Service Reliability Measurement Framework using Smart Card Data: Application to the London Underground

By:
David Louis Uniman

Bachelor of Science in Industrial Engineering and Operations Research
University of California at Berkeley, 2005

Submitted to the Department of Civil and Environmental Engineering and
the Department of Urban Studies and Planning
in Partial Fulfillment of the Requirements for the Degrees of

MASTER OF SCIENCE IN TRANSPORTATION
MASTER IN CITY PLANNING

at the
Massachusetts Institute of Technology
June 2009

Author ……………………………………………………………………………………………
Department of Civil and Environmental Engineering
Department of Urban Studies and Planning
May 21, 2009

Certified by ……………………………………………………………………………………
John P. Attanucci
Lecturer of Civil and Environmental Engineering
Thesis Supervisor

Certified by ……………………………………………………………………………………
Rabi G. Mishalani
Visiting Associate Professor of Civil and Environmental Engineering
Thesis Supervisor

Certified by ……………………………………………………………………………………
Christopher P. Zegras
Assistant Professor of Urban Studies and Planning
Thesis Reader

Accepted by ……………………………………………………………………………………...
Joseph Ferreira
Chairman, Master in City Planning Committee
Department of Urban Studies and Planning

Accepted by ……………………………………………………………………………………...
Daniele Veneziano
Chairman, Departmental Committee for Graduate Students

Service Reliability Measurement Framework using Smart Card Data:
Application to the London Underground

By:
David L. Uniman

Submitted to the Department of Civil and Environmental Engineering and the
Department of Urban Studies and Planning on May 21, 2009
in Partial Fulfillment of the Requirements for the Degrees of:

MASTER OF SCIENCE IN TRANSPORTATION
MASTER IN CITY PLANNING

Abstract

Service reliability is an important dimension of performance to transit passengers, affecting not only their perceptions of service quality, but their travel behaviour as well. The ability of transit operators to understand and improve reliability relies on their ability to measure it. Until recently, efforts to quantify this attribute of service from the perspective of passengers were limited by the small sample sizes obtained from manual surveys, or the use of supply-side data to indirectly capture the passenger experience. With the emergence of data from automated fare media, it becomes possible under certain conditions to directly observe travel times experienced by passengers and obtain improved estimates of the reliability of transit service.

A framework is developed to estimate service reliability on heavy rail transit systems with both entry and exit fare control using data from Automated Fare Collection systems. A methodology is proposed as part of the framework to classify performance into incident-related and recurrent conditions in order to both gain new insight into the contribution of the different causes of unreliability as well as develop more robust measures of service reliability.

The classification methodology is validated against incident log data corresponding to three origin-destination (O-D) pairs on the London Underground. Subsequently, the proposed framework is used to characterize the reliability of 800 Underground O-D pairs representing the highest-volume journeys on the system. Furthermore, models are estimated to quantify the effects of journey length, interchanges, and incident-related disruptions on reliability.

Two practical applications of the framework are also developed for the Underground. First, an extension of the existing service quality measurement system is proposed in order to quantify reliability as part of routine performance monitoring efforts. An application of this extension to the Victoria line during the morning peak reveals that reliability is an important part of service quality, with a contribution to total perceived travel times comparable to that of various average travel time components. Second, a way to provide passengers with reliability information through Transport for London's trip planning software is presented, in order to mitigate the negative impact of uncertain journey times. The potential benefits from the provision of this additional information are found to be appreciable relative to the current ability of the trip planning software to reduce uncertainty for Underground passengers.

Thesis Supervisor: John P. Attanucci
Title: Lecturer of Civil and Environmental Engineering

Thesis Supervisor: Rabi G. Mishalani
Title: Visiting Associate Professor of Civil and Environmental Engineering

Thesis Reader: Christopher P. Zegras
Title: Assistant Professor of Urban Studies and Planning

"Oh very young
What will you leave us this time
You're only dancing on this earth for a short while"

– Cat Stevens (1948 -   )

Acknowledgements

This thesis is the culmination of three years of mentorship from some of the most dedicated, giving, rigorous and passionate individuals I have ever come across. They have taught me to be a stronger professional and made me more determined to make a difference in the world through my career.

I would like to begin by thanking Professor Nigel Wilson, who through his persistent drive towards excellence, taught me to take pride in everything I do. My sincere gratitude also goes to John Attanucci, for believing in me and giving me the chance to thrive. I would also like to express my deepest appreciation for Rabi Mishalani, not only for the unmatchable work ethic which I had the privilege to witness first hand, but for the sincere level of caring for my work and ideas he showed throughout the writing process.

I would also like to express my gratitude to Mikel Murga, for being an infallible source of inspiration, meditation, and wisdom; to Fred Salvucci for making sure that the transit research group at MIT remains alive and well; and to Chris Zegras, for helping me stay in tune with my priorities. Also, I want to give a special thanks to Ginny Siggia for keeping me (and my plants) alive while at MIT, and to Transport for London, for making this research possible.

To my friends at MIT – Princi-Klein, Valerie, Harvey, Andre, Julian, Hazem, DCBS, Lavanya, Yossi, Tony, Jared, Liz, Mini, Frumix, Amey, Pete, Candy, Joanne, Zhan, Jinhua, Margarini, Travis, and all sorts of 1st, 2nd and 3rd years – thank you for being my tears, my laughs, my panic attacks, and my triumphs. I also want to thank Nene, for having shared with me the light of her love when I needed it the most.

Finally, I would like to say thank you from the bottom of my heart to the people who are at the center of everything that I accomplish: my Mom, my Dad, and my family – this degree is as much yours as it is mine. I love you.

Thanks.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Apart from efficiently using limited resources, a central concern of transit agencies is to ensure that the quality of service, or the performance of the system from the passenger's perspective, remains competitive relative to other modes (Kittelson & Associates et al., 2003b). This concern has led to the development and study of service performance measures that are appropriate for representing the passenger experience and can be used to evaluate strategies aimed at improving the quality of the service.

The gradual emergence of Automatic Data Collection (ADC) systems over the past few decades has allowed transit agencies to improve not only the efficiency and cost-effectiveness of traditional functions of data collection related to service quality monitoring, but also to enhance their ability to represent performance from the passenger's standpoint. These developments have taken place largely through the use of vehicle location information such as traditional signal data for rail and more recently through Automated Vehicle Location (AVL) technologies for bus modes, which coupled with data on passenger counts either at station gates or through Automatic Passenger Counter (APC) systems, have allowed analysts to estimate service quality at more disaggregate levels than previously possible through manual data collection methods (Furth et al., 2006b; Fattouche, 2007).

Data from Automated Fare Collection (AFC) systems is another important source of information that has become available to transit agencies in recent years. More specifically, the proliferation of Smart Cards has opened up a range of applications benefiting from the high resolution of this source of data, capturing information at the individual passenger level over time (Chan, 2007). One important application of this source of data is in the area of service quality monitoring, where Smart Cards have the inherent advantage over traditional methods based on vehicle-location data of directly capturing the performance of the system as experienced by passengers.

This thesis explores the potential of AFC Smart Card data as a way to quantify and gain insight into performance from the perspective of passengers. In particular, it centers on the topic of service reliability, taking advantage of the characteristics brought forth by this type of data, as well as some of the theoretical work done in this area, to propose a practical framework for quantifying this aspect of service quality and developing applications to aid transit agencies in their efforts to improve it. A set of reliability measures is proposed as part of the framework that can be used to help breakdown performance and identify the contribution of different factors to unreliability.

The framework is used to the develop applications for the London Underground; an ideal setting for this research given the characteristics of the system (e.g. exit as well as entry fare control) and the characteristics of the Oyster Smart Card (e.g. high trip penetration). Specific applications include a reliability addition to the existing passenger-oriented performance measurement system and improvements to the web-based trip planning software provided by Transport for London (TfL), illustrating how these and other applications can be developed for a broader range of contexts and used by transit agencies for improving service reliability.

1.1: Motivation

The reliability of transit service is important to operators not only because it is directly related to the cost-efficiency of providing service, but also because it is important to their customers.

From the perspective of transit operators, reliability is often defined as adherence to schedule, and is associated with having a more efficient use of available resources (Abkowitz et al., 1978; Strathman et al., 2000). From the perspective of the passenger, however, reliability is usually associated with its effects on the probability of on-time arrival at one's destination, leading to changes in passenger behaviour due to the increased disutility of travel derived from increased travel time uncertainty (Abkowitz et al., 1978; Bates et al., 2001). The most direct consequence of unreliable service is on the departure time decision, where travelers seek to minimize the generalized cost of travel in the face of variable journey times. Other decisions affected by changes in reliability are route choice and even mode choice, possibly leading at the extreme to a decrease in ridership and revenues from fares (Evans et al., 2004).

These effects on the disutility of travel and the travel decisions of passengers are also important to transit agencies because of their implications for transit planning and management. An accurate estimation of the behaviour of passengers with respect to reliability can lead not only to improved ridership forecasting, but also to general improvements in mode and route choice prediction and transport model assignment (Recker et al., 2005). Additionally, knowledge on the value travelers place on reliability and the ability to quantify changes in this attribute of service are important for the evaluation of different policies and investment strategies aimed at improving service quality. Researchers also benefit from this study as this research contributes to the growing, yet still modest, body of applications in this area devoted to public transport, adding to the base of empirical knowledge on which to develop future studies (and applications) on the measurement and understanding of transit service reliability. In addition, an improved understanding of service reliability can lead to a better use of this type of data for overall service quality monitoring and evaluation.

Finally, a large portion of the motivation for this study derives from the benefits to Transport for London and the Underground in particular. At a strategic level, TfL has proposed to improve the consistency and reliability of journey times for the network among its ten key objectives (TfL, 2007a). Currently, its primary service quality monitoring regime focuses on average performance and could therefore benefit from this work in terms of explicitly measuring service reliability. At a more specific level, this research is especially relevant to travel in the Underground because of the high proportion of journeys that are work-related on this mode, which is one of the segments of passenger trips where reliability has been found to be most important. Moreover, this work is possible largely due to the fact that unlike other modes and heavy rail systems, the Underground requires both exit and entry fare validation, providing researchers with valuable data on passenger travel times.

1.2: Research Objectives

The overarching goal of this thesis is to explore the potential for using Smart Card data to quantify reliability from the perspective of passengers by developing a practical framework that can be used by transit agencies, including the London Underground, to improve this aspect of service.

The ability to quantify reliability can be used to shed light on simple yet important questions such as “How reliable is the London Underground?” More specifically, the framework can be used to answer additional questions, including:

- How does one *think* about and *define* reliability from the perspective of passengers?
- How does one operationalize this definition in order to *quantify* reliability?
- How does one gain *insight* into the causes of unreliability and their relative contribution?
- How can reliability and overall service quality be *improved*?

In order to achieve its primary goal, and in the process help answer some of the broader questions posed by it, this thesis sets out to achieve the following objectives:

1. Identify the characteristics of Smart Card data relevant to performance monitoring applications focused on service reliability.

2. Review the literature on the effects of reliability on transit passengers to provide a theoretical basis on which to develop measures of reliability.

3. Propose a set of reliability measures that reflects passenger perceptions, makes use of the characteristics of Smart Card data, and is complementary to the existing service quality measurement system currently in place at the London Underground.

4. Develop a methodology based on the set of reliability measures for breaking down performance and identifying the different factors and their contribution to the reliability of the system.

5. Develop a set of practical applications based on the proposed framework that can be used by transit agencies as part of their routine delivery of service to improve reliability and overall service quality.

1.3: Research Approach & Thesis Structure

Given that the aim of this thesis is to take a step away from existing methods for quantifying reliability and move towards a reliability framework that takes advantage of Smart Card data, a blank slate is used as a starting point on which to build a theoretical basis for the way passengers define reliability, and how they are affected by it. Chapter 2 achieves this through a comprehensive review of the literature on the topic of transit service reliability.

Chapter 3 presents information on the specific context in which this research was developed, starting with a review of public transport in London and the existing service quality measurement system in place at the Underground, and finalizing with a description of the characteristics of the Oyster Smart Card data used in this study.

The framework is developed in Chapter 4, starting with a definition of the proposed set of reliability measures. Subsequently, the measures are used to develop a methodology for breaking down performance into different categories that relate to the causes of unreliability, which is in turn used to develop robust extensions to the initial set of reliability measures. A theoretical presentation of the framework is followed, complemented by empirical examples from the performance of the London Underground to illustrate the key concepts.

The framework is used in Chapter 5 to characterize the reliability of the Underground and quantify the contribution of different factors to performance using Smart Card data from two four-week periods in 2007. First, the reliability of three origin-destination pairs in the Underground is quantified in order to provide a more detailed view of reliability, serve as an example for how the framework is applied, and validate an important element of the methodology. Second, the contribution of various factors to reliability is estimated through a regression analysis of the 800 highest-volume O-D pairs in the system.

Two practical applications are developed in Chapter 6 for improving reliability in the London Underground based on the proposed framework. First, an extension of the Underground’s existing service quality monitoring system is presented. This extension makes it possible to quantify impact of unreliability on service quality in a way that is compatible with current monitoring efforts at the agency. The second application puts forth an approach for providing reliability information to Underground users through one of the existing passenger information systems, Transport for London’s web-based trip planning software.

Finally, Chapter 7 provides a summary of the key findings from the study, as well as recommendations for future research directions.

Chapter 2: Literature Review

There is a substantial body of work devoted to the topic of transit reliability dating as far back as the late 1960's and continuing through today. This on-going study of reliability is testimony not only to the importance transit agencies place on it, but to its complexity, given that it is still far from being resolved.

Within the literature, there is the seminal Transit Reliability Study which presented a broad overview of the topic, much of which still holds to this day (Abkowitz et al., 1978). Specifically, the authors proposed a formal definition of transit service reliability, which is taken as a useful starting point for this research. Reliability was defined as *the invariability of service attributes which influence the decision of travelers and transportation providers*. This general definition provides two key insights into the study of reliability.

First, reliability is defined within the context of the service concept, referring to the consistency of the attributes that result from the provision of transit service. This meaning separates itself from another equally important definition of reliability, which refers to the availability of equipment and other parts of the physical infrastructure for providing service. The availability of vehicles for service provision is linked to the final outcome in terms of service performance, as will be discussed in section 2.1.2 (causes of unreliability). However, measures of this type of reliability tend to be separate from measures focused on the execution of the operating plan, often being quantified as the amount of resources used as a percentage of the total time they were scheduled for use (Kittelson & Associates et al., 2003a). This thesis is more concerned with the reliability of the service provided, which is also the intention of the definition put forth by Abkowitz et al.

Second and more importantly, this definition of reliability makes the key distinction between the perspective of the transit provider and the transit patron. This is a useful way of categorizing subsequent studies of reliability into two general streams of work: reliability from a supply-side perspective and reliability as experienced by passengers. Each approach has had not only its own definition of reliability, but also its own focus. There is also a considerable difference in the degree of application of these two streams of work by transit agencies in their day-to-day operation, with the former group receiving the bulk of the attention. This can be attributed to, among other factors, the availability of data, where the existence of vehicle location data from signaling systems for rail and more recently for bus modes through AVL technology has made it possible to gather the large amounts of information necessary for a detailed analysis of reliability from a supply-side perspective (Furth et al., 2006b). With the appearance of Smart Card data from AFC systems, however, this picture has changed, allowing agencies to contemplate the possibility of measuring reliability directly from the perspective of passengers in a cost-effective way. The focus of this thesis is in this particular area, aiming to develop a framework for quantifying reliability directly from the perspective of passengers, taking advantage of the proliferation of this new type of data in cities like London.

This chapter reviews the literature relevant for developing this type of application, starting in section 2.1 with a general overview of previous reliability studies from the perspective of operators. Section 2.2 provides an in-depth review of the literature concerning reliability and passenger perceptions in order to provide both an understanding of how the two perspectives differ, but mainly as a foundation on which to develop a set of reliability measures later on. Section 2.3 contributes to this by providing a review of previous work directed at developing reliability measures from the perspective of passengers. Finally, section 2.4 provides concluding remarks that serve as inputs for the reliability framework proposed later on in Chapter 4.

2.1: Overview of Supply-side Reliability Studies

Reliability from the perspective of operators is usually defined in terms of schedule adherence (Abkowitz et al., 1978). This approach tends to focus on the variability of the attributes of service around the operating plan, such as vehicle running times, headway regularity, adherence to schedule at specific stops (particularly for long-headway service), and the evenness of vehicle loadings (Strathman et al., 2000).

There has been a considerable amount of work devoted to reliability from the perspective of operators, especially in recent years with the appearance of AVL systems for bus that made it possible to obtain the large data sets, previously infeasible to collect through manual surveys, needed to estimate the distribution of the service attributes (Furth et al., 2006b). The work directed towards reliability from the perspective of transit providers has focused on four main areas: the effects of reliability, the selection of reliability measures, the causes of unreliability, and the strategies to address reliability problems (Cham, 2006). Work in these four areas is reviewed in this section for three reasons. First, it serves as a point of comparison for distinguishing between supply-side and passenger-focused reliability studies. Second, an understanding of the causes of unreliability and the strategies to address them is used as an input into the development of the reliability framework in Chapter 4. Third and foremost, the stream of work focused on supply-side measures of reliability has had a high degree of application by current transit service providers, and in order to develop applications of the proposed framework, it is important to first understand the current state-of-the-practice. The four main areas that have been the focus of previous studies within this stream of work are discussed next.

2.1.1: The Effects of Unreliability

The effects of unreliability on transit operators are twofold. On the one hand, a high degree of variability in the service attributes makes it more difficult for resources to be used efficiently in the provision of service, leading to increased costs of operation. On the other hand, unreliability directly impacts service quality, potentially leading to changes in ridership and revenues from fares.

The first type of effects are well-studied and are related to the variability of the attributes of service that hinder the ability of transit operators to provide the desired service according to plan, while making efficient use of limited resources. Strathman et al. (2000), for example, point out that a decrease in the variability of arrival times at terminal points would allow schedulers to decrease any excess running time from schedules, creating the potential for a reduction in the number of vehicles needed to provide a certain frequency of service. Additionally, improved reliability of arrival times at stops reduces the problem of vehicle bunching, which can lead to problems of utilization of resources as seen through the overcrowding of some vehicles and spare capacity in others. More generally, Abkowitz et al. (1978) point out that improvement in the reliability of the attributes of service can lead to two types of benefits for transit operators. First, operators can reduce the amount of resources needed to provide a given level of service, enabling them to either save the costs of operating them, or to schedule their use towards providing service along a different line or route. Second, operators can maintain the existing level of resources and provide a higher level of service than before.

