $\tau'$ is the actual duration of the previous trip.

Half-cycle time dispatching is closely related to schedule-based dispatching, since schedules are often based on half-cycle times. The difference is that while a specified schedule could be inappropriate for actual running times, in half-cycle dispatching a schedule naturally emerges from the number of vehicles, the specified headway, and the running time distributions at different times of the day. From a practical point of view, the advantage is that schedules do not need to be an input for simulation. However, the start time of the first trip of each vehicle must be specified.

### Target Headway Dispatching

With target headway dispatching, the notion of a schedule, be it directly specified or implied, is abandoned in favor of obtaining target headways. This strategy is better suited for high-frequency operations, where passengers are more sensitive to service regularity than schedule adherence (assuming they arrive randomly). Passengers arriving randomly would not be able to tell the difference between a situation in which all vehicles are on time and another in which all vehicles are ten minutes late.

With this strategy, holding time upon arriving at a terminal depends exclusively on the headway leading the vehicle: the vehicle holds if the headway is shorter than specified, or begins the next trip immediately if the headway is longer than specified. Mathematically,

$$\eta = \max\left(0, h_{\text{target}} - h_{\text{leading}}\right) \tag{3.5}$$

where $h_{\text{target}}$ and $h_{\text{leading}}$ are the target and leading headways.

This strategy naturally accommodates changes in headway throughout the day, as holding time depends exclusively on leading headway. However, the leading and trailing headways at the terminal need not be balanced, and this can be exacerbated later (under certain demand situations) through the bunching mechanism discussed in Sections 1.1 and 3.6.1.

Figure 3-3 illustrates *leading* and *trailing* headways. Vehicle $a$ is at the terminal. The leading vehicle is $b$ and the trailing vehicle is $c$. The leading headway is the time it will take $a$ to arrive at $b$'s current position if it departs immediately, and the trailing headway is the time it will take $c$ to arrive at the terminal, where $a$ is currently. With target headway dispatching, the layover time of $a$ depends only on the leading headway, while with even headways dispatching (discussed next) it depends on both the leading and the trailing headways. Neither of the headways are a concern in dispatching strategies based on schedule or half-cycle times.

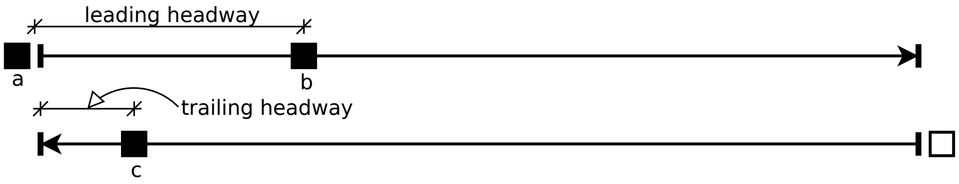


**Figure 3-3:** Leading and trailing headways for vehicle $a$.

### Even Headway Dispatching

Even headway dispatching abandons not only the schedule, but also the target headway. The operational goal in this strategy is to hold vehicles at terminals until the headways leading and trailing the vehicle are balanced. Thus, the likelihood of bunching is decreased. In the absence of curtailments and assuming random passenger arrivals, this strategy tends to yield the lowest expected waiting times.

When this strategy is applied strictly, changes in the number of active vehicles (as well as abrupt changes in running times) can trigger an undesirable chain of stand times. This is best seen through an example. Consider an idealized cyclical service on the verge of a transition from headway $h_0$ to a shorter headway $h_1$ and from $n_0$ vehicles to a greater $n_1$ vehicles. Suppose that the $n_0$ vehicles are perfectly distributed in time and a new vehicle will be inserted. The even headway strategy inserts the vehicle at the midpoint of an existing headway. When the next vehicle arrives at the terminal, it will have a leading headway of $h_0/2$ and a trailing headway of $h_0$. Therefore, it will hold until both headways are $3h_0/4$. Likewise, the next vehicle holds until both headways are $7h_0/8$, etc. A backward-propagating wave of stands is induced, which can be prevented by temporarily using the target headway dispatching strategy when changing the number of active vehicles.

### Considering Terminal Capacity

An explicit terminal capacity can be specified in the form of a maximum limit on the number of vehicles waiting at a terminal. With this, a vehicle may arrive at the terminal and stand (as dictated by the dispatching strategy) if the limit is not exceeded. If the arrival of the vehicle causes the limit to be exceeded, it is taken out of service or the first vehicle in the queue is dispatched immediately to vacate the necessary space.

## 3.3 Simulation

Based on the general architecture for transit operation simulation models, a simple example is now developed, which models a single high-frequency bus service operating cyclically. Vehicles in the model move one segment at a time from the beginning to the end of the run in each direction of the route. Upon reaching the end, active vehicles either begin their trip in the opposite direction immediately or hold at the terminal and subsequently begin their next trip. Segments are defined as portions of a run in one of the directions extending from one stop to another (not necessarily consecutive) downstream stop.

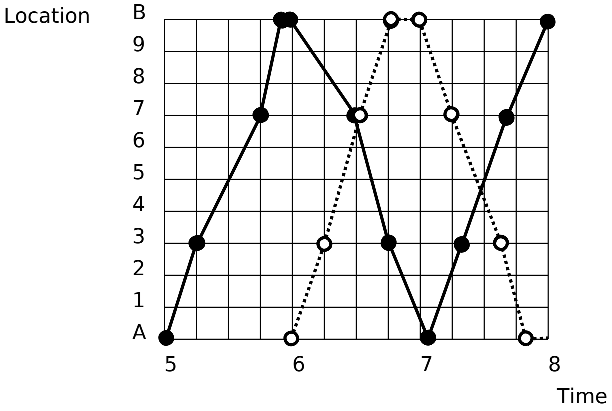


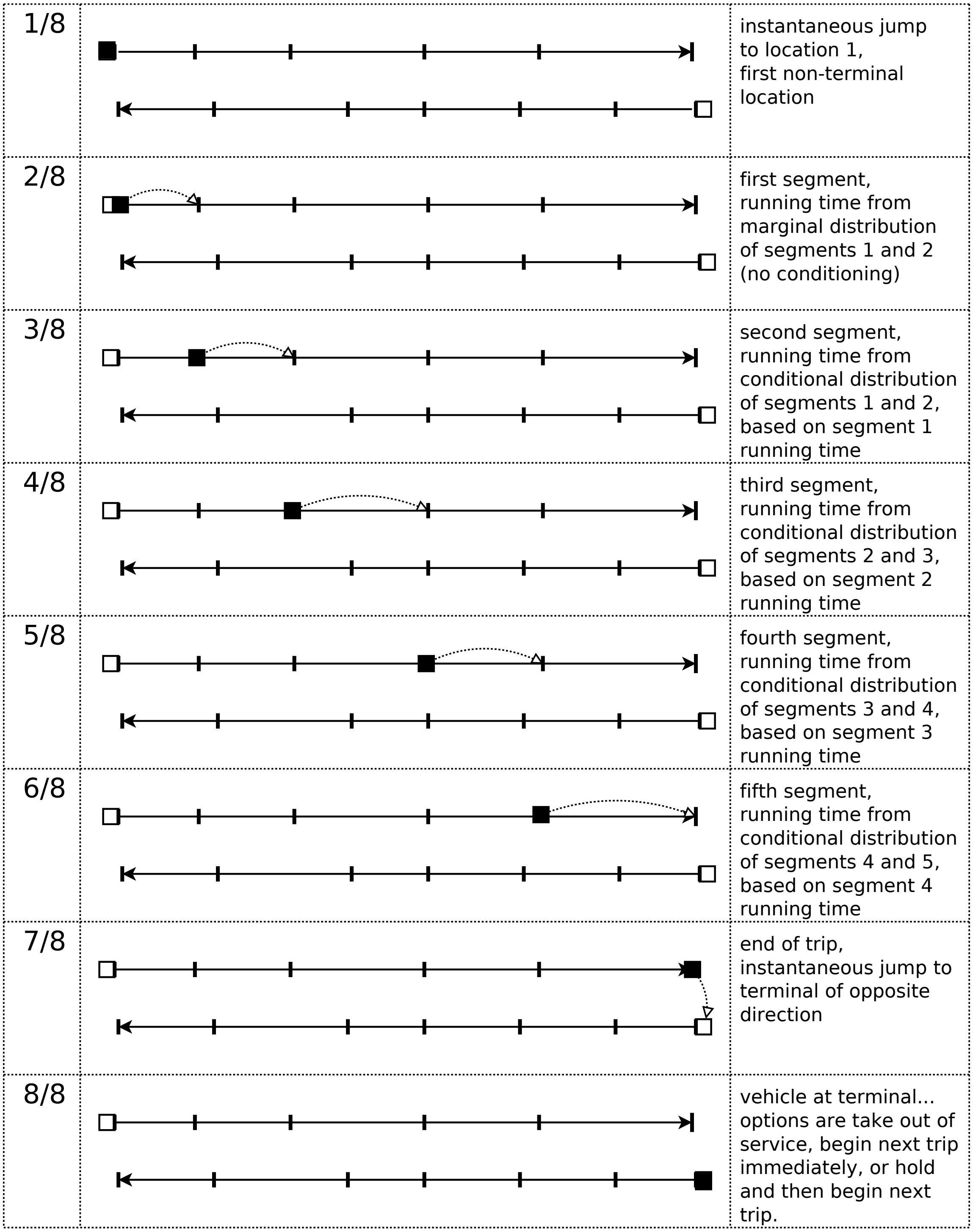
**Figure 3-4:** Space-time diagram of two vehicles on a hypothetical route

Figure 3-4 shows an illustrative space-time diagram of two vehicles as they advance in the simulation model. Location is on the vertical axis and time on the horizontal axis. One vehicle is represented with solid lines and the other with dashed lines. Vehicles travel from one terminal to the other, advancing one segment at a time. Segments go from terminal A to stop 3, stop 3 to stop 7, and stop 7 to terminal B. When vehicles arrive early at a terminal, they stand to recover their schedule or headway.

Figure 3-5 provides a walk-through of the movement of a vehicle in the simulation from the start to the end of a trip. Only one vehicle is shown for clarity, but note that there are usually more vehicles, and events are processed in chronological order, independent of which vehicle they refer to.

The simulation is based on the model architecture described in Section 3.2. Since the architecture is rather flexible, a discussion of the specifics of the model is in order. Routes in the simulation are modeled individually, so interactions with other services are disregarded. Route specification includes locations for terminals and timing points, not necessarily all stops. Therefore, running times are modeled at the segment level, specifically defined as the time elapsed from the departure at one location to the departure at another location, including all dwell times in the segment. For segments ending each direction, running times are defined from departure to arrival rather than departure to departure. Bivariate running time distributions are used to capture correlations between running times of the same vehicle on adjacent segments, and different distributions are used for different times of the day.

Vehicles are modeled without specific times and locations for removal. Insertions and removals are based on a vehicle profile: a vehicle is introduced for every step up in the vehicle profile, and removed for every step down. The removal of vehicles occurs at terminals with the first vehicle to arrive after a decrement in the vehicle profile. Passenger arrivals at stops, boardings, and alightings are not included in the model, but boarding rates at stops are used for aggregation of performance measures to weigh higher-ridership stops more heavily. Operations control, beyond stand at the terminals, is also disregarded; control options such as holding at an intermediate stop or curtailing trips are not considered. The dispatch strategy at terminals is based on target headways as discussed in Section 3.2.3.



| | |
|---|---|
| 1/8 | instantaneous jump to location 1, first non-terminal location |
| 2/8 | first segment, running time from marginal distribution of segments 1 and 2 (no conditioning) |
| 3/8 | second segment, running time from conditional distribution of segments 1 and 2, based on segment 1 running time |
| 4/8 | third segment, running time from conditional distribution of segments 2 and 3, based on segment 2 running time |
| 5/8 | fourth segment, running time from conditional distribution of segments 3 and 4, based on segment 3 running time |
| 6/8 | fifth segment, running time from conditional distribution of segments 4 and 5, based on segment 4 running time |
| 7/8 | end of trip, instantaneous jump to terminal of opposite direction |
| 8/8 | vehicle at terminal... options are take out of service, begin next trip immediately, or hold and then begin next trip. |

**Figure 3-5:** Walk-through of events for a single vehicle.

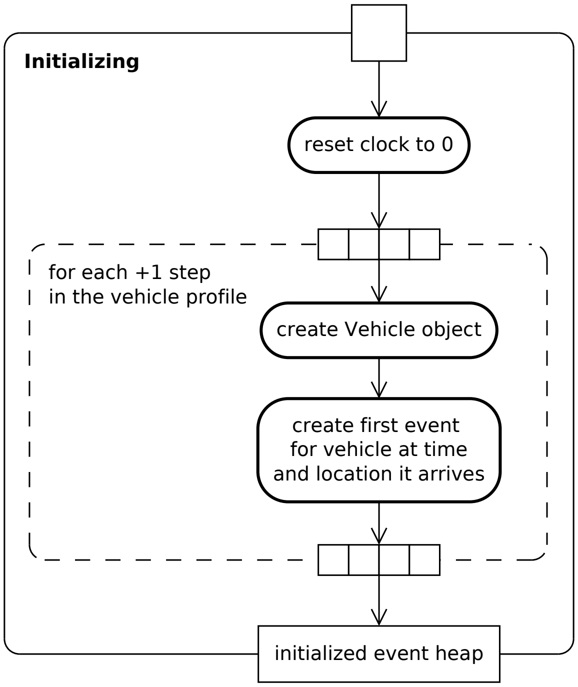

Initializing
reset clock to 0
for each +1 step in the vehicle profile
create Vehicle object
create first event for vehicle at time and location it arrives
initialized event heap

**Figure 3-6:** Activity diagram of the initialization phase of each replication.

Each replication consists of two phases: *initialization* and *data collection*. The former prepares the model to run, while the latter simulates a day of operation and collects performance data.

### 3.3.1 Initialization

Figure 3-6 shows an activity diagram of the initialization phase. The following tasks are performed:

1. The clock is initialized to 0.

2. A *vehicle* object is created to represent each vehicle that will operate at any moment of the day. This is done for each unit increment of the vehicle profile.

3. A start-of-service event is created for every vehicle and added to the event heap. For each one of these events $e$, the location $s_e$ is set to the specified insertion location and the time $t_e$ is set to the time of the increment in the vehicle profile, which represents the time at which the vehicle is ready to operate.

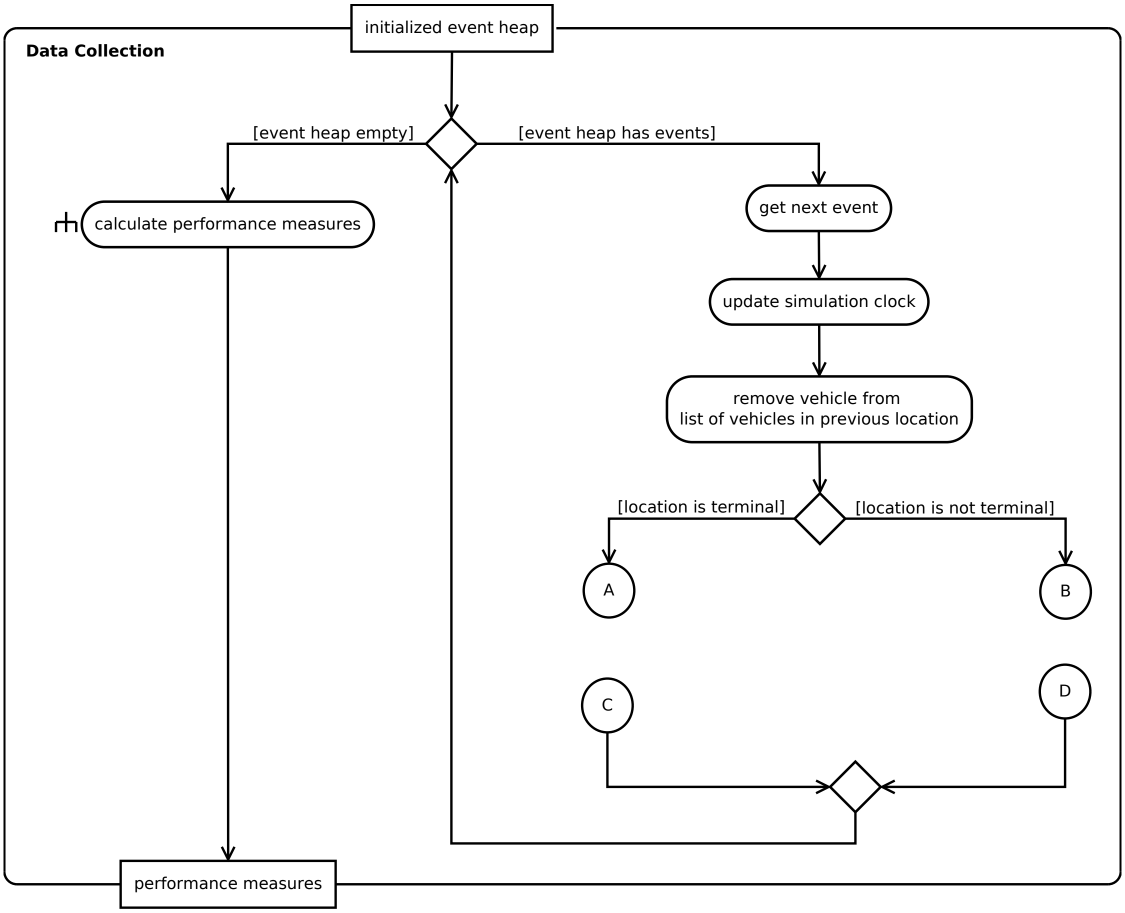


**Figure 3-7:** Activity diagram of data collection phase.

### 3.3.2 Data Collection

The data collection phase begins immediately after initialization. Refer to the activity diagrams in Figures 3-7, 3-8, and 3-9. Events in the event heap are processed in chronological order. Each event $e$ has a time $t_e$, a location $s_e$, and a vehicle $v_e$. Every time an event is processed, the following steps are followed:

1. The simulation clock is updated to the time of the event.

2. The vehicle $v_e$ is removed from the list of vehicles at its previous location.

3. This step is only carried out if the vehicle is not at a terminal.

   (a) The times of vehicle arrival at and departure from the location are recorded in the location object. Since in this version of the model dwell times are included in running times rather than modeled separately, arrival and departure times are equal.

   (b) The vehicle is added to the list of vehicles at the current location $s_e$.

   (c) If the current location $s_e$ is the last location in the current direction, the event

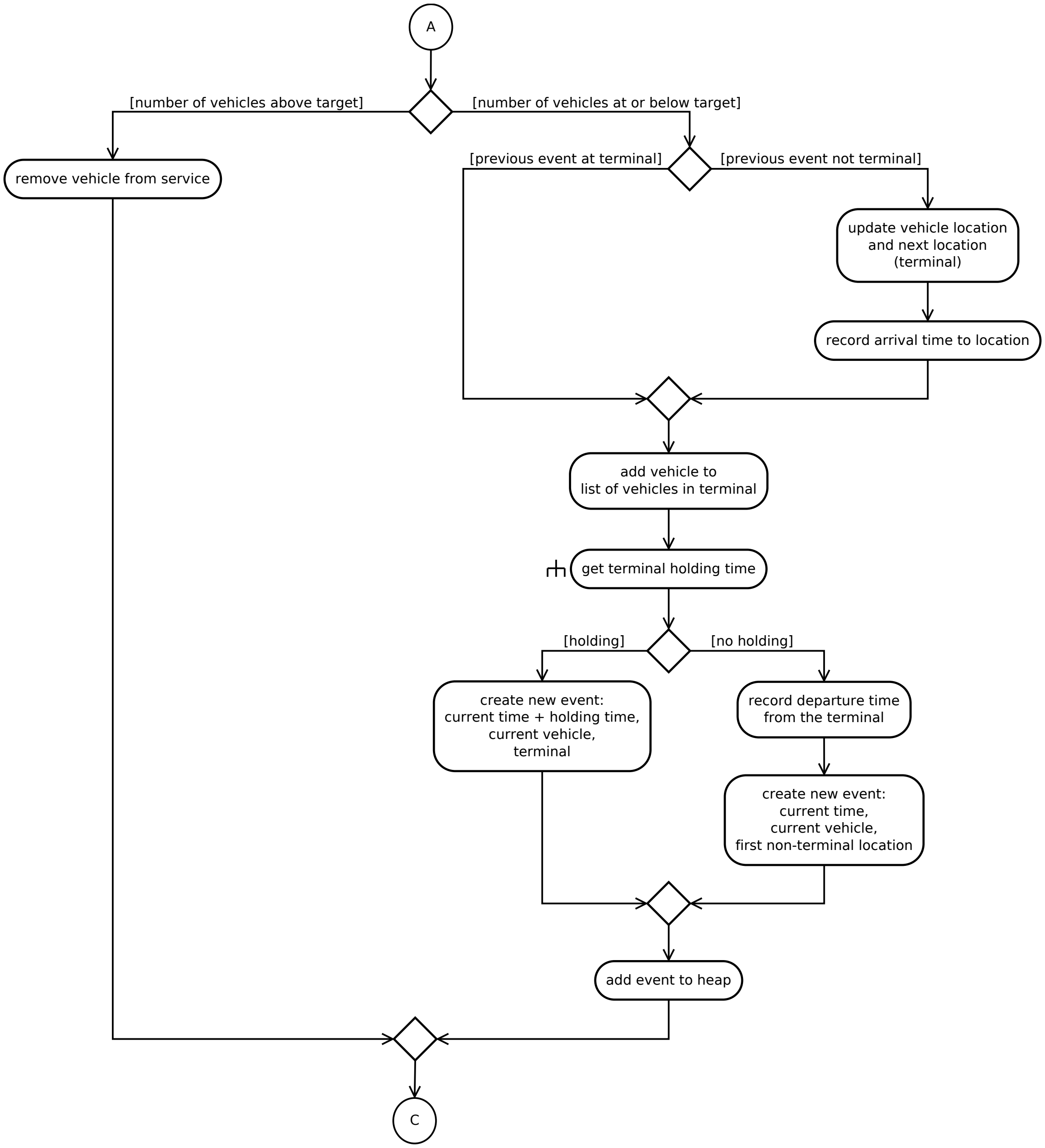

A
[number of vehicles above target]
[number of vehicles at or below target]
remove vehicle from service
[previous event at terminal]
[previous event not terminal]
update vehicle location and next location (terminal)
record arrival time to location
add vehicle to list of vehicles in terminal
get terminal holding time
[holding]
[no holding]
create new event: current time + holding time, current vehicle, terminal
record departure time from the terminal
create new event: current time, current vehicle, first non-terminal location
add event to heap
C

**Figure 3-8:** Activity diagram of data collection phase: when vehicle is at a terminal

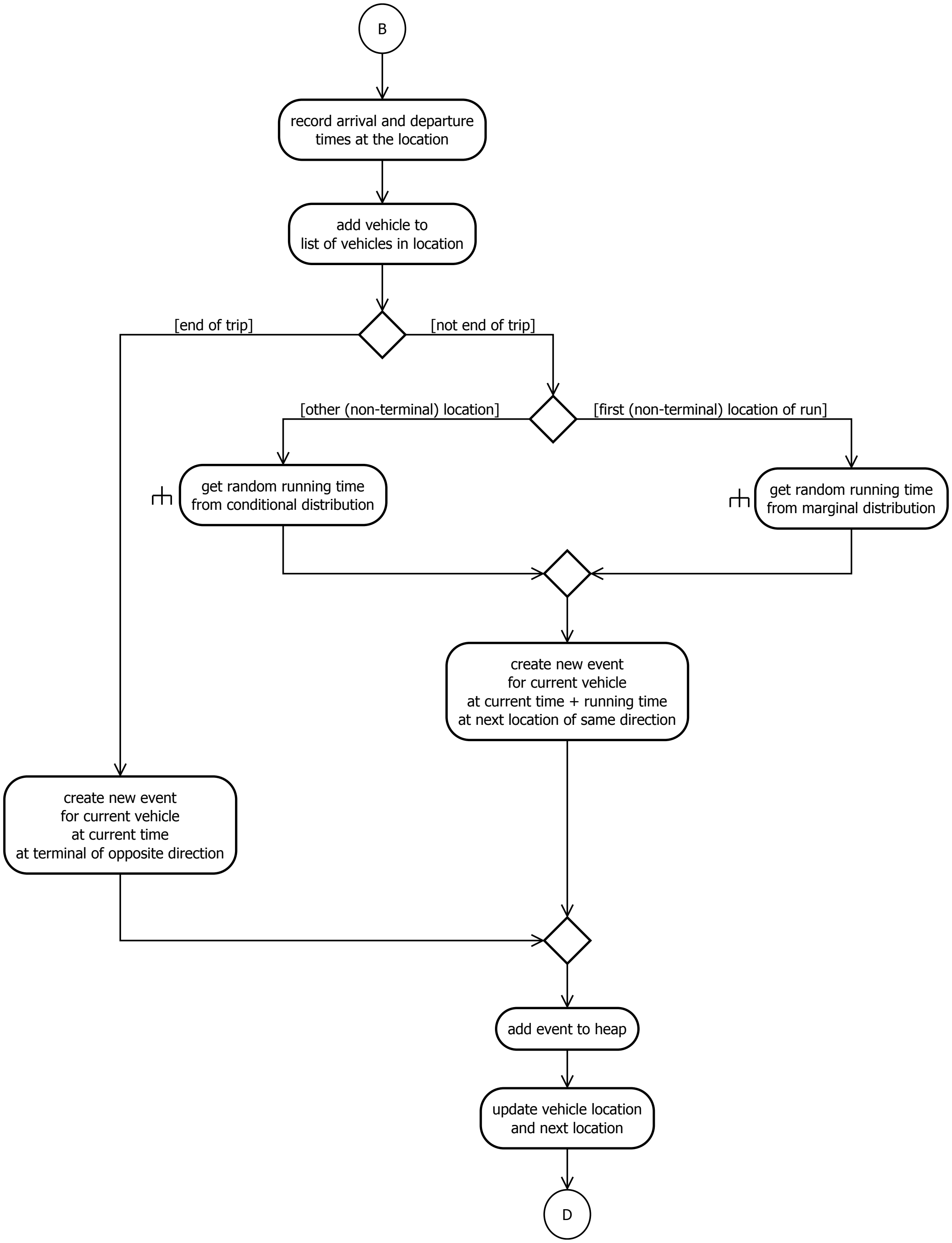

B
record arrival and departure times at the location
add vehicle to list of vehicles in location
[end of trip]
[not end of trip]
[other (non-terminal) location]
[first (non-terminal) location of run]
get random running time from conditional distribution
get random running time from marginal distribution
create new event for current vehicle at current time + running time at next location of same direction
create new event for current vehicle at current time at terminal of opposite direction
add event to heap
update vehicle location and next location
D

**Figure 3-9:** Activity diagram of data collection phase: when vehicle is at a stop

marks the end of a trip. Therefore, a new event $e'$ is created for this vehicle, at the terminal in the opposite direction and at the current time $t_e$ (since travel between an endpoint and a terminal is assumed to be instantaneous).