The second type of effect of unreliability is on the quality of service experienced by passengers. This is of concern to transit agencies because of its effects on ridership and revenues from fares. Historical surveys point to the importance of reliability as a part of service quality, showing that in the case of the journey to work, passengers rated it as being equally if not more important than more traditional measures of travel time, such as the average speed (Abkowitz et al., 1978 citing Payne et al., 1976). More recent empirical studies confirm this, finding that under certain circumstances passengers value an improvement in reliability more heavily than an improvement in the mean time spent for a journey (De Jong et al., 2004). As will be covered in section 2.2, a decrease in reliability can lead to a significant increase in the disutility of travel perceived by passengers and changes in their travel behaviour, with a possible loss of ridership in extreme cases (Evans et al., 2005). An improvement in reliability benefits transit agencies by increasing the competitiveness of transit relative to private modes, leading to sustained ridership and possibly additional journeys.

The concern with the effects of reliability on operational quality as well as service quality has led to the use of measures that try to capture both types of effects. These measures of operational quality also tend to be used as proxies for service quality, as discussed in the following section.

2.1.2: Reliability Measures

The measures of reliability commonly used by transit agencies quantify this attribute of service quality indirectly through the use of supply-side data from Automated Vehicle Location systems and rely on the relationship between operational quality and the passenger experience to make inferences on the level of service being provided. This differs from an approach where reliability as experienced by passengers is directly *measured*, as opposed to *estimated*, and has proliferated due to the availability and cost-effectiveness of data collection from AVL systems. Wilson et al. (1992) point out that though it is possible in theory to gather information on the individual passenger experience directly through surveys, this is difficult to support as part of routine performance evaluation and recommend its use only periodically to validate the operations-based proxies for the passenger experience.

It is because of the widespread use of operations-based reliability measures that they are reviewed next. The first two measures of reliability, On-Time Performance (OTP) and Headway Regularity (HR), are commonly used by transit agencies and act as proxies for the passenger experience. The last two measures, Mean and Excess Passenger Wait Time, are more elaborate measures of reliability, deriving the average passenger experience from vehicle location data using mathematical relationships.

On-Time Performance and Headway Regularity

Furth et al. (2006b) distinguish between two useful set of measures of service reliability: those related to schedule adherence, and those related to the distribution of vehicle headways. Schedule adherence measures can be described as the percentage of vehicle departures that take place in a defined on-time window at a specific location. These types of measures are important for infrequent users, timed transfers, and low frequency service (Strathman et al., 2000). The last item, low frequency service, is especially relevant given that passengers are expected to try to minimize their wait time at the station by coordinating their arrivals with the scheduled service departures, in the case of long-headway service, leading to a mean wait time that is less than half the headway (Turnquist & Bowman, 1981). On-time Performance is the measure of schedule adherence most commonly used by transit agencies, measuring the percentage of journeys that departed up to 1 minute early and 5 minutes late from the scheduled departure time (Kittelson & Associates et al., 2003a). Other less commonly used measures of schedule adherence exist, with variations in their definition that make them, at the expense of losing their simplicity, more robust for transit agencies to use. For an interesting discussion of OTP and potential improvements, see Henderson et al. (1991a).

For high-frequency service, often defined as headways of ten minutes or less, variation in vehicle headways becomes a more relevant measure (Furth et al., 2006b). This is because at short service intervals, passengers are expected to arrive at transit stops randomly, and maintaining equally spaced vehicles becomes important for minimizing the aggregate wait time of passengers; not to mention to maintain balanced loadings across vehicles and ensure high levels of utilization of capacity (Strathman et al., 2000). Headway Regularity, defined as the percentage of headways that fall within a specified range from the scheduled headway, is also a reliability measure frequently used by transit agencies and research studies (Cham, 2006). New York's MTA-NYCT, for example, considered a vehicle's headway as reliable if it deviated by +/- 50% from the scheduled headway (Kittelson & Associates et al., 2003a). Other interesting extensions of these measures exist which also provide useful insight into performance for transit operators. For example, the Transit Capacity and Quality of Service Manual assigns different levels of service based on the values of the coefficient of variation of vehicle headway deviations[1] (Kittelson & Associates et al., 2003b). For an interesting discussion on another useful extension of measures of headway consistency, see the Headway Regularity Index based on Gini's ratio in Henderson et al. (1991a).

These two types of measures represent the operational quality of the transit system more than the actual passenger experience. There are, however, other more elaborate measures that act as better proxies for the way passengers experience reliability. Two of these measures are discussed next.

[1] Coefficient of variation of headway deviations = standard deviation of headway deviations / mean scheduled headway.

Mean Passenger Wait Time and Excess Passenger Wait Time

Like the previous set of reliability measures, this set also estimates the passenger experience indirectly from vehicle location data and measures of operational performance. However, these types of measures more accurately estimate the actual passenger experience from operations data through mathematical relationships, allowing transit agencies and researchers to capture service quality with greater fidelity. Wilson et al. (1992) discuss two key measures of service quality, the first being the commonly cited relationship between vehicle headways and the mean passenger wait time for high-frequency service shown in Equation 2-1 (Osuna & Newell, 1972):

$$\overline{w} = \frac{\overline{h}}{2}[1 + \text{cov}^2(h)] \qquad [2\text{-}1]$$

Where $\overline{w}$ = mean passenger wait time,
$\overline{h}$ = mean headway, and
$\text{cov}(h)$ = coefficient of variation of the headways.

This relationship is an important one for service quality evaluation, not only because of the high value passengers place on this component of travel time, but because it is straightforward to estimate using operational data already being collected (i.e. for measures of Headway Regularity). This method of estimating the mean passenger wait time is also regarded as being representative of the actual passenger experience, being used in more recent studies to quantify the costs of alternative schedules to passengers in terms of wait time (Fattouche, 2007).

Wilson et al. (1992) also discuss another similar measure of reliability, Excess Wait Time. It is defined as the difference between the actual expected passenger wait time and the expected wait time that would result from perfect adherence to schedule, and is noted for being useful for comparing service quality across routes with different headways. This type of measure is currently used by Transport for London as part of its Journey Time Metric system (see section 3.3.1), estimating passenger wait time from train signal data and comparing it to an ideal or scheduled value.

Given the high cost of data collection through manual surveys and the advantages of data from AVL systems, these measures are not unreasonable proxies for quantifying the passenger experience. However, they are still indirect measures of the service as experienced by passengers, with a focus ranging from a purely operational perspective to mathematical derivations of the actual passenger experience. This leaves room for improvement, particularly with the availability of data obtained from Automated Fare Collection systems, which can be used to directly measure passenger travel times (in the presence of entry- and exit-validation), and service quality overall.

2.1.3: Causes of Unreliability and Strategies to Address Them

The ability to quantify reliability allows analysts to study the factors that contribute to unreliability and to determine the strategies that can be used to improve the service. The causes of unreliability can be categorized along two dimensions that relate the nature of the different causes of unreliability to strategies that mitigate their effects (Abkowitz, 1983).

The first dimension distinguishes between causes of unreliability that are *intrinsic* to the service concept, and those that are *exogenous* to it. The first can be understood as those factors that can be controlled directly by the transit agency, and can themselves be grouped into three general categories: route-related factors, maintenance related activities, and human factors. Route-related factors are related to route and network design (e.g. stop spacing) and service planning (e.g. schedule accuracy and service frequency). Maintenance related activities are those factors related to vehicle maintenance and availability, including the maintenance plan and the quality of the vehicle stock. The last category, human factors, has to do with the contribution to unreliability by staff. This includes the contribution of staff to unreliability (e.g. no-shows by drivers), as well as the contribution of poor communication across departments and between management and their staff. The sum of these three categories makes up the set of intrinsic factors.

Exogenous factors, on the other hand, pertain to the environment in which the delivery of service takes place and are beyond the direct influence of the transit provider. They include weather-related problems, accidents, unpredictable events, and unpredictable variation in demand (e.g. length of tennis match). For bus systems, street congestion and road construction also apply.

The second dimension is related to the types of strategies that can be applied to deal with the various reliability factors. On one end of the spectrum are those factors that affect the system repeatedly over time, or that offer some level of predictability of occurring. These factors relate strongly to long-term strategies for dealing with them and often involve fundamental changes in the structure of the service that are designed to address predictable sources of unreliability – i.e. service planning. Strathman et al. (2000) point out that through modifications in the schedule, recurring delays such as traffic congestion for bus systems can be integrated into the operating plan and mitigate its impact on reliability.

On the other end are those factors that must be dealt with on a real-time basis given that they are unpredictable by nature in terms of when they occur, as well as the way they will impact the service. These factors include small and large incidents such as driver absenteeism and vehicle breakdowns, which can be mitigated through corrective or remedial measures as they occur. The strategies directed towards this type of factors are short-term by nature, and less on the correction of structural inefficiencies in the delivery of service and more on bringing the operation in-line with the plan. Table 2-1 illustrates where several causes of unreliability fall within these two dimensions.

Table 2-1: Factors influencing service reliability (adapted from Abkowitz, 1983)[2]

	Long-term/Planning	Short-term/Real-Time
Controllable/ Intrinsic	• route and network design • schedule planning • time of day and seasonal variation in demand and running time • stop frequency and location • driver behaviour – predictable aspects • route length • overloaded buses • high service frequency • missed runs – predictable	• vehicle breakdowns • poor dispatching from terminal • late pullouts from depot • absenteeism – unpredictable • driver behaviour – unpredictable • missed runs – unpredictable
Exogenous/ Environmental	• demand variation – predictable • weather	• fires and accidents • passenger-related incidents • unpredictable events (e.g. length of tennis match) • large demand variation - unexpected

The strategies related to each of the quadrants of Table 2-1 differ in terms of the immediacy in which they are applied and the type of approach taken. Starting with the most directly implementable by a transit agency are those strategies related to factors that are controllable and that can be dealt with through long-term planning solutions. These strategies involve recurring problems that can be corrected through better schedule and route design. The next group, factors that are controllable but must be dealt with in real-time, is slightly harder to approach yet still within the agency's ability to respond. The strategies in this group are mainly operations control strategies, or bringing the service back to the planned schedule, and can arguably be considered more difficult to identify and address than longer-term problems. Next are the causes that are exogenous to the transit agency but that are repeatable in nature and therefore can be dealt with through planning strategies. The contribution to unreliability by this group can be mitigated by taking these sources of delays into account in the operating plan (e.g. predictable variation in demand), or through coordination with external agencies capable of dealing with these sources of delay (e.g. large events). Finally, the most difficult category of factors to address includes those that are unpredictable and beyond the control of the transit agency. These factors must be addressed through mitigation strategies to respond to delays in real-time, which ultimately add a large amount of complexity to the routine delivery of service.

Finally, another strategy for improving service reliability that is more related to the perception of passengers is providing updated and improved wait and travel time information. This especially applies to riders with little previous experience riding the system since they depend on published information to make their travel decisions (Abkowitz, 1983).

[2] There is some overlap of factors that occur randomly by nature but offer some degree of predictability, such as missed runs and driver absenteeism.

2.1.4: Limitations of Current Approach

The limitations of the current approach towards quantifying reliability are well known and discussed in different parts of the literature. Many of these limitations, however, are due not to the poor development of service quality measures, but rather because of the practical constraints that come with using supply-side data and the tradeoff between the cost-efficiency of data collection with the level of resolution desired for evaluating performance. Wilson et al. (1992) point out that realistic goals for these types of performance measurement systems are to (a) measure average service quality, (b) compare actual service quality with an ideal standard, and (c) measure the percentage of passengers receiving good service. One can say that these goals are met by the current approach, providing transit agencies with a cost-effective way of estimating reliability and average service quality.

However, there are limitations to the current approach. Furth et al. (2006a) begin their analysis of reliability measurement by pointing out that current measures, such as those covered in the preceding section, are focused more on operational quality than on the passenger experience due to their focus on average values as opposed to extreme values of travel time – considered central to the way passengers experience reliability. Henderson et al. (1991b) second this perspective and raise the issue of interpretation of results, where a 10 percent increase in On-Time Performance is not easily translated into quantifiable benefits to passengers. This problem is exacerbated by the focus on vehicle trips, where often times reliability is measured at remote stations and terminals which might not be representative of where the majority of passengers are active, or during operating periods as opposed to shorter time windows reflective of passengers' typical commute times.

In addition, measures that evaluate service quality using the scheduled service as a benchmark can be misleading in terms of benefits/costs accruing to passengers (Abkowitz et al., 1978). This is because a change in the schedule might artificially lead to measured improvements in performance, without any changes perceived by passengers. This is also true because existing reliability measures might not accurately reflect the perceptions of passengers. Strathman et al. (2000) point out the case where a service that consistently departs a station two minutes late is considered more reliable by passengers than a service that arrives one minute early on some days and four minutes late on others.

Several of these concerns can and have been in some cases addressed with the increasing use of AVL and Automated Passenger Counter systems, providing large sample sizes to both measure performance at higher levels of resolution (e.g. O-D level, shorter time windows) and move beyond average performance to focus on extreme values. However, in order to accurately represent the passenger experience, it is important to step away from traditional operations-based measures of performance and begin with an understanding of how passengers define and experience reliability. These perceptions are explored in the following section.

2.2: Reliability and Passenger Behaviour

The objective of this section is to establish an understanding for the way passengers perceive reliability and are affected by it, so that more effective measures of service quality can be developed as part of the framework. Specific questions that are explored include:

- How do passengers define reliability and how does this relate to specific service attributes?
- What are the effects of unreliable service on traveler behaviour and how are journey decisions used to counteract the effects of uncertainty?
- What value do passengers place on reliability and what factors influence this perception?
- How much do passengers' perceptions of reliability deviate from the actual reliability of the system, and what role does trip information and passenger experience play in this?

Insight into these questions is provided starting in section 2.2.1 with a review of studies on passenger attitudes and perceptions of reliability, leading to a qualitative definition of this attribute of service. Section 2.2.2 provides an overview of some of the primary quantitative theoretical frameworks used to explain how reliability influences passengers' travel costs and behaviour. Lastly, section 2.2.3 covers some of the empirical findings that support the theoretical frameworks discussed previously.

2.2.1: Passenger Views on Reliability

Two types of studies provide insight into the way passengers define and experience service reliability. The first are attitudinal surveys, which try to determine the importance passengers place on certain service attributes relative to others. The second are market research surveys, which try to capture anecdotal evidence for the way passengers view different aspects of the service.

Passenger Attitudes towards Reliability

Much of the historical evidence showing that reliability is an important part of service quality for passengers comes from attitudinal surveys (Bates et al., 2001). These surveys present various attributes of service to passengers under varying contexts and have them rate them on a scale of importance (e.g. 1 to 7). Some of the attributes identified as most related to reliability include:

- Enables me to arrive at my destination at the intended time
- Inconveniences me by leaving ahead of schedule
- Has a travel time that varies a lot from day-to-day

- Is likely to suffer from a vehicle breakdown

These results have found the impact of reliability-related attributes on service quality to be significant, with those related to “arrival at intended time” appearing to be, in the case of work-related travel, more important than savings in average travel times or increases in speed and only second to “arrival without accident” (Paine et al., 1976). These attributes provide some insight into the relationship between specific service characteristics and passengers’ definitions of reliability.

Passenger Perceptions of Reliability

A study commissioned by London Buses titled “The Meaning of Reliability” explored reliability as defined by passengers as well as the key drivers behind these perceptions, through the use of ten focus groups composed of passengers of the system (Love & Jackson, 2000) [3]. The study was conducted on commuters as well as leisure travelers across various bus routes in the network. Some of the passenger testimonies from the survey are included for illustration.

The first part of the study discusses the general perception of passengers when asked to define the reliability of bus service. The concept of reliability was described by passengers both in terms of the specific attributes of the service, as well as the outcomes associated with reliable service. Specific attributes of the service included adherence to a printed timetable, the consistency of vehicle arrivals at the stop across multiple days, the absence of vehicle break-downs while en-route, not being left behind because of capacity issues (pass-ups), clean interior, as well as adequate seating and other aspects related to comfort. Responses related to the outcome of reliable service were mainly regarding on-time arrival at one’s destination and the ability to trust/depend on the information provided to passengers.

> *“Reliability means being able to get to work on time.”*
> *(Tottenham peak time bus user)*
>
> *“Reliability means being able to trust what they tell you.”*
> *(Acton leisure bus user)*

Bus users were also asked to rate the bus service, revealing some insight into the factors that influence their perceptions. These included individual experiences, the purpose of the journey, and their concern with the service adhering to a strict schedule. Users with fixed time appointments tended to rate the service lower than users traveling for leisure. The study also emphasized the difference between objective reliability and subjective reliability, attributing the majority of passenger responses to a definition of reliable service that was their own.

[3] Due to the passenger-centric nature of the study, its findings are still useful for understanding the definition of reliability for passengers of heavy rail services like the London Underground.

"It is not the easy journeys you remember, it is the frustrating journeys that stay with you."
(Acton commuter)

It was noted that passengers expressed their perceptions of the reliability of the service at the route level (as opposed to buses as a mode in general), comparing between the reliability of specific routes, and in some cases by direction (e.g. morning vs. evening commute). Overall, however, the general feeling of passengers (especially those with fixed-time appointments) towards the meaning of reliability was based on the bus arriving and reaching its destination on-time. The inconsistency of both the vehicle arrival and overall travel time forced passengers to adjust their personal schedules to adapt to the vagaries of the service.

"Not having to leave home ¾ hour early to be at work on time."
(Tottenham commuter)

"It depends if you are a gambler. Half hour extra in bed, you risk it."
(Tottenham commuter)

Another important issue raised by passengers was the timetable. The authors of the study pointed to the difference between a printed timetable and a virtual timetable, or the schedule internalized by passengers from experience. Passengers often dismissed the printed timetables, relying more on their "folk" knowledge about the routes. Other passengers simply provided for an accepted range from which they considered whether the service was reliable or not, usually building their allowances (i.e. virtual timetable) based on previous experiences. Perceptions of Countdown, the vehicle arrival time information system, also varied, with the majority of passengers expressing approval but also recognizing that it was not the information that took the brunt of the blame, but rather the buses if they did not arrive as promised.

"I would be more inclined to ask a local than go by the bus timetable."
(Hammersmith fixed-time leisure)

"So long as you allow about ten minutes either way you should be safe."
(Stratford commuter)

An interesting discussion arose concerning the tradeoffs between service frequency and service reliability. The opinions were divided, with some passengers expressing more concern with the frequency of arrivals at stations, than with looking at printed schedules or a specific number of minutes. Others, however, expressed their preference for reliable service, mentioning that they would be willing to have 15-minute to half-hour service intervals on some routes, in exchange for schedule adherence.

"If it is every half hour, that is OK. I don't have to hang around for half an hour, I know that 29 minutes later there will be another bus. That is what I call reliability, it is not the number of buses, it is the fact that you know what time they will get there."
(Acton commuter)

Finally, besides issues related to the timeliness of the service, passengers expressed concern for the quality of the journey, emphasizing the comfort and level of safety experienced. The authors close with a discussion of appropriate ways for measuring objective reliability and subjective reliability, noting that ultimately the latter score is what will count when evaluating passengers' perceptions of the service.

Passenger-oriented Definition of Reliability

Beyond any anecdotal evidence for the meaning passengers give to reliability, a more formal definition of reliability is still needed in order to translate passenger perceptions into something that can be systematically measured based on the attributes of service. The following discussion, based largely on the discussion in Bates et al. (2001) and Noland & Polak (2002), attempts to provide such a definition by breaking down the concept of reliability into its different components.