(d) Otherwise, the trip continues in the same direction.

i. If the current location is the first non-terminal location in the direction, a random running time $x_{s_e}$ is drawn from $f_{X_{s_e}}$, the marginal distribution of the first segment of the bivariate distribution with running time pairs of the first and second segments.

ii. Otherwise, a random running time $x_{s_e}$ is drawn from $f_{X_{s_e}|x_{s_e-1}}$, the conditional distribution of the second segment of the distribution with running time pairs of this and the previous segments, conditioned on the running time realized for the previous segment.

A new event $e'$ is created for this vehicle, at the next location in this direction, for time $t_{e'} = t_e + x_{s_e}$.

(e) The new event $e'$ is added to the heap.

(f) The vehicle's location and next location are updated in the vehicle object.

4. This step is only carried out when the vehicle is at a terminal.

(a) If the number of vehicles in service exceeds the quantity specified by the $n(t)$ profile for this time $t$, the vehicle is taken out of service. No new event is created for this vehicle.

(b) Otherwise, the vehicle continues in service.

i. If the vehicle's previous location was not this terminal, then its location is updated to this terminal and its arrival time is recorded in the terminal location object.

ii. The vehicle is added to the list of vehicles at the terminal.

iii. Stand time $\eta$ at the terminal is determined based on the terminal dispatch strategy chosen for the model (see Section 3.2.3).

iv. The next step depends on whether the vehicle is held or not.

A. If there is holding ($\eta > 0$), the vehicle holds at the terminal. A new event $e'$ is created for the vehicle at the terminal, at time $t_{e'} = t + \eta$.

B. If there is no holding ($\eta = 0$), the vehicle is set to depart the terminal to begin its next trip. The departure time is recorded in the terminal location object. A new event $e'$ is created for the vehicle at the first non-terminal location of the direction, at the current time (since movement to and from the terminal is instantaneous).

v. The new event $e'$ is added to the event heap.

The above steps are repeated until the event heap is exhausted, which happens when all vehicles have completed their final trips and have been removed from service.

## 3.4 Obtaining and Modeling Input Parameters

One of the advantages of simulation as an analytical technique is the ability to directly use vehicle location and fare collection data from automated data collection systems, without relying on distributional assumptions and possibly capturing complex correlation structures. Automated data collection systems generate such large amounts of data that it is intractable to explore it to its full depth and richness without the aid of computer systems.

The simulation model described in Section 3.3 requires a set of parameters that characterize the service being modeled and the way it operates. These include route structure, running time distributions, target headways, boarding rates, and a vehicle profile. The source of this information is specific to each situation, but it is likely that much of it can be obtained, derived, or inferred from AVL and AFC databases of a modern transit agency or service provider. This section discusses how to obtain and model input parameters with the London Buses data model.

### 3.4.1 Route Structure

Route structure consists of an ordered list of locations for each direction of a route. Table 3.1 shows an example of the type of information available in internal London Buses databases concerning route structure. The simulation model needs a unique identifier for each location, which can be the *Stop Id* or *Stop Code*. *Stop Sequence* is used to retrieve the stop list in order. No other fields are necessary for the simulation.

**Table 3.1:** Sample Route Specification

Route W15 from William Morris School towards Pembury Road

| Stop Sequence | Stop Id | Stop Code | Stop Name |
|---|---|---|---|
| 1 | 2546 | BP2540 | William Morris School |
| 2 | 2557 | 10330 | Lawrence Avenue |
| 3 | 2558 | 10334 | Higham Hill Road |
| ... | ... | ... | ... |
| 51 | 692 | 129 | Amhurst Road / Hackney Downs Station |
| 52 | 2665 | BP1104 | Pembury Road |

The analyst can represent all stops as locations, or pick a subset of them. Modeling all stops leads to the most accurate vehicle loads, since demand does not have to be aggregated by segment. However, picking fewer stops leads to shorter computation time and decreases potential accumulation of segment running time errors. (If there is an error $\epsilon$ associated with segment running times, the total error for end-to-end running times is the number of segments times $\epsilon$. Picking fewer (longer) segments reduces total error in comparison to picking more (shorter) segments, because GPS-based AVL stop detection errors are independent of segment length.) For purposes of this research, the arbitrary choice is to model only stops that are time points, which is typically one of every four or five stops.

### 3.4.2 Running Times

Running times are also required by the simulation model, in the form of bivariate probability distributions for adjacent segments. Running time data is often available in AVL databases. Table 3.2 shows a sample of AVL data for route W15 found in the London Buses *iBus* database.

**Table 3.2:** Sample AVL Data for Route W15

| Date | Trip Id | Stop Id | Direction | Stop Sequence | Scheduled Time | Departure Time | Arrival Time |
|---|---|---|---|---|---|---|---|
| 2011–03–04 | 134826 | 2546 | 1 | 1 | 23:20:00 | 23:18:55 | 23:18:55 |
| 2011–03–04 | 134826 | 2557 | 1 | 2 | 23:22:00 | 23:20:46 | 23:20:46 |
| 2011–03–04 | 134826 | 2558 | 1 | 3 | 23:23:00 | 23:21:01 | 23:21:01 |
| . . . | . . . | . . . | . . . | . . . | . . . | . . . | . . . |

AVL records are filtered for the date range and day type of interest (for example, weekdays from 2011–03–05 to 2011–04–01). Next, trips are identified as sets of stop visits. End-of-route stop detection problems are addressed with simple heuristics. For example, trips without a time stamp at the last stop can be discarded, or the mean stop-to-stop running time between the penultimate and the last stop can be used to infer the missing time stamp. Outliers (in terms of end-to-end trip duration) are discarded using empirical rules, such as discarding values located at the extremes, e.g. 1.5 times the interquartile range above the third quartile or below the first quartile. (See Section 1.3.2 in Kottegoda and Rosso, 2008 for an explanation of this type of outlier detection method.)

Construction of bivariate distributions follows, one pair of adjacent segments and one time period at a time. Refer to Figure 3-10 for the corresponding activity diagram. First, a set of running time pairs is associated with each segment pair (for example, the first and second segments in the first direction), and populated with running times from trips that (1) visited the first stop of the first segment in the time period specified and (2) visited all stops in both segments in revenue service. Because of curtailments, some trips may not cover both segments.

Second, observations are classified according to their running time. Kottegoda and Rosso (2008) present a formula, credited to Freedman and Diaconis (1981), that gives an appropriate number of classes $n_c$ to represent data in a histogram:

$$n_c = \frac{rn^{1/3}}{2\,(\mathrm{IQR})} \tag{3.6}$$

where $r$ is the range (the difference between the maximum and the minimum observation), $n$ is the number of observations, and IQR is the inter-quartile range (the difference between the third and the first quartile). The result is rounded to the nearest integer. The total number of classes used is the product of $n_c$ of the two segments. Observations falling in each two-dimensional class are counted; the greater this number, the more likely it is for a running time in simulation to fall in this class.

Figure 3-11 illustrates a bivariate running time probability function built this way. The axis labels indicate the range of running times, in seconds, that define each class. The

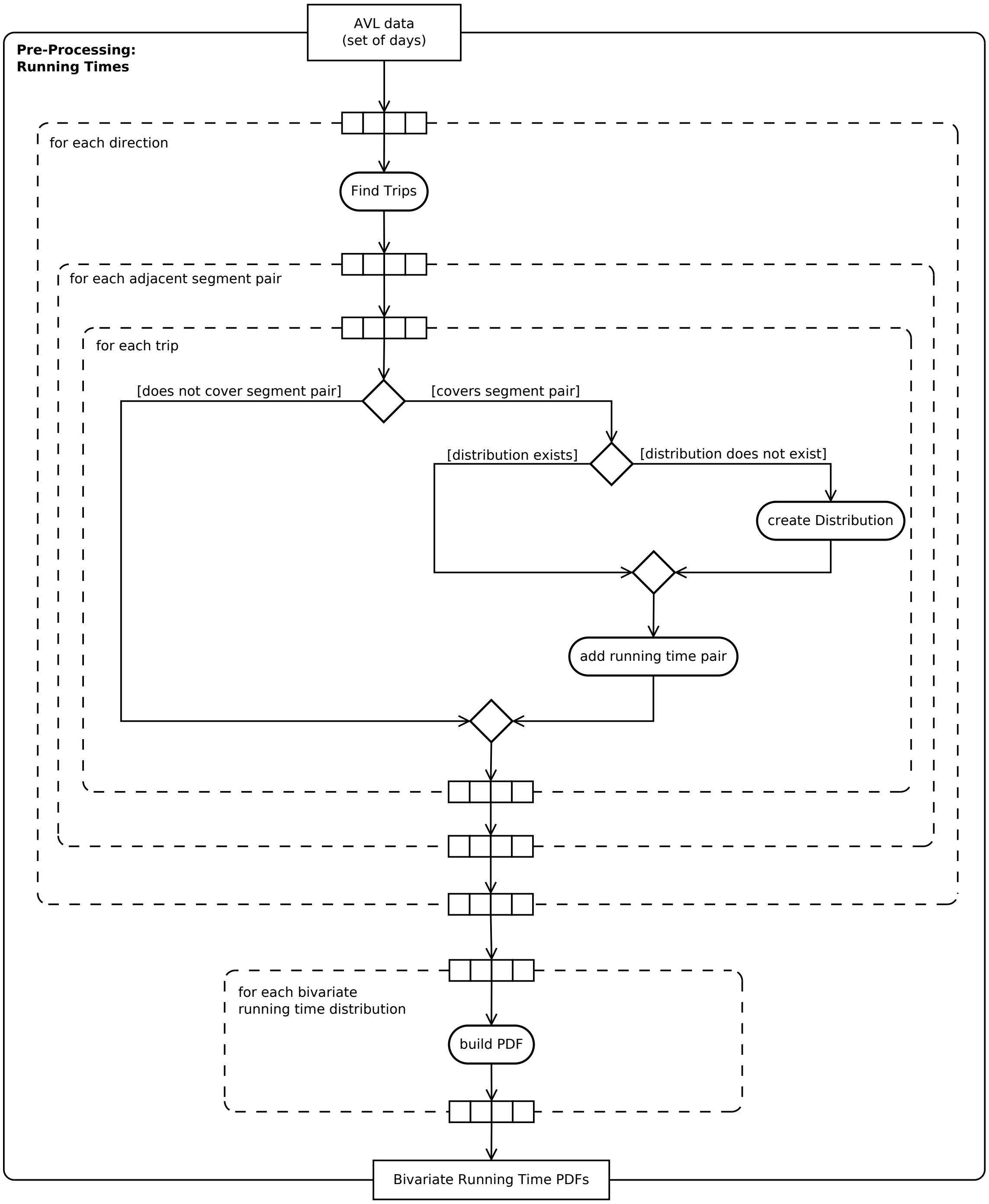
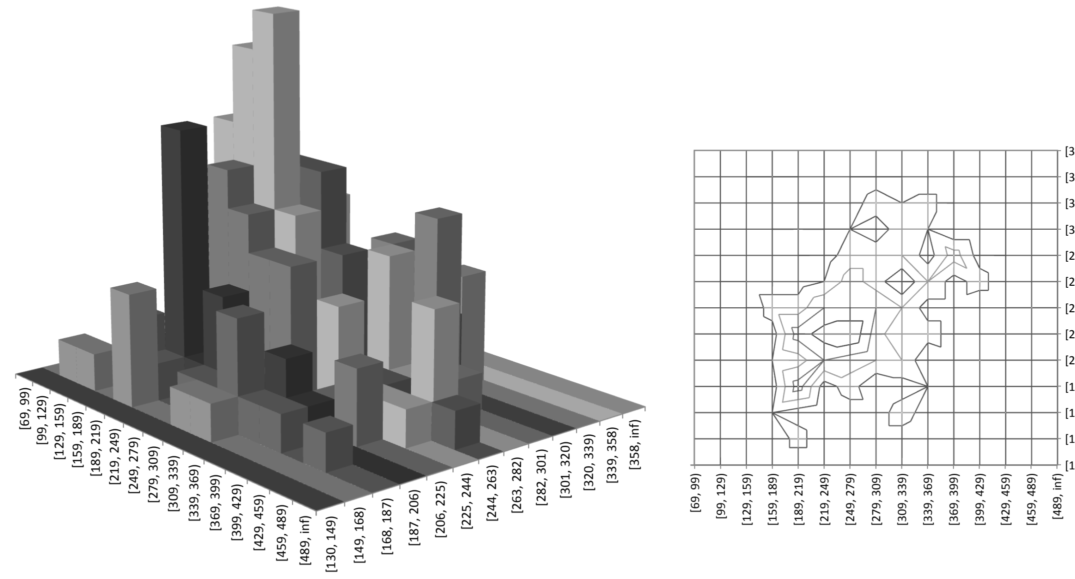


**Figure 3-10:** Activity diagram for pre-processing running times



**(a)** 3-d plot

**(b)** contour plot

**Figure 3-11:** Bivariate segment running time distributions

same bivariate running time distribution is shown in a 3-d plot and a contour plot. Notice there is some degree of positive correlation between the running times of the two adjacent segments. In other words, there is a tendency for greater running times on the second segment to follow greater running times on the first segment, and vice versa.

This process is carried out individually for each time period. For purposes of generating running time distributions, time periods are defined systematically to cover the entire service day in short intervals. For example, time periods thirty minutes long shifted every thirty minutes may be used if there are sufficient observations to characterize the probability distributions. For instance, one distribution may model running times of the first two segments of a direction between 7:30 and 8:00.

### 3.4.3 Target Headways

Target headways $h_s(t)$ by time of day $t$ at each location $s$ are used in simulation for headway-based control. While target headway terminal dispatching only requires target headways at terminals, target headways at stops could be used to model headway-regulating real-time control strategies. Modeling target headways with a single function for all stops would be simpler but unrealistic, because transitions from one scheduled headway to another occur at terminals (when vehicles are dispatched or removed from service) and then gradually propagate (with a lag depending on running times) to the rest of the route.

Target headways can be calculated from the London Buses *iBus* (AVL) database, since *Scheduled Time* is given for each planned stop visit, as shown in Table 3.2. Looking only at a single day of AVL data, scheduled stop visit times are obtained and sorted by time. Headways are obtained by taking the difference between consecutive scheduled visits to each stop. Refer to Figure 3-12 for an activity diagram of target headway pre-processing.

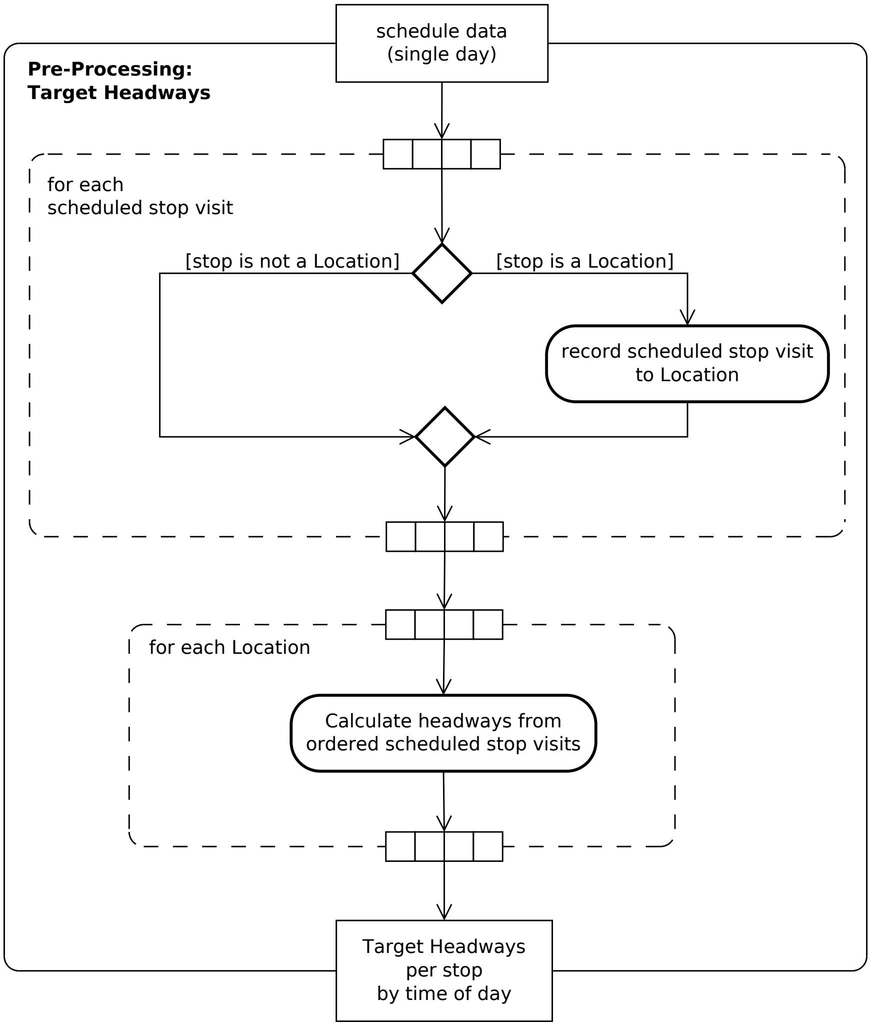


**Figure 3-12:** Activity diagram for obtaining target headways from schedule data

### 3.4.4 Boarding Rates

Boarding rates per stop $\rho_s(t)$ are also read from a database, either directly if they have been previously calculated or in the form of electronic ticketing machine (ETM) records if they have not. Table 3.3 shows a sample of ETM data for route W15. Although ETM data does not specify boarding location, it can be inferred from *Transaction Date* (which is essentially a time-stamp corresponding to a fare card tap or other interaction with the fare box) and vehicle positions from the AVL database. For example, the three boardings at 8:07 in Table 3.3 probably occurred at the third stop of the direction, since the vehicle arrived at the third stop at 8:07 according to the AVL database.

**Table 3.3:** Sample ETM Data for Route W15

| Route Id | Transaction Date | Trip Number | Trip Date |
|---|---|---|---|
| W15 | 2011-03-04 8:04 | 40 | 2011-03-04 7:50 |
| W15 | 2011-03-04 8:04 | 40 | 2011-03-04 7:50 |
| W15 | 2011-03-04 8:05 | 40 | 2011-03-04 7:50 |
| W15 | 2011-03-04 8:05 | 40 | 2011-03-04 7:50 |
| W15 | 2011-03-04 8:05 | 40 | 2011-03-04 7:50 |
| W15 | 2011-03-04 8:07 | 40 | 2011-03-04 7:50 |
| W15 | 2011-03-04 8:07 | 40 | 2011-03-04 7:50 |
| W15 | 2011-03-04 8:07 | 40 | 2011-03-04 7:50 |
| W15 | 2011-03-04 8:23 | 40 | 2011-03-04 7:50 |
| ... | ... | ... | ... |

Since this simulation model does not model riders, boarding rates are not used in the simulation. Instead, they are used in post-processing to weigh stop-level performance measures, such that, for instance, expected waiting time at a very important stop in the route is considered more heavily than that at a stop where few passengers board. Although from the outside observer's perspective all stops are equal, passengers are more likely to experience waiting at stops with high boarding rates. This is akin to the phenomenon where most riders of a transit service experience crowded vehicles even when average crowding levels seem reasonable; a more passenger-centric measure of crowding comes from weighing by vehicle load to reflect average passenger experience.

Figure 3-13 illustrates the activity diagram of boarding rates pre-processing from disaggregate ETM records. Boarding rates by stop are aggregated by segment and associated with the *location* beginning the segment. If the simulation model were extended to explicitly model rider events, boarding rates could also be used to generate random passenger arrivals at stops. In the latter case, AFC data and inference techniques could be used to find boarding rates by origin-destination pair rather than just origin, as discussed in Section 3.2.2.