The concept of reliability evokes parallels with other attributes such as dependability or predictability. Inherent in all these concepts is the idea that a (transit) service will consistently produce outcomes that are close to passengers' expectations of what it *should* do, which may or may not be based on what is promised through published timetables and other information. Hence, one way to think about reliability is as an indicator for how much and how often outcomes deviate from the planned and/or expected outcome. For the case of transit, this can be related to the consistency of comfort and safety aspects, but more typically to attributes related to travel times.

There is a similarity between travel time deviations from their expected value and travel time variability in the statistical sense. Variance, for example, measures the deviations of observations from their mean or expected value. Thus, it would seem reasonable that when quantifying reliability, one could use measures of the variation of travel times *about* their expected value, with the definition of the latter becoming central to the results one will obtain.

Bates et al. (2001) illustrate this point with a passenger who travels to work every day at the same time during the morning peak (e.g. 8:00am). If that same passenger were to make the same journey but at off-peak times, she would experience a different travel time than usual, caused possibly by changes in the schedule (e.g. lower service frequencies), road congestion, or by the changing likelihood of non-recurring events (e.g. incidents). In this case, to consider a different travel time as pertaining to an unreliable journey would be unreasonable given that the trip was not made under comparable conditions. Therefore, when measuring reliability about an expected value, similar conditions, if not repeatable, that a passenger could come to *expect* should be considered. Noland & Polak (2002) point out that deviations from expectations could be measured at various levels, including vehicle-to-vehicle variations (e.g. due to different drivers), time-of-day variations as in the example above, or, more commonly, variations across days. The key point being the separation of predictable performance as much as possible from unpredictable performance, capturing the former through an expected value one measures deviations from. Bates et al. (2001) adds to the discussion by noting that predictable performance can be thought of in terms of non-random outcomes, whereas deviations from this expectation would pertain more to random effects, such as incidents or other day-specific conditions (see Table 2-1).

Another issue with the selection of the expected value used to measure deviations in performance is related to the sources of information that passengers base their expectations on. A transit patron who frequently makes the same journey might be compelled to rely solely on his previous experiences if he finds the printed timetables to be a poor reflection of reality. Conversely, a passenger making a journey for the first time would be forced to rely on the information provided in, for example, a trip planner tool. This difference in expectations could help explain why a service arriving consistently late at the destination by five minutes might not be viewed by the repeat rider as unreliable (because variation around the five minutes is zero), whereas the rider traveling for the first time might be irritated by the delay vis-à-vis the printed schedule. This highlights two important aspects of reliability. First, it is understandable why for scheduled transit services, as opposed to travel in private modes, the idea of *punctuality* is so strongly related to reliability, acting as a half-way point between the expectations of passengers and transit operators. Second, the important role traveler learning and passenger information play in efforts to improve reliability becomes evident. Section 2.2.2 discusses these concepts further and provides a formal explanation for the way passengers are affected by deviations of the service from expectations.

2.2.2: Theory of Reliability Effects on Travel Behaviour

There is an important amount of work done to establish a theory for the way passengers are affected by unreliability dating back several decades. This work has focused largely on identifying the specific costs of uncertainty, and the effects these have on the travel choices of passengers. The theoretical underpinnings of such work are summarized in this section, followed by a brief sample of the empirical findings in this area.

Theoretical Models of the Cost of Travel Time Uncertainty

The relationship between travel time reliability and the (dis)utility of travel of passengers is captured through two general approaches, both rooted in discrete choice analysis and utility maximization theory.

The first approach tries to capture the effects of unreliability directly under the assumption that there is an inherent burden to passengers caused by uncertainty (similar to the burden of mean travel time). It is commonly referred to as the "mean-variance" approach because of the use of a term representing travel time variability explicitly in utility model specifications as shown in Equation 2-2:

$$U = \alpha \cdot (Travel\_Time) + \beta \cdot (Std\_Deviation) \qquad [2\text{-}2]$$

Where U = utility derived from a journey with a particular travel time distribution,
Travel_Time = average of trip travel time distribution, and
Std_Deviation = standard deviation of trip travel time distribution.

Typical measures of variability used by empirical studies include the standard deviation and variance (Hollander, 2006). However, a recent study by Lam and Small (2001) found the difference of the 90th percentile and the median travel time (itself found to be a better predictor of choice than the mean) to have better explanatory power using this same approach.

The second approach is based on "schedule delay" models. It captures the effects of travel time variability on passenger *indirectly* through its changes in passenger scheduling decisions[4]. The underlying theory behind these models is that the primary mechanism passengers have to respond to uncertainty in travel times, given a situation where there is a preferred time of arrival, is by modifying their departure time. This is because of the tradeoff that exists between the time of departure and the chances of on-time arrival at destination, with earlier departures offering a lower chance late arrival at the expense of a greater allocation of total time for traveling. Variables that capture the level and probability of delay caused by travel time uncertainty for various departure time choices are used, making it possible to estimate the time of departure that minimizes the expected disutility of travel. Hollander (2006) points out that models of the "mean-variance" form are also used to model departure time choice, though these do not make use of variables related to passenger schedules.

Empirical results tend to lean in favor of the "schedule delay" approach, with some studies concluding that the use of scheduling variables makes it unnecessary to include an additional term for the explicit variability of travel times (Noland & Polak, 2002; Hollander, 2006). Other authors have found that although applicable to private modes, one cannot readily discard the use of a direct measure of travel time variation in the case of public transportation (Bates et al., 2001). However, there appears to be no clear consensus on the appropriateness of one model over the other. Despite the fact that the "mean-variance" approach has received more attention due to the ease of obtaining adequate data compared to "schedule delay" approaches, several authors conclude that the inclusion of scheduling variables is necessary to capture the full effects of unreliability (Hollander, 2006). It is argued that the truth probably lies somewhere in between, where under certain conditions the "mean-variance" approach can be shown to be a special case of the "schedule delay" structure (Bates et al., 2001). The approach based on schedule delays is reviewed next and is used as a starting point for subsequent discussions on the role of passenger information on perceptions and the impact of unreliability on service quality.

Schedule Delay and Travel Time Variability

Scheduling considerations are taken as central to travelers' choice of departure time, taking into account the tradeoffs between the preferred arrival time (PAT) of a journey and the travel decisions (and costs) required to achieve it. The cost of unreliability can be captured through changes in the departure time used by passengers to compensate for increased travel time variability. These scheduling penalties are observed through the sub-optimal arrival time with

[4] For the remainder of this section, the term "scheduling" refers to the travel decisions of passengers related to their desired time for carrying out a given activity, and should not be confused with the term as used in transit operations.

respect to the PAT that would be required on average in order to achieve a certain level of certainty of on-time arrival. The work by Noland & Small (1995) presents one of the most complete models in terms of representing the effects of uncertain travel times on passengers. It is based on previous work by Small (1982) and other authors that earlier authors that explored the idea of a "slack" or buffer time in the form of an earlier departure, being used by passengers to counteract the effects of travel time variability (Noland & Polak, 2002; Bates et al., 2001).

Small (1982) conceived of a model for departure time choice under travel times that varied across the day, yet were deterministic. This closely resembled the effects of peak period congestion where travelers would have to trade off between travel during the most congested period in order to arrive at their destination at their PAT, or shift to an earlier or later departure times to experience lower total travel times at the cost of an early or late arrival, with the possibility of an additional fixed penalty for arriving late. In the context of utility maximization, a traveler would be expected to choose the departure time, t_h that would provide the lowest cost of making the journey. The utility obtained from a given departure time could be expressed as:

$$U(t_h) = \alpha \cdot T + \beta \cdot SDE + \chi \cdot SDL + \delta \cdot D_L \qquad [2\text{-}3]$$

Where $U(t_h)$ = utility derived from a journey with a departure time t_h,
T = total travel time for the journey for departure time t_h,
SDE = schedule delay early = Max[0, PAT-(t_h + $T(t_h)$)],
SDL = schedule delay late = Max[0, (t_h + $T(t_h)$ - PAT)], and
D_L = 1 if late arrival (i.e. (t_h + $T(t_h)$) > PAT) and 0 otherwise.

These terms simply capture the cost associated with the total journey time, the penalty for arriving before the PAT, the penalty for arriving after the PAT, and the fixed cost associated with arriving late per se, for a particular departure time. In a system with unlimited capacity (i.e. no congestion), users would be expected to select a departure time that was T units before their PAT, thus guaranteeing a minimal disutility of travel (with schedule delay equal to zero). This is unrealistic, however, as travelers often face changes in T with changes in t_h, (e.g. peak period congestion). In this case, Small (1982) hypothesized that users would, in order to achieve their PAT, tradeoff between longer travel times at the peak of the peak with changes schedule delays, depending on the relative burden of these delays the cost of being late. These expectations were confirmed by empirical results in the work by the same author. Figure 2-1 illustrates these tradeoffs by showing the level of disutility obtained for different arrival times relative to a PAT.

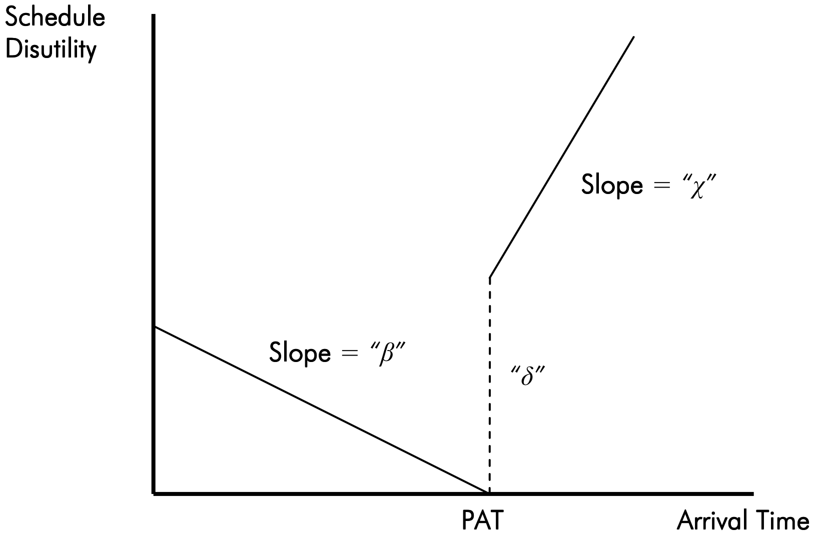


Figure 2-1: Small's schedule utility function (adapted from Bates et al., 2001)

In Figure 2-1, the relative values between the coefficients in Equation 2-3 are illustrated, where $\beta > \alpha > \chi$. From these values it is implied that travelers preferred time spent early at the destination to additional time spent traveling, but that they preferred a minute of extra travel time to a minute of late arrival, and finally (and as expected) they preferred a minute of early arrival to a minute of late arrival.

Another key contribution to the theory was made by Noland & Small's (1995) model, which included the effects of travel time variability to the schedule delay framework developed previously by Small (1982). This addition separated travel times into three components:

$$T = T_f + T_x + T_r \qquad [2\text{-}4]$$

Where T = total travel time for a particular journey,

T_f = travel time under free flow conditions (i.e. the minimum travel time),

T_x = additional travel time due to recurrent delays such as those caused by congestion in the peak period, and

T_r = travel time caused by non-recurrent delays due to factors such as incidents and obtained from a *distribution*.

This specification of T made it possible to incorporate the effect of variations in travel time into the model, in this case attributed to non-recurring delays[5] since repeated delays are foreseen by passengers and are not expected to contribute to uncertainty. Because of the probabilistic nature of the outcomes for each departure time alternative, the expected utility is maximized (MEU) for each departure time and can be expressed as in Equation 2-5:

[5] Recurring delays were assumed to be deterministic, which as will be seen in Chapter 4, does not have to be the case – there can be a recurring travel time distribution underlying the overall performance.

$\mathrm{E}[U(t_{\mathrm{h}})] = \int_0^{\infty} U(t_{\mathrm{h}})\, \mathrm{f}(T_{\mathrm{r}})\, \mathrm{d}T_{\mathrm{r}}$. [2-5]

Because $\mathrm{f}(T_{\mathrm{r}})$ is now a distribution (i.e. stochastic), all the variables in the original utility formulation become random and an expected utility can be estimated as follows:

$\mathrm{E}[U(t_{\mathrm{h}})] = \alpha^*\mathrm{E}[T] + \beta^*\mathrm{E}[SDE] + \chi^*\mathrm{E}[SDL] + \delta^*\mathrm{E}[P_{\mathrm{L}}]$, [2-6]

Where P_{L} = the probability of late arrival (replacing the deterministic binary variable).

The formulation above does not make any assumptions of the shape of the distribution $\mathrm{f}(T_{\mathrm{r}})$. However, it is possible to see how the expected schedule delay would change when $\mathrm{var}(T_{\mathrm{r}}) = 0$ (i.e. deterministic travel times), compared to one where variability is included in the time of arrival at destination. In Figure 2-2 the expected schedule delay is illustrated for a hypothetical journey assuming $T_{\mathrm{x}} = 0$ and $T_{\mathrm{f}} = 20$ minutes (a journey of at least 20 minutes of duration), for various departure times shown as minutes before the PAT.

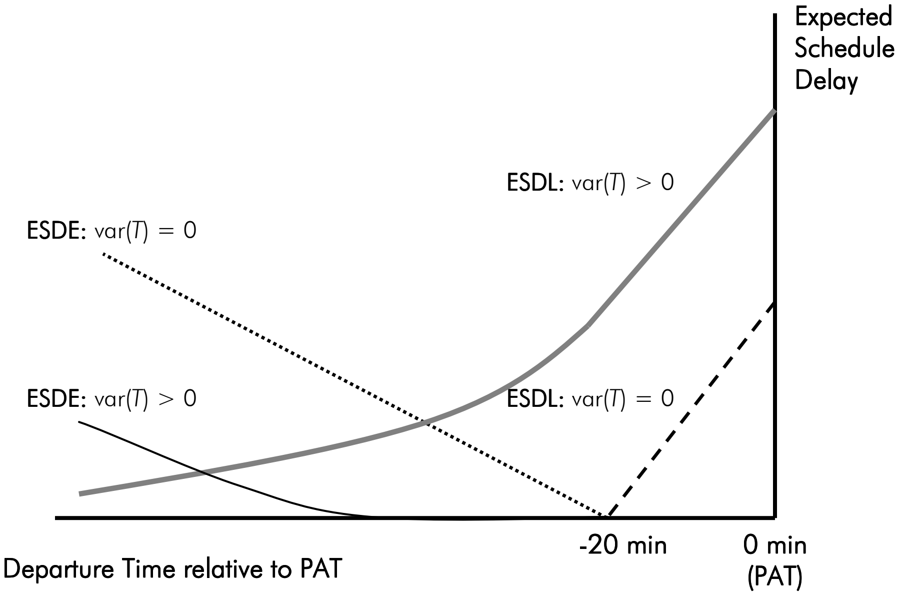


Figure 2-2: Expected schedule delay with travel time variation (adapted from Bates et al., 2001)

Figure 2-2 includes the case when travel times are deterministic, where a departure time before 20 minutes from the PAT would always lead to a late arrival. This is due to the minimum time required to make the journey, with each additional minute of late departure adding one minute to the expected schedule delay late (ESDL). Conversely, for every minute earlier in the departure time before the 20 minutes prior to the PAT, one minute of expected schedule delay early (ESDE) is added linearly. The optimal departure time would then lie exactly 20 minutes before the PAT, leading to an arrival exactly at the desired time of arrival (assuming no congestion effects, or $T_{\mathrm{x}} = 0$). When variation is added to overall travel times, however, the expected schedule delay, hence the optimal departure time, is not as straightforward to determine. For the expected schedule delay curve for stochastic travel times (i.e. $\mathrm{var}(T_{\mathrm{r}}) > 0$), every minute that a passenger departs before the 20 minute mark, rather than having a guaranteed early arrival as before, there would still be a non-zero probability of arriving after the PAT. For every minute of earlier departure, the ESDL would gradually be reduced as the probability of late arrival diminished (i.e. the tail of the travel time distribution went to infinity). In the case of early arrivals, for every minute of earlier departure before the 20 minute time mark, the ESDE would also increase similarly to the deterministic case, but at a slower rate. This is because even if the journey allows for the minimum travel time needed to complete the journey, there is still a chance of late arrival, thus reducing the ESDE compared to the deterministic case (where the chance of early arrival would be 100%). As a result, there is a range of departure times where passengers must tradeoff between the ESDE and ESDL, with both values being non-zero. This tradeoff will determine, in addition to the probability of late arrival (and the fixed penalty associated with it) and any congestion effects ignored for the moment, the optimal departure time for passengers with a PAT.

The literature points out that though it is not possible to solve this minimization (of disutility) problem analytically, several studies have made simplifying assumptions that have enabled an approximate solution (Bates et al., 2001). Of these, the model by Noland & Small (1995) assumes that the distribution T_{r} is fixed with departure time (i.e. non-recurrent delays are independent of departure time – which the literature seems to support), and that Δ, the rate of change of the profile of recurrent delays, be set to $\Delta = -T'_{\mathrm{x}} / (1+T'_{\mathrm{x}}) < 1$ to ensure first-in-first-out conditions for departure times (Bates et al., 2001). More importantly, Noland & Small's (1995) model assumes an exponential distribution with parameter "k" being the mean (and variance) for T_{r}. Using this distribution, they derive the optimal time of departure that minimizes the expected disutility of travel t^*_{h}, expressed as:

$$t^*_{\mathrm{h}} = [\mathrm{PAT} - (T_{\mathrm{f}} + T_{\mathrm{x}})] - \mathrm{k} \cdot \ln[(\delta + \mathrm{k}(\beta+\chi))/(\mathrm{k}(\beta-\alpha\Delta))]. \qquad [2\text{-}7]$$

The head-start time a passenger would leave under deterministic travel times (e.g. due to recurrent congestion) is captured by the first term on the right hand side of Equation 2-7. The second term, which is a function of the variability of the travel times, "k", captures the additional "buffer" time a passenger would leave because of the uncertainty of travel times. Inserting the optimal departure time into the generalized cost equation for passengers (see Equation 2-6), the minimum expected disutility becomes:

$$\mathrm{E}[U(t^*_{\mathrm{h}})] = \alpha(T_{\mathrm{f}} + T_{\mathrm{x}} + \mathrm{k}) + \delta P_{\mathrm{L}} + \mathrm{k}\ [\ \beta\cdot\ln[(\delta+\mathrm{k}(\beta+\chi))/(\mathrm{k}(\beta-\alpha\Delta))] - [(\delta(\beta-\alpha\Delta))/(\delta + \mathrm{k}(\beta+\chi))] - \alpha\Delta\ ]]. \qquad [2\text{-}8]$$

The first term on the right of Equation 2-8 captures the contribution of the expected travel time, with the traditional components of travel time augmented by the expected travel time from T_{r}, or the mean "k". The second component captures the disutility derived from a probability of late arrival, P_{L}. The third term captures the contribution of the expected schedule delay (early or late). The importance of this result is that the variability of travel times can have an important contribution to the disutility of travel by passengers, leading to changes in scheduling which will often result in a slack or "buffer" time being built into ones' travel budget for time, to compensate for the uncertainty of the time to complete the journey.

The connection with the "mean-variance" approach has been derived as a special case of the general framework for the scheduling delay problem (Noland & Polak, 2002). Assuming $d = \Delta = 0$, that is, that there is no fixed cost for late arrival (only a cost for the magnitude of the lateness as captured by ESDL), and that there is no change in the profile of recurrent delays across different departure times, the minimum expected disutility becomes:

$$E[U(t^*_h)] = \alpha\,(T_f + T_x + k) + k \cdot \beta \cdot \ln(1 + \chi\ /\ \beta). \qquad [2\text{-}9]$$

This leads to the expected cost being expressed as a function of both the mean travel times, but also the variance of the travel time distribution, k. This finding, however, relates specifically to the case where an exponential travel time distribution is assumed, which might not always apply. The authors point out that even though this is an important result, especially given some of the findings from other studies that support its validity, the inclusion of a variable to account for the probability of late arrival and the effects of time-varying recurrent delays add explanatory power to most of the empirical models to date and need to be included. Thus, estimating the effects of uncertain travel times on passengers indirectly through its impact on scheduling considerations has shown to be more convincing than a direct estimation of the impact of variability per se.

Extension of Framework to Public Transportation

The framework described above was developed in the context of continuous departure times, which pertain more to journeys made by car than by scheduled public transport services due to their discrete departure times or service intervals. Another complexity that can be introduced to capture the full effects of reliability on transit riders is the focus on the variability of other travel time components besides the total travel time (assumed previously), such as the wait time at the station. Passengers might also be irritated from poor punctuality, such as with infrequent travelers that could base their travel decisions on the published timetable as opposed to previous experience with the system. This requires that variables that consider the impact of schedule adherence on travel time decisions and on the disutility of making a particular journey be taken into account. Finally, the possibility of a transfer in a journey adds to the complexity of the study of the effects of travel time uncertainty, where passengers would now have to consider the variability of both legs of the journey with respect to an arrival time.

The study by Bates et al. (2001) addresses some of these concerns by showing how, at the expense of increased complexity, they can be fit into the general framework presented above. However, the work by Abkowitz et al. (1978) suggests that this added complexity might not be necessary for the case of high frequency services with travelers familiar with the system (e.g. commuters using the London Underground).

Abkowitz et al. (1978) proposed a typology, framed within the same link between travel time variability and scheduling considerations, for explaining the relationship between the arrival time at stops by travelers (related to the departure time decision) and the arrival times of the service at the transit station. The authors postulated the following three factors that would help determine the arrival patterns of passengers at stops:

- Whether it is important for travelers to reach their destination by a certain time
- Whether or not travelers are familiar with the system/have experience making that journey
- The arrival patterns of vehicles at the station, in terms of the service frequency and predictability to transit users.

Travelers concerned with on-time arrival (e.g. commuters) are hypothesized to select an arrival time at the transit station t_{total}, defined as the latest arrival time at the station which ensures that the risk of arriving late at the destination is not too great. These travelers might also be concerned with minimizing the expected time spent waiting at the station for the desired service, and could also try to arrive at the station at time t_{wait}. It is reasonable to expect travelers to trade off between these two possible arrival times at the station, with t_{total} being later than or equal to t_{wait} by definition. On the one hand, if a passenger has a preference for minimizing the mean and/or variability of the time spent waiting, she would time her arrival with the service, at the risk of possibly arriving earlier than desired at her destination. On the other hand, if a passenger is indifferent with respect to wait times, either because of personal preference (highly unlikely), or because they do not perceive the service arrivals to be predictable enough to be able to coordinate their arrivals, she will arrive at time t_{total} at the station. The latter type of behaviour would apply to cases where either vehicle arrivals at the station are unreliable, or when the service is considered high frequency and passenger arrivals are random (thus obeying considerations of total travel time more than wait time). The decision of when to arrive at the station when the two times are equivalent is trivial for passengers, producing an all-around optimal departure time. Table 2-2 summarizes this typology.

Table 2-2: Theoretical relationship between vehicle and passenger arrivals at transit stations (adapted from Abkowitz et al., 1978)

Are passengers concerned with on-time arrival?	Are passengers familiar with the system?	Service Arrival Patterns at Station	Passenger Arrival Patterns at Station
Yes	Yes	Headway < 10-min, where passengers view arrivals as unpredictable	Arrival times based solely on t_{total}
Yes	Yes	10-min < Headway < 20-min, where passengers view arrivals as predictable	Arrival times between t_{total} and t_{wait}, based on passenger preferences
Yes	Yes	Headway > 20-min	Arrival times based solely on t_{wait}
Yes	No	All Headways	Arrival times based solely on perceived value of t_{total}
No	Yes	Headway < 10-min, where passengers view arrivals as unpredictable	Arrival times generally based on considerations not related to bus level of service attributes
No	Yes	10-min < Headway < 20-min, where passengers view arrivals as predictable	Arrival times generally based on considerations not related to bus level of service attributes
No	Yes	Headway > 20-min	Arrival times based solely on t_{wait}
No	No	All Headways	Arrival times generally based on considerations not related to bus level of service attributes

The typology in Table 2-2 shows how AM Peak travelers on the London Underground would fall into the first category from top-to-bottom (see section 3.1). That is, they are likely to be concerned with on-time arrival at their destination and to be making a repeat journey. In addition, the short service intervals make it difficult and unnecessary (due to the short maximum wait time) for passengers to time their arrival at the station with a particular departure. This implies that the departure time decision of Underground passengers is likely based on considerations of the probability of late arrival at their destination. This is similar to the conditions presented in the general framework of reliability and schedule delays, where passengers' concern with total travel time variability was emphasized. Also, scheduled departures are semi-continuous in high-frequency service, reducing the need to capture the effects of discrete departure times. This shows how, through reasonable assumptions, the schedule delay framework can be used to understand the way passengers react to total travel time variability in the face of a possible late arrival at their destination.

2.2.3: Passenger Learning and Travel Time Information

The previous section explained the way passengers counteract the effects of travel time variability by modifying their departure time decision. This assumed, however, that passengers' selection of an optimal departure time was based on their perfect knowledge of the actual distribution of travel times. As Bates et al. (2001) point out, there exist important differences in the objective reliability of a system versus the subjective or perceived reliability, with the latter depending not only on personal preferences (which will be put aside for simplicity), but on the information available to passengers on travel conditions. The effects of misperceptions of travel time variability on passenger behaviour and travel disutility are reviewed next.

The Effects of Passenger Perceptions of Travel Time Variability

Assuming that there is an objective travel time distribution for a journey, T_o, that is perfectly known to a particular passenger, this individual will select the optimal departure time t^*_h depending on her value of the schedule delay, lateness, and travel time components. Figure 2-3 illustrates this optimal departure time where, for simplicity, it assumed that $d = \Delta = 0$.

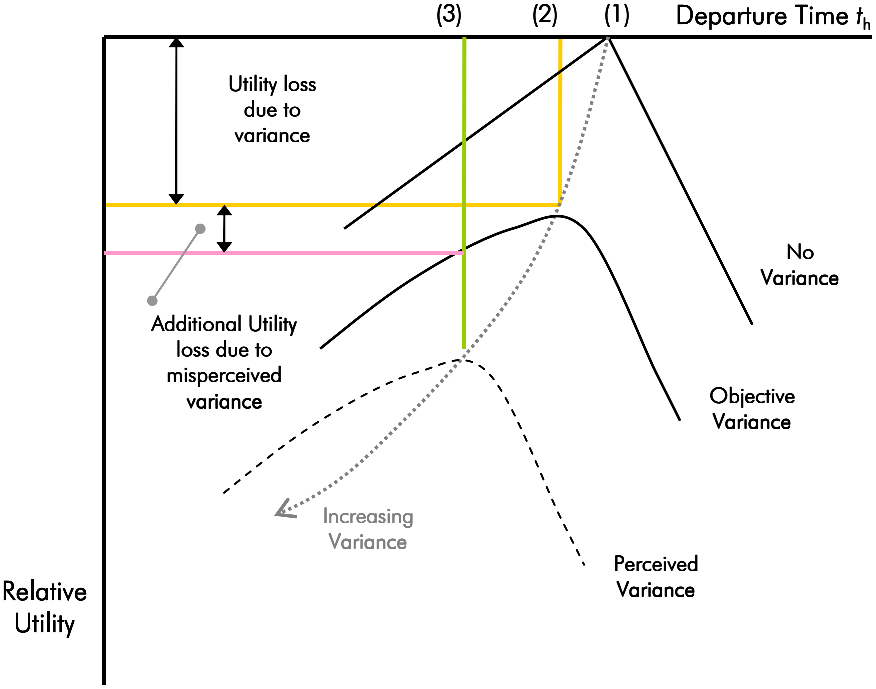


Figure 2-3: Utility loss due to misperception of variance (adapted from Bates et al., 2001)

In Figure 2-3, departure time (1) is the optimal departure time when travel times are deterministic. In this particular case, there is a linear increase in disutility with every minute away from the optimal departure time at the rate of the coefficients β and χ, for early or late departure respectively. In addition, the disutility would theoretically be zero at the point of optimal departure because of an assumed arrival to the destination at the PAT. If the same passenger were traveling under stochastic travel times, she would trade off between the probability of early and late arrival and shift her departure time earlier (because here it is assumed that a late arrival is worse than early arrival, or $\beta > \chi$, and β and $\chi < 0$). This earlier departure would cause a disutility greater than the case of zero variance even at the new optimal departure time (2). Finally, a different passenger with a subjective perception of the travel time distribution (but holding the value of schedule delay and travel times constant) would be expected to choose a different departure time, leading to a higher level of overall disutility. This is illustrated by assuming that this passenger tends to overestimate the variability of travel times, seeing the service as more unreliable than it really is. Since this passenger would base her departure time decision on her perceptions of travel time variability, she would choose the optimal departure time (3). On the one hand this means that an increased variability of travel times would theoretically lead to additional disutility through a larger safety margin (i.e. earlier departure time), with the dashed line representing the optimal departure times for increasing levels of travel time variability. On the other hand, the increase in disutility experienced by this is due to misperceptions of the true value of this variance and the variability of travel times.

This discussion illustrates the way misperceptions of travel time variability can lead to increases in the disutility of travel by hindering passengers' ability to make optimal departure time decisions. Naturally, the amount and quality of the information available to passengers when making a journey will play a crucial role in reducing these misperceptions, and improving the level of reliability experienced by passengers overall. This topic is further discussed in section 6.2.1.

2.3: Reliability and Service Quality Measurement

There is a substantial body of work devoted to the measurement of reliability from the passenger's perspective additional to studies focused on supply-side measures of performance (see section 2.1). This work is reviewed starting in section 2.3.1 with a theoretical discussion on the qualities and requirements for adequately representing passenger views on reliability. Both the seminal study by Abkowitz et al. (1978) and a synthesis of the different types of reliability measures used in other studies are covered. Section 2.3.2 reviews the work by Furth et al. (2006a) and Chan (2007), where passenger-focused reliability measures were proposed based on the schedule-delay framework, using AVL and Smart Card data, respectively.

2.3.1: Theoretical Discussion of Service Reliability Measurement

The study by Abkowitz et al. (1978) is a useful starting point because it makes recommendations on the specific form reliability measures should adopt in order to accurately reflect the passenger view, and because unlike a large proportion of the work in this area, it is focused on public transport users.

The transit reliability study begins by distinguishing between the effects that unreliability has on operators and passengers, recommending that separate measures of performance be developed for each actor. For transit passengers, the variability of total travel time, waiting time, in-vehicle time, and seat availability, were identified as the attributes of importance that need to be considered. The authors also point out, however, that the ultimate goal of these measures is to allow transit agencies to (i) identify and understand reliability problems, (ii) measure actual improvements in reliability, (iii) relate such improvements to particular strategies, and (iv) modify these strategies, methods, and designs to obtain greater reliability improvements. This simply reflects the need to develop measures that balance their representativeness of the passenger experience with their ability to provide operators with meaningful information useful for improving service quality.

One of the main concerns raised by the study was the fact that many of the existing measures of reliability measured deviations around a scheduled value, as opposed to the compactness of the attribute's distribution. The former has two weaknesses. First, these measures are susceptible to the effects of faulty schedules (e.g. a service can have a compact distribution, yet be off from the scheduled travel time). Second, these measures might not accurately represent the perspective of passengers, who in the case of repeat travelers, might be inclined to ignore schedules and go with their experience (especially if the schedules are faulty). Therefore the authors recommend the use of measures of compactness for quantifying reliability from the perspective of passengers. Note, however, that this is different from not measuring performance against an ideal standard, which is an important part of any service-quality monitoring program as identified by Wilson et al. (1992). Rather, it implies that the measured compactness of the travel time distribution could be compared with an ideal compactness, as will be the case in the reliability framework developed in Chapter 4.

Several criteria are then identified for an appropriate measure of reliability. First, the measure must reflect the concerns of passengers. For example, is it appropriate that early arrivals at the destination are valued equally to late arrivals by using the standard deviation or variance, or should the measure account for differences between these? Additionally, measures of compactness must be appropriately chosen to control for the skew that most of these travel time distributions have (presence of both minimum travel times and very long delays). Another recommended criterion is the measures' ability to compare across different services (e.g. by normalizing the measure of compactness by the mean travel time to control for differences in trip length). Finally, the authors recommend having a separate measure, independent of the measure of compactness, for capturing the likelihood of extremely long delays. This would conceptually represent for passengers the case where a change in mode or route should be considered, and for operators, the chances of a system failure.

Along this criteria, various measures of compactness are compared, leading to a recommendation that the passenger experience be captured by a measure of the mean of the distribution, a measure of compactness such as the coefficient of variation (measured after removing the most delayed journeys, or with travel times greater than 2.32 standard deviations greater than the mean), and a measure of extremely long delays, measured as the percentage of observations taking longer than the upper limit of the measure of compactness (which serves as the threshold for an "acceptable" travel time for passengers).

Three Approaches towards Measuring Reliability

There are a number of additional studies on the development of reliability measures, with a large proportion of them focused on users of private modes. However, a review of these studies finds important parallels with measures of reliability focused on transit passengers because both consider the variability of total travel time at the individual user level. A summary of the different types of reliability measures provided in these studies is provided.

The study by Lomax et al. (2003) identified three types of reliability measures: measures of statistical range or variability, measures of additional budgeted travel time, and measures of extremely long delays or "tardy trips". The first type of measure relates to the idea of the "mean-variance" approach discussed in section 2.2, and typically focuses on statistically measuring variability of travel times about some value. These measures typically include a "central" value such as the mean or median, and the range provided by measures of statistical deviation from the center of the distribution. The underlying concept behind this type of measure is that variability has an inherent disutility for passengers and excludes any effects of unreliability through schedule delays. The second type of measure is related to the "slack" or "buffer" time that is added by passengers through a shift in their departure time. This additional time can be expressed as a percentage of the average trip duration or as an absolute value additional to some expected travel time. The third type of measure relates to the concept of schedule delay, and finds the likelihood that a passenger will arrive at their destination *unacceptably* late. This is estimated by determining a threshold for what is considered to be an "unacceptable" travel time for passengers, either as a percentage of the typical travel time, or a certain fixed time in minutes. In a sense, the study by Abkowitz et al. (1978) covers the first and third type of measures, focusing on measures of compactness and extremely long delays.

Other studies such as that by De Jong et al. (2004) reviewed the different measures used in studies of the valuation of reliability, and found three distinct approaches as well. These can be summarized as measures of (i) mean-variance, (ii) percentiles, and (iii) of schedule delay. The first and second approaches fall under the first approach identified previously by Lomax et al. (2003), measuring the compactness of the distribution. The distinction between the use of percentiles and other measures of variability comes from the fact that in other studies, the median has been found to be a more effective measure than the mean, and the difference between the 90th percentile and the median to be a more effective measure of travel time variability, than traditional measures like the standard deviation (Lam & Small, 2001). Lastly, the third approach in De Jong et al. (2004) simply looks at the effects of variability on schedule delays, most often attributed to increases in the extra "buffer" time allowed by travelers to reach their destination on-time.

Each of the discussed approaches have strengths and weaknesses for quantifying reliability, leaving it up to the analyst to select the most appropriate type of measure for a specific application.

2.3.2: Review of Passenger-oriented Service Reliability Measures

The work by Furth et al. (2006a) and Chan (2007) developed passenger-focused measures under different contexts. The first study developed reliability measures for bus passengers using AVL data, focusing on the distribution of passenger wait times caused by variability in service headways. The second study proposed a measure of reliability focusing on the distribution of total travel times based on Smart Card data. This study by Chan (2007) serves as a starting point for the reliability framework proposed in Chapter 4.

Study on Service Reliability and "Hidden Waiting Time"

The study by Furth et al. (2006a) argues that the effects of unreliable (bus) service are not accurately represented by traditional measures. The authors point out that measures such as on-time performance or the coefficient of headway variation capture operational quality more than the way passengers are affected by irregular service. Additionally, the authors point out that one of the traditional measures of the service quality, mean passenger wait time (see section 2.1.2), underestimates the effects of unreliability on passengers, thus undervaluing any improvements made towards improving reliability. This is because of the idea captured in section 2.2, where unreliability is expected to force passengers to allocate an additional amount of time to complete a journey through shifts in their departure time decision. The authors argue that in the case of high-frequency bus service, this also affects the amount of wait time passengers include in their travel schedules, because if passengers only allowed for the mean wait time when making their trip, they would arrive late around half the time (ignoring any in-vehicle delay). Therefore, it is proposed that a better measure of the effects of reliability on passengers would be the 95th percentile wait time, or the amount of wait time budgeted in order to complete a journey by the desired arrival time. In addition, the difference between the budgeted wait time and the mean wait time can be considered as the *potential wait time*, or the time that would have potentially been used for waiting, except that in most cases it would be spent at the destination after an early arrival. Or, as the title of the paper suggests, potential waiting time can be thought of as a hidden cost of unreliable service.

In order to estimate the 95th percentile wait time, the passenger wait time distribution is derived from the distribution of vehicle headways assuming that passengers arrive randomly at the stop (appropriate for high-frequency service), and that passengers can board the first vehicle (no pass-ups). The passenger wait time density for an assumed uniform headway distribution U[h_{min}, h_{max}] is shown in Figure 2-4.