### 3.4.5 Vehicle Profile

A profile specifying the number of vehicles available for operations at any particular time of day, denoted by $n(t)$, can also be obtained from the AVL database. The procedure is similar to obtaining target headways in that a single day of schedule data is used. (Unless there

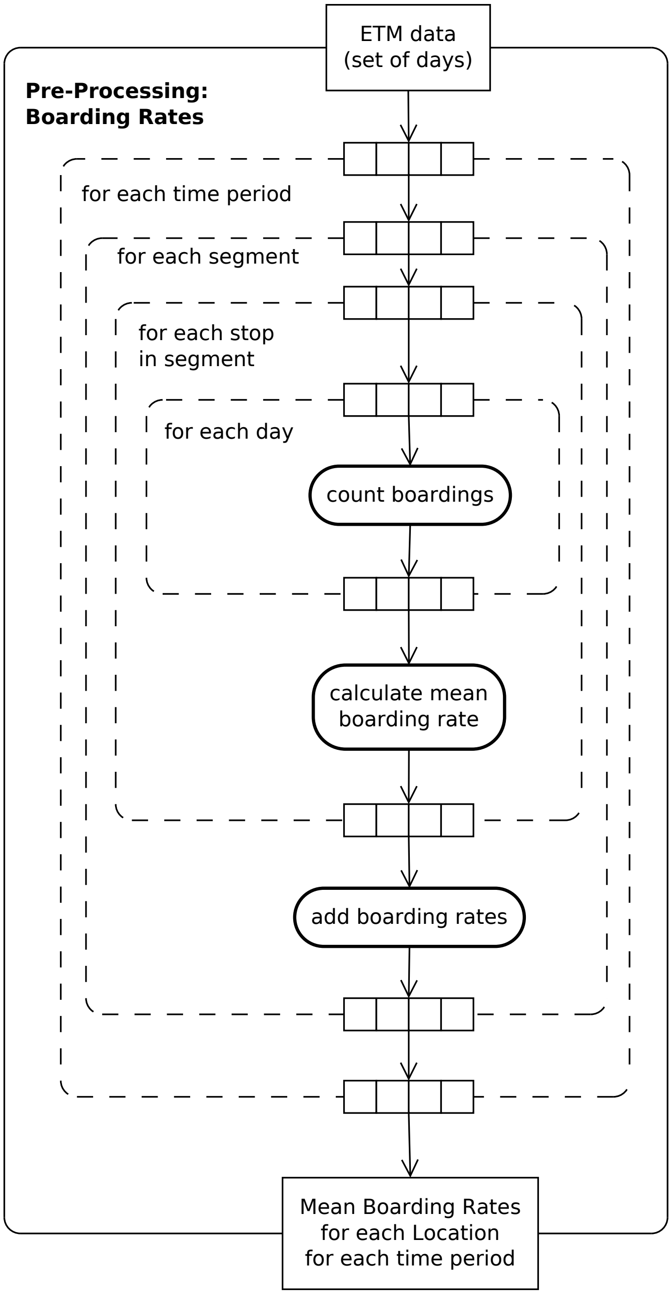


**Figure 3-13:** Activity diagram for calculating mean boarding rates from ETM data

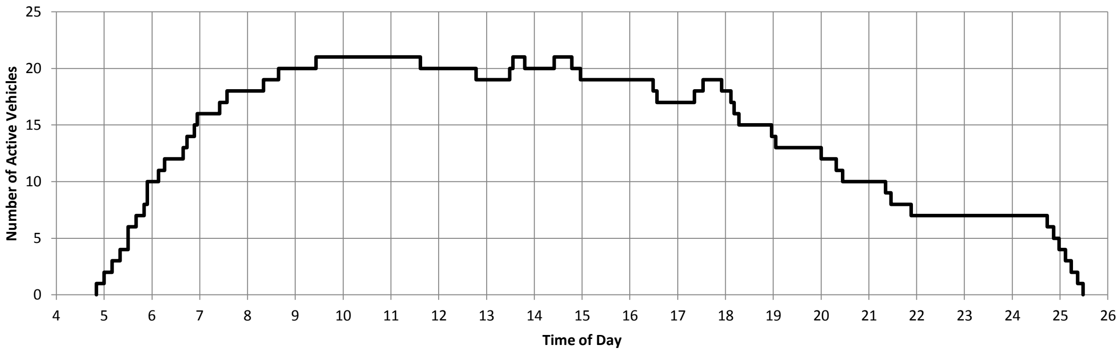


**Figure 3-14:** Example of a vehicle profile

are changes to the schedule, the profile should be the same every day.) However, instead of looking at visits by stop, this procedure looks at visits by vehicle to determine how many vehicles operate simultaneously at any given time of day.

A tolerance long enough for running time between stops and stand time at terminals must be used to avoid incorrect fluctuations in the profile. For example, if the time between the last scheduled visit of a trip and the next visit of the vehicle is more than 60 minutes, the vehicle might be taken out of service and brought back to service in time for the next trip. Conversely, if this time is 15 minutes, the vehicle stays in service and is considered to be standing by, but active. This guesswork can be bypassed if information on scheduled stand times is available.

The profile begins with zero vehicles before the scheduled start of service and ends with zero vehicles after the last scheduled trip ends. An example is shown in Figure 3-14, for a service that starts shortly before 5:00, ramps up to a peak vehicle requirement of 21 vehicles shortly before 9:30, exhibits small fluctuations in active number of vehicles during the day, and gradually ramps down to zero vehicles (in two stages).

Since the objective of simulation in this research is to forecast how a route's service quality changes as vehicles are added or removed, the vehicle profile must be modified to reflect the level of resource available in each hypothetical scenario. Several techniques to alter the current profile are discussed in Chapter 4.

## 3.5 Post-Processing

### 3.5.1 Aggregating Observations

Upon completing all replications, there are a set of disaggregate observations of performance, which may include arrival times of vehicles at stops, lengths of stand time at terminals, and end-to-end running times. If passengers are modeled as individual entities, there may also be observations of the waiting time, in-vehicle travel time, and in-vehicle crowding experienced by each passenger. These observations must now be aggregated and converted to performance measures with meaning to the analyst using the simulation model.

Aggregation consists of pooling observations from each replication and grouping them by time period. Some performance measures are simply a statistic of one of the data items (for example, mean running time), while others are derived. One of the principal derived performance measures of the simulation model described in this chapter is expected waiting time $\hat{w}(t_i, t_f)$ at the route level for any time-of-day interval beginning at time $t_i$ and ending at time $t_f$. (This period should have a single target headway.) Mathematically, it is the weighted average of stop-level expected waiting times $w_s(t_i, t_f)$, weighted by estimated mean passenger arrival rates $\rho_s(t_i, t_f)$.

$$w(t_i, t_f) = \frac{\sum_s \rho_s(t_i, t_f) w_s(t_i, t_f)}{\sum_s \rho_s(t_i, t_f)} \tag{3.7}$$

The stop-level expected waiting time of the $s^{\text{th}}$ stop (assuming no capacity constraints and random passenger arrivals) is calculated based on the mean $\mu_s$ and variance $\sigma_s^2$ of the headways observed at $s$ in the interval $(t_i, t_f)$.

$$w_s(t_i, t_f) = \frac{\mu_s}{2}\left(1 + \frac{\sigma_s^2}{\mu_s^2}\right) \tag{3.8}$$

### 3.5.2 Establishing Confidence Intervals

Variability in the input parameters leads to variability in observations from one trip to another, so performance statistics like mean and standard deviation of running time are estimated with some degree of uncertainty. The accuracy of estimates improves as the sample size grows, which in simulation is achieved by running a larger number of replications, just like for analysis of real data it is achieved by including observations for a greater number of days.

Confidence intervals can accompany statistical estimates to give a sense of accuracy. Consider a hypothetical scenario where a mean headway is estimated to be 6.7 minutes for one route and 5.9 minutes for another. When figures are reported without confidence intervals, one might conclude that the second route has a lower mean headway. Now suppose that the estimates are reported with their confidence intervals: 6.7 ± 0.5 minutes for the first route and 5.9 ± 0.7 minutes for the second, at 90% confidence level. Loosely speaking, one is 90% confident that the mean headway is in the range of 6.2–7.2 minutes for the first route and 5.2–6.6 minutes for the second. Because there is overlap between the two ranges, one cannot claim that either is greater than the other with 90% confidence.

In general, if the performance measure of interest is $x$, the objective is to provide an estimator $\hat{x}(t_i, t_f)$ at the route-level for any time-of-day interval $(t_i, t_f)$. When the performance measure of interest is the mean of a group of observations in $(t_i, t_f)$ (for example, mean running time), a confidence interval for the point estimate can be established by invoking the Central Limit Theorem, which states that when $x_i$ are independently and identically distributed (i.i.d.), the distribution of $\bar{x}(t_i, t_f)$ is asymptotically normal, regardless of the distribution of $X$. (In practice, this usually also holds under mild violations of the i.i.d. assumption.) Mathematically, as the number of observations $n$ tends to $\infty$, the standard error tends to

$$\sigma_{\bar{x}} = \frac{\sigma_X}{\sqrt{n}} \tag{3.9}$$

Since the true standard error $\sigma_{\bar{x}}$ is unknown, it must be estimated as the sample standard deviation $\hat{S}_X$ and the Students $t$ distribution should be used. However, this distribution approaches the normal distribution for large sample sizes (for example, $n \geq 120$), so the normal distribution may be used in practice. Therefore, the 95% central two-sided confidence interval for the true mean of $X$ is

$$\bar{x}(t_i, t_f) \pm 1.96 \frac{\hat{S}_X}{\sqrt{n}} \tag{3.10}$$

Confidence interval (3.10), based on the Central Limit Theorem, is valid only when the statistic being calculated is the arithmetic mean of a set of observations. Unfortunately, there are many relevant performance measures that are not the mean of a set of observations; for example, the standard deviation of running times, reliability buffer time (the difference between the 95th and 50th percentile running times), and expected waiting time (a nonlinear function of both the mean and the standard deviation of headway) are relevant performance metrics, but confidence intervals for them cannot be established with (3.10). Computational non-parametric statistical methods exist to overcome this difficulty. One commonly used method is called *bootstrapping*. The discussion of its use lies beyond the scope of this thesis; see Kottegoda and Rosso (2008) for a quick description of the method and Efron (1981) for a more thorough discussion.

If a confidence interval is not established for a performance measure, at least its convergence error at the replication level should be known to the analyst. This provides an informal notion of the accuracy of estimates that may be enough to decide if the number of replications is appropriate. The convergence error from replication to replication is the difference between estimates of a performance measure with and without the observations of an additional replication. To illustrate, Figure 3-15 shows the convergence of expected waiting time for route W15 between 7:30 and 9:30 over the course of 500 replications. In this case, the mean expected waiting time is quite unstable for the first 50 replications, and then gradually stabilizes to 337.3 seconds. Over the last 10 replications, the estimate oscillates between 337.2 and 337.3 seconds, so there is an informal notion that the accuracy is sufficient.

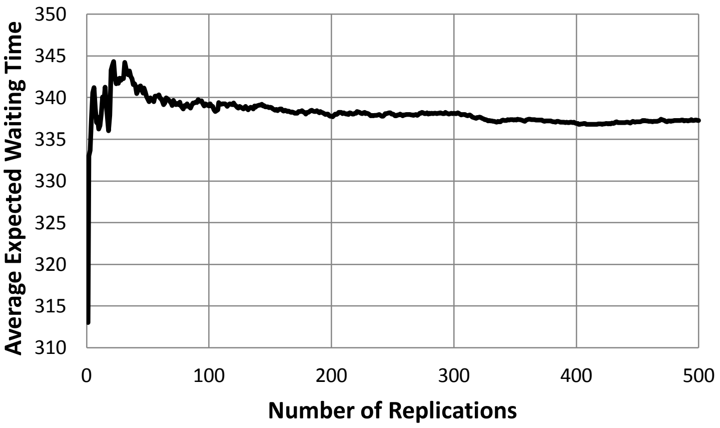


**Figure 3-15:** Convergence of expected waiting time for route W15

## 3.6 Limitations

### 3.6.1 Headway Endogeneity

The segment running times used by this model include dwell times, which are a function of historical operating conditions. Real operating conditions, such as headways and loads, may be very different from those pertaining to the vehicle in the simulation. Since bus loads and people waiting at stops are not modeled, these factors are being treated as exogenous when in reality they are endogenous in determining running times (because running times include dwell time).

For example, consider the dwell time model below, which ignores capacity constraints and assumes Poisson arrivals for different OD pairs:

$$T_{\text{dwell}} = \sum_{i=1}^{n_{\text{stops}}} x_i \left(T_{\text{arrive}} + T_{\text{depart}} + \max\left(n_{a,i} T_{\text{alight}}, n_{b,i} T_{\text{board}}\right)\right) \tag{3.11a}$$

$$x_i = \min(1, \max(n_{a,i}, n_{b,i})) \tag{3.11b}$$

$$n_{a,i} = \sum_{j=1}^{i} n_{b,j,i} \tag{3.11c}$$

$$n_{b,i} = \sum_{j=i}^{n_{\text{stops}}} n_{b,i,j} \tag{3.11d}$$

$$n_{b,i,j} = p_{i,j}(h_i) \tag{3.11e}$$

$$P_{i,j}(h_i) \sim \text{Poisson}(\lambda_{i,j} h_i) \qquad \text{i.e. } f_{P_{i,j}}(p_{i,j}, \lambda_{i,j}, h_i) = \frac{(\lambda_{i,j} h_i)^{p_{i,j}}\, e^{-\lambda_{i,j} h_i}}{p_{i,j}!} \tag{3.11f}$$

In this formulation, upper-case letters are used to denote random variables, while lower-case letters are used to denote deterministic quantities or particular realizations of the random variables. $T_{\text{arrive}}$ denotes a random variable for the time it takes the bus to slow down, arrive at a stop, and open the doors. $T_{\text{depart}}$ denotes a random variable for the time it takes the bus to close the doors, merge with traffic, and accelerate. $T_{\text{alight}}$ and $T_{\text{board}}$ denote random variables for the time it takes a single passenger to alight and board a vehicle, respectively. The number of passengers alighting and boarding at stop $i$ are denoted by $n_{a,i}$ and $n_{b,i}$, respectively.

From this model we can observe the following:

1. A vehicle is more likely to stop at a bus stop as the leading headway $h_i$ increases, because the probability that a passenger will be waiting at the stop increases. This increases dwell time by increasing the probability that $x_i = 1$.

2. A vehicle is more likely to stop at a bus stop as the number of passengers on board increases, because the probability that a passenger in the bus requests a stop increases. This increases dwell time by increasing the probability that $x_i = 1$.

3. Conditional on the event that the bus stops at a bus stop, the likelihood of a higher number of passengers boarding and alighting increases with increasing leading headway, which in turn increases dwell time through the boardings and alightings term in

(3.11a).

In reality there are more effects than those captured by the model, such as an increase in boarding and alighting times when vehicles are highly loaded, but we can nevertheless reason that these factors may lead to bunching when we consider vehicles in succession. For instance, if at some random time we observe a lightly loaded vehicle followed by a heavily loaded vehicle, the latter will probably experience greater dwell times and the gap between the two vehicles will widen. Similarly, if we observe a sequence of three equally-loaded vehicles, and the gap between the first and the second is much larger than the gap between the second and the third, we expect that the second will slow down and the third will catch up with it. Headways can be regulated with real-time control strategies such as holding at timing points.

Since the simulation model ignores these effects, it could systematically underestimate (downward bias) expected waiting time. The magnitude of the bias will depend on the level of correlations between running times and leading headways.

### 3.6.2 Operator Control

Operator control beyond holding at the terminal is not modeled, but in reality operators can hold at intermediate time points, curtail trips, and deadhead vehicles. If employed carefully, these real-time control strategies can improve headway balance and help attain lower expected waiting times. Therefore, excluding this aspect of operations could result in a systematic overestimation (upward bias) of waiting times.

### 3.6.3 Vehicle and Crew Constraints

Operators must handle vehicle and crew scheduling constraints. For example, a driver might be instructed to end a trip early and deadhead back to the terminal in order to satisfy a maximum number of continuous hours worked before a break, or simply to avoid paying overtime. In the simulation model, which is an abstraction of reality, vehicle and crew constraints are not modeled because it is impractical to obtain accurate crew and vehicle schedules for every hypothetical scenario being tested in the simulation.

This leads to an important difference between real and simulated operations with respect to the criteria for removing vehicles from service. In reality, vehicles might be removed when a driver is nearing the end of his shift. Since the model lacks a notion of drivers and shifts, vehicles in the model are pulled out of service at the first opportunity after the number of active vehicles exceeds the specified vehicle profile at that time. In other words, when there is a decrement in the profile, the first vehicle to reach a terminal is removed. In reality it might be the second or third to arrive at the terminal, since the governing criteria are vehicle and crew schedules rather than a vehicle profile. A real vehicle could be taken out of service before the corresponding decrement in the vehicle profile, if the end of a driver's duty is imminent. Further study is required in order to understand the direction and magnitude of the resulting bias.

## 3.7 Implementation

The simulation model discussed in Section 3.3 was implemented in VisualBasic.NET for the Microsoft .NET Framework 4.0. Random number generation was provided by the built-in Random() .NET library. A run of 1000 replications simulating London Buses route W15 (with the current vehicle profile) was timed on a computer with an Intel Core i5 2.4 GHz CPU and 4 GB of RAM running Windows 7 64-bit. Computation time per replication was estimated to be 0.513 ± 0.002 seconds at 95% confidence level. At this rate, 117 replications can be completed per minute. Of course, the duration of each replication depends heavily on the number of events processed per replication, which in turn depends on the fleet size and the number of locations. Simulations with more vehicles and more locations will take longer.

## 3.8 Verification

In the context of simulation modeling, *verification* is the process that ensures that the algorithm works as intended (Law, 2007). Tests on the individual components of the simulation model were conducted prior to more general verification.

At the heart of the simulation algorithm is the process of drawing random numbers from distributions. A series of tests were conducted to verify the correctness of this component. First, a random variable $X_1$ was defined with the following probability mass function (pmf):

$$f_{X_1}(x) = \begin{cases} 0.10 & x = 1 \\ 0.50 & x = 3 \\ 0.30 & x = 4 \\ 0.10 & x = 5 \\ 0.00 & \text{otherwise} \end{cases} \tag{3.12}$$

One million random draws were made from $X_1$, and the output pmf closely matched (3.12), with sample probabilities matching input probabilities at least to 2 decimal places.

Next, a bivariate pmf was defined as follows,

$$f_{X_2,Y_2}(x,y) = \begin{cases} 0.10 & x = 1, y = 1 \\ 0.50 & x = 3, y = 3 \\ 0.30 & x = 4, y = 4 \\ 0.10 & x = 5, y = 5 \\ 0.00 & \text{otherwise} \end{cases} \tag{3.13}$$

and one million random draws were made from the marginal pmf of $X_2$. Once again, the output distribution closely matched the input probabilities. Another test on the bivariate pmf made one million random draws from (3.13), conditioning on $x = 4$. Every resulting draw had $y = 4$, as expected. Similar tests were conducted with other univariate and bivariate distributions to test the correctness of the algorithmic implementation of the distribution classes within the simulation model, and all were passed.

Further component tests confirmed the following:

```
2012-04-03 21:34:26:  Created simulation instance.  Route 24
2012-04-03 21:34:26:  Loading route geometry.
2012-04-03 21:34:26:  Picked dates:  Period 13 (2011) & April 1, Workdays
2012-04-03 21:34:26:  Loading AVL...
2012-04-03 21:34:41:  Building running time distributions...
2012-04-03 21:34:41:  Getting headways...
2012-04-03 21:34:41:  Getting vehicle profile...
2012-04-03 21:34:41:  Getting vehicles (simple)...
2012-04-03 21:34:41:  Done pre-processing.
2012-04-03 21:34:41:  Starting replication 0
Processing event.  t=17101 v=8292 l=2-0
No vehicle ahead; no holding.
Processing event.  t=17101 v=8292 l=2-1
First stop.  Marginal distribution.  Running time is 347.
Processing event.  t=17448 v=8292 l=2-9
Conditional distribution.  Running time is 198.
Processing event.  t=17646 v=8292 l=2-12
Conditional distribution.  Running time is 168.
Processing event.  t=17814 v=8292 l=2-14
Conditional distribution.  Running time is 136.
Processing event.  t=17950 v=8292 l=2-17
Conditional distribution.  Running time is 365.
Processing event.  t=18001 v=8301 l=2-0
Hold vehicle for 26.  Target headway is 900.  Nearest vehicle 8292 is 874 ahead.
```

**Figure 3-16:** Sample entries of a simulation log file used for verification.

- The event heap returned events in chronological order.
- Bivariate and Univariate distribution classes calculated statistics such as mean, variance, count, and percentiles correctly.
- The StepFunction class, which was used to store the vehicle profile and return the target number of vehicles at any given time of day, worked correctly.
- The correct distributions were being used, given a location and a time.
- The vehicles ahead of, and behind, a given vehicle were identified correctly, and the headways between them, which were calculated based on mean segment running times and linear interpolation, were reasonable.

Following the verification of individual components, more general tests were conducted. The simulation model was verified with an interactive debugger, a detailed log file, and animated playback. The interactive debugger allows running the algorithm line by line, verifying the values of variables and the procedural sequence. The detailed log file is a record of actions and decisions executed during runtime. An example is shown in Figure 3-16. Animated playback shows vehicles moving on the screen throughout the day at an accelerated rate. Figure 3-17 shows a snapshot of animated playback. The three verification tools work together: animated playback allows an analyst to see an entire replication in a reasonable time and detect potential errors, while debugging and reading the log file allow him to understand the root of any suspicious behavior.

The tests outlined in this section indeed helped identify malfunctioning parts. In an early

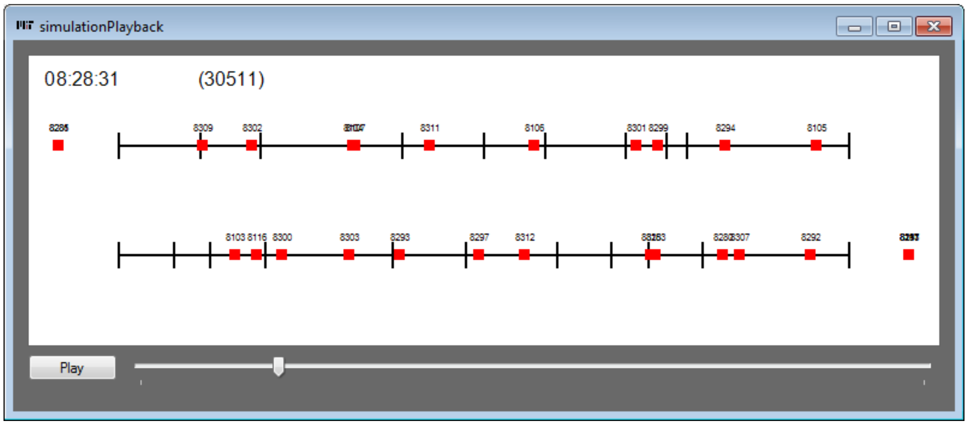


**Figure 3-17:** Verification using animated playback

version of the algorithm, a programming error was causing a small (but statistically significant, and therefore detectable) bias in mean end-to-end running times. This had to do with the way .NET handles casting from double to integer types: it rounds rather than truncates, but the latter had been incorrectly assumed. This was corrected by changing the operation to one that truncates.

In another situation, the model was dispatching two vehicles simultaneously from the terminal in a very few cases. A thorough investigation of this issue found that when two vehicles arrived at the terminal at exactly the same time, neither was ahead of the other for purposes of headway determination, so the next vehicle ahead was used to determine the holding time for both vehicles. This issue was resolved by preventing simultaneous arrivals of two or more distinct vehicles to the same location; whenever such conflict is detected, the arrival time of one of the vehicles is incremented by one second, so it is always clear which vehicle is ahead of the other. Cycles of verification tests, debugging, and issue resolution were repeated until all verification tests were passed.

## 3.9 Validation

Even if the simulation is running as intended, it may not be a sufficiently realistic for the target applications. *Validation* is the process of determining if a model represents the actual system being studied accurately enough for the target applications (Law, 2007). In this research, the validation process compared real performance of a route to simulated performance with the same vehicle profile in terms of end-to-end running times, segment running times, and headways. After observing some discrepancies, further validation tests were conducted with a conceptual route.