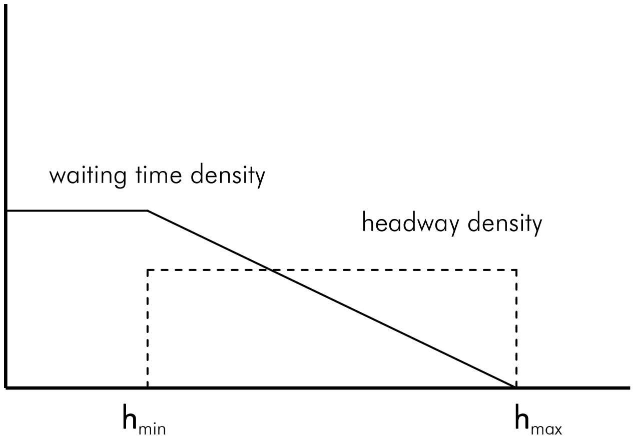


Figure 2-4: Uniform headway distribution and resulting waiting time distribution (adapted from Furth et al., 2006a)

The wait time density in Figure 2-4 is flat before the minimum headway, h_{min}, since waiting times below this value can occur with any headway length. Also, as vehicle headways gradually increase, the proportion of passengers waiting that much longer diminishes. This is due to passengers arriving at the beginning of a service interval having to wait the full amount of a longer headway (additional to the wait time for shorter headways). General formulas for the density of passenger wait times are derived, showing how, in the general case of non-uniform headway distributions, reliability measures such as the budgeted wait time (i.e. the 95th percentile wait time) can be estimated.

In addition to the effects of unreliable service on budgeted wait times, the authors also argue that changes in the coefficient of variation of headways affect the expected passenger wait time and the budgeted passenger wait time at different rates. Table 2-3 shows the theoretical effects of changes in the coefficient of variation of headways (Headway CV), assuming a normal distribution for this attribute, on the expected wait and 95th percentile wait times.

Table 2-3: Relationship between headway variation and passenger waiting time (adapted from Furth et al, 2006a)

Headway CV	0	0.15	0.25	0.35	0.45
E[wait]	0.50	0.51	0.53	0.56	0.60
95th percentile wait	0.95	1.02	1.12	1.24	1.37

For example, an increase in the Headway CV from 0 to 0.35 leads to a 30% increase in the 95th percentile passenger wait time, while the expected wait time increases by only 12%. This difference in the impact of service variability on the expected wait time and the budgeted wait time highlights the inability of traditional measures to fully capture the impact of unreliability on passengers.

In order to evaluate the benefits of reliability on passenger wait time, Furth et al (2006b) propose a set of reliability measures that separate the effects of planning and operations by comparing the actual passenger experience with a benchmark that assumes zero headway variability. The generalized cost of waiting time due to unreliability is also found by converting the value of time for potential waiting time into units of mean wait time, under the hypothesis that potential wait time has a lower value than actual wait time because it is seldom fully realized (i.e. spent as an early arrival).

This particular study illustrates how the impact of unreliability on passengers can be captured through the use of vehicle-location data. A natural extension of this approach would be to weigh performance estimates by passenger loads obtained from APC systems. Given that many agencies are not equipped with Smart Card data for tracking individual passenger movements, but might have AVL and APC systems already in place, this type of analysis could help agencies quantify the effects of unreliability more accurately.

The Journey Time Reliability Factor

Within the study by Chan (2007) on the applications of Smart Card data for rail transit systems, a passenger-focused measure of reliability was proposed and applied to the specific case of the London Underground. In her criteria for a measure of reliability, Chan identified the following:

- *Ease of interpretation*: the measure should be easy to estimate and interpret – e.g. providing results in units of minutes

- *Representative of performance*: a minimum sample size should be specified to address estimation errors caused by individual behaviour effects, leading to measures that are representative of the experience of the majority of passengers

- *Independent from schedule*: This criterion was suggested primarily because that same study proposed another measure that captured the deviation of the average journey times from the scheduled values, but also because it is important for representing the passenger experience (see section 2.3.1).

- *Comparability across Lines*: reliability should be comparable across O-D pairs with different characteristics, such as journey length and service frequencies, so that spatial aggregation would be possible (e.g. line-level measures of performance).

Based on this set of criteria, Chan proposed a measure of reliability, the Reliability Factor (RF), defined as the difference between the Nth percentile and the median travel time from the total travel time distribution for an O-D pair. The RF was essentially a measure of compactness that followed several of the recommendations discussed in section 2.3.1. Among the measures' strengths was its independence from schedules and its separation of the effects of extremely long delays through the proper setting of N. In this respect, the study concluded that in the particular case of the London Underground, the effects of extremely long travel times attributable to individual passenger behaviour than to system performance, were observed beyond the 99th percentile.

However, since the reliability factor was also conceptually the time a passenger would need to allocate in order to arrive on-time 99% of the time, the value of N was scaled back to the 95th percentile. This upper limit implied a chance of one late arrival per month, and thus seemed a reasonable representation of passenger concerns. The author also suggests that that the value of N could be redefined based on three factors: (i) the shape of the travel time distribution, (ii) the desired sensitivity of the results, and (iii) the service standards of the agency applying the measure. The RF measure is summarized by the following, noting that for sample sizes between the minimum 20 and 200, N was adjusted to control for the effects of small sample size:

$$RF_{\mathrm{OD}} = \begin{cases} 95^{th}\ Percentile\ JT_{OD} - Median\ JT_{OD}, & \text{if } SampleSize_{OD} \geq 200 \\ \mathrm{Max}(95^{th}\ Percentile\ JT_{OD}, Second\ Max\ JT_{OD}) - Median\ JT_{OD}, & \text{if } 20 \leq SampleSize_{OD} < 200 \end{cases} \qquad [2\text{-}10]$$

Where RF_{OD} = Reliability Factor for an O-D pair "OD",
JT_{OD} = the journey time from the travel time distribution for an O-D pair "OD", and
$SampleSize_{OD}$ = number of journeys for an O-D pair "OD" during the observation period.

Because one of the stated requirements for the RF was its ability to measure performance across O-D pairs with different characteristics, Chan (2007) included an initial estimation of the effects of journey length and service frequency. It was found that journeys shorter than 10 minutes and longer than 40 tended to be more unreliable (i.e. higher RF). This was explained by the higher influence on overall variability that access and egress times had in the case of shorter journeys because they represented a larger proportion of overall travel time. In the case of longer journeys, the chance of encountering larger in-vehicle delays was higher. The study also found, not surprisingly, that journeys with lower service frequencies tended to have higher Reliability Factor values. Because, unlike the non-trunk parts of the system with varying service frequencies, the trunk portions across lines were similar, and were proposed as the basis for a line-level measure. Only same-line (i.e. non-interchange) journeys were included. The formulation for the line-level measure of reliability was given by:

$$RF_{\text{LINE}} = \frac{\sum_{\text{OD in Trunk of Line}} RF_{\text{OD}} * Demand_{\text{OD}}}{\sum_{\text{OD in Trunk of Line}} Demand_{\text{OD}}} \qquad [2\text{-}11]$$

Where RF_{LINE} = Reliability Factor for a line "LINE", and
$Demand_{\text{OD}}$ = passenger demand for O-D pair "OD".

An initial application of the RF to the London Underground compared the performance of five lines for each direction, finding important directional imbalances in reliability. Also, preliminary findings suggested that both journey distance as well as the inability to board the first vehicle at crowded stations had an effect on the variability of total travel times.

2.4: Conclusions from the Literature Review

Despite the existence of a broad body of work focused on quantifying reliability from the passenger perspective, there is still considerable room for improvement in current performance monitoring approaches used in practice. As noted in section 2.1.4, many of the weaknesses of existing methods to quantify reliability come from the limited amount of disaggregate data. Several of these weaknesses have been addressed with the appearance of AVL and APC systems, but there is still room for improvement in the degree to which passenger concerns related to reliability are represented. Namely, measures currently used in practice fail to account for the effects of travel time variability on scheduling considerations and uncertainty per se, which empirical studies suggest can be as important as the contribution of average travel times to service quality.

Several alternative approaches for quantifying service reliability have been identified. These approaches differ not only in mathematical form and presentation, but also in their relative strengths and weaknesses, depending on the requirements of their specific application. Therefore, this study develops its own set of considerations when selecting and developing an appropriate family of reliability measures to be used as part of the proposed framework.

Three considerations are presented as input for subsequent chapters. First, it is clear from the literature that any measure should be based on either a "mean-variance" approach, or one related to the effects of unreliability on scheduling, or both. Second, when selecting a set of measures, it is important to take into account the context in which it will be applied. This includes sensitivity to the characteristics of the data, the type of service being offered, and any existing performance measurement systems already in use by the particular transit agency being studied. Third, the set of measures used as part of the reliability framework must balance their ability to represent the passenger point of view with their ability to provide useful information to their primary user, the transit provider, for improving overall service reliability. Because of these considerations, the work by Chan (2007) is taken as a useful starting point for the work to follow.

Chapter 3: London Public Transport and Oyster Smart Card Data

This chapter provides background information on London's public transport system and the Oyster Smart Card system in order to understand the contextual requirements for the proposed reliability measurement framework. It begins in section 3.1 with a brief description of public transportation in London and the particular transport market served by the London Underground. Section 3.2 discusses the importance of reliability as part of the city's broader transport strategy and reviews high-level indicators used to measure progress in this area. This is followed in section 3.3 by an explanation of the existing passenger-oriented performance measurement system at the London Underground, serving as the context for some of the applications developed in Chapter 6. Lastly, section 3.4 describes the Oyster Smart Card system and the characteristics of the data obtained from it, followed by a summary of the data sources and specific samples used in this research.

3.1: Public Transport in London

Since 2000, transport in the London area is organized under Transport for London (TfL), the umbrella organization charged with managing Greater London's road network and public transport modes, including London Buses, the London Underground (since 2003), Croydon Tramlink, the London Overground (since 2007), the Docklands Light Rail (DLR), and the London River services (GLA Transport Home Page, 1999). TfL is controlled by the mayor, who in turn is in responsible for setting TfL's budget and fares and the selection of new projects. The mayor's office also determines the long-term vision for the city, which TfL is responsible to support through its transport strategy.

As of January 2007, public transport mode share in London enjoyed its highest levels since 1993, with a continuous decrease in the percentage of journeys being made by private modes (TfL, 2007b). Within public transport, the largest proportion of the average number of daily journeys was made by bus, fueled in recent years by a combination of reduced fares and improved services. Journeys on the Underground made up over a quarter of all public transport journeys, closely followed by trips on National Rail services serving the greater London area (see Figure 3-1).

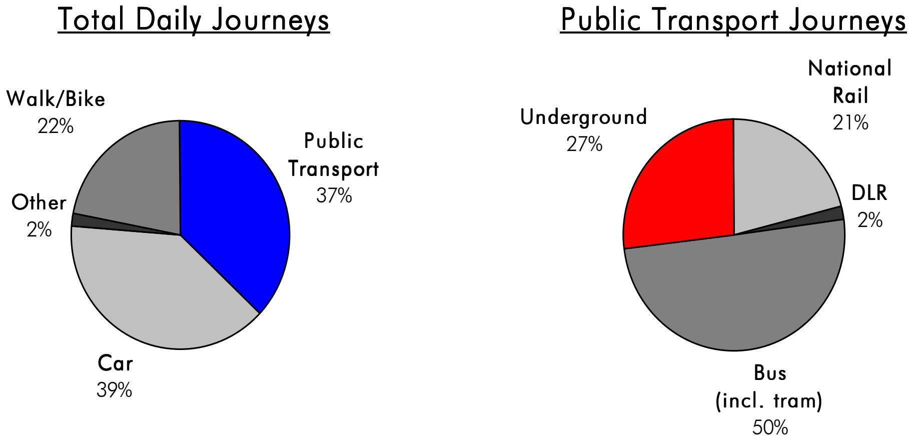
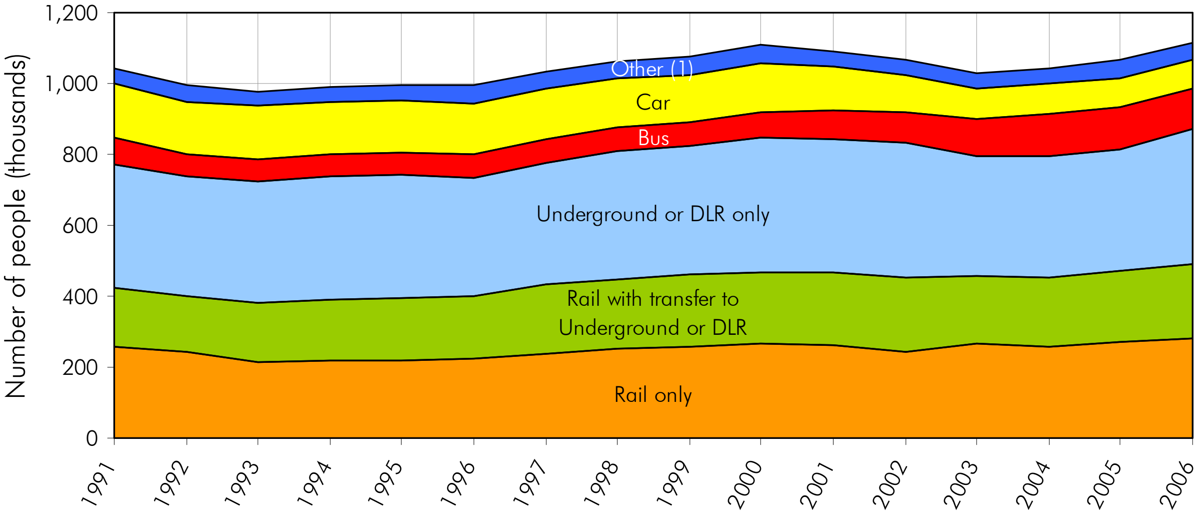


Figure 3-1: Modal split for daily journeys in London (adapted from TfL, 2007b)[6]

The proportion of journeys served by each mode varies greatly by location within the city. Public transport trips tended to dominate travel to and from the central areas in 2007, with a decreasing influence towards the periphery. In particular, over 89% of all journeys into Central London, roughly represented as fare zone 1 in Appendix A, are made by public transport. Travel on Bus dominates the proportion of journeys within outer areas, where over a third of the journeys made on this mode start and end within fare zones 4, 5, and 6.

Time-of-day also had an important effect on the distribution of travel across public transport modes. In the case of the Underground, only 10% of all journeys were carried by this mode throughout Greater London. During the morning peak, however, over half (53%) of all trips into Central London used this mode, highlighting the important role it plays in serving commute trips. Figure 3-2 shows the number of people entering Central London during the morning peak yearly by mode since 1991.

[6] London Overground did not start service until the end of 2007, and is therefore not included.



1. Includes coach/minibus, taxi, two wheeled motor vehicle and cycle.

Figure 3-2: Passengers entering Central London during the AM Peak by mode since 1991 (adapted from TfL, 2007b)

In addition, the majority of journeys served by the Underground were commute trips during the peak hours of the day. This is reflected in Figure 3-3, and by the fact that over 69% of journeys served by this mode were work or education related in 2007.

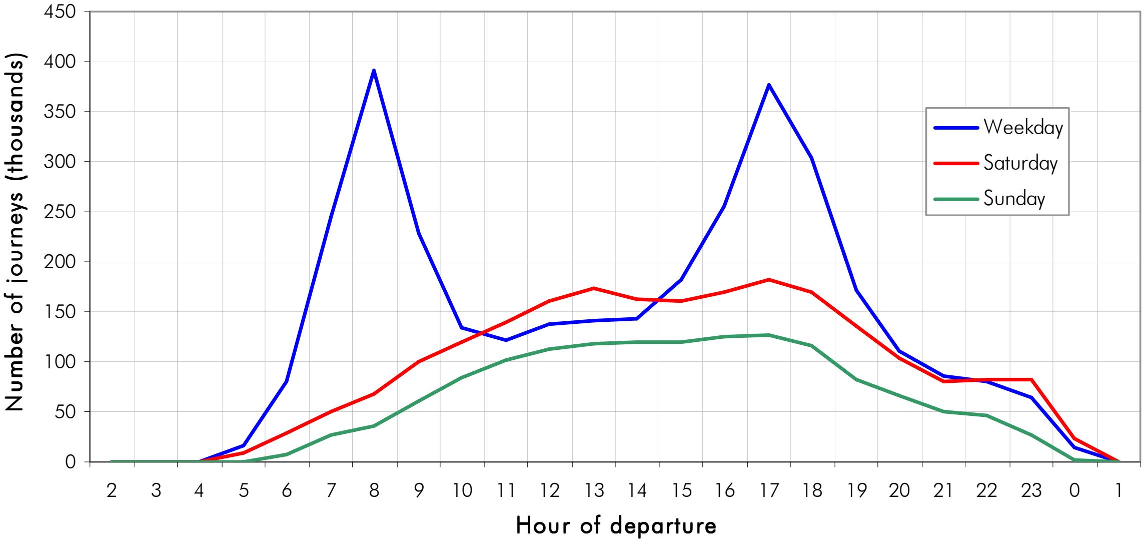


Figure 3-3: Weekday and Weekend passenger entries into the London Underground by time-of-day – 2005/6 (adapted from TfL, 2007b)

As a consequence, a large proportion of the passenger market served by the London Underground can be expected to be concerned with on-time arrival at their destination and with the reliability of the service. These considerations, together with on-going work at TfL to improve service quality, make the London Underground an ideal setting for this research.

3.2: Reliability and London Transport Strategy

In support of the mayor's London Plan, TfL is charged with carrying out the long-term transport plan, Transport 2025, which outlines the agency's role in achieving three overarching goals. These three goals are (TfL, 2006):

- *Support economic development and growth*: achieved through increases in capacity and the enhancement of service quality through refurbishment of infrastructure and better management of the transport network

- *Tackle climate change and enhance the environment*: pursued through reductions in carbon dioxide emissions from a shift in travel to more environmentally efficient modes

- *Improve social inclusion*: attained by improving access by all to the various opportunities the city has to offer, including economically disadvantaged groups and those physically impaired

In December 2007, one year after Transport 2025 was published, TfL put forth its updated business plan outlining the specific actions the agency would take by 2010 in support of the mayor's vision. The business plan included information on the financial aspects of delivering on-going projects, and specified concrete objectives to support the three overarching goals established by the mayor. Progress towards these objectives would be evaluated based on a set of Key Performance Indicators (KPI), estimated from more disaggregate measures of transport performance. These objectives included (TfL, 2007a):

- Improve door-to-door journey time and reliability across all modes
- Ensure that the movement of freight and services within London is efficient and reliable
- Reduce CO_2 emissions from ground transport and improve the energy efficiency of its operators
- Improve the local environment and urban realm
- Influence a shift towards more sustainable modes of transport
- Support sustainable growth and regeneration
- Improve the economic, physical and spatial accessibility of transport networks
- Operate a safe and secure transport network
- Engage Londoners in the delivery of TfL's plans and provide timely and relevant information
- Deliver value for money

The first objective called for improvements in journey time reliability as a central part of TfL's transport strategy. More specifically, the business plan stated that this objective would be achieved "by reducing the time taken to travel across the transport network and improving the consistency and predictability of the expected time for those journeys." This understanding of reliability clearly alludes to the definition of reliability from the previous chapter, providing further motivation for this research. Currently, the first objective is captured by a set of indicators, as discussed further below.