### 3.9.1 Validation of Simulations of Real Routes

For validation tests with real routes, simulation runs with 300 replications were conducted for route W15 in London. Means and standard deviations of end-to-end running times, segment running times, and headways (at all timing points) estimated by the simulation model were compared to the same statistics obtained from real operations during weekdays from March 7 to April 1 of 2011. The following time periods were used:

1. morning peak, 7:30–9:30

2. mid-day, 9:30–15:30

3. early afternoon peak, 15:30–16:30

4. afternoon peak, 16:30–18:30

5. evening, 18:30–20:30

Figure 3-18 compares mean segment running times in these time periods for all segments of route W15. Each dot represents a segment during one of the time periods. Dots located near the diagonal line have their simulation mean segment running time closely approximate the sample mean of real observations. As indicated by the general closeness of dots to the diagonal and a root mean squared error (RMSE) of 0.1 minutes, there was good agreement between the simulation and real observations in terms of mean segment running times. (Root mean squared error is an aggregate measure of the difference between two sets of observations. In this case, it measures the estimation error of the simulation with respect to real observations.)

Figure 3-19 compares the standard deviations of segment running times in a similar fashion. It shows good agreement between standard deviations of simulation-estimated segment running times and of real observations, with a RMSE of 0.1 minutes. Low errors were expected for both the means and the standard deviations of segment running times, since segment running times in the simulation are drawn from the distributions of observed segment running times, and an appropriate number of replications was used.

Figure 3-20 compares mean end-to-end running times in the same time periods, with a dot for every direction-period combination. Overall there was good agreement between simulation and observation means. The RMSE is 2.2 minutes, which is relatively low but not negligible. Figure 3-21 compares the standard deviations of end-to-end running times in the same fashion. An RMSE of 0.5 minutes indicates a similar level of agreement between the simulation and real operations in terms of end-to-end running time variability. End-to-end running times in the simulation are generated from the convolution of segment running times, in a process that captures correlation only between adjacent segments. A potential source of error in simulation estimates is neglecting higher-order or more complex correlation structures.

Figure 3-22 compares mean headways in the five time periods. A dot is shown for each period and timing point. While there is positive correlation between simulation and observation mean headways, many headway means are overestimated by the simulation model. The RMSE is 2.4 minutes, which is considerable given that observation mean headways are between 7 and 12 minutes. Figure 3-23 compares the standard deviations of headways at all timing points. In this case there is a slightly stronger positive correlation between the two sets, and no obvious bias. However, the distance of the dots from the diagonal and an RMSE of 2.1 minutes (with respect to observations in the 1–9 minute range) indicate poor agreement between simulation and observation standard deviations.

In developing the simulation model being validated here, no special effort was made to model operator behavior realistically. Real-time control actions for regulating headways

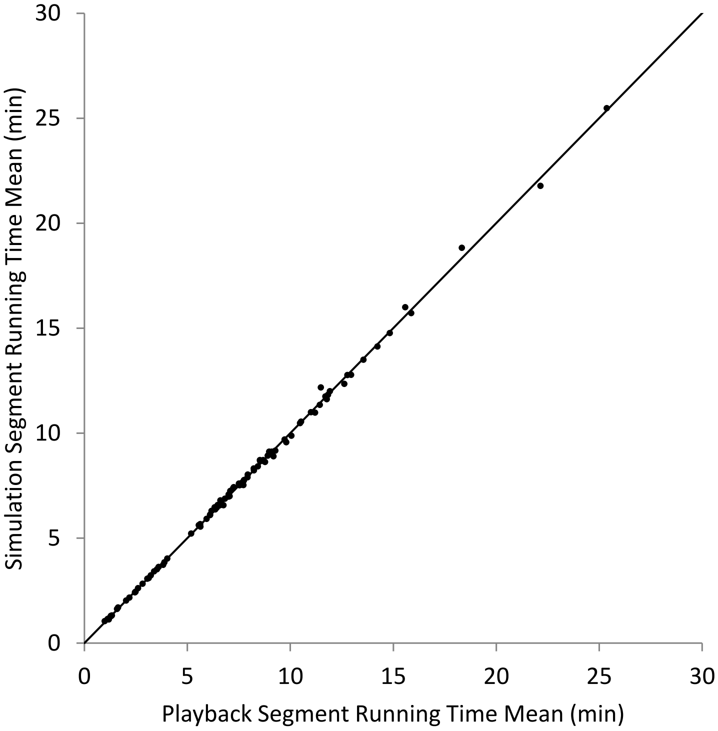


**Figure 3-18:** Validation of segment running time means

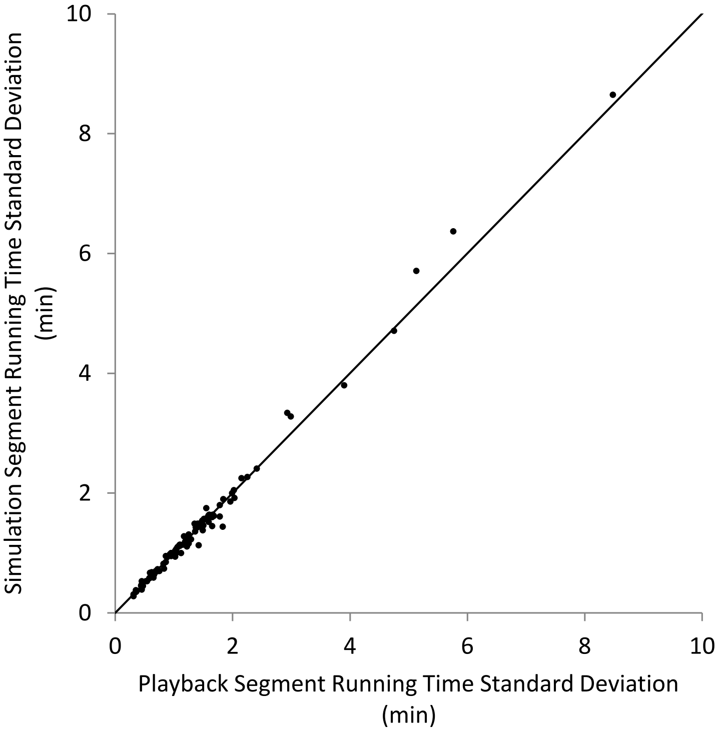


**Figure 3-19:** Validation of segment running time standard deviations

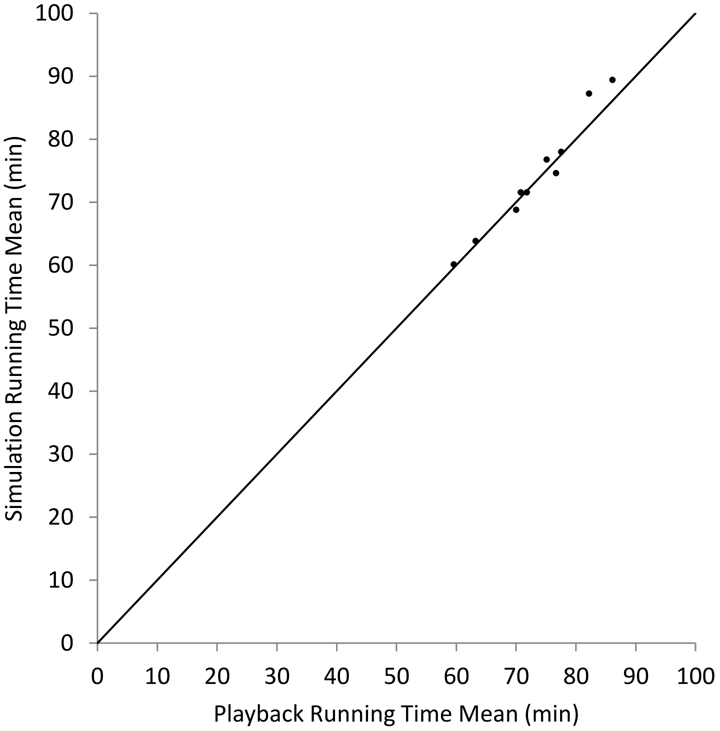


**Figure 3-20:** Validation of end-to-end running time means

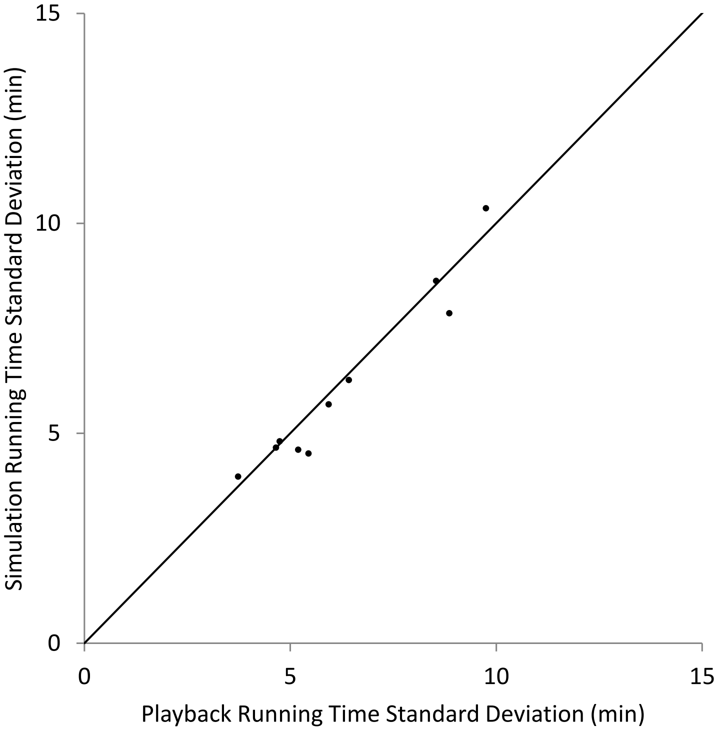


**Figure 3-21:** Validation of end-to-end running time standard deviations

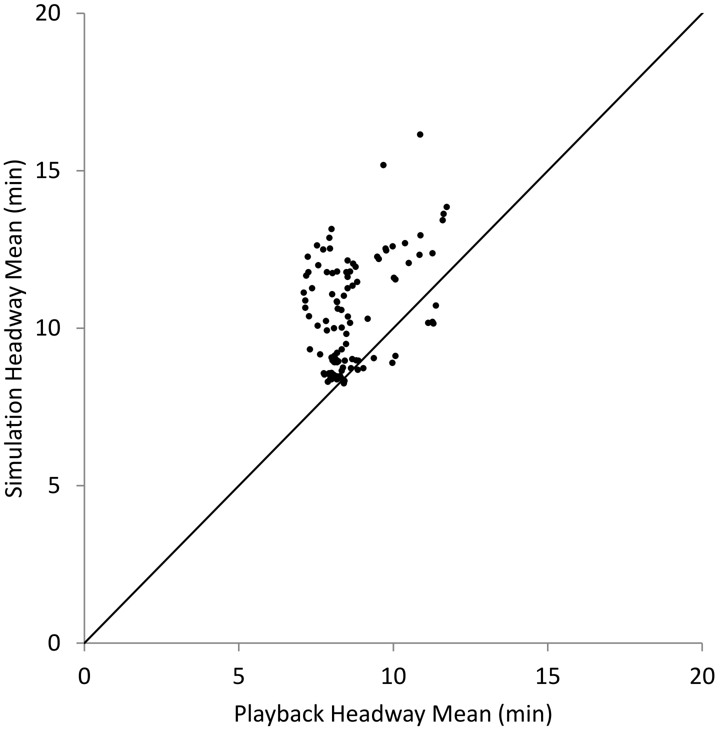


**Figure 3-22:** Validation of headway means

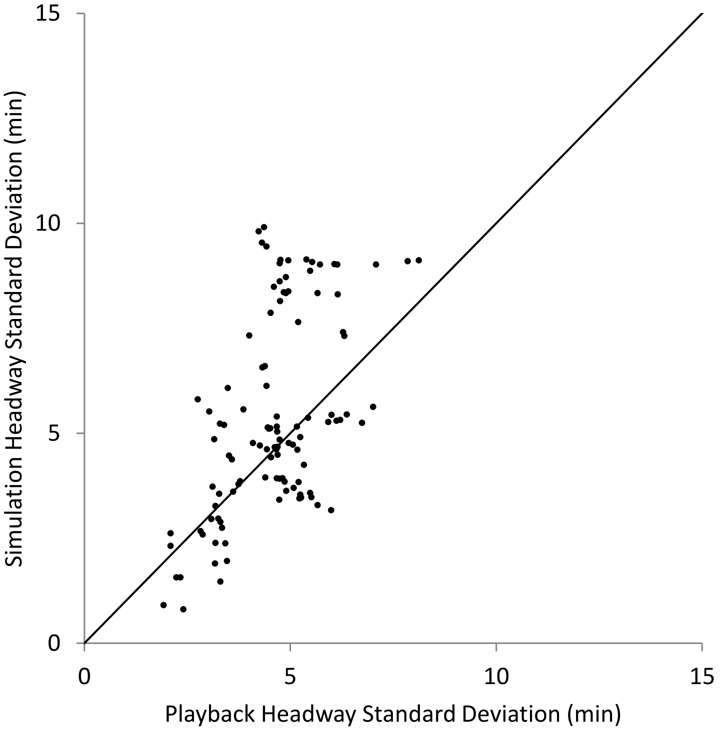


**Figure 3-23:** Validation of headway standard deviations

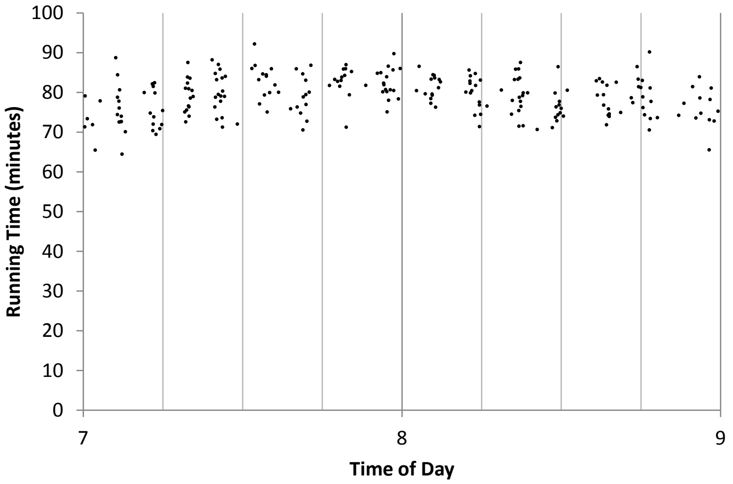


**Figure 3-24:** Route W15 morning running times and dispatch times

(such as holding and curtailing trips) are absent from the model. Moreover, the decision to base dispatching on target headway was not informed by an analysis of alternative dispatch strategies. AVL Playback (discussed in Appendix B) could have been used to pick the best dispatch strategy from those discussed in Section 3.2.3 or to devise a new one. Figure 3-24 shows a scatter plot of running times and dispatch times for one of the directions of London route W15 between 7:00 and 9:00. Clustering of dispatches by time of day is apparent across all weekdays from March 7 to April 1 of 2011, which suggests the operator may be attempting to dispatch vehicles on schedule. If this were the case, the errors seen in estimating headway means and standard deviations could decrease after changing the model's dispatch strategy to reflect this. This example highlights the importance of modeling operator behavior (and other aspects of bus operations) realistically.

Table 3.4 summarizes the validation results of tests on real routes.

**Table 3.4:** Summary of validation results for real routes

| Item | Statistic | Result | RMSE (min) |
|---|---|---|---|
| segment running times | mean | good agreement | 0.1 |
| | standard deviation | good agreement | 0.1 |
| end-to-end running times | mean | good agreement | 2.2 |
| | standard deviation | good agreement | 0.5 |
| headways | mean | overestimated | 2.4 |
| | standard deviation | poor agreement | 2.1 |

### 3.9.2 Validation with a Conceptual Route

Further validation tests were conducted on a conceptual route in order to find potential sources of error. For these tests, the operation of the conceptual route was simulated (running 1000 replications for each case) using generated AVL data, and the resulting means and standard deviations of running times and headways were compared to their AVL counterparts. This technique allowed testing the simulation with input data meeting more of the underlying simplifying assumptions, introducing complexity one step at a time.

The conceptual route had three segments per direction and a target headway of 10 minutes throughout the day. For the first test, the segment running times were exactly 10 minutes throughout the day, which led to deterministic end-to-end running times of 30 minutes. Six vehicles were available throughout the day, which led to cycle times of exactly 60 minutes, without layover time at terminals.

The second test introduced running time variability. Segment running times in the AVL file were drawn independently from a normal distribution with a mean of 10 minutes and a standard deviation of 1 minute. The fleet size was increased to 8 vehicles, which led to cycle times of 80 minutes. For purposes of generating artificial AVL data, stand time at terminals ensured half-cycle times of 40 minutes in each direction.

The third test built upon the second by introducing positive correlation between running times of adjacent segment pairs. Segment running times in the AVL file were drawn from a trivariate normal random variable (one variate per segment) with mean segment running times of 10 minutes in every segment and the following covariance matrix $\mathbf{\Sigma}$ (in squared minutes):

$$\mathbf{\Sigma} = \begin{bmatrix} 1.0 & 0.5 & 0.0 \\ 0.5 & 1.0 & 0.5 \\ 0.0 & 0.5 & 1.0 \end{bmatrix} \tag{3.14}$$

Since only terms off the diagonal by one are non-zero, there is correlation only between running times of adjacent segments. In this scenario, the same trivariate distribution was used to generate running times throughout the day. The fleet size remained at 8 vehicles and the target headways at 10 minutes.

The fourth and final test built upon the previous by introducing peaks. Segment running times were still generated from a trivariate normal random variable with correlation $\mathbf{\Sigma}$, but the random variates were modified by adding 2 minutes of running time to each segment during the peak morning and afternoon hours, 7:30–9:30 and 16:30–18:30, respectively.

Table 3.5 presents results of validation tests on the conceptual route for the four tests described above. The mean and standard deviation of end-to-end running times, segment running times, and headways were measured from the generated AVL data and from the simulation model. The differences between the simulation and AVL statistics (i.e. errors) were calculated by statistic and time period. Minimum and maximum errors are shown in the table to indicate the range of errors. Statistics for segment running times in the first segment and headways at the first stop (for each direction) are shown in addition to those for all segments and all stops.

Simulation results in the deterministic scenario match playback results exactly, with identical means and standard deviations of running times, both at the trip and segment level, and equal headway means and standard deviations. In contrast, the three stochastic scenarios reveal differences between simulation and playback results.

Errors in means and standard deviations of running times in the first segment are relatively small in the three stochastic scenarios. In the simulation model, these segment running times are drawn from the distribution of AVL running times, with no conditioning. Errors

**Table 3.5:** Summary of validation results for conceptual route

| Scenario | Data Item | Statistic | Min Error (min) | Max Error (min) |
|---|---|---|---:|---:|
| deterministic | segment running time, first | mean | 0.0 | 0.0 |
| | | standard deviation | 0.0 | 0.0 |
| | segment running time, all | mean | 0.0 | 0.0 |
| | | standard deviation | 0.0 | 0.0 |
| | end-to-end running time | mean | 0.0 | 0.0 |
| | | standard deviation | 0.0 | 0.0 |
| | headways, first | mean | 0.0 | 0.0 |
| | | standard deviation | 0.0 | 0.0 |
| | headways, all | mean | 0.0 | 0.0 |
| | | standard deviation | 0.0 | 0.0 |
| normal i.i.d. | segment running time, first | mean | 0.0 | 0.1 |
| | | standard deviation | –0.1 | 0.0 |
| | segment running time, all | mean | –0.1 | 0.1 |
| | | standard deviation | –0.2 | 0.0 |
| | end-to-end running time | mean | –0.2 | 0.3 |
| | | standard deviation | –0.2 | 0.0 |
| | headways, first | mean | 0.2 | 1.7 |
| | | standard deviation | 0.6 | 4.4 |
| | headways, all | mean | 0.2 | 1.7 |
| | | standard deviation | –0.7 | 4.4 |
| normal correlated, no peaks | segment running time, first | mean | 0.0 | 0.0 |
| | | standard deviation | –0.1 | 0.0 |
| | segment running time, all | mean | –0.1 | 0.1 |
| | | standard deviation | –0.1 | 0.0 |
| | end-to-end running time | mean | –0.1 | 0.1 |
| | | standard deviation | –0.1 | 0.0 |
| | headways, first | mean | 0.2 | 1.7 |
| | | standard deviation | 0.8 | 4.5 |
| | headways, all | mean | 0.2 | 1.7 |
| | | standard deviation | –0.5 | 4.5 |
| normal correlated, with peaks | segment running time, first | mean | –0.1 | 0.1 |
| | | standard deviation | –0.1 | 0.0 |
| | segment running time, all | mean | –0.3 | 0.3 |
| | | standard deviation | –0.2 | 0.2 |
| | end-to-end running time | mean | –0.1 | 0.2 |
| | | standard deviation | –0.5 | 0.0 |
| | headways, first | mean | 0.1 | 1.5 |
| | | standard deviation | 0.6 | 4.2 |
| | headways, all | mean | 0.1 | 1.5 |
| | | standard deviation | –0.6 | 4.2 |

in means and standard deviations of segment running times in all segments are also small for the first two stochastic scenarios, and slightly larger for the last scenario, which features peaking. In the simulation model, running times of segments after the first are drawn from a conditional distribution in order to capture possible correlations between running times of adjacent segment pairs. Peaking introduces dynamic higher-order correlation in segment running times. For example, in the transition to the afternoon peak, there is positive correlation between running times of the first and third segments, since both are increased by 2 minutes. This additional correlation, which is not captured in the model, may be leading to larger errors.

Simulation end-to-end running times arise from mixing of segment running times, so it is expected that errors in the means of end-to-end running times are of similar magnitude to those of segment running times. This appears to be the case in the scenario with normal correlated segment running times with no peaks. The scenario with normal i.i.d. segment running times exhibits a wider range of errors in end-to-end running times with respect to its segment-level errors. In contrast, the scenario with normal correlated segment running times with peaks exhibits a narrower range of errors at the direction level than at the segment-level. These differences could stem from how errors are accumulated. For example, if an error of +0.1 minutes per segment is accumulated, the error at the direction level can be +0.3 minutes, as in the normal i.i.d. scenario. If the segment-level errors have different signs, then the end-to-end error might be smaller, as might have happened in the last scenario.

The range of errors for end-to-end running time standard deviations in the two stochastic scenarios without peaking correspond well to their segment-level counterparts. A difference is seen only in the last scenario, which features morning and afternoon peaks. The range of errors in this case goes from −0.5 minutes to 0.0 minutes, indicating a consistent underestimation of running time variability by the simulation. Since this is evident only in the scenario with peaking, systematic transitions in running times are a likely cause because they introduce higher-order correlations that the simulation model is not designed to capture. In fact, the −0.5 minute error corresponds to the early afternoon peak period, which has gradually increasing running times toward the afternoon peak period.

The range of errors for headways in all three stochastic scenarios are relatively wider than for running times, suggesting that headways are not accurately represented in the simulation. The AVL data used for this validation was generated based on on-schedule terminal dispatching, without any strategy for en route headway regulation. Because dwell times are not explicitly modeled in the dataset, headways are exogenous, and the endogeneity problem discussed in Section 3.6.1 does not have an effect. (One of the benefits of validation with a conceptual route is the ability to separate sources of potential errors.) Given these properties, errors in headway estimation shown in Table 3.5 are most likely due to the headway-based dispatching strategy used in simulation vs. the schedule-based dispatching in AVL data. Reinforcing this argument, maximum errors in mean and standard deviation, as well as minimum errors in mean for the first stop are the same as for all stops, suggesting that the largest errors in simulation headways are seen right after dispatch.