Key Performance Indicators

The KPI corresponding to each of the ten key objectives in the TfL Business Plan are used to evaluate performance at the system level, and are therefore derived from more disaggregate performance measures. These indicators are also typically broken down by mode, reflecting the relative performance of each element of the transport system with respect to the objectives. In particular, the four Key Performance Indicators that correspond to the first objective as applied to the London Underground are summarized in Table 3-1.

Table 3-1: Key Performance Indicators for Objective # 1 – London Underground (adapted from TfL, 2007a & Wainberg, 2008)

Performance Indicator	Description [units]	Measurement
Weighted Excess Total Journey Time	The time to complete an average journey over and above the expected time, weighted by customer values of time for each component of the journey [minutes].	This measure is derived from the system level results of the Journey Time Metric system. It is estimated from surveys, simulation models, and train movement data, which allow for the estimation of the travel times of each component of a journey.
% of Peak Cancellations due to Operator Not Available (ONA)	The number of scheduled trains that are cancelled at the key measurement times (Mon-Fri, 9:00-18:00) due to no operator available, as a percentage of the total expected trains in service [%].	This is estimated from the CUPID system, logging the number of missed runs and their causes
% Scheduled Kms Operated[7]	The number of scheduled train kilometers operated for customer service as a percentage of the scheduled train kilometers [%].	Data is taken from both line controller's failure and delay summaries (actual train Kms) and from working timetables (scheduled train Kms).
PPP Availability – Lost Customer Hours	The number of lost customer hours due to incidents attributed to infrastructure problems (InfraCo's) [1,000's hours].	Estimated from the CUPID (incident log) database using NACHs tables by time of day, location, etc. (see section 3.3.1).

[7] This is a measure of service provision more than of service reliability as the others are, but is included because it is related to the first objective in the TfL Business Plan.

Of the four measures used for quantifying improvements in journey time and reliability, the second and third capture the availability of service as opposed to focusing on service quality directly. The fourth measure, PPP Availability – Lost Customer Hours, takes a step further by estimating the impact of unavailable service on passengers in terms of delays. The key performance indicator that directly attempts to estimate improvements in reliability as it relates to passenger valuations of travel time is the Weighted Excess Total Journey Time. This measure estimates the level of average delays passengers experience as the difference between actual average journey times and their scheduled average. However, it focuses primarily on average travel time delays, leaving room for improvement in its ability to represent the effects of reliability on passengers.

The next section presents an overview of the existing passenger-oriented performance measurement system at the Underground, which serves as the primary input for the overall Weighted Excess Total Journey Time KPI.

3.3: Passenger-oriented Performance Measurement System

In 1997 the London Underground Marketing and Planning department introduced the Journey Time Metric (JTM) as a primary means of assessing service performance from the passenger's perspective. JTM results are published at the end of every four-week period and are used by line managers to evaluate the performance of the part of the system under their control, in addition to its use by TfL as a primary input for some of their KPI (see section 3.2).

This section reviews the main aspects of JTM and identifies some of the requirements for the reliability extension of the metric proposed in Chapter 6. The definitions and estimation methodology used in JTM are introduced in section 3.3.1, followed by a discussion in section 3.3.2 of the strengths and weaknesses of the approach.

3.3.1: Journey Time Metric Definitions and Methodology

The Journey Time Metric compares the estimated actual journey times experienced by passengers during a particular four-week revenue accounting period with the scheduled travel time for that same journey. The difference is the Excess Journey Time, and it is used as an indicator for how well the system performed from the passenger's perspective compared with everything going according to plan. The polarity of this measure is negative, meaning that a higher Excess Journey Time is associated with worsening performance.

Journey times are broken down into five components, each reflecting a different aspect of the journey. For each component there is an actual and a scheduled value. These values can be affected by both capital improvements such as station refurbishment, and non-capital initiatives such as tighter dwell time management at stations that could impact the observed performance of the system, and possibly in the case of long-term sustained improvements, the scheduled values. Each component is estimated differently, using various data sources ranging from manual surveys to data obtained through ADC systems.

Journey Time Components

Under JTM, the total duration of a journey is broken down into five components. The definition and estimation procedure for each are summarized below, as well as their typical contribution to overall journey time[8].

1. Ticket Purchase Time (TPT)

TPT represents both the time in queue waiting to purchase a ticket and the time to complete the transaction at the ticket office window or automated ticket machines. Scheduled values are derived from the previous years' observed transaction time (e.g. 90% of previous year). This component typically represents around 1% of the total journey time, and is estimated through manual surveys.

2. Access, Egress, and Interchange Time (AEI)

AEI captures the time required to traverse a station when entering or leaving the Underground or when transferring between services. Access and Egress times are defined as the time taken to walk from the station entrance to the mid-point of the platform and vice versa. Interchange (transfer) times are measured as the time to walk between the mid-points of the two platforms within the same station. Actual travel times are measured using a combination of periodic surveys (around 12 per four-week period) for 27 stations representing nearly half of the system's demand, and pedestrian path modeling for the remaining stations. Ideal AEI times are defined for the purposes of JTM as the time needed to walk through a station under free-flow conditions, and deviations from them are attributed to three factors: train service regularity, the proper functioning of lifts and escalators, and increases in demand. This component of the journey represents around 21% of the total travel time for a typical journey, which explains the large effort devoted to its measurement.

3. Platform Wait Time (PWT)

This part of the journey is defined as the time from a passenger's arrival at the mid-point of a platform to the time of departure (i.e. wheel start) of the boarded train. In its estimation, passengers are assumed to board the first train that goes to their destination, or when there is infrequent direct service, they are assumed to board the first train and interchange at a later point. Scheduled values for passenger wait time are taken to be half the headway, where perfectly regular service and the ability of all passengers to board the first service are assumed. Estimation of the actual time spent waiting on the platform is derived from data on service headways, estimated from line signaling data for all of the lines except the District (where time on-platform surveys (TOPS) are used instead). This estimation takes into account the effects of

[8] Results taken from JTM Period 13 – 2007/08 Actual (unweighted) Journey Time data at the Network level (results may vary over time and level of aggregation).

both the regularity of the service, as well as the throughput provided (% of scheduled service) through the following relationship for the average platform wait time:

$$\overline{PWT} = \frac{\sum H^2}{2 \cdot \sum H} \tag{3-1}$$

Where $\overline{PWT}$ = average passenger platform wait time, and
H = observed individual vehicle headway.

In the case where passengers are unable to board the first vehicle departure due to on-train crowding (defined as Left Behinds in JTM), an additional platform wait time is estimated and assigned based on three factors: link loads, gate counts, and the regularity of service intervals. The scheduled value for Left Behinds is zero. For a typical journey, this component represents around 13% of total time spent travelling, and is considered to be one of the most onerous to passengers in terms of their value of time (see Table 3-2).

4. On-train Time (OTT)

The time spent on a train is measured from the start of a train's movement to the opening of doors at the destination. Similar to the estimation of PWT, data for this component is derived from the train signaling systems where available, and for the District and parts of the Hammersmith, Metropolitan and Circle lines, signal cabin box sheet data are used. The scheduled component of OTT is derived from the working timetables. This aspect of the journey can be one of the largest components of total travel time, representing over half (64%) of a typical journey.

5. Closures

Closures, which contribute around 1% to the average actual journey time over the network (see Table 3-2), are categorized by their level of duration and predictability. Three types of closures are estimated as part of JTM, with their impact subdivided at the station or line level:

- *Unplanned (short-term) Closures and Service Disruptions*: These closures capture the effect of service disruptions that have a duration of 30 minutes or more, beyond the scheduled headway (e.g. due to incidents). The nature of these disruptions is unpredictable, and therefore their impact on customer delays is estimated after the occurrence using the Nominally Accumulated Customer Hours (NACHs) system. This system uses equivalency tables to estimate, for a specific location and time where the closure or disruption occurred, the additional travel time incurred by passengers in units of hundreds of customer hours of delay (NAX). This type of closure makes up about 18% of the delays produced by all non-scheduled closures.

- *Planned Closures*: Planned closures differ from the previous type in that they are known sufficiently far in advance that passengers can be informed, and remedial action can be taken (e.g. rail-replacement bus services). However, they are still short-term and thus their impact on passenger travel times is still considered a delay above the scheduled time and is also estimated through NACHs. These represent the majority of non-scheduled closures (82%).

- *Scheduled (long-term) Closures*: These closures differ from Planned closures in that they are usually of longer duration (around 4 weeks or more), and are known well in advance. In case of a scheduled closure, passengers must be notified at least 2 weeks before it takes place, allowing them to modify their travel plans for the duration of the disruption. Because of this, the impact of scheduled closures is not considered to be excess journey time. Instead, its impact is estimated and added to the scheduled journey time.

Weighted and Unweighted Excess Journey Time

In order to better characterize service performance from a passenger's perspective, JTM weighs each journey component by value of time (VOT) in units equivalent to on-train travel time under uncrowded conditions (i.e. VOT_{OTT} = 1).

The values of time differ not only across the five main components, but also between the different parts of each component. For example, within the AEI component, the VOT of walking horizontally is half that of walking *up* stairs, which is itself more onerous that walking *down* stairs. Delays from both planned and unplanned closures are weighted by the same VOT at the station and line levels, and are weighted at twice the value of uncrowded on-train times.

The weights are applied both to actual and scheduled journey time components to make them comparable for estimating a weighted Excess Journey Time. In the case of the OTT, a crowding penalty is estimated from the proportion of Left Behinds, and is used to increase the VOT for passengers from 1 to 2.48. In addition, the scheduled time for this component includes the level of crowding expected when service runs perfectly to timetable. Table 3-2 summarizes the proportion of the total *perceived* journey time changes for a typical journey when travel time components, and more specifically their individual parts, are weighted by their respective VOT.

Table 3-2: Proportion of total actual journey time for each component: weighted & unweighted by VOT

Journey Time Component	Value of Time Weighting [range]	Percentage of Actual Typical Journey (Unweighted)	Percentage of Actual Typical Journey (Weighted)
Ticket Purchase Time	2.5-3.4	1%	2%
Access, Egress, and Interchange	2.0-4.0	21%	31%
Platform Wait Time (Left-Behinds)	2.0 (3.0)	13%	16%
On-Train Time (as a function of crowding)	1.0-2.48	64%	49%
Closures	2.0	1%	2%

Table 3-2 shows that when the actual journey time components are weighted to represent the generalized cost of passengers in units of uncrowded on-train travel time, the proportion of time spent during each phase of the trip changes, with AEI and OTT having the largest increases and decreases, respectively.

Estimation Methodology

The amount of Excess Journey Time is estimated at various levels of aggregation, leading to a Network level measure of system performance that is used as an input for the primary KPI for journey time and reliability.

For each of the five individual components of a journey, an actual value of travel time is compared to a scheduled or "ideal" travel time. The actual travel time, however, is estimated at the line section level, where links with similar service frequencies are grouped together to estimate a platform wait and on-train time, with station-related components (e.g. AEI, TPT) naturally measured at the station level. Both the actual and scheduled journey times (i.e. excess journey times) are then weighted by passenger volumes to find an average passenger excess journey time at the line level. The performance of stations that allow for an interchange between lines is attributed to the line that is responsible for the station's management. Using the results for each component at the line level, it is possible to aggregate even further and estimate the performance of the overall Network in the form of a weighted (by line volumes) average. However, because the line-level results only capture the experience of same-line journeys, the network level OTT and PWT components are scaled up by a Journey Leg Factor (JLF) of 1.39. This is to account for the fact that around 40% of the journeys in the system involve an interchange, and are expected to experience at least one additional wait time and in-vehicle time component. Finally, the impact of Closures and Incidents is then added onto the scaled up travel time values to produce the final Network Excess Journey Time, both for weighted and unweighted (by VOT) Excess Journey Time.

An important characteristic of JTM is the use of a static demand matrix when weighing passenger journey times by volume to separate between the effects of changes in service quality and changes in passenger volumes on the Excess Journey Time. This way, any increases in average journey times are due to changes in service performance and not because of increased demand along a particular corridor. At the end of each four-week period, changes in each of the journey time components and their causes are reported.

3.3.2: Strengths and Weaknesses of JTM

In her work, Chan (2007) identified the main strengths and areas for improvement of JTM.

Strengths of Current Approach

- JTM is **an effective management tool** through its breakdown of passenger journeys into their different components and its easy-to-interpret summary statistics aggregated at the line level, allow for diagnosing the performance of the Underground and attributing changes to the relevant line and station managers.

- The current approach **captures performance from the perspective of passengers** by representing not only the entire duration of a journey, including ticket purchase time and the effects of incidents and closures, but also the value of time placed by passengers on each of the journey components. Additionally, by weighing journey times by demand, the JTM emphasizes performance at key stations and for those origin-destination pairs carrying higher passenger volumes.

- Two important JTM components are **estimated using reliable data sources.** Namely On-Train Time and Platform Wait Time are estimated from train run time data (100% sample except for District line), improving the accuracy of the results.

Weaknesses of Current Approach

- The **attribution of station performance to a single line** can lead to unfair comparisons between lines. This convention, intended to simplify the responsibility for changes in station performance to a single line manager, can bias performance against parts of the network with a large proportion of highly congested stations. Chan (2007) cites the example of Oxford Circus station, which serves both the Victoria and Central lines, but is under Bakerloo line management, therefore contributing only to the latter's Excess Journey Time through its AEI.

- JTM's **reliance on a fixed demand matrix**[9] can be a potential source for misrepresentation of the current passenger experience as travel patterns change over time (e.g. the Jubilee line extension). This can lead not only to the inaccurate weighting of performance by demand, but to outdated estimates of link loads and crowding levels. Additionally, the weekend demand matrix is based on weekday estimates, leading to further discrepancies between estimates and the actual passenger experience.

- The use of manual surveys for estimating AEI and TPT times leads to **costly data collection**, as well as possible estimation errors derived from small sample sizes (as compared to the volume of data available through ADC systems).

Through her proposal for an Oyster-based Excess Journey Time metric, Chan (2007) addressed some of the weaknesses of the existing system, while preserving its main strengths. In particular, the use of Oyster Smart Card data, which captures journey duration from gate entry until gate exit for approximately 70% of the journeys using the Underground as of March 2007, provides a cost-effective way of capturing what previously required various sources of data. Other aspects of the proposed Oyster-based metric improved upon some of JTM's weaknesses, such as through the redefinition of scheduled platform wait time to be the full scheduled headway (as opposed to half of the scheduled headway), in order to take into

[9] Referring to the O-D Matrix obtained from the Rolling Origin-Destination Survey (RODS) estimated in 2002.

account the longer waiting times occurring under perfectly reliable service. One of the main contributions of Chan, however, was to point out that the current methodology for measuring journey times as experienced by passengers did not adequately capture the reliability of the service, or more specifically, the variability of journey times (see section 2.3.2).

Reliability and the Journey Time Metric

An important weakness JTM that this research attempts to improve is the degree to which the impact of unreliability on passengers is captured. Currently, the main mechanism through which unreliability effects are represented is in passenger wait time estimates. That is, as vehicle arrivals increase in variability, so does the average passenger wait time. As discussed in section 2.3.2, there is a "hidden cost" of travel time variability that is not captured by focusing on average performance, leading to an underestimation of the true effects of unreliability on passengers. This problem was initially raised by Chan (2007) and is the motivation for the first application of the reliability framework in section 6.1, where a reliability extension to JTM is proposed.

3.4: Automated Fare Data and Oyster Smart Cards

Since its appearance during the early 1990's, Smart Card technology has been increasingly adopted by transit agencies given the benefits it provides above existing magnetic stripe cards and non-automated fare collection methods (Multisystems Inc., 2003). One advantage comes from the ability to have complex fare structures that support improved fare strategies (e.g. zonal charges, transfer pricing and policy, and peak/off-peak differentials), payment options (e.g. period pass and stored value), and pricing options (e.g. discounts for prepaid alternatives). Transport for London has reaped some of these benefits with the introduction of its Oyster Smart Card.

Another important advantage is the large amount of disaggregate data on individual passenger journeys produced by Smart Cards that can be used for a variety of applications, including service quality measurement as explored in this research. This section describes the Oyster Smart Card system, starting in section 3.4.1 with an overview of the current fare structure and penetration rates of different ticket types. Section 3.4.2 goes on to discuss the characteristics of Oyster Smart Card data as with regards to their use for measuring service quality. The section ends in 3.4.3 with a summary of the specific data obtained from the Oyster system, as well as other data sources used for this study.

3.4.1: London Fare Structure and Smart Card Usage

The fare structure in London is distance-based with travel divided into six circumferential fare zones (see Appendix A). It has become increasingly more sophisticated since the introduction of the Oyster contactless Smart Card in 2003. Though initially only available to the public on Annual Travelcards, the incremental transition of the remaining travel passes and individual trip tickets from magnetic stripe cards to Oyster, as well as changes in fares favoring use of the Smart Card, has led to a continuous increase in this fare media's trip penetration rate. As of 2007, over two-thirds of journeys on the Underground were made using Oyster, with increases in this proportion expected to continue as National Rail services are integrated into the system. Figure 3-4 reflects this trend by plotting the proportion of the various ticket types sold at Underground stations over time.

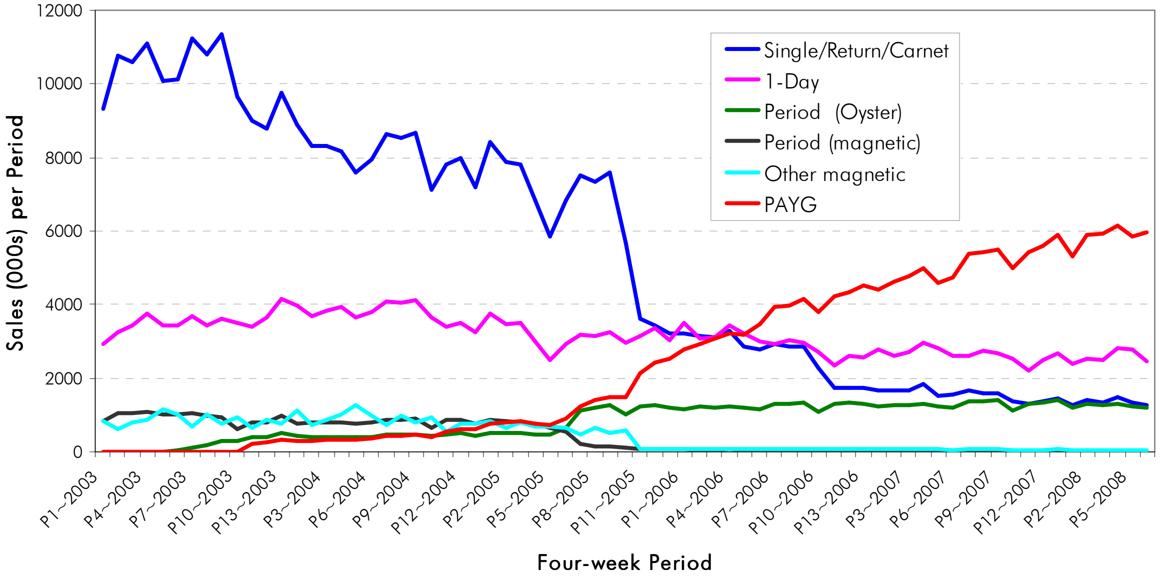
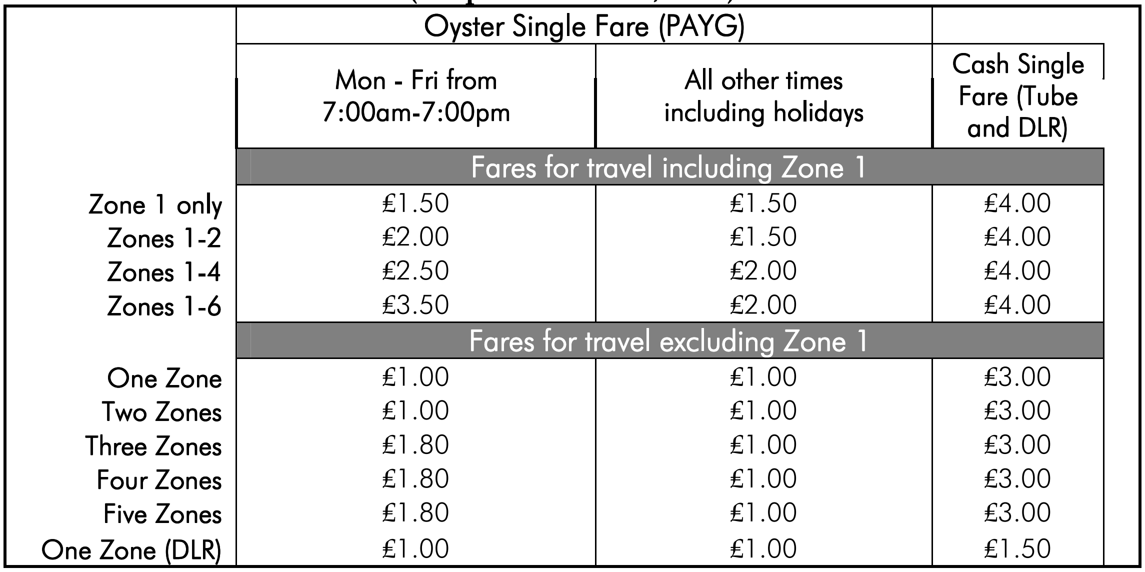


Figure 3-4: Ticket sales at Underground stations by type of fare media (adapted from Wood, 2008)

The introduction of a stored value or Pay-As-You-Go (PAYG) ticket in 2004, which can be used as the primary payment option, or in parallel with the time-based Travelcard passes, was an important milestone for the Oyster system. This combination allows users who purchase a Travelcard to travel outside their normal zone with the same card and simply pay for the journey through their card's stored value. Another key strategy has been price differentiation, where travelers are currently given a steep discount on the cost of a single journey when using Oyster PAYG compared to purchasing a single ride on magnetic tickets. Other strategies that also segment the cost of travel have included the introduction of Off-peak fares in 2004, the availability of a Bus-only travel pass, as well as more complex mechanisms such as "daily capping" and "maximum fare". The former simply ensures that the cost of travel on Oyster PAYG for a day is less expensive than the cost of a daily paper Travelcard pass. This upper bound is different depending on the combinations of travel mode, time of day and day of week. The latter, Maximum Fare, was introduced in late 2006 to discourage passengers from not validating upon either entry or exit (such as when arriving at non-fully gated station in the outer zones of London). To achieve this, passengers are charged the maximum cash fare when starting a journey at select stations (e.g. £4-5), and are then refunded the overcharge when they validate upon exit. Table 3-3 illustrates the different fare options for travel under various Transport for London modes.

Table 3-3: Fares in 2008 for London Underground, DLR, and Overground Services (adapted from TfL, 2008)



	Oyster Single Fare (PAYG)		
	Mon - Fri from 7:00am-7:00pm	All other times including holidays	Cash Single Fare (Tube and DLR)
	Fares for travel including Zone 1		
Zone 1 only	£1.50	£1.50	£4.00
Zones 1-2	£2.00	£1.50	£4.00
Zones 1-4	£2.50	£2.00	£4.00
Zones 1-6	£3.50	£2.00	£4.00
	Fares for travel excluding Zone 1		
One Zone	£1.00	£1.00	£3.00
Two Zones	£1.00	£1.00	£3.00
Three Zones	£1.80	£1.00	£3.00
Four Zones	£1.80	£1.00	£3.00
Five Zones	£1.80	£1.00	£3.00
One Zone (DLR)	£1.00	£1.00	£1.50

Presently, Oyster card access is valid for all modes under TfL management, including the Underground, Bus, Tram, Overground, and an increasing proportion of stations within the National Rail network.

3.4.2: Characteristics of Oyster Smart Card Data

One advantage of Smart Card technology is the wealth of disaggregated data that it produces. This data has already been used to develop applications in several areas, including fares and ticketing, planning, and management (for an example of O-D Matrix estimation using Smart Card data, see Chan, 2007). The application developed in this research is within the area of performance measurement, and is centered on the ability to estimate a large number of individual passenger journey times from the data. This ability to use Oyster data to measure reliability is affected by both the type of information obtained from it, itself dependent on the "format" of the data, as well as its penetration rate.

Oyster Data Format and Reliability Measurement

The format of the data obtained from Oyster includes both the specific fields of information collected as well as their level of resolution. In order to obtain individual passenger journey times, at least five fields of information are required. The first is a unique card identifier, which makes it possible not only to link the various journeys made by a single passenger, but more importantly to match an individual's entry and exit transactions. The second and third fields record the location of these entry and exit transactions, where in the case of the Underground represent individual gatelines at stations. This information can be useful when trying to determine the mode taken by a passenger in cases where more than one type of service is available (e.g. Underground, National Rail, and DLR). The last two required fields are naturally the times that a passenger enters and exits the system. The resolution of this information is a key determinant of the level of analysis and accuracy possible from the data.

In the case of Oyster data, the time of entry and exit is recorded at the whole minute level (e.g. 7:30am as opposed to 7:30:42am). This high level of disaggregation makes it possible to construct travel time distributions for a particular O-D pair during very short time intervals (e.g. for a particular 15 minutes within the day). It also, however, implies a maximum level of accuracy for estimates of total travel time. As Chan (2007) pointed out, because time of entry and exit are measured at the minute level, there is a margin for error of +/- 59 seconds. Figure 3-5 illustrates this margin of error in the passenger journey times derived from Oyster data.

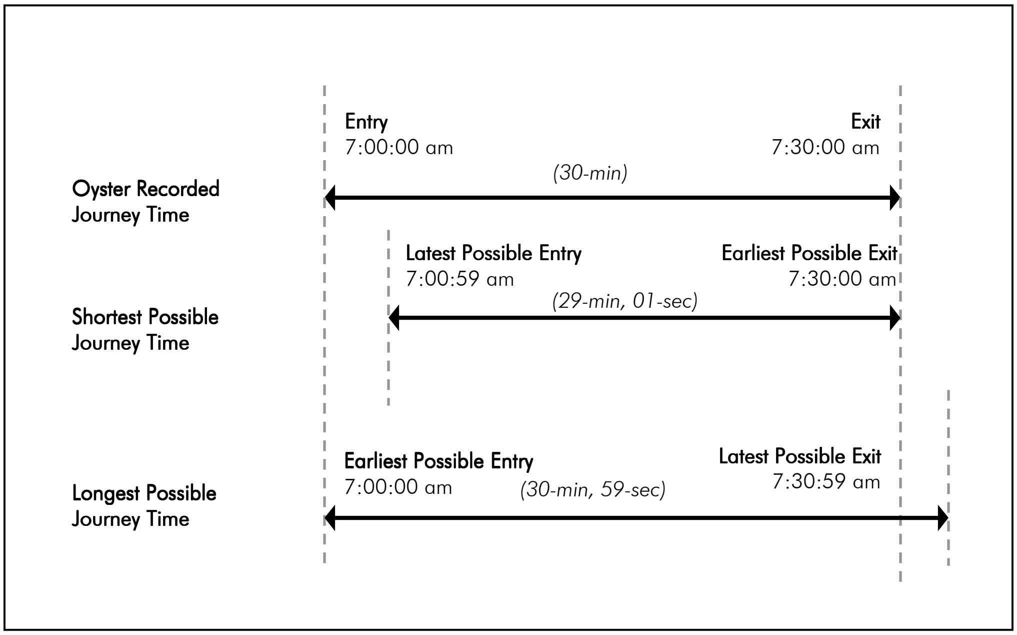


Figure 3-5: Oyster journey time discrepancy (adapted from Chan, 2007)

The example in Figure 3-5 shows how for entry and exit transactions recorded as 7:00am and ending at 7:30am, respectively, there is a shortest possible actual journey time of 29 minutes and 01 seconds, and a longest possible actual journey time of 30 minutes and 59 seconds. This range is due to the truncation of the seconds in the transaction times, and could be ameliorated through their rounding.

Other sources of discrepancy can appear when estimating statistics to describe the travel time distribution for a particular O-D pair. In this research, discrete percentiles are used to construct the proposed set of reliability measures and the resulting margin of error is identified.

The definition of a percentile used in this study is the lowest value that is greater than *or equal to* a certain percentage of the observations in a sample[10]. In some cases, a percentile value might not be represented by one of the values in the sample, in which case a continuous

[10] This definition reflects that of the software used to estimate them: SQL Developer v. 1.2.1 (2007), which uses the same algorithm for estimating continuous percentiles as Microsoft Excel.

percentile is estimated. A simple illustration of this is with the set of numbers {1,2,3,4,5,6}, where the 50th percentile would be 3.5. In the case where discrete percentiles is desired, the first value with 50% *or more* of the observations beneath it would be selected (in this example the number 4). In the particular case where the quotient of the desired percentile and the sample size is not a whole number, this would lead to a small margin of error representative of the difference between the continuous and discrete percentile values for a particular distribution. The size of the error would depend on whether there exist consecutive observations surrounding the value returned for a particular percentile. Figure 3-6 provides an example of this, comparing two travel time distributions that report the 95th percentile as being 18 minutes (because it is the travel time value that ensures *at least* 95% of all observations are below it). In the case where there are observations with a value of 17 minutes (1 minute before the reported discrete percentile), the maximum overestimation of the continuous percentile would be 1 minute. However, in the case where observations are not consecutive, or the previous observation is found at 16 minutes, the margin of error has a maximum of 2 minutes, and so forth.

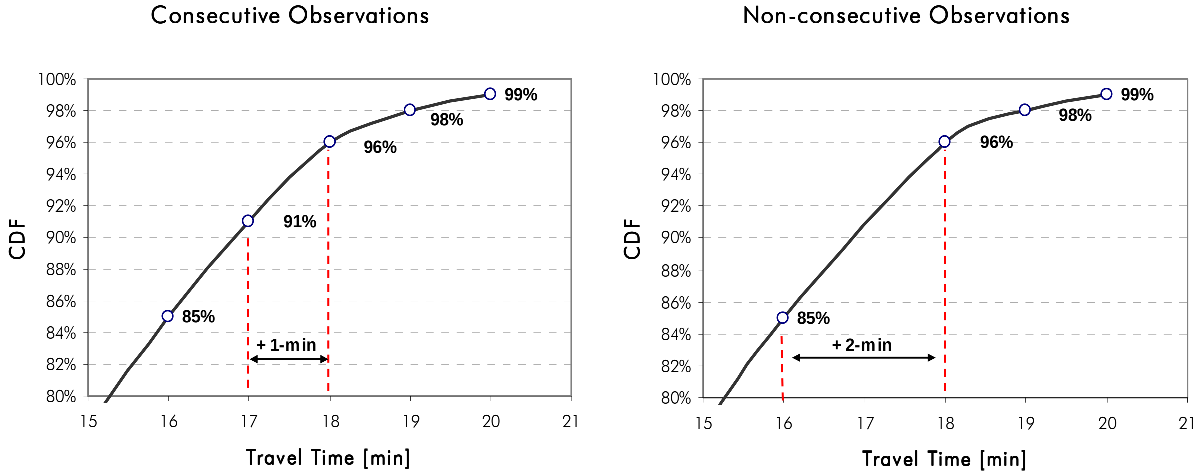


Figure 3-6: Margin of error for percentile estimation from discrete travel time data

The probability of having consecutive observations surrounding the estimated percentile will depend on two things: the percentile being estimated and the size of the sample used for constructing the discrete travel time distribution. The former has an impact because as higher percentiles are estimated (e.g. 90th as opposed to the 50th), it is more likely that "jumps" caused by outliers or values in the tail of the distribution will be found. The latter affects this probability because as the size of the sample gets smaller, each observation represents a greater proportion of all the observations. For example, if the sample size was 19 observations or less, estimation of the 95th discrete percentile would always be the maximum observation. If the sample size increased from 20 to 39, the 95th percentile would now correspond to the 2nd highest observation, and so on. This would make it possible to decrease the probability of seeing non-consecutive observations (i.e. the margin of error) by increasing the sample size used to obtain the travel time distribution. Figure 3-7 shows how the percentage of times the discrete 95th percentile of a 30-minute travel time distribution for the 50 largest O-D pairs in the Underground during the morning peak changes with the sample size of the given distribution.

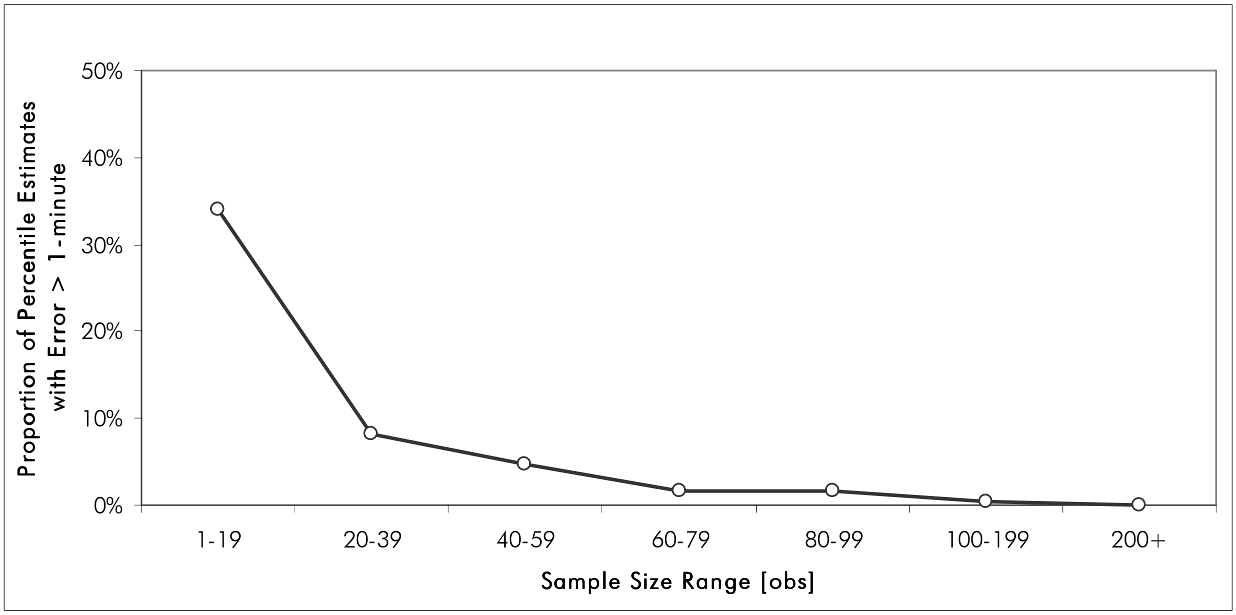
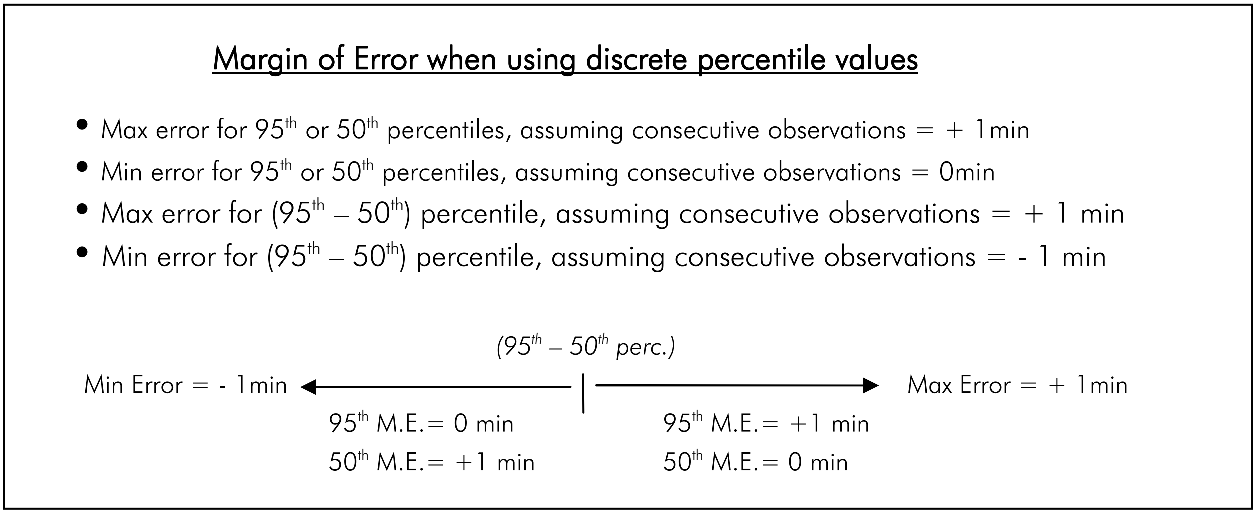


Figure 3-7: Proportion of 95th percentile estimates with a margin of error greater than 1 minute for the 50 highest-volume Underground O-D pairs – AM Peak, Feb. 2007

Figure 3-7 shows how for every increase of 20 observations in the sample size used to construct the travel time distribution for a particular O-D pair, the likelihood of not having consecutive values around the 95th percentile decreases, thereby minimizing the chance that the margin of error would exceed 1 minute. This also informs the decision by Chan (2007) to recommend a minimum sample size of 20 observations when estimating the 95th percentile as part of her proposed measure of reliability (see section 2.3.2).

In the specific context of this research, it is also important to consider the margin of error resulting from the difference between two percentiles, since this particular operation is used to define the metrics of reliability later on. Assuming that a discrete percentile will overestimate the actual travel time value by at most 1 minute, **the difference between two percentiles has a margin of error of +/- 1 unit**, and would depend on the size of the discrepancy for each percentile relative to one another. Figure 3-8 illustrates this by showing the margin of error of the difference between the 95th and the median travel time for a particular distribution.