All errors for mean headways are positive in the three stochastic scenarios, indicating a consistent overestimation of mean headways in the simulation model. It is possible that, despite verification efforts, there are remaining implementation errors causing this. Errors for standard deviation of headways in the first stop are also all-positive. Aside from a possible undetected implementation error in the simulation model dispatch controller, this upward bias suggests that schedule-based dispatching may yield better headway regularity than target-headway dispatching in cases where a conservative amount of recovery time is provided at the terminals. Further investigation could confirm or refute this.

## 3.10 Conclusion

Performance analysis tools discussed in Chapter 2 help characterize current operating conditions and identify opportunities to improve service quality through addition of vehicles. Having recognized the need to carry out cost-benefit analyses to study proposals of resource changes to a route, this chapter discussed the potential role of simulation models in estimating how service quality may improve or worsen after addition or removal of vehicles.

A general framework for simulation modeling of transit operations was developed. Although it was devised with bus service in mind, it is flexible enough to model various bus lines operating together or rail transit. Special emphasis was placed on the use of automatically collected transit data to obtain input parameters for the model. A simple simulation model of bus service operating between two terminals was developed as an example. This model did not use many of the capabilities of the general framework; for instance, it did not consider real-time control strategies such as holding and curtailments. However, it was capable of capturing pairwise correlation of segment running times.

The simulation model was then subjected to verification and validation tests. Verification tests on individual components and the overall simulation algorithm did not reveal errors. Validation of simulation of a real bus route was based on comparing headway and running time statistics calculated from real AVL data and the simulation model. Running time means and standard deviations at the segment and direction level were estimated with relatively small errors. Headway means were largely overestimated, and there was a wide range of errors in estimates of headway variability. To study the matter further, simulations for a conceptual route were carried out based on artificially generated AVL data. Once again, there was generally good agreement between AVL and simulation running time statistics, and larger discrepancies in headway statistics.

The mismatch of headway means and standard deviations was most likely due to inaccurate modeling of terminal dispatching strategies. Headway endogeneity in running times and real-time control mechanisms aimed at regulating headways may also have contributed to errors in the case of the simulation of the real route. In spite of successful verification tests and extensive code revisions, it is possible that errors in the algorithm design or programming led to these differences. This highlights the importance of modeling operator behavior accurately and verifying and validating simulation models before using them to estimate performance measures under hypothetical scenarios.

# Chapter 4

# Service Performance and Resource Allocation

## 4.1 Introduction

Chapter 3 introduced a simulation model designed to evaluate performance of bus routes under hypothetical scenarios, in particular scenarios that vary the number of vehicles available to operate a service at a given headway. This provided a convenient way of using automatically collected data to estimate the performance improvements gained by an additional vehicle, or conversely, the performance degradation resulting from removing a vehicle. By combining model forecasts with knowledge of cost (or savings) associated with such a change, it becomes possible to conduct cost-benefit analysis to support decisions regarding resource allocation.

Nevertheless, the scope of analysis was limited to one route, and there was no consideration of a budget. Expanding the scope to a set of (independent) bus routes, and considering an operating budget in the form of a total fleet size, the following question arises: what resource allocation maximizes service performance?

Most likely, the management of a transit agency such as London Buses has, through careful service planning and performance monitoring, set the fleet size on most routes at an appropriate level. However, the nature of this ongoing task places more emphasis on problem routes; for instance, routes that are suffering delays due to roadwork or routes that are heavily crowded. Management will receive more complaints from passengers using these services, or it might have the initiative to revise resource levels in anticipation of such problems. In the meantime, there may be undetected opportunities to save resources on routes that are performing well and have excess resources, as well as to improve service performance of routes that could greatly benefit from added resources but have not yet received special attention. This chapter addresses the problem of assigning vehicles to routes of a bus network, given a target headway for each route and a fixed total fleet size, with the objective of maximizing total service performance.

There are a number of factors that affect service performance. Some are in control of the operator at the time of operation, such as the dispatching discipline used at terminals and the set of strategies used to regulate headways. Others, like infrastructure, congestion, and demand for service are considered fixed in the short term. Still others, like scheduled cycle time, are in control of management at the service planning stage. Greater cycle times result in greater stand times at terminals, thereby making it easier to dispatch vehicles regularly. In the service planning stage, increasing cycle times is an effective strategy to manage running time variability.

Cycle time $c$ is determined by target headway $h$ and fleet size $n$, through the relationship $c = nh$. Greater cycle times can be achieved by increasing fleet size, increasing headway, or a combination of both. Increasing headway while holding fleet size constant increases cycle time without additional operating cost, but it decreases the passenger carrying capacity of the service, making vehicles more crowded, in addition to increasing passenger waiting times. Although the scheduled cycle time will have increased, dwell times at stops will increase and become more variable, so the net effect could very well be lower service quality if ridership is high. When ridership is relatively low and the operator is struggling to dispatch vehicles at regular intervals, the additional recovery time provided by an increased headway can improve service reliability.

An alternative way to increase cycle times is to increase fleet size while holding the target headway constant. This does increase operational cost, but it adds slack to running times without directly affecting waiting times or loads. If the operator uses the additional resources correctly, headway regularity can improve, which in turn decreases waiting times and balances loads from vehicle to vehicle. Balanced loads can, in turn, reduce running time variability through less variable dwell times. The extra slack can be used at the terminals to improve dispatch regularity, but also en route in the form of holding or slowing vehicles strategically to maintain regular headways. In the latter case, running time variability decreases at the expense of higher typical running times. Increasing fleet size is perhaps the most effective way of dealing with running time variability at the service planning stage.

Of all the factors that affect service quality, only fleet size will be considered in this chapter. Headway determination responds to a wide range of factors, including policy and network effects. Simultaneous consideration of headway and fleet size would certainly be a more complete approach to optimizing resource allocation, but also a much more complex one falling outside of this thesis's scope. The premise of this chapter is that the target headway of each route has been appropriately set based on service delivery policy and demand for service. The objective is to allocate resources in a way that maximizes service quality in the current set of routes, without altering the service plan.

This chapter is organized as follows: Section 4.2 presents a framework to optimize resource allocation, and Section 4.3 discusses a specific optimization method. The problem is solved separately for each time period (e.g. first the morning peak, then the mid-day, then the afternoon peak, etc.) One of the challenges faced is the systematic variation of running times and headways within each of these time bands. An operator will vary the fleet size of a route according to this, providing the peak vehicle requirement of each time period only when it is needed. Since schedule data is not available for the hypothetical scenarios with smaller or larger fleets, vehicle profiles must be estimated. Section 4.4 discusses this further and proposes heuristics to overcome the obstacle. Finally, a simplified case study is presented in Section 4.5 and Section 4.6 closes with a summary and concluding remarks.

## 4.2 Framework

Increasing the fleet size of a bus service can enhance an operator's ability to maintain regular headways, thereby improving service performance. It is in everyone's interest to have good performance, but adding resources is feasible only up to the overall fleet size (or budget) constraint. Optimizing resource allocation involves identifying where resources are needed the most. The objective is to find the resource allocation that attains the best overall service performance under an existing service plan, present operating conditions, and specified overall fleet size.

### 4.2.1 Service Performance

Service performance is a general term that can capture many quantifiable aspects of bus operations. The primary factors relate to service quality as perceived by passengers. For example, performance may include measures of waiting time and crowding. These may be expressed in absolute terms or relative to a performance standard. For example, the average number of minutes of wait experienced by passengers is an absolute measure of waiting time, while excess waiting time is expressed relative to the headway specified in the service plan. Performance measures based on vehicles (rather than passengers), such as the standard deviation of headways, can also be included. Moreover, service performance can capture general aspects of operations, such as congestion at terminals due to excess number of vehicles laying over (i.e. standing).

When multiple criteria are driving the optimization process, the relative importance of each factor must be considered. There may be a situation in which adding a vehicle to one route improves one aspect of service performance considerably while not significantly affecting a second aspect. Adding the vehicle to a different route instead may considerably improve the second aspect of service performance while not affecting the first very much. In such cases, the trade-off between the two aspects must be considered in deciding to which route the vehicle should be allocated.

The performance improvement resulting from adding a vehicle to a route will decrease as fleet size increases. When a route is severely under-resourced, the addition of one vehicle can have a profound impact on service performance. In contrast, when a route has excess resources, the addition of one vehicle might not have visible effects on performance. Figure 4-1 illustrates the conceptual relationship between fleet size and the performance of a route. There are four regimes in this relationship, separated by dotted vertical lines. At least one vehicle must be allocated to provide service. Once service is being provided, additional vehicles will bring about performance improvements, albeit at a decreasing rate. The rate will continue to decrease until adding a vehicle does not result in visible performance improvements. While there are no capacity problems as a result of a large fleet size, continuing to add vehicles will not affect performance. However, having an excess number of vehicles can degrade performance if, for instance, there is insufficient capacity at terminals for vehicles to lay over (i.e. stand) and vehicle congestion at terminals is one of the performance factors under consideration. The exact shape of the performance function depends on the factors composing it and the way operators manage the different levels of resource. However, it is hypothesized that many practical performance functions will be concave, as shown in Figure 4-1.

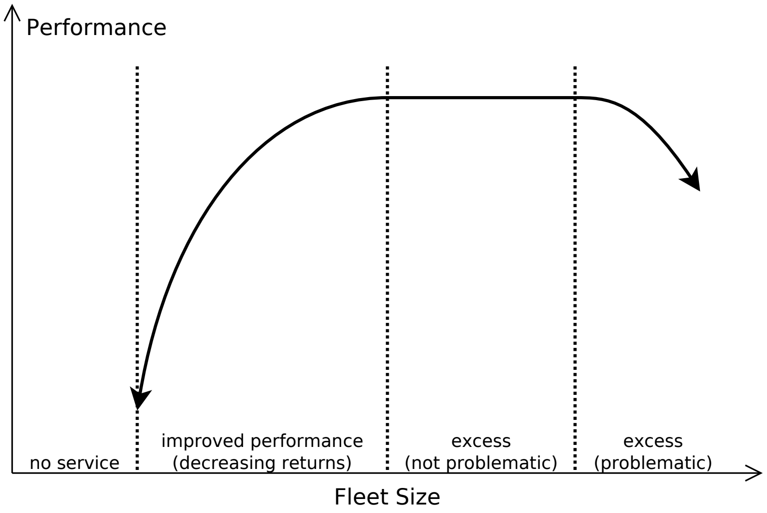


**Figure 4-1:** General relationship between performance and fleet size

Factors such as waiting time and loads should decrease as resources increase, as long as the operator behavior is consistent across allocations. These factors are "negative" in the sense that, for example, higher waiting times imply lower performance. If a "positive" performance function is desired, these factors must be transformed somehow (for example, by taking negative waiting time as a measure of performance). If all aspects of performance being considered are "negative", an alternative is to include them untransformed in the performance function. In this case, the performance function will be convex rather than concave, with lower values of it signifying better performance.

Were the objective to maximize performance of a single route, the optimal allocation would add vehicles until performance stops improving or until all available vehicles have been allocated, whichever comes first. In the absence of resource constraints, the same logic applied independently to each route would yield optimum performance over a set of routes. However, the resulting allocation may not be feasible when total available resources are constrained. In this case, resources should be first allocated where they yield the largest performance benefit, until the total available resources are exhausted.

Figure 4-2 provides an illustrative example. Route A is shown with a solid line and Route B with a dashed line. The current allocation, marked with boxes on each route, is 6 vehicles to Route A and 7 to Route B. In a scenario where total fleet size cannot increase, removing a vehicle from Route B and adding it to Route A improves total performance. The optimization method presented in Section 4.3 makes systematic changes of this type until an optimal solution is reached.

### 4.2.2 Existing Service Plan and Operating Conditions

As stated earlier, the objective is to find the optimum resource-constrained allocation under an existing service plan and operating conditions. Since the matter in question is how performance responds to changes in fleet size, fleet size will be varied while other factors remain controlled. Route alignment, span of service, and specified headway by time of day

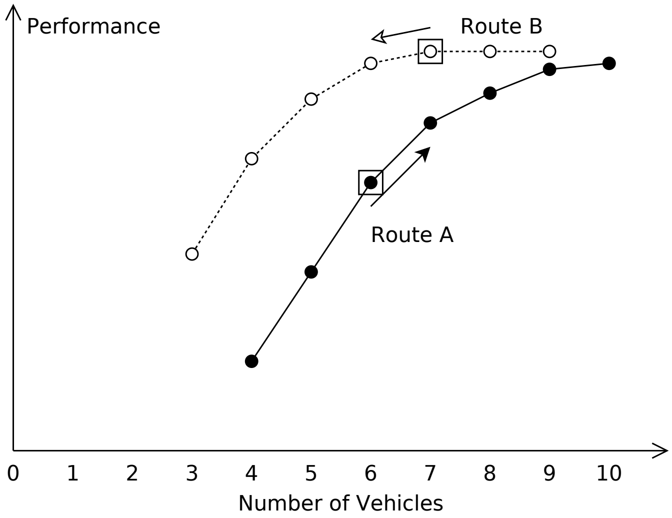


**Figure 4-2:** Illustrative performance curves and optimization mechanism

(all of which form part of the service plan) will be held constant. Moreover, operating conditions such as levels of traffic, road geometry, and ridership patterns will be assumed constant. The implication is that running times (possibly excluding dwell times) will remain constant, independent of fleet size.

The effect of changing the fleet size of a route will be partially governed by operating behavior. For instance, if dispatching at terminals is based on a target headway, the mean headway will not fall below the specified headway, regardless of how much resource is allocated. On the other hand, if terminal dispatching is based on even headways, the specified headway does not come directly into play. In this case, observed mean headway may decrease below the specified headway if the resource allocation is high enough. This implies, in effect, a simultaneous change in fleet size and operating headway, which is not the intent of this framework.

Allocations that result in a mean observed headway below the specified headway might be considered excessive. Accordingly, maximum fleet size limits for each route can be imposed to prevent this from occurring. These can be set based on the percentile of the running time distributions implied by a given fleet size and the specified headway. For example, allocations that imply a percentile above 95 might be regarded as excessive and not considered.

A key assumption of this optimization framework is that the current operating conditions will not change as a result of changes in fleet size. This assumption may be valid only under small changes in resource allocation and in the short run. Large changes in fleet size may attract new ridership or drive away current passengers to alternative routes or modes, especially if a common corridor is shared among multiple routes. It would be possible to capture these effects with corridor- or network-level ridership models, but this falls outside the scope of this thesis. Instead, a limit on the maximum change in fleet size per route can be imposed. For example, adjustments of up to two vehicles added or removed could be considered.

### 4.2.3 Optimization Model Formulation

The problem is formulated as an optimization problem. Its structure takes the following form:

$$\text{maximize} \sum_{r \in R} Q_r(x_r) \tag{4.1a}$$

$$\text{s.t.} \sum_{r \in R} x_r = n \tag{4.1b}$$

$$x_r \in \mathbb{Z}^+ \quad \forall r \in R \tag{4.1c}$$

where $R$ denotes the set of routes under consideration, $Q_r(x_r)$ is the performance function of the $r^{\text{th}}$ route, $x_r$ is the number of vehicles assigned to the $r^{\text{th}}$ route, and $n$ is the total fleet size.

$Q_r(x_r)$ in (4.1a) captures the service performance of route $r$ when it operates with $x_r$ vehicles, using a linear combination of performance measures. The function can represent performance measures such as expected waiting time and in-vehicle crowding, with weights reflecting the relative importance of each. For example, it could take the following form:

$$Q_r = -\left(5 \cdot \text{EWT} + \text{EL}\right) \tag{4.2}$$

where EWT is the route-level excess waiting time in minutes and EL is the route-level excess load, some measure of in-vehicle crowding measured in units of passengers. The 5 preceding EWT means that an excess load of five passengers is as onerous as one minute of excess waiting time. The sign of the terms is reversed to make them "positive", such that higher values of $Q_r$ indicate better performance.

The set of routes in $R$ can range from one route to all routes in a network. It can contain only routes belonging to the same operator, or routes based on the same depot, in case it is desirable to maintain the current total resource levels on an operator or depot basis.

Constraint (4.1b) ensures that the sum of individual route fleet sizes equals the total available fleet size, while constraint (4.1c) limits vehicle allocations to (strictly) positive integers. More constraints can be added to enrich the model formulation.

A limit on the maximum change in fleet size of each route can be imposed. Such a constraint is necessary for the assumption of a constant operating environment independent of fleet size to remain valid, as discussed in Section 4.2.2. To allow a maximum change of $\pm\Delta_{\max}$ vehicles, the following constraint is added:

$$|x_r - x_{r,0}| \leq \Delta_{\max} \tag{4.3}$$

where $x_{r,0}$ denotes the original allocation of vehicles assigned to route $r$ and $\Delta_{\max}$ is the maximum allowable change. Setting $\Delta_{\max} = 2$ limits changes to ±2 vehicles.

As discussed in Section 4.2.2, it may be desirable to avoid allocating an excessive amount of vehicles to a route given the running time distribution and specified headway $h_r$, since this may allow an operator to serve the route at a frequency higher than specified (which implies simultaneous changes in fleet size and frequency). Scheduled cycle time can be limited to

$c_{r,\max}$ by imposing the following constraint:

$$x_r \leq \frac{c_{r,\max}}{h_r} \tag{4.4}$$

The cycle time limit can be determined based on a maximum allowable percentile of running times. For example, $c_{r,\max}$ could be set to the sum of 95th percentile running times in each direction.

In a similar fashion, it may be desirable to impose a minimum acceptable performance level. To enforce a minimum of some aspect of performance $q$, a constraint of the form

$$q_r \geq L_r \quad \forall r \in R \tag{4.5}$$

is included in the formulation, where $q_r$ is a particular component of $Q_r$ and $L_r$ is its lower limit. This is primarily applicable to "positive" factors of performance. For "negative" factors, an upper limit is appropriate:

$$q_r \leq U_r \quad \forall r \in R \tag{4.6}$$

where $U_r$ is an upper limit of $q_r$. For example, a constraint of the form $\mathrm{EWT}_r \leq 5$ would prevent an excess waiting time above five minutes. This same type of constraint can be used in cases where there is concern that the addition of vehicles to a route given by the optimal solution (in terms of service quality criteria) leads to a congestion problem at the terminal because the physical limit on the amount of vehicles that can be standing simultaneously is exceeded. This is achieved by letting $q_r$ in (4.6) be some measure of congestion.

It is possible to penalize some aspect of service performance without explicitly constraining it. For example, a penalty can be imposed on exceeding the capacity of a terminal for standing vehicles. To do so, the objective function takes the following form:

$$\text{maximize} \sum_{r \in R} \left(Q_r(x_r) - P_r(x_r)\right) \tag{4.7}$$

where $P_r(x_r)$ is a penalty due to exceeding terminal capacity. For instance, $P_r(x_r)$ might be the duration in minutes of instances where more vehicles than the specified limit are standing simultaneously, multiplied by a coefficient to weigh the penalty against the different performance criteria. This type of penalty is sometimes called a "soft" constraint.

## 4.3 Optimization Method

### 4.3.1 Solution Framework

Three processes work together towards the optimal solution:

1. an *optimizer* allocates $x_r$ vehicles to route $r \in R$, subject to a set of constraints, and adjusts this allocation in search of better performance;

2. a *vehicle profiler* determines vehicle profiles (i.e. the number of available vehicles by time of day) given the allocation specified by the *optimizer*; and

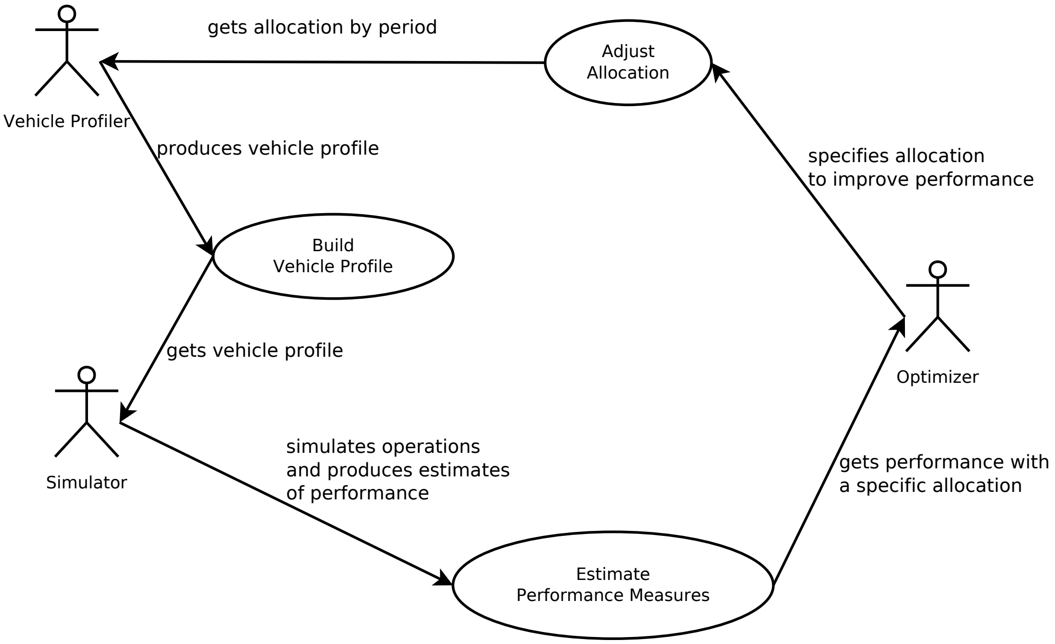

gets allocation by period
Adjust Allocation
Vehicle Profiler
produces vehicle profile
specifies allocation to improve performance
Build Vehicle Profile
gets vehicle profile
Optimizer
Simulator
simulates operations and produces estimates of performance
gets performance with a specific allocation
Estimate Performance Measures

**Figure 4-3:** The three algorithms involved in optimizing vehicle allocation to routes

3. a *simulator* models operations of route $r$ with a given vehicle profile and estimates the performance measures with which the objective function is evaluated.

A use case diagram showing the roles these processes play in the general optimization methodology can be found in Figure 4-3. Arrows are used to indicate the sequence in which the three processes interact.

This chapter focuses on the optimizer process, and assumes that auxiliary *vehicle profiler* and *simulator* algorithms as described above are available. The simulation model presented in Chapter 3, or an extended version of it, can be used as the *simulator*, since it estimates performance measures given a vehicle profile. The model should be as realistic as possible in terms of running times, dwell times, operator behavior, and ridership patterns. Simulation gives estimators of performance measures, which come from a stochastic environment. A different estimate will be obtained from each replication, so each run should have a tight confidence interval (by way of an appropriate number of replications). See Chapter 3 for further details. Devising a realistic *vehicle profiler* to translate $x_r$ to a vehicle profile is a non-trivial aspect of the problem, which will be discussed further in Section 4.4.

The allocation problem is solved independently for each time period, so it is required that a single set of common time periods is used for all routes. A few reasonably long time periods should be used rather than many short ones; this is because the resulting number of vehicles assigned to each route can vary from time period to time period, and a solution would not be practical if this fluctuated, for instance, every 30 minutes. (This is less of a concern with interlined operations, in which case vehicles are assigned to multiple routes simultaneously and it becomes possible to have more frequent variation of resources by route. However, interlining is not typically seen in high-frequency bus operations, so it is not considered in this thesis.) The allocation problem should be solved for the first time period first, then for the second, etc, because previous time periods dictate the initial conditions of the time period being analyzed.

The objective is to maximize total performance, which in the formulation is the sum of performance functions of all routes. In the absence of restrictions on the shape of performance functions, their sum can take on an arbitrary shape, and the only way of finding the (globally) optimal allocation is to enumerate all feasible solutions. It was argued in Section 4.2.1, however, that most practical specifications of performance functions are concave. Assuming this is the case, the objective function is also concave, so it has only one local maximum, which is also the global maximum[1]. This has an important implication: if at some point in the optimization process there is no way to improve the solution by taking a local step, that solution is guaranteed to be globally optimal. The concavity assumption can be checked during the optimization process, and if a violation is detected, a more thorough analysis can be made.

The optimization procedure is broken down into two stages. The first stage of the procedure deals with obtaining a solution that uses exactly $n$ vehicles. The second stage moves vehicles from one route to another until there are no further changes that can improve service quality. Both stages employ a greedy algorithm that takes discrete local steps of one vehicle. An activity diagram of the method is shown in Figure 4-4. The result is globally optimal under the global concavity assumption discussed previously.

### 4.3.2 Achieving the Target Total Fleet Size

1. An initial solution is obtained, either by setting each $x_r$ to the current values or by choosing positive values arbitrarily. The solution need not meet constraint (4.1b) but must be otherwise feasible.

2. $Q_r(x_r)$ is evaluated with the current assignments $x_r$.

3. This step is carried out only if there are less than $n$ vehicles assigned.

   (a) $Q_r(x_r + 1)$ is evaluated for all routes.

   (b) The marginal benefit of adding a vehicle to route $r$, denoted by $\delta_r^+$, is determined for every route.

   $$\delta_r^+ \leftarrow Q_r(x_r + 1) - Q_r(x_r) \tag{4.8}$$

   Let $r_*^+$ denote the route with maximum $\delta_r^+$.

   $$r_*^+ = \text{argmax}_{r \in R}\{\delta_r^+\} \tag{4.9}$$

   (c) One vehicle is added to the route with greatest marginal benefit.

   $$x_{r_*^+} \leftarrow x_{r_*^+} + 1 \tag{4.10}$$

[1] A mathematical property of concave functions is that the sum of multiple concave functions is concave. By extension, the sum of multiple convex functions is convex. See Hillier and Lieberman (2010) or Bertsimas and Tsitsiklis (1997) for details.

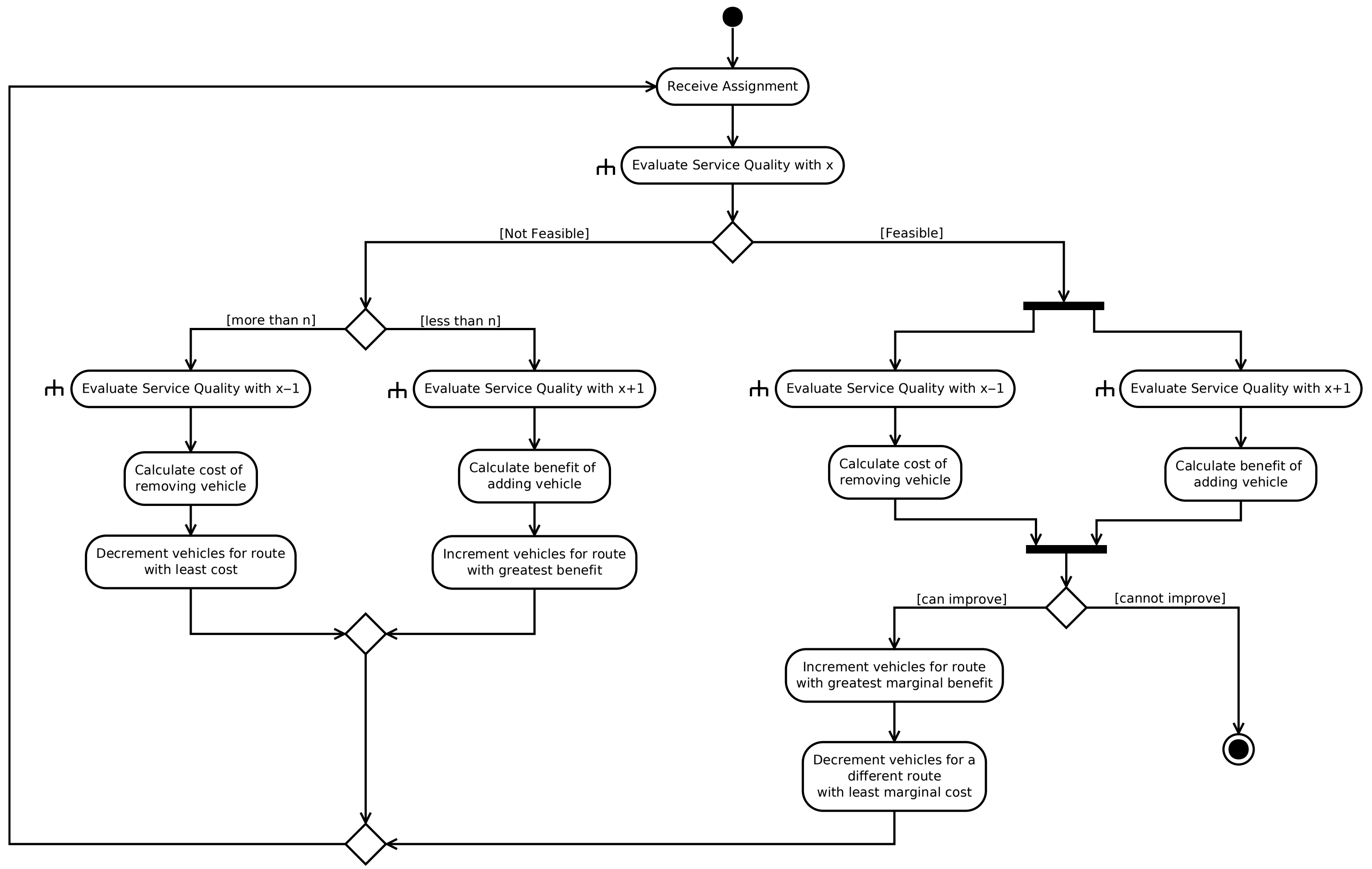

Receive Assignment
Evaluate Service Quality with x
[Not Feasible]
[Feasible]
[more than n]
[less than n]
Evaluate Service Quality with x−1
Evaluate Service Quality with x+1
Calculate cost of removing vehicle
Calculate benefit of adding vehicle
Decrement vehicles for route with least cost
Increment vehicles for route with greatest benefit
Evaluate Service Quality with x−1
Evaluate Service Quality with x+1
Calculate cost of removing vehicle
Calculate benefit of adding vehicle
[can improve]
[cannot improve]
Increment vehicles for route with greatest marginal benefit
Decrement vehicles for a different route with least marginal cost

**Figure 4-4:** Activity Diagram of optimization method

(d) Step 3 is repeated until the number of vehicles assigned equals the target fleet size $n$. At this point, the solution is feasible and the algorithm continues with the steps of Section 4.3.3.

4. This step is carried out only if there are more than $n$ vehicles assigned.

   (a) $Q_r(x_r - 1)$ is evaluated for all routes.

   (b) The marginal cost of removing a vehicle from route $r$, denoted by $\delta_r^-$, is determined for every route.

   $$\delta_r^- \leftarrow Q_r(x_r) - Q_r(x_r - 1) \tag{4.11}$$

   Let $r_*^-$ denote the route with minimum $\delta_r^-$.

   $$r_*^- = \text{argmin}_{r \in R}\{\delta_r^-\} \tag{4.12}$$

   (c) One vehicle is removed from the route with the least marginal cost.

   $$x_{r_*^-} \leftarrow x_{r_*^-} - 1 \tag{4.13}$$

   (d) Step 4 is repeated until the number of vehicles assigned equals the target fleet size $n$. At this point, the solution is feasible and the algorithm continues with the steps of Section 4.3.3.

### 4.3.3 Obtaining the Optimal Solution

Once a feasible solution is obtained, the following steps are followed to arrive at an optimal solution:

1. Both $Q_r(x_r - 1)$ and $Q_r(x_r + 1)$ are evaluated for all routes.

2. The marginal benefit of adding a vehicle is determined for every route.

   $$\delta_r^+ \leftarrow Q_r(x_r + 1) - Q_r(x_r) \tag{4.14}$$

   Likewise, the marginal cost of removing a vehicle is determined for every route.

   $$\delta_r^- \leftarrow Q_r(x_r) - Q_r(x_r - 1) \tag{4.15}$$

   Let $r_*^+$ denote the route with greatest marginal benefit of adding a vehicle and $r_*^-$ denote a different route with least marginal cost of removing a vehicle.

   $$r_*^+ = \text{argmax}_{r \in R}\{\delta_r^+\} \tag{4.16a}$$
   $$r_*^- = \text{argmin}_{r \in R \setminus r_*^+}\{\delta_r^-\} \tag{4.16b}$$

3. If $\delta_{r_*^+}^+ > \delta_{r_*^-}^-$, a vehicle is removed from $r_*^-$ and added to $r_*^+$. This keeps the total

number of vehicles at $n$ and increases the objective function value.

$$x_{r^-} \leftarrow x_{r^-} - 1 \tag{4.17a}$$
$$x_{r^+} \leftarrow x_{r^+} + 1 \tag{4.17b}$$

4. Steps 1–3 are repeated until $\delta^+_{r^+_*} \leq \delta^-_{r^-_*}$. At this point, the solution is optimal and the algorithm terminates.

### 4.3.4 Maintaining Feasibility

Feasibility must be maintained throughout the optimization process. If the formulation includes constraints imposing a maximum allowable change in fleet size, a maximum implied cycle time, or lower or upper bounds on some aspect of service performance, only fleet sizes meeting these constraints can be considered. For example, adding a vehicle to route $X$ will not be considered if this causes unacceptable congestion at terminals, even if this allocation increases total performance the most. The next best route becomes the candidate to receive the additional vehicle.

## 4.4 Building a Vehicle Profile

The optimization algorithm presented in Section 4.3 makes reference to a *simulator* that evaluates service performance of routes under different vehicle allocations. For reasons described in Section 4.2, the simulation model developed in Chapter 3, or more likely an extension of it, is appropriate for this purpose. Unfortunately, while the optimization problem deals with a single $x_r$ per period, real bus operations may exhibit within-period variation in fleet size. The simulation model is capable of capturing this, but it requires a vehicle profile as an input.

In London, TfL specifies the target headway by time of day and a minimum performance standard (measured in terms of excess waiting time for high-frequency service). It is then up to the (private) operator to devise a vehicle schedule to meet the specification based on their analysis of running time data, and to defend it during the contract award process (Barry, 2012). Within the limits of the specification, the operator will try to minimize the operating cost associated with the schedule in order to submit a competitive bid. Hence, vehicle and crew constraints known only to the operator are built into the operations plan, and the number of active vehicles varies accordingly. It is unlikely that an operator will plan to operate with the peak vehicle requirement at times during the peak period that require a lower number of vehicles. Although this variation will be known to TfL with the submittal of the vehicle schedule, hypothetical schedules for situations with a smaller or larger fleet size are not available.

Faced with this reality, we must devise a way of building a vehicle profile given $x_r$. The approach used here modifies an existing vehicle profile. There are multiple ways to do this, each with its implications on estimates of service performance depending on the operating strategies modeled. Three heuristics are described below and illustrated in Figure 4-5. In the illustration, $x_r$ is one vehicle less than the number of vehicles currently used. The

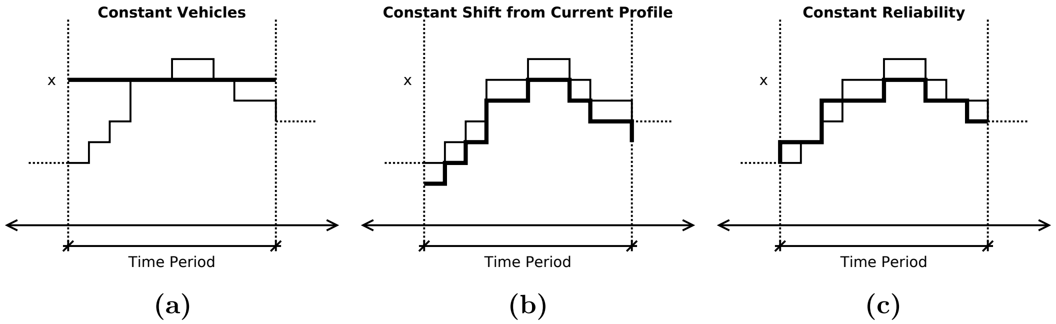


**Figure 4-5:** Three methods for obtaining a vehicle profile from $x_r$

current vehicle profile is shown with a thin solid line and the vehicle profile assumed under the new allocation $x_r$ is indicated by a thick solid line.

1. The simplest strategy is to make all $x_r$ vehicles available throughout the entire period. If $x_r$ is the peak vehicle requirement for the period, it is likely that the model will operate with more vehicles than in reality (outside this within-period peak), and performance will be overestimated. With even headway dispatching, this means lower waiting times. With target headway dispatching, this leads to a higher probability of perfect headway adherence at terminals, which leads initially to lower waiting times and eventually to excess vehicles queuing at terminals to begin trips.

2. A strategy that tries to incorporate variation of active vehicles within a period is to shift the current vehicle profile by a constant amount throughout the period, namely the adjustment in the peak vehicle requirement. For example, if the peak vehicle requirement being tested is one less than the base case, the entire vehicle profile is shifted down by one vehicle. (Of course, the profile must remain strictly positive at all times.) The result may be different from the operator's plan in response to the change in peak vehicle requirement, but it is not clear if performance is over- or underestimated in this case.

3. An alternative strategy that also tries to capture within-period variability in the active route fleet size is to calculate end-to-end running time reliability obtained with $x_r$ in the portion of the period with peak vehicle requirement in the base profile, and then calculate the fleet size required to deliver the same running time reliability during the rest of the period. For example, an allocation of 20 peak vehicles might be made for the morning peak period of a route, but perhaps all 20 vehicles are only used during the busiest half hour of the period. Given the specified headway $h$ and the fleet size $n$, the cycle time is $c = nh$, which implies, for instance, $90^{\text{th}}$ percentile running times in the peak of the peak. Then the fleet size during the remainder of the morning period is set so that it achieves the same level of reliability. In other words, it is calculated with $n \approx c/h$, with the cycle time determined the $90^{\text{th}}$ percentile of the running time distributions at other times in the period. This strategy may lead to changes in the shape of the base vehicle profile. If the operator bases new operational plans on running time distributions, this strategy may lead to a realistic vehicle profile. This is not necessarily the case, however, since crew and vehicle constraints are not considered.

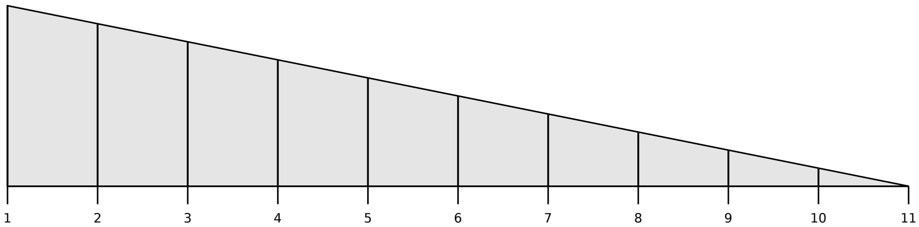


**Figure 4-6:** Spatial distribution of mean boarding rate

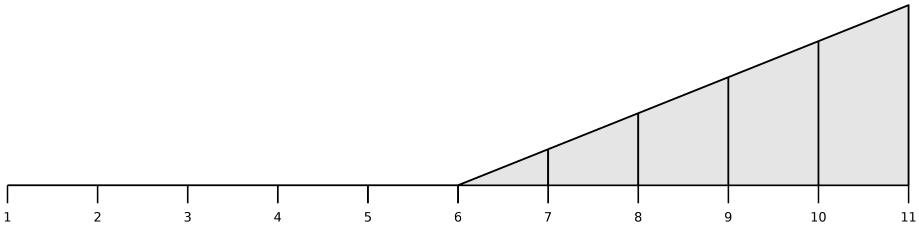


**Figure 4-7:** Spatial distribution of mean alighting rate

An alternative approach is to build an operating plan that meets level-of-service specifications and minimizes operating cost. An optimization package could be used to build vehicle and crew schedules once the time table has been generated. This approach does not depend on an existing vehicle profile, but requires modeling vehicle and crew costs and constraints, a task that falls outside the scope of this research. Moreover, this approach would generate a vehicle profile for the entire day, not just the time period of interest, so the optimization process would not be as separable by time periods as proposed here.

## 4.5 Simplified Case Study

This section illustrates the process of optimizing resource allocation through a simple hypothetical scenario. Consider two routes A and B operating independently. Both routes have 11 stops per direction and a target headway of 8 minutes in the morning peak, which runs from 7:30 to 9:30. Vehicles on both routes can carry up to 40 seated passengers and 20 standees, for a maximum capacity of 60 passengers. The demand pattern is the same for both routes; passengers arrive randomly following a Poisson process. The overall morning peak mean passenger arrival rate is 375 passengers per hour in the peak direction and 150 passengers per hour in the off-peak direction. These overall rates are spatially distributed as shown in Figure 4-6, with the highest arrival rate at the first stop. Likewise, the alighting rate for passengers on board a vehicle at a particular stop is distributed as shown in Figure 4-7 (showing stop 6), with the highest alighting rate at the last stop.

The only difference between the two routes is their running time distributions. Route A has no running time variability. Segment running times are a deterministic 3 minutes for all ten segments in both directions. In contrast, segment running times for Route B are stochastic, following a shifted Erlang distribution of degree 2 and a rate parameter of 1. The shift is of one minute, which makes the mean segment running time 3 minutes like in Route A, and the standard deviation is 1.41 minutes. The probability density function

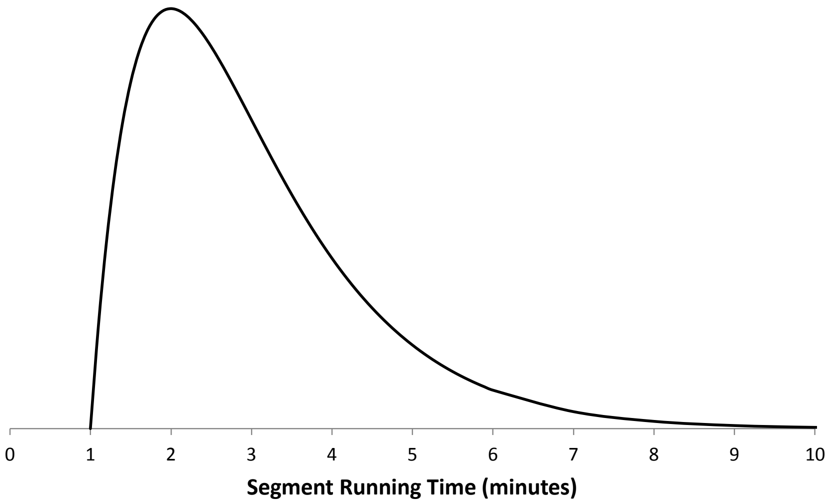


**Figure 4-8:** Route B Segment Running Time PDF

of segment running times for Route B is shown in Figure 4-8. For the sake of simplicity, running times of different segments are independent and include dwell times. In addition, fleet size remains constant throughout the morning peak.

Initially, each route has a fleet of seven vehicles, but there are newly available resources to add up to three vehicles in total. The problem is to determine how these three extra vehicles should be allocated among the two routes to obtain the greatest improvement in service performance. The optimization framework presented in this chapter can be used to find the optimum allocation of resources.

Before proceeding, the objective function (4.1a) must be defined in terms of performance measures. In order to keep the example simple, performance is defined based only on excess waiting time (EWT) experienced by passengers. Since the scheduled headway during the morning peak is 8 minutes, the expected waiting time without variability for a passenger arriving at random is 4 minutes. EWT is the difference between the mean waiting time and the base waiting time of 4 minutes.

A simulation model based on the framework of Chapter 3 is used to estimate excess waiting time with the initial fleet size (7 vehicles) for each route. This simulation model includes demand representation, so it is possible to estimate vehicle loads (the load limit of 60 is enforced) and passenger-based waiting times, which are used to calculate EWT. Based on 1000 replications, the EWT for route A is 0.53 minutes and 1.87 for Route B.

The total fleet size must be increased from 14 to 17. In order to determine to which route the additional vehicles should be added, simulations with a fleet of 8 vehicles are run for each route. EWT under these conditions are 0.33 minutes for Route A and 1.40 minutes for Route B, representing improvements of 0.20 and 0.47 minutes, respectively. Based on this criterion, resources are better allocated on Route B, so the fleet size of Route B is increased to 8. This process is repeated twice more, increasing the fleet size of each route by one vehicle, for a total of 8 vehicles on Route A and 9 vehicle on Route B.

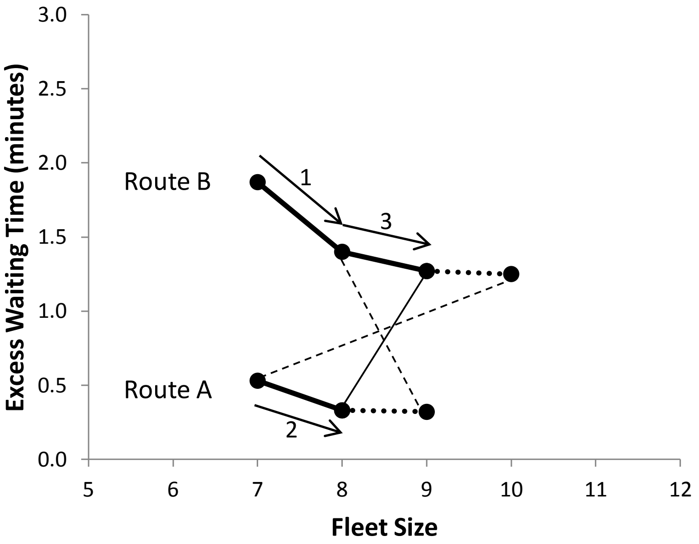


**Figure 4-9:** Finding optimal fleet size allocation for two routes based on excess waiting time

At this point the solution has the target total fleet size of 17 vehicles. The EWT is 0.33 minutes for Route A and 1.27 minutes for Route B. If there were other routes in the set, the second stage of the optimization algorithm would consider removing a vehicle from a Route C and giving it to Route A or Route B. Since in this example there are no other routes, the current solution is optimal and the optimization process terminates.

A graph of EWT for both routes is shown in Figure 4-9, with arrows showing the sequence of modifications made to fleet size and diagonal lines indicating feasible allocations (i.e. those that add up to 17 vehicles). The shape of the performance curves for each route agrees with the description of Section 4.2; although they are convex rather than concave (since a lower EWT implies a higher performance), there are diminishing returns to adding vehicles to a service with target headway dispatching.

Two other things are apparent in Figure 4-9. First, the two routes are identical except for their running time variability, and the one with higher running time variability requires more resources. Second, adding more resource to any of the routes will not improve service quality significantly, since the slope of the rightmost segments is almost horizontal. This implies that, in absence of increased frequency or real-time control actions to regulate headways, the route with stochastic running times will never perform as well as the other, irrespective of resource level.

The optimization problem can be solved with more complex objective functions as well. For example, it may be desirable to consider both waiting time and in-vehicle comfort. The excess waiting time measure captures waiting time relative to the performance standard, in this case the target headway of 8 minutes. A similar measure is required to capture in-vehicle comfort. The simulation model measures the load each passenger encounters upon boarding a vehicle, so load statistics are available. For example, suppose the standard is that there should be no more than 5 standees in 95% of the cases. Then, assuming capacity to seat 40 passengers, the 95th percentile loads exceeding 45 passengers can be regarded as excess load (EL).

The two criteria (EWT and EL) must be combined into a single measure of performance, using weights according to their relative importance. For example, having an excess load of 10 passengers might be considered equivalent to having one minute of excess waiting time. In this case, the objective is to minimize the weighted combination of EWT and EL:

$$\text{minimize } 10 \cdot \text{EWT}_A + \text{EL}_A + 10 \cdot \text{EWT}_B + \text{EL}_B \tag{4.18a}$$

$$\text{s.t. } x_A + x_B = 17 \tag{4.18b}$$

$$x_A, x_B \in \mathbb{Z}^+ \tag{4.18c}$$

It turns out from simulation estimates that EL is 1 on Route B with 7 vehicles, and 0 with 8 vehicles or more, and 0 on Route A with 7 vehicles or more. Therefore, the optimal allocation and the optimization steps are the same as before. This is not surprising, given that both EL and EWT arise from headway variability and the demand pattern is the same on both routes. However, in a situation with different demand patterns on each route, the optimum allocation may change to reflect the new multi-criteria objective.

## 4.6 Conclusion

The overall budget-constrained service performance of a bus network can improve with the optimization of resource allocation. The strategy is to systematically move vehicles from routes in which they provide less benefit to routes in which they provide more benefit. It is not expected that the changes suggested by such an optimization routine will be dramatic; instead, they will most likely be small adjustments to an already reasonable allocation of resources.

Three algorithms work together towards the optimal solution: an *optimizer*, a *simulator*, and a *vehicle profiler*. In order to carry out the optimization, an objective function must be defined. It may contain one or more quantifiable aspects of service quality, each weighted according to their relative importance. Examples include performance measures such as expected waiting time and excess load. The *optimizer* acts on the objective function to decide where to remove and add resources. Since there are no real-life observations of performance under the hypothetical resource levels being investigated, estimates of these performance measures must be obtained; a simulation model based on the formulation presented in Chapter 3 is well-suited for the task. One of the challenges faced is the need to translate a single measure of fleet size into a realistic time-varying vehicle profile. Three simple strategies presented in Section 4.4 accomplish this by modifying the existing vehicle profile. It is also be possible to build a vehicle profile based on hypothetical vehicle and crew schedules prepared with an optimization package.

Several simplifying assumptions are made to make the problem tractable. First, it is assumed that target headways (and other service planning parameters) for each route under consideration have been set and will remain constant. Only fleet size will be manipulated. Second, it is assumed that operating conditions, including traffic, road geometry, and ridership are independent of fleet size and remain constant. It is known that improvements in reliability and comfort can attract new ridership (and conversely, that degraded reliability and comfort can drive away riders to other routes or modes), especially when a corridor is shared among multiple routes. Moreover, improved headway regularity can lower running times. All of these effects are ignored. It is possible to incorporate ridership models into the optimization framework, but this introduces complex network level effects. Ridership shifts between routes serving a common corridor could be captured by considering multiple routes simultaneously in optimization and simulation, but this is left for future work. In order to limit potential consequences of these simplifications, changes in fleet size can be limited, say, to adding or removing up to two vehicles from any route.

The mechanism by which resource allocations become operational is also greatly simplified. The optimization framework manipulates a single fleet size for each route by globally defined time periods, but real vehicle profiles may exhibit considerable variation in the number of active vehicles within these periods. The heuristics proposed in Section 4.4 do not consider vehicle and crew constraints or costs that govern the creation of real vehicle profiles.

The optimization algorithm discussed in Section 4.3 assumes that the functions modeling performance of each route according to fleet size are concave. This assumption is consistent with the notion that adding resources will improve performance at a decreasing rate, and that, in the case of target headway dispatching discipline, there comes a point where adding more resources does not improve service performance further. This assumption allows finding an optimal solution with a greedy algorithm. There may be some forms of performance curves, strategies of building vehicle profiles, or types of operator behaviors that introduce violations to the concavity assumption. If this were detected, it would still be possible to find the global optimum with other algorithms not discussed here, for example, systematic enumeration of all solutions. In most practical situations where only small adjustments are being considered, this should not be a critical issue.

A simplified case study was presented in Section 4.5, which optimized the allocation of three additional vehicles among two hypothetical routes. The routes were identical in all respects except their running time variability: one route had deterministic running times and the other stochastic. Two objective functions were tested. The first was based only on excess waiting time, while the second was based on a weighted combination of excess waiting time and excess load. In both cases, the optimal allocation gave two additional vehicles to the route with running time variability, and one to the route with deterministic running times. This result reinforces the fact that more resources are necessary to operate routes with higher running time variability.

Only a fully verified and validated simulation model should be used to make the performance estimates that drive the optimization process. As such, the specific simulation model used in Chapter 3 may not be adequate to make recommendations on a real bus network. An extension of this model, based on the framework of Chapter 3 and validated in terms of running times, headways, and loads is ideal for real-world applications. Once this is available, a resource optimization program based on the framework presented here could run in a bus agency's server, processing the latest performance data and detecting opportunities to improve service or save wasted resources.

Although the optimization process is based on a preset objective function, detailed, disaggregate performance estimates can be part of the output in order to assist analysts in following up on automatically-generated suggestions.

# Chapter 5

# Conclusion

## 5.1 Summary

This research explores the relationship between running time variability and resource allocation in high-frequency bus operations. Emphasis is placed on developing performance analysis tools that take advantage of the large amounts of automatically collected data increasingly common in transit, including Automated Vehicle Location (AVL) and Automated Fare Collection (AFC).

Running time variability is the a-priori uncertainty of a future trip's duration, which affects a route's service performance and resource requirements. Sources of variability include traffic, other public transportation vehicles, different levels of driver aggressiveness, fluctuations in demand, and operator control (or lack thereof). These factors interact in complex ways to determine a route's total variability.

Running time variability negatively affects the service performance of high-frequency transit by requiring riders to budget "buffer" time to compensate for the uncertainty in waiting time and in-vehicle travel time. Since most passengers do not time their stop arrivals on high-frequency routes, passengers arrive randomly, with a higher probability of arriving during a long headway than during a short one. For this reason, running time variability increases mean waiting time. Headway imbalance ensues, resulting in uneven crowding levels and dwell times, which feeds back to increase running time variability. Operators must actively intervene to maintain even headways, which requires slack (i.e. buffer) in running times achieved by basing cycle times on conservative high-percentile running times. Vehicles are held at terminals and en route timing points when they are early (in schedule-based operations) or when their leading headway is short (in headway-based operations). Higher-percentile choices for cycle time imply higher reliability but larger fleet size and higher cost.

The first step toward analyzing running time variability is measuring it in a consistent and regular manner, in order to establish a record of the routes, places, seasons, and times exhibiting higher running time variability. Services with higher variability may require more attention during planning and monitoring to maintain reliable service, so it is important, especially in large networks, to identify them.

The basis for variability measures is the quantification of running time variability in homogeneous periods, either in absolute terms or relative to typical running times, using either mean-based or percentile-based statistics. Absolute measures, which express variability in units of time (e.g. minutes) are directly related to the amount of extra resources required to provide service with a certain level of dispatch reliability. Relative measures, which express variability without units, can be used to compare variabilities of items with different running times (e.g. a short route vs. a long route). Mean-based measures (standard deviation and coefficient of variation) are appropriate for modeling running times with fitted theoretical distributions, especially with small sample sizes. Percentile-based measures (introduced as *spread* and *normalized spread*) are appropriate for applications in which large quantities of running time data are available (so sample size is not a concern), because they provide an intuitive notion of the range of running times at a particular time of day without knowing the shape of the running time distribution.

Running times vary randomly within time periods and also transition systematically across homogeneous time periods. Systematic trends can be addressed in service planning by adjusting the fleet size throughout the day. In contrast, random variation follows no pattern (practically speaking) and is dealt with by providing a buffer time so that lateness does not propagate across trips. This buffer time requires resources in the form of additional vehicles and crew. There is practical value in quantifying the random component of variability for arbitrary (heterogeneous) periods, for instance, to compare running time variability across routes during the day. An aggregation method was developed to address this need. It measures variability separately in successive, short, overlapping time periods (referred to as *windows*) covering the time period being analyzed, and reporting the mean variability across windows. Assuming that running times in each window are approximately homogeneous, this measure captures random variability while being relatively insensitive to systematic variation. When *spread* is used to measure the variability of each window, the aggregate *mean spread* (MS) variability measure is obtained. This measure can be used on one or both directions of a route, and for any period of time during the day.

The *diurnal mean spread* (DMS) was developed as a standard measure of overall daytime random variability for both directions of a route. Even though the general variability measures and the aggregation technique are flexible, DMS was defined more precisely so that consistent measurements of variability can be made on different routes and over time. The algorithm used to compute DMS is outlined in Appendix A. Detailed measures, at the direction, time period, and segment level, can be combined with DMS to provide a complete profile of running time variability for a route.

Visual representations of running time variability make it easier for analysts to identify where and when variability is strongest without browsing through numerical reports. Two visual analysis tools were presented in Chapter 2. The first consists of running time scatter plots for each direction of a route, showing running times throughout the day and moving 10th, 50th, and 90th percentile lines to indicate typical running times and running time spread. The second illustrates variability on a map, with segments color-coded according to aggregate spreads or any other performance measure.

Understanding the factors that determine typical running times and running time variability can help service planners anticipate the resource requirements of a proposed service as well as the changes in resource requirements following a route modification. Various linear models were estimated to explore general patterns present across a sample of bus routes in

London. Results suggest that running time variability varies greatly from route to route, that typical running times are lower in the summer and higher in the fall relative to spring, and that routes with higher typical running times also tend to have more variable running times. Routes entering central London have higher and more variable running times. Distance, number of stops, and ridership all tend to contribute to higher and more variable running times. Routes exhibiting greater operational speeds have slightly lower running time variabilities.

Performance analysis tools discussed in Chapter 2 help characterize current operating conditions. In some cases it is necessary to adjust the level of resource allocated to a route, adding resources where it is difficult to manage running time variability and lowering it where excessive resources are allocated, because it can be used more efficiently elsewhere. It is desirable to inform this adjustment process with a cost-benefit analysis that estimates the improvement in service quality attained when vehicles are added, and the performance deterioration following a fleet size reduction. Simulation modeling is well suited for the task because, in contrast with analytical models, it is adaptable and extensible, it can capture interaction effects naturally, and it lends itself well to data-driven analysis.

A general framework for simulation modeling of transit operations was developed in Chapter 3. Although it was devised with bus service in mind, it is flexible enough to model various bus lines operating together or rail transit. Emphasis was placed on the use of automatically collected transit data to obtain input parameters for the model. A simple simulation model of bus service operating between two terminals was developed as an example. This particular model does not use many of the capabilities of the general framework; for instance, it does not consider real-time control strategies such as holding and curtailments. However, it is capable of capturing pairwise correlation of segment running times.

The simple simulation model was validated using data from several TfL routes and artificially generated data from a conceptual route. The validation was based on a comparison of headway and running time statistics calculated from AVL data and the simulation model. Running time means and standard deviations at the segment and direction level were estimated with relatively small errors. Headway means were generally overestimated, and there was a wide range of errors in estimates of headway variability. The results indicate generally good agreement between AVL and simulation running time statistics, and larger discrepancies in headway statistics.

The mismatch of headway means and standard deviations in the simple simulation model is most likely due to a simplified representation of terminal dispatching strategies. Headway endogeneity in running times and real-time control mechanisms aimed at regulating headways may also contribute to errors when simulating real service. This highlights the importance of modeling operator behavior accurately before using them for decision support.

The overall budget-constrained service performance of a bus network can improve with the optimization of resource allocation. The strategy is to systematically move vehicles from routes where they provide less benefit to routes where they provide more benefit. It is not expected that the changes suggested by such an optimization routine will be dramatic; instead, they will most likely be small adjustments to an already reasonable allocation of resources.

The resource allocation approach consists of three processes: an *optimizer*, a *simulator*, and a *vehicle profiler*. The optimizer is based on an objective function containing one or more quantifiable aspects of service quality, each weighted according to their relative importance. Examples include performance measures such as expected waiting time and excess load. The *optimizer* acts on the objective function to decide where to remove and add resources. Since there are no real-life observations of performance under hypothetical resource levels, estimates of these performance measures must be obtained; a simulation model based on the formulation presented in Chapter 3 is well-suited for the task.

Several simplifying assumptions are made to make the problem tractable. First, it is assumed that target headways (and other service planning parameters) for each route under consideration have been set and remain constant. Only fleet size is manipulated. Second, it is assumed that operating conditions, including traffic, road geometry, and ridership are independent of fleet size and remain constant. As such, some potentially important effects are ignored. For example, improvements in reliability and comfort can attract new ridership (and conversely, that degraded reliability and comfort can drive away riders to other routes or modes). Moreover, improved headway regularity can lower running times. It is possible to incorporate ridership models into the optimization framework, but this introduces complex network level effects. In order to limit potential consequences of these simplifications, only small changes in fleet size are considered.

The mechanism by which resource allocations become operational is also greatly simplified. The optimization framework manipulates a single fleet size for each route by globally defined time periods, although real vehicle profiles may exhibit considerable variation in the number of active vehicles within these periods. Three heuristics proposed in Chapter 4 capture this variation by modifying an existing vehicle profile, without considering vehicle and crew constraints or costs that might govern the creation of real vehicle profiles.

The optimization algorithm discussed in Chapter 4 assumes that the functions modeling route performance as a function of fleet size are concave. This assumption is consistent with the notion that adding resources improves performance at a decreasing rate, and that, in the case of target headway dispatching discipline, there comes a point where adding more resources does not further improve service performance. This assumption allows finding an optimal solution with a greedy algorithm. There may be some forms of performance curves, strategies of building vehicle profiles, or types of operator behaviors that introduce violations to the concavity assumption. For such cases, it is still possible to find the global optimum with other algorithms. In many practical situations where only small adjustments are being considered, this is not a critical issue.

A simplified case study was presented in Chapter 4, which allocated additional vehicles among two hypothetical routes. The routes were identical in all respects except their running time variability: one route had deterministic running times and the other stochastic. Two objective functions were tested. The first was based only on excess waiting time, while the second was based on a weighted combination of excess waiting time and excess load. In both cases, the optimal solution allocated two additional vehicles to the route with running time variability, and one to the route with deterministic running times. This result reinforces the fact that more resources are necessary to operate routes with higher running time variability.

A resource optimization program based on the framework presented in Chapter 4 could be used to process performance data and detecting opportunities to improve service or save resources.

## 5.2 Limitations and Future Work

There are many opportunities to expand on the analysis examples presented in this thesis. The exploration of general patterns of typical running times and running time variability presented in Chapter 2 is based on a limited data set, so exploring new sources of data could lead to better models. Chapter 3 presents a very simple simulation model that does not use the full capabilities of the simulation architecture and framework discussed earlier in the same chapter. The optimization algorithm of Chapter 4 assumes fixed headways, ignores possible vehicle and crew constraints, and solves the problem independently for each time period.

Real time control strategies such as holding, curtailing, and controlling the speed of trips have been mentioned but not analyzed in this thesis. Real time control is an important aspect of high-frequency transit operations, and it can have dramatic effects on running times, dwell times, in-vehicle comfort, and other aspects of service performance. Simulation-driven analysis of the different strategies could improve our understanding of their effect.

Specific areas of future work are identified in the following sections.

### 5.2.1 Improved Running Time Models

Models that predict typical running time and running time variability by time of day, given a set of route and operating environment characteristics, can assist service planners in evaluating resource requirements of new routes, and also to predict changes in requirements as a consequence of route modifications. As a first step, several models were specified and estimated in Chapter 2, and some general patterns were identified. However, the statistical significance of these models (especially variability models) was not very high.

The predictive power of the models was limited by the datasets they were based on. Some improvement could result from increasing the size of the dataset from a sample of routes to all routes of a bus network, because more observations improve the ability to discern the contribution of each factor in a model's specification. Further improvement could result from using less aggregate data. For example, instead of capturing the effect of ridership on running time variability with the total number of boardings at the route level, stop-level boarding and alighting data could be used, recognizing that some stops contribute more to running time variability than others. The inference and expansion techniques discussed in Appendix B could be used for this purpose.

New sources of data could also be explored. Traffic and road characteristics (for example, number of lanes, availability of bus lanes, and density of intersections), which together govern congestion levels, might play a major role in determining running times, but their effect is not explicitly captured in the models of Chapter 2 simply because the data was not available when the work was being carried out.

Characterizing running times by segment (instead of by route) may also help forecast median running times and running time variability with greater accuracy. The way in which stops, ridership, and traffic contribute to running time variability may vary by segment, especially when different segments have different characteristics or are located in different zones of a city. This type of analysis could be used to categorize segments by type, and route-level variability could be estimated according to the mix of segments composing it.

### 5.2.2 Improved Simulation Capabilities

The simple simulation model implemented in Chapter 3 does not have the full capabilities of the simulation architecture presented earlier in the same chapter. The validation process revealed that the model represents running times fairly accurately, but not headways. This outcome is reasonable given the effort to represent running times accurately, capturing correlation between adjacent segments, while a comparable effort was not made for headways. A simulation model should be validated in all regards before its output is trusted for decision-making.

A more robust simulation model can be built, based on the very flexible and scalable architecture presented in Chapter 3. Operator behavior can be represented more accurately by inserting operator behavior models into the simulation model. These models can be heuristic or probabilistic, based on observations of real operations. The data playback capability discussed in Appendix B can aid in building a dataset of how real dispatches occur.

The first step towards a more realistic headway representation is capturing terminal dispatching and en route headway regulating control actions. Whether the dispatching discipline for a given route in the morning peak is based on schedule, headway, or a combination depends on what happens in the field. Behavior may be modeled with fuzziness to capture stochasticity and "imperfections" inherent in operator decisions. Aside from dispatching discipline, models of holding and curtailing behavior can also be developed, estimated based on observations, and inserted into the simulation model. Devising realistic operator behavior models can be challenging because decisions can be based on criteria for which there is no data, such as vehicle or crew constraints. Additionally, there may be a delay between decision time and when the related action occurs; only the latter is observed in the data.

The simulation model presented in Chapter 3 uses running time distributions that include dwell time, which does not allow distinguishing vehicle movement time from dwell time components. Simulated dwell times, therefore, depend on headways, loads, and ridership of real service in the past, which may be very different from simulated conditions. A particular realization of a segment's running time may be especially long if it corresponds to a heavily loaded vehicle with a long leading headway (in AVL data), but this running time may be applied in the simulation model to a lightly-loaded vehicle with a short leading headway. This leads to an exogenous representation of an operating environment that is inherently headway-endogenous, as discussed in Chapter 3.

Explicit modeling of passengers and dwell times may resolve this issue. In order to do this, segment running time distributions must include only vehicle movement time, without the dwell time component. Dwell times can be reinserted during simulation using a dwell time model, such that dwell time realizations depend on the circumstances of each situation. Factors such as the number of boardings, the number of alightings, and in-vehicle load jointly govern dwell times. A dwell time model from the literature can be used (for example, see Milkovits, 2008a), or a new model can be estimated using collected dwell time data. The separation of dwell time from vehicle movement time is generally not a straightforward process, and there are various approaches to it. (For example, Robinson (2013) presents a heuristic approach.)

Aside from modeling dwell times realistically, modeling passengers explicitly in the simulation can help estimate passenger-centric performance measures. For example, instead of calculating excess waiting time based on headway observations alone, it can be calculated from the distribution of waiting times experienced by (simulated) passengers. Likewise, in-vehicle comfort could be measured based on the loads experienced by each passenger during the entire time on board a vehicle. These disaggregate observations can be combined to obtain an aggregate performance measure.

The simulation model architecture supports modeling operations of several routes simultaneously. This allows modeling groups of routes sharing stops, in order to capture network effects. This can be important when a significant portion of passengers can take more than one route to their destination. (See Viggiano, 2013 for more on this topic.) The architecture also supports modeling multi-branch transit services, as well as rail transit systems (which may restrict where vehicles can overtake each other and their approach speeds to other vehicles).

### 5.2.3 Generalized Optimization Formulation

The optimization formulation presented in Chapter 4 can be generalized in several ways. It assumes that routes are independent and that their ridership is fixed, so network effects are disregarded. In the case of a pair of routes that share a common corridor with significant ridership, reducing the fleet size of one route may encourage some passengers to switch to the other. These changes may occur in different time horizons. Simple route choice models can distribute ridership in a corridor among the different routes serving it. More sophisticated models may consider entire multi-leg journeys, such that altering the service quality of one leg affects ridership on all legs of the journey, including those on rail. In order to begin capturing network effects in passenger route choice, the quality of service of a route should be a function of not only its fleet size, but also the fleet sizes of other related routes. The objective, then, is to maximize the total service quality of a group of routes given the complete allocation:

$$\text{maximize} \sum_{r \in R} Q_r(\mathbf{X}_R) \tag{5.1}$$

where $\mathbf{X}_R$ is a vector of fleet sizes of all routes being considered. This formulation explicitly allows the performance of route A to vary in response to a change in the fleet size of route B, which may serve a corridor in common with Route A.

The formulation in Chapter 4 assumes ridership remains constant as changes in fleet size are made. The maximum fleet size change for each route is limited to keep the consequences of this assumption small. Changes in the characteristics of service in one route may encourage riders to alter their path choice in the network, to choose another mode, to change the times at which they travel, and even to change their origins and destinations. This may bring new ridership to the network, or there can be a net ridership loss. Inserting a network-level ridership model to estimate changes in ridership as a function of service quality would make the formulation more realistic. This can be done at different levels of sophistication (and complexity). Simpler models may be based on elasticities, while more sophisticated models may consider the activities that generate trips in a metropolitan region.

The service quality of a route responds to headway in addition to fleet size. Target headways are fixed in the formulation of Chapter 4, under the premise that they have been appropriately set by the service planning team according to ridership and policy. Notwithstanding the validity of this premise in many practical situations, there may be other situations in which the goal is to optimize fleet size and headways simultaneously. Constraints on headways can be added to ensure that the resulting solution meets policy requirements. The formulation of the objective would take the following mathematical form:

$$\text{maximize} \sum_{r \in R} Q_r(\mathbf{X}_R, \mathbf{H}_R) \tag{5.2}$$

where $\mathbf{X}_R$ and $\mathbf{H}_R$ are fleet size and target headway vectors.

The optimization problem presented in Chapter 4 is solved independently for each time period. Doing this greatly simplifies the optimization task, but requires time periods to be defined. A single set of time periods applies to all routes under consideration. In reality, optimal time period definition can be different across routes. Moreover, resource requirements vary within periods, an issue that Chapter 4 addresses with heuristics to alter a given vehicle profile. It is possible to generalize the formulation to solve the problem for a day at a time, without defining time periods. One way of proceeding is to model the flow of vehicles across time and routes, constraining the maximum flow at any instant in time to meet the target fleet size. Sophisticated algorithms may be required to solve the generalized problem.

### 5.2.4 Further Applications of Transit Simulation

Besides expanding on the analysis examples presented in this thesis, there are opportunities to use simulation models based on the framework presented in Chapter 3 for purposes other than characterizing the response of service quality to fleet size. One area with research potential is simulation-based evaluation of real-time control strategies.

Scientific literature on real-time control in bus transit has presented evaluations of holding strategies. (For examples, see Abkowitz and Lepofsky (1990), Eberlein et al. (2001), Chandrasekar et al. (2002), Daganzo (2009), Delgado et al. (2009), Furth and Muller (2009), Cats et al. (2011), and Bartholdi and Eisenstein, 2012.) Less emphasis has been placed on other strategies, such as signal priority (for example, see Chandrasekar et al., 2002), speed control (for example, see Daganzo and Pilachowski, 2011), and boarding limits (for example, see Delgado et al., 2009). Besides the early work of Bly and Jackson (1974), Jackson (1977), and Andersson et al. (1979), little has been said about real-time short turning in the context of bus transit, although some work has been published in the context of rail transit. (See Shen and Wilson, 2001 for an example of short turning as an incident management strategy in rail transit.)

Real-time short turning has not been methodically evaluated as a real-time service quality improvement strategy. The action of turning a trip short can be beneficial or detrimental to service quality, depending on the loads and headways of vehicles on the route and the number of passengers waiting to board the next vehicle. One of the reasons the strategy has not been studied in great detail is that its effect depends on many factors and interactions, making it difficult to study with analytical models. However, a simulation model based on the framework of Chapter 3 is capable of capturing the necessary level of detail to help researchers identify the circumstances under which real-time short turning is beneficial, the frequency with which these circumstances arise in typical transit operations, and the magnitude of performance improvements that can be expected from including short turns in the menu of real-time control actions available to the operator.

# Appendix A

# Diurnal Mean Spread (DMS)

Often it is desirable to summarize the running time variability of a route with a single figure. The *diurnal mean spread* (DMS) is defined for this purpose. It is simply the *mean spread*, as presented in Chapter 2 and defined in (2.9), calculated for the daytime hours of both directions of a route. DMS is too aggregate to set resource levels by time period, but it is useful to classify routes by their running time variability and to track general variability over time.

To ensure that the variability measures are consistent for comparisons, DMS should always be calculated with the same parameters. For example, it may be calculated on observations of 15 successive workdays from 7:30 to 19:30 (a span of 12 hours covering the morning and afternoon peaks), using observation windows spaced 15 minutes apart, each 30 minutes wide.

The following steps outline the proposed algorithm for calculating the *diurnal mean spread* based on the afore-mentioned parameters.

1. Define observation windows 30 minutes wide shifted every 15 minutes, starting at 7:30 and ending at 19:30. Start with the 7:30 window, which includes observations from 7:15 to 7:45. Continue with the 7:45 window, which includes observations from 7:30 to 8:00, etc. End with the 19:30 window, which includes observations from 19:15 to 19:45. There are a total of 49 windows.

2. Calculate the 90th and 10th percentiles of running times for every window for each direction.

3. Compute the spread of each window over both directions:

$$S_w = p_{90}(T_{1,w}) - p_{10}(T_{1,w}) + p_{90}(T_{2,w}) - p_{10}(T_{2,w}) \tag{A.1}$$

4. DMS is the arithmetic mean of the spreads, which is sum of the spreads in the previous step divided by 49.

$$\text{DMS} = \frac{1}{49} \sum_{w \in W} S_w \tag{A.2}$$

# Appendix B

# Performance Analysis and Monitoring with Data Playback

## B.1 General Concept

This thesis focuses on the relationship between running time variability, service performance, and resource allocation in high-frequency bus operations. Since adjustments in resource allocation affect service for relatively long periods of time, ranging from a few weeks to entire seasons and even years, the analysis tools developed help evaluate service performance over multiple days; indications of what happened on individual days or trips are not part of the output. Nevertheless, the origins of running time variability can only be studied at a microscopic level. Individual dwell times, operator decisions, signals, and congestion each contribute to the specific running time of a trip, often interactively. Although decisions about resource allocation should not be made based on studying what happened one day, this type of analysis can help us understand the factors leading to running time variability, from which better models can be developed to improve how resource allocation problems are solved. This appendix introduces the *data playback* technique, and two tools based on it, to aid in studying the more microscopic aspects of bus operations and how they relate to running time variability and overall service quality.

Data playback is a technique in which events, associated each to one element (for example, a vehicle), are played back chronologically, irrespective of the element they refer to, so that interactions between the different elements are observed. (This is different from the analysis tools of Chapter 2, in which the particular sequence of running times and the relationship between them was not a concern.) In the context of AVL data, each observation specifies the location and time of a vehicle. By playing back these observations chronologically and with all the vehicles of a route at once, features such as the distance between vehicles can be observed, although these are not found in AVL data per se. AVL Playback is discussed in Section B.2. Multiple types of data can be played back simultaneously. For example, playing back AVL and AFC data together facilitates analyzing the role passengers play in running time variability. Combined AVL and AFC playback is discussed in Section B.3.

Two playback analysis tools were developed. The first is a *playback visualizer*, which displays an animated sequence of events of a particular day (or portion of day). Much can be learned about how a transit service operates by simply observing vehicles move in a schematic of the route. Summarized performance measures (like those of Chapter 2) are sometimes not detailed enough, and studying a large set of detailed data would be tedious and time consuming. However, thousands of events can be displayed in animation form at a quick rate, and the analyst is capable of identifying patterns, anomalies, or specific events of interest. The form of these might be too subtle or complex to capture with generic performance measures.

The second analysis tool is a *playback event recorder*, which is used to create specialized datasets of events of interest and their characteristics. For instance, suppose the *visualizer* displays several curtailments in the afternoon peak of a particular day and route.[1] This might motivate a detailed investigation of curtailments over a longer period of time. Rather than watching many animated playbacks and manually making a record of each curtailment, the *event recorder* can be configured to automatically identify curtailments and record a set of relevant characteristics as it processes many event sequences (of possibly many routes and days). In the case of curtailments, relevant characteristics might include route, operator, date, time of day, last stop of the curtailed trip, first stop of the next trip, and load on the vehicle at the time of curtailment.

## B.2 AVL Playback

AVL playback consists of playing back AVL data chronologically, either to generate an animation with the *visualizer* or to generate specialized datasets with the *event recorder*. (A similar concept was used as a verification tool in Chapter 3, but based on artificially generated data of the simulation model instead of observations from an AVL database.) Figure B-1 shows an example of an animation, in this case of London route W15 for March 7, 2011. At 15:48 that day, vehicle 903 was laying over at one of the terminals, and three vehicles were being curtailed. It is possible to use route geometry information to render the animation in map format.

The *playback event recorder* can be used to generate AVL-derived datasets tailored to specific applications. An example application, identified in Chapter 5 as future research work, is to identify the factors that lead to curtailments. It is hypothesized that operators use curtailments both to improve headway regularity and to reduce overtime compensation to drivers. In order to test this hypothesis, AVL playback is used to collect relevant characteristics of situations in which the operator could choose to curtail a trip, along with the decision of the operator (to curtail the trip or not). This information can be used to estimate parameters for behavior models that explain curtailment decisions as a function of explanatory variables.

Similarly, dispatch behavior models could be developed by recording leading and trailing headways along with scheduled departure times for each observed dispatch. This could help identify which of the dispatching strategies presented in Chapter 3 (e.g. on schedule, on even headway, or on target headway), is a better model of actual dispatch behavior to

[1] A curtailment occurs when an operator ends a trip before reaching the end of the run. It is a form of short-turning that happens as a real-time control decision rather than a planned variation in the schedule. In London, operators are allowed to make these curtailments without explicit penalties, as long as the reason for doing so is external (for instance, traffic).

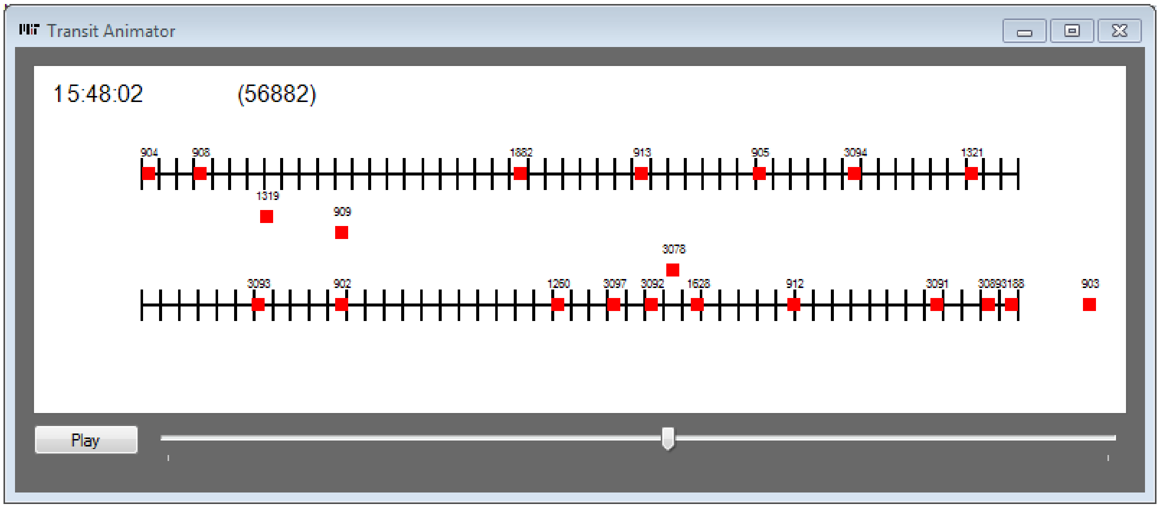


**Figure B-1:** Animated AVL playback

be used in the simulation of a bus route's operations. The results of such research could make simulation models more realistic, taking them a step closer to what is required for practical applications of simulation-based optimizations of resource allocation.

## B.3 Combined AVL and AFC Playback

The playback technique is not limited to data from one source; perhaps the most interesting and useful applications of it involve multiple datasets. A logical extension of AVL playback is combined AVL and AFC playback. In its simplest form, vehicle positions and fare collection data are obtained as input, and the combined sequence of events is played back chronologically.

The usefulness of combined AVL and AFC playback depends on the nature of the AFC data. AVL, almost universally, gives the times vehicles visited each stop or enough information to determine them. In contrast, there are many different AFC systems, some of which provide more information than others. With a few exceptions, bus AFC systems only record when passengers board vehicles, and not when they alight. Only boardings are recorded in London Buses. In rail there are systems that record only when passengers enter stations, and others that record when passengers leave stations as well; in London Underground, both entries and exits are recorded.

Inference techniques can greatly enhance AFC datasets. Based on previous research, Gordon (2012) developed algorithms that infer origins, destinations, and transfers (or interchanges) of passengers who pay their fares with smart media. This greatly enhances raw AFC data by providing in many cases the alighting time and stop of bus trips, and potentially approximate passenger arrival times to origin stops when transferring from rail. It is not possible to do this with all passengers, but with a reasonably large portion of them. Expansion heuristics can be applied to the uninferred portion. AFC datasets enhanced in this manner should be used for playback when possible.

When AVL and AFC data are combined, the *visualizer* displays an animation of vehicle movement with the number of people boarding at each stop and estimated vehicle loads; a scale of colors is used to paint vehicles according to their loads. This additional layer of information can help explain, for instance, what may be driving an operator's decision to hold or curtail a trip. When the AFC dataset has been enhanced with inference procedures as described above, both origins and destinations are known. (For some passengers, only origins are known, but expansion techniques can be used to approximate destinations collectively for load estimates. See Gordon (2012) for details.)

The *playback event recorder* can be a very useful analysis tool when AVL and AFC data are combined. For example, it can be used to obtain distributions of waiting times, in-vehicle travel time, and combined waiting and travel time for passengers traveling from a stop $O$ to a stop $D$, bearing in mind that multiple routes may serve the $O$-$D$ pair. This is demonstrated by Schil (2012). Relying once again on AFC data enhanced by inference (without expansion in this case), the headway preceding a vehicle's arrival at a stop (where a passenger whose origin and destination are known boards) can be determined from the sequence of stop visits made by vehicles serving both stop $O$ and stop $D$, even when multiple routes serve the pair of stops. In the absence of additional information, passengers boarding that vehicle could have arrived any time in that headway with equal probability, which gives a uniform distribution. The distributions of each passenger can then be combined into one general distribution of waiting time specific to an $O$-$D$ pair and time period. Likewise, in-vehicle travel times of each passenger (which in this case are known and deterministic) can be combined to obtain a distribution for a given $O$-$D$ pair and time period. Finally, the two can be combined to obtain a distribution of total time (waiting and traveling) for a given $O$-$D$ pair and time period. This information can be used not only in service planning and performance monitoring, but also to provide the riding public with both typical travel time and reliability buffer time for planned journeys through trip planner applications.

A similar application is being developed by Viggiano (2013) with focus on route choice modeling in multi-route bus corridors. In these corridors there may be a large portion of passengers who choose not to board the first vehicle serving their $O$-$D$ pair, perhaps because it is too crowded, but also because there may be limited-stop services with lower in-vehicle travel time. In other words, passengers may be deciding to wait more in exchange for a more comfortable or faster ride to their destination. The *playback event recorder* has made it possible to obtain a set of vehicle arrival times relevant to each observed boarding: (a) the arrival time of the boarded vehicle, (b) the arrival time of the vehicle prior to the boarded vehicle that also serves the $O$-$D$ pair, regardless of route, and (c) the arrival time of the vehicle prior to the boarded vehicle that also serves the $O$-$D$ pair, of the same route of the boarded vehicle. A dataset of this nature can then be used to estimate models of route choice, and a good route choice model can be used for ridership forecasting in multi-route corridors.

While the above applications concern one route or a small group of routes, the *playback event recorder* may also help answer questions about a bus network in general. For example, it is possible to estimate the percentage of bus trips that can potentially be made in more than one route. Vehicle arrivals observed close to the boarding time of each passenger (whose destination has been inferred) are examined. If arrivals of more than two routes serving the $O$-$D$ pair are present close to the recorded boarding time (say, within 15 minutes of the recorded boarding time), the trip is counted as one that can be made in more than one route. Otherwise, it is counted as a trip that is served by only one route. Once all passengers have been processed, the desired fraction can be estimated, albeit with some degree of uncertainty because the trips for which it was not possible to infer a destination are not used. This process was carried out with data of the London Buses network in the morning of October 17, 2011. The AFC dataset included journey segments beginning between 7:30 and 9:00 for which it was possible to infer a destination, a total of 673,007 fare transactions. The AVL dataset included stop arrival information for over 700 routes and almost 19,000 bus stops. With the threshold to include potential vehicle stop arrivals set at ±15 minutes with respect to recorded boarding time of each passenger, it was estimated that approximately 40% of the trips could have been made in more than one route. This result motivates further research in network-level analysis, planning and operations.

# Bibliography

Abkowitz, M., Josef, R., Tozzi, J., and Driscoll, M.K. Operational Feasibility of Timed Transfer in Transit Systems. *Journal of Transportation Engineering*, 113(2):168–177 (1987).

Abkowitz, M.D. and Lepofsky, M. Implementing Headway-Based Reliability Control on Transit Routes. *Journal of Transportation Engineering*, 116(1) (1990).

Altun, S.Z. and Furth, P.G. Scheduling Buses to Take Advantage of Transit Signal Priority. *Transportation Research Record*, (2111):50–59 (2009).

Andersson, P., Hermansson, A., Tengvald, E., and Scalia-Tomba, G.P. Analysis and Simulation of an Urban Bus Route. *Transportation Research Part A: General*, 13A:439–466 (1979).

Balcombe, R., Mackett, R., Paulley, N., Preston, J., Shires, J., Titheridge, H., Wardman, M., and White, P. The Demand for Public Transport: A Practical Guide. Technical Report TRL593, Transport Research Laboratory (2004).

Barnett, A. On Controlling Randomness in Transit Operations. *Transportation Science*, 8(2):102–116 (1974).

Barry, J. London Buses Head of Network Development (2012). Private interview.

Bartholdi, J.J. and Eisenstein, D.D. A Self-Coördinating Bus Route To Resist Bus Bunching. *Transportation Research (Part B: Methodological)*, 46:481–491 (2012).

Bates, J., Polak, J., Jones, P., and Cook, A. The Valuation of Reliability for Personal Travel. *Transportation Research Part E: Logistics and Transportation Review*, 37:191–229 (2001).

Benn, H.P. Bus Route Evaluation Standards. Technical Report TCRP Synthesis 10, Transit Cooperative Research Program (1995).

Bertsimas, D. and Tsitsiklis, J.N. *Introduction to Linear Optimization*. Athena Scientific and Dynamic Ideas, Belmont, MA (1997).

Bly, P. Depleted Bus Services: The Effect of Rescheduling. Technical Report LR699, Transport and Road Research Laboratory (1976).

Bly, P. and Jackson, R. Evaluation of Bus Control Strategies by Simulation. Technical Report LR637, Transport and Road Research Laboratory (1974).

Carey, M. Ex Ante Heuristic Measures of Schedule Reliability. *Transportation Research Part B: Methodological*, 33:473–494 (1999).

Cats, O. *Dynamic Modelling of Transit Operations and Passenger Decisions.* Ph.D. thesis, KTH - Royal Institute of Technology (2011).

Cats, O., Larijani, A.N., Koutsopoulos, H.N., and Burghout, W. Impacts of Holding Control Strategies on Transit Performance: Bus Simulation Model Analysis. *Transportation Research Record*, 1:51–58 (2011).

Cham, L.C. *Understanding Bus Service Reliability: A Practical Framework Using AVL/APC Data.* Master's thesis, Massachusetts Institute of Technology (2006).

Chandrasekar, P., Cheu, R.L., and Chin, H.C. Simulation Evaluation of Route-Based Control of Bus Operations. *Journal of Transportation Engineering*, 128(6) (2002).

Daganzo, C.F. A Headway-Based Approach to Eliminate Bus Bunching: Systematic Analysis and Comparisons. *Transportation Research Part B: Methodological*, 43(10):913–921 (2009).

Daganzo, C.F. and Pilachowski, J. Reducing Bunching with Bus-to-Bus Cooperation. *Transportation Research (Part B: Methodological)*, 45(1):267–277 (2011).

Delgado, F., Muñoz, J.C., Giesen, R., and Cipriano, A. Real-Time Control of Buses in a Transit Corridor Based on Vehicle Holding and Boarding Limits. *Transportation Research Record*, 2090:59–67 (2009).

Desaulniers, G. and Hickman, M.D. Public Transit. In C. Barnhart and G. Laporte, editors, *Handbook in OR & MS*, volume 14, chapter 2, pages 69–127. Elsevier (2007).

Eberlein, X., Wilson, N., and Bernstein, D. The Holding Problem with Real-Time Information Available. *Transportation science*, 35(1):1–18 (2001).

Efron, B. Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods. *Biometrika*, 68(3):589–599 (1981).

Ehrlich, J.E. *Applications of Automatic Vehicle Location Systems Towards Improving Service Reliability and Operations Planning in London.* Master's thesis, MIT (2010).

Furth, P. and Muller, T. Service Reliability and Optimal Running Time Schedules. *Transportation Research Record*, 2034:55–61 (2007).

Furth, P. and Wilson, N. Setting Frequencies on Bus Routes: Theory and Practice. *Transportation Research Record*, (818):1–7 (1981).

Furth, P.G. Data Analysis for Bus Planning and Monitoring. Technical Report Synthesis 34, Transit Cooperative Research Program (2000).

Furth, P.G., Hemily, B., Muller, T.H., and Strathman, J.G. Using Archived AVL-APC Data to Improve Transit Performance and Management. Technical Report TCRP113, Transit Cooperative Research Program (2006).

Furth, P.G. and Muller, T.H.J. Optimality Conditions for Public Transport Schedules with Timepoint Holding. *Public Transport*, 1(2):87–102 (2009).

Gordon, J.B. *Intermodal Passenger Flows on London's Public Transport Network: Automated Inference of Full Passenger Journeys Using Fare-Transaction and Vehicle-Location Data.* Master's thesis, MIT (2012).

Hillier, F.S. and Lieberman, G.J. *Introduction to Operations Research.* McGraw-Hill, 9th edition (2010).

Jackson, R. Evaluation by Simulation of Control Strategies for a High Frequency Bus Service. Technical Report LR807, Transport and Road Research Laboratory (1977).

Kittelson & Associates, KFH Group, Parsons Brinckerhoff Quade & Douglass, and Hunter-Zaworski, K. Transit Capacity and Quality of Service Manual, 2nd Edition. Technical report, Transit Cooperative Research Program (2003).

Kottegoda, N. and Rosso, R. *Applied Statistics for Civil and Environmental Engineers.* Wiley-Blackwell, 2nd edition (2008).

Larrain, H., Giesen, R., and Muñoz, J.C. Choosing the Right Express Services for Bus Corridor with Capacity Restrictions. *Transportation Research Record: Journal of the Transportation Research Board*, 2197:63–70 (2010).

Law, A.M. *Simulation Modeling and Analysis.* McGraw-Hill, New York, NY, 4th edition (2007).

Li, Z., Hensher, D.a., and Rose, J.M. Willingness to Pay for Travel Time Reliability in Passenger Transport: A Review and Some New Empirical Evidence. *Transportation Research Part E: Logistics and Transportation Review*, 46(3):384–403 (2010).

Liao, C.F. Data-Driven Support Tools for Transit Data Analysis, Scheduling and Planning. Technical report, Intelligent Transportation Systems Institute, Center for Transportation Studies, Minneapolis, MN (2011).

London Buses. London's Bus Contracting and Tendering Process (2009).
**URL:** *http://www.tfl.gov.uk/tfl/businessandpartners/buses/tenderresults/lbsl-tendering-and-contracting-feb-09.pdf*

Marguier, P.H.J. *Bus Route Performance Evaluation Under Stochastic Considerations.* Ph.D. thesis, MIT (1985).

Milkovits, M.N. Modeling the Factors Affecting Bus Stop Dwell Time: Use of Automatic Passenger Counting, Automatic Fare Counting, and Automatic Vehicle Location Data. *Transportation Research Record*, 2072:125–130 (2008a).

Milkovits, M.N. *Simulating Service Reliability of a High Frequency Bus Route Using Automatically Collected Data.* Master's thesis, MIT (2008b).

Moses, I.E. *A Transit Route Simulator for the Evaluation of Control Strategies Using Automatically Collected Data.* Master's thesis, MIT (2005).

Newell, G. Control of Pairing of Vehicles on a Public Transportation Route, Two Vehicles, One Control Point. *Transportation Science*, 8(3):248–264 (1974).

Osuna, E. and Newell, G. Control Strategies for an Idealized Public Transportation System. *Transportation Science*, 6(1):52–72 (1972).

Peeta, S. and Ziliaskopoulos, A.K. Foundations of Dynamic Traffic Assignment: The Past, the Present and the Future. *Networks and Spatial Economics*, 1:233–265 (2001).

Pratt, R.H., Park, G., Texas Transportation Institute, Cambridge Systematics, Parsons Brinckerhoff Quade & Douglas, SG Associates, and McCollom Management Consulting.

Traveler Response to Transportation System Changes. Technical Report B-12, Transit Cooperative Research Program (2000).

Robinson, S. Measuring Bus Stop Dwell Time and Time Lost Serving Stop Using London Buses iBus AVL Data (2013). Submitted to Transportation Research Board.

Schil, M. *Measuring Journey Time Reliability in London Using Automated Data Collection Systems*. Master's thesis, MIT (2012).

Shen, S. and Wilson, N.H.M. An Optimal Integrated Real-Time Disruption Control Model for Rail Transit Systems. *Computer-Aided Scheduling of Public Transport*, pages 335–363 (2001).

Strathman, J.G. and Hopper, J.R. Empirical Analysis of Bus Transit On-Time Performance. *Transportation Research*, 27A(2):93–100 (1993).

Trompet, M., Liu, X., and Graham, D.J. Development of Key Performance Indicator to Compare Regularity of Service Between Urban Bus Operators. *Transportation Research Record*, (2216):33–41 (2011).

Uniman, D.L., Attanucci, J., Mishalani, R.G., and Wilson, N.H. Service Reliability Measurement Using Automated Fare Card Data: Application to the London Underground. *Transportation Research Record*, 2143:92–99 (2010).

van Oort, N. *Service Reliability and Urban Public Transport Design*. Ph.D. thesis, Technische Universiteit Delft (2011).

Viggiano, C. Analysis of Bus User Behavior in a Multi-Route Corridor (2013). In-progress master's thesis, MIT.

Xuan, Y., Argote, J., and Daganzo, C.F. Dynamic Bus Holding Strategies for Schedule Reliability: Optimal Linear Control and Performance Analysis. *Transportation Research Part B: Methodological*, 45(10):1831–1845 (2011).