Margin of Error when using discrete percentile values

- Max error for 95th or 50th percentiles, assuming consecutive observations = + 1min
- Min error for 95th or 50th percentiles, assuming consecutive observations = 0min
- Max error for (95th – 50th) percentile, assuming consecutive observations = + 1 min
- Min error for (95th – 50th) percentile, assuming consecutive observations = - 1 min

Figure 3-8: Margin of error when estimating percentiles from discrete travel time data

Establishing this margin of error is important not only for correctness of interpretation, but also when making recommendations in terms of minimum sample size requirements and their direct relationship to the level of aggregation possible (i.e. resolution) for monitoring performance on a regular basis (see *Oyster Data Coverage and Penetration Rates* next).

A final point related to the measurement of travel times using Oyster data important to this study has to do with the physical distance between the gatelines at a station and the platform. This is particularly important when using Oyster travel time measurements alongside estimates obtained from JTM, which define the total journey distance differently. Chan (2007) points to the fact that JTM defines access and egress time as starting from the station entrance, whereas Oyster measurements occur between the gatelines (already within the station). Figure 3-9 illustrates this discrepancy in measurement.

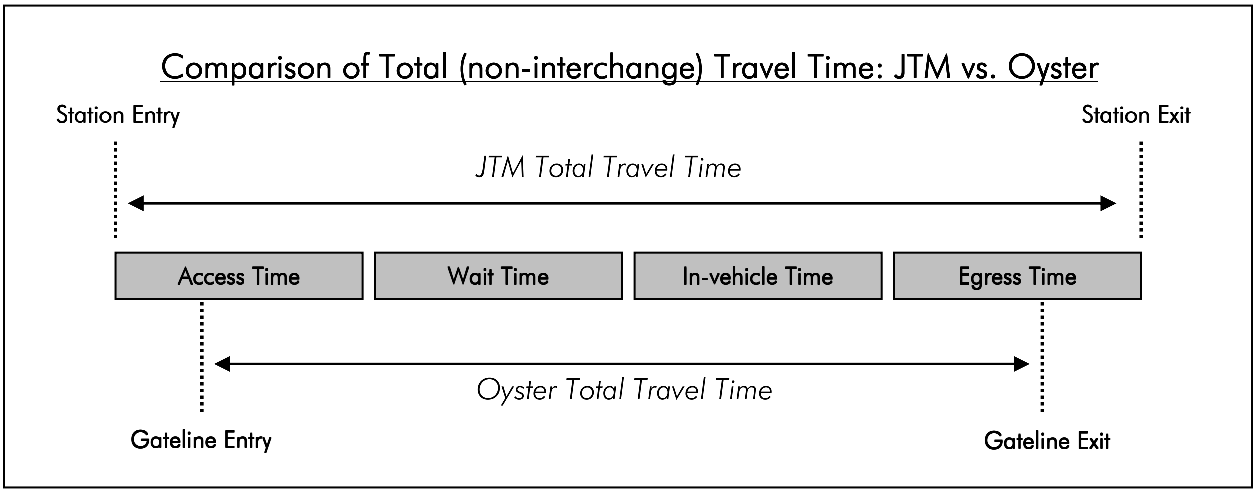


Figure 3-9: Comparison of the total travel time for Oyster and JTM (adapted from Chan, 2007)

This can be addressed in the short-term through corrections, such as using 85% of the JTM access and egress time to represent the time not captured by Oyster when walking from station entrance to the gates, on average. However, in the long-run, the compatibility of Oyster with the existing performance measurement systems for the different modes under TfL will depend to a large extent on whether travel times are capturing the same physical journey.

Oyster Data Coverage and Penetration Rates

A second consideration important for this study is the level of penetration of the Oyster Smart Card in the Underground's travel. A high level of use of the card will affect the sample sizes and coverage of the system that can be attained using Oyster data.

The level of spatial coverage of Oyster depends mainly on the availability of gates allowing this form of payment at particular stations, and their configuration. Chan (2007) estimated the proportion of O-D pairs in the Underground that were covered by Oyster card data to be around 56%, or 77% if O-D pairs with low numbers of journeys are included[11]. This level of comprehensiveness is a good starting point for using this data for performance measurement as journeys are increasingly made on this single fare media.

In addition, it is important to have a sufficiently large sample size for the O-D pairs being evaluated so that performance can be accurately measured. The sample sizes available from Oyster data at a given point in time depend on two factors: the rate at which transactions are incomplete, and the resolution at which performance is to be estimated.

Incomplete transactions, or the failure of a passenger to either leave a record of entry or exit from the system, make it impossible to estimate the total travel time for a journey. There are several reasons for incomplete transactions, including the presence of non-fully gated stations, assistance by station staff in passing through the gateline barriers, or even fare evasion. This data is excluded from the total sample available for analysis, representing around 9% for the particular periods used in this study[12].

Also affecting sample size is the level of resolution desired for measuring performance. Specifically, as the time interval for which a travel time distribution is estimated decreases (e.g. moving from a 3-hour to a 15-minute distribution of trips), the sample size would naturally decrease. Insufficient sample sizes will lead to inaccuracy in the results, and lower system coverage if minimum sample sizes are respected. Figure 3-10 illustrates this effect by showing the number of days required for each of the 800 highest-volume O-D pairs in the system to achieve the minimum sample size of 20 trips for four levels of temporal aggregation within the AM Peak – 3-hour, 1-hour, 30-minute, and at the 15-minute level.

[11] The number of O-D pairs covered by Oyster transactions after rounding to the nearest integer was 41,901 for January 2007. Before rounding, the number was 57,407, or 77% of the total number of O-D pairs possible in the Underground, 74,256 (Chan, 2007).

[12] The percentage of Oyster journeys missing either an entry or exit stamp (incomplete) for the Underground, averaged over 5 weekdays in mid-November 2007 (12th-16th) was around 9%.

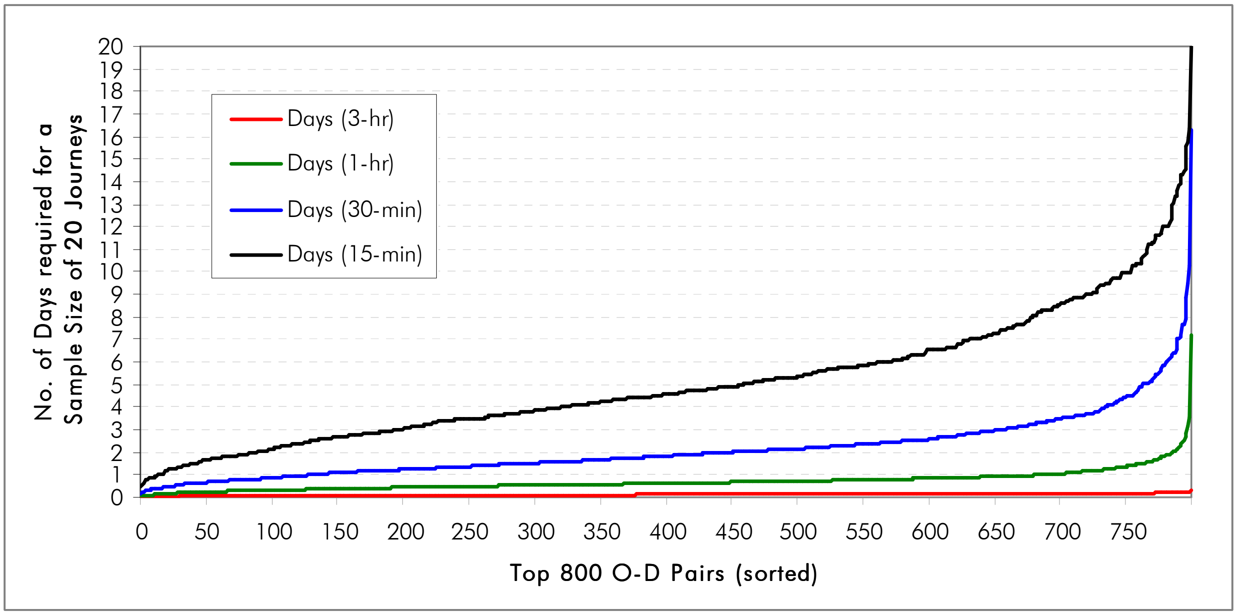


Figure 3-10: Number of days required to achieve the minimum sample size across levels of temporal resolution for 800 O-D pairs– AM Peak, Feb. 2007

Figure 3-10 shows how the number of O-D pairs that require more days to achieve the minimum sample size of 20 journeys increases with higher levels of temporal resolution. For example, aggregating journeys into a 3-hour travel time distribution ensures that all of the highest-volume 800 O-D pairs in the Underground require pooling only 1 day of journeys to achieve at least 20 observations. At the other extreme, aggregating journeys into a travel time distribution at the 15-minute level leads to only 14 of the 800 highest-volume O-D pairs (2%) as having enough observations with only 1 day of journeys being pooled together (intersection point of y-axis = 1 and 15-minute aggregation line).

The relationship between the temporal resolution and sample sizes found for travel on Oyster in the Underground indicates that at the AM Peak (3-hour) level, the 800 highest-volume O-D pairs have sufficient sample size to measure performance when using a single peak period (1 day). This means that when journeys across four weeks are pooled to estimate reliability, as will be done in subsequent chapters, Oyster data is a viable option for measuring system performance on a routine basis.

3.4.3: Data Sources and Samples used in this Research

The analysis in this study was largely based on data from the Oyster Smart Card system. Specifically, a sample of 100% of all AM Peak journeys in the Underground across four weeks from 4 February to 3 March 2007 and four weeks from 28 October to 24 November 2007 was used. These two four-week periods correspond to Period 12/06-07 and Periods 8 and 9/07-08 in the Underground's calendar, respectively. Journeys that included a segment of their travel on other modes in addition to the Underground, such as National Rail and DLR, were also included in the sample.

Other sources of data were used complementary to data from Oyster. Results from the Journey Time Metric system at the network and line levels from Period 1/03-04 to Period 1/08-09 in the Underground's calendar were used. In addition, incident log data obtained from the NACHs system for Period 12/06-07 (matching the first 4 weeks of Oyster data) provided the records of incidents that occurred on every line in the system during the morning peak. Finally, data from the Journey Planner trip planning software, downloaded from its website during December 2008 and May 2009 for various O-D pairs in the system (TfL, 2009).

Chapter 4: Service Reliability Measurement Framework

This chapter proposes a practical framework for quantifying reliability from the perspective of passengers using Smart Card data. It is designed to serve as the basis for the development of applications that will help transit agencies to both increase their understanding of reliability and develop strategies to improve service quality.

The framework can be broken down into three layers that are woven together in this chapter. The first layer (section 4.1) brings together several of the concepts from the literature review to put forth a set of reliability measures that effectively capture the passenger experience and can be measured using Smart Card data. The second layer (section 4.2) proposes to classify performance into recurring and non-recurring system conditions, and shows how this distinction can be used to gain insight into the causes of unreliability. Both the conceptual underpinnings of the approach, as well as the estimation methodology are discussed. Finally, section 4.3 presents the third layer of the framework, where the initial set of reliability measures is extended and made more robust.

4.1: Reliability Buffer Time Metric

The foundation for the reliability framework is developed in this section. First, general criteria are identified in section 4.1.1 and used for guiding the development of the proposed reliability measures. Section 4.1.2 defines both conceptually and mathematically an initial set of reliability measures and provides a rationale for the parameter values proposed in this research. Lastly, section 4.1.3 discusses some of the advantages of the proposed measures with respect to the established criteria.

4.1.1: Requirements for the Measures of Reliability

In this section two sets of criteria are determined that are used to define the reliability measures used in the framework. The first set presents conceptual guidelines to follow when developing the measures of reliability and builds on the lessons from the literature review. The second set of criteria provides recommendations for the appropriate mathematical form the measures should adopt.

Conceptual Guidelines for the Measures of Reliability

Five general criteria are used to develop a set of reliability measures. They balance the need to accurately represent the passenger experience with the desire to produce straightforward measures that can be applied by transit agencies as part of their routine monitoring of performance. These criteria build on the work by Chan (2007), and consist of:

1. **Representative of the Passenger Perspective** – One of the primary drawbacks of existing measures of reliability is their focus on operational quality as opposed to the

passenger experience. An appropriate set of reliability measures should take into account the effects of travel time variability on passengers by focusing on extreme values (as opposed to average performance), and should be sensitive to the way passengers build their expectations of service. The measure should also be representative of the experience of most passengers, and not be overly susceptible to the effects of individual traveler behaviour or day-specific events that could skew the characterization of performance.

2. **Straightforward Estimation and Interpretation** – A successful measure should balance between technical/mathematical complexity and ease of use and implementation. That is, the results should be easily interpreted not only by analysts but also by decision-makers and passengers alike. On the other hand, the reliability measure should not be so simple as to lose its value for analysts and transit operators. Data availability and estimation costs should also be factored in as part of the degree to which estimation of a particular measure is straightforward, and the likelihood that it is implementable from the perspective of the transit agency.

3. **Comparability and Aggregation of Results** – It is important that the output of any particular measure be comparable across different spatial and temporal levels. Spatially, the ability to estimate results for particular portions of the system that are meaningful when placed side-by-side is useful for both comparative analyses, and the aggregation of results into higher spatial units (e.g. moving from the O-D pair level to line level estimates of performance). This is particularly important in the context of JTM, where network level measures of performance are derived from more disaggregate estimates (see section 3.2). The ability to estimate performance across smaller time intervals allows for different types of analysis (e.g. study of crowding *within* the morning peak), and makes the aggregation of results into time periods appropriate for reporting performance (e.g. 3-hour morning peak period).

4. **Useful as input to the Evaluation of Performance** – There are three aspects that should be covered by the set of reliability measures in order for them to be useful in the context of monitoring and evaluation of performance. First, the information provided by the measure must be useful in setting specific performance goals (i.e. setting a standard for performance) and evaluating progress made towards them (i.e. determining the proportion of passengers receiving good and bad service). The second aspect is precision and accuracy. Namely, the results must be both repeatable and take into account sampling errors (see section 3.4.2). The third aspect that is important to keep in mind is the compatibility of the proposed measures with the existing service quality measures in place at the transit agency. The proposed measures should be sensitive to the characteristics of current measures (e.g. units of measurement) and resource availability (e.g. staff and data related constraints) already in place at the transit agency.

5. **Provides Insight into Causes of Unreliability** – The measure should ideally help analysts identify and quantify the contribution of different causes of unreliability. Through meaningful feedback obtained from the measures of reliability strategies

can be selected to target the specific causes of unreliability and improve performance.

Recommendations on the Mathematical Form for the Measures of Reliability

The work by Abkowitz et al. (1978), reviewed in section 2.3.1, as well as earlier work by Martland (1972), provided several recommendations for the appropriate functional form that reliability measures should adopt in order to be most effective. In this research the following four recommendations are used:

1. Measures of reliability should focus on capturing the compactness of the travel time distribution, as opposed to measuring deviations from scheduled values.

2. The mathematical form of the measure should take into account the shape of the travel time distribution, controlling for the effects of skew in biasing the results.

3. It is recommended that the measure consist of three separate parts:
   - A measure of compactness of the distribution
   - A measure of the center of the distribution such as the median
   - A *separate* measure for the likelihood of extremely long delays (i.e. system failure)

4. There should be a distinction between the variability of travel times that would lead to a late as opposed to early arrival. A focus on "lateness" is considered more representative of the passenger view, since it is considered to be more onerous than the penalties of variability resulting in an early arrival at the destination.

In addition to these recommendations, the work by Lomax et al. (2003) and De Jong et al. (2004) discussed earlier in section 2.3.1 provides a useful summary for the different functional forms that are typically used in passenger-oriented measures of reliability. Specifically, this research finds useful the second approach described in each study (i.e. the "percentile-based" approach from the former and the "slack time" approach from the latter) and adopts them in the next section to propose a measure of reliability: the Reliability Buffer Time (RBT).

4.1.2: Definition of the Reliability Buffer Time Metric

The general form for an initial set of reliability measures is proposed in this section, starting with a general description of the measure, followed by a discussion of each of its components.

Reliability Buffer Time Metric Definition

Measures described under the "percentile-based" approach typically quantify the compactness of the travel time distribution as the difference between a percentile in the upper quintile and one representing either the center of the distribution (i.e. median) or a lower quantile, as opposed to using traditional measures like the standard deviation. The "slack time" approach captures the effects of unreliability by measuring the amount of time passengers are required to allocate in order to complete a journey on-time with high probability. The proposed measure follows these related ideas. Specifically, the compactness of the travel time distribution is measured as the difference between an upper percentile, N, and an intermediate or lower percentile, M, and this value is defined as the additional "buffer" time that would be required of passengers in order to be N-percent sure of on-time arrival at their destination. The general structure of the measure is given by:

$$\textbf{Reliability Buffer Time} = (\mathbf{N^{th}\ percentile\ travel\ time - M^{th}\ percentile\ travel\ time})_{\textit{O-D, Time Interval, Sample Period}} \qquad [4\text{-}1]$$

The Reliability Buffer Time is applied to the total travel time distribution obtained from Oyster data for a particular O-D pair at different levels of temporal aggregation, denoted by the subscripts in Equation 4-1. The subscript *O-D* denotes the physical journey that is being represented, and is fixed at this level in order to be conceptually meaningful. *Time Interval* represents the level of aggregation made within a day in terms of the time of departure of the journeys, ranging anywhere from 15 minutes to the duration of the AM Peak (3 hours for the London Underground). The third subscript, *Sample Period*, represents the number of days of observations from which the measure is derived (e.g. a 1-weekday sample, a 20-weekday sample, etc.).

A discussion of each of the parts of the measure and their strengths relative to the criteria specified previously is discussed next, in addition to the parameters and their recommended values.

Indicator of the Typical Journey Time: M = 50th Percentile Travel Time

This part of the measure captures the travel time that a passenger travelling frequently on the system would come to base his/her expectations on, as opposed to basing the buffer time on scheduled travel times. Therefore, M represents the "typical" duration of a journey over the long-run. It is advantageous to use the 50th percentile travel time instead of the mean of the distribution because it is not sensitive to outliers or the tail of the distribution (which as will be discussed shortly, tends to be skewed to the left). Additionally, other reliability studies have found the median to be a better predictor of passenger path choice than the mean (Lam & Small, 2001).

Indicator for the Threshold of Certainty for Reliable Service: N = 95th Percentile Travel Time

This value represents the travel time required by passengers to arrive on or before their desired arrival time (i.e. on-time) at their destination with a certain level of certainty; here set at 95% for reasons discussed below. However, this figure is ultimately a policy decision for the transit agency, which is forced to balance between three major considerations: relevance to passengers, realistic for operators, and data feasibility.

From the perspective of passengers, N represents the probability of on-time arrival. It can also be thought of in terms of the odds of arriving on-time given repeat journeys, where passengers would be willing to accept a late arrival every certain period of time as part of the buffer time they budget into their schedules (Henderson et al., 1991b). When selecting a value that is representative of the concerns of passengers, it is important to keep in mind the non-linear change in the odds of on-time arrival with respect to a linear change in the upper percentile of the RBT. Figure 4-1 summarizes this effect by comparing the number of days a commuter would arrive on-time before she would experience a late arrival at her destination for a buffer time with an upper percentile value N.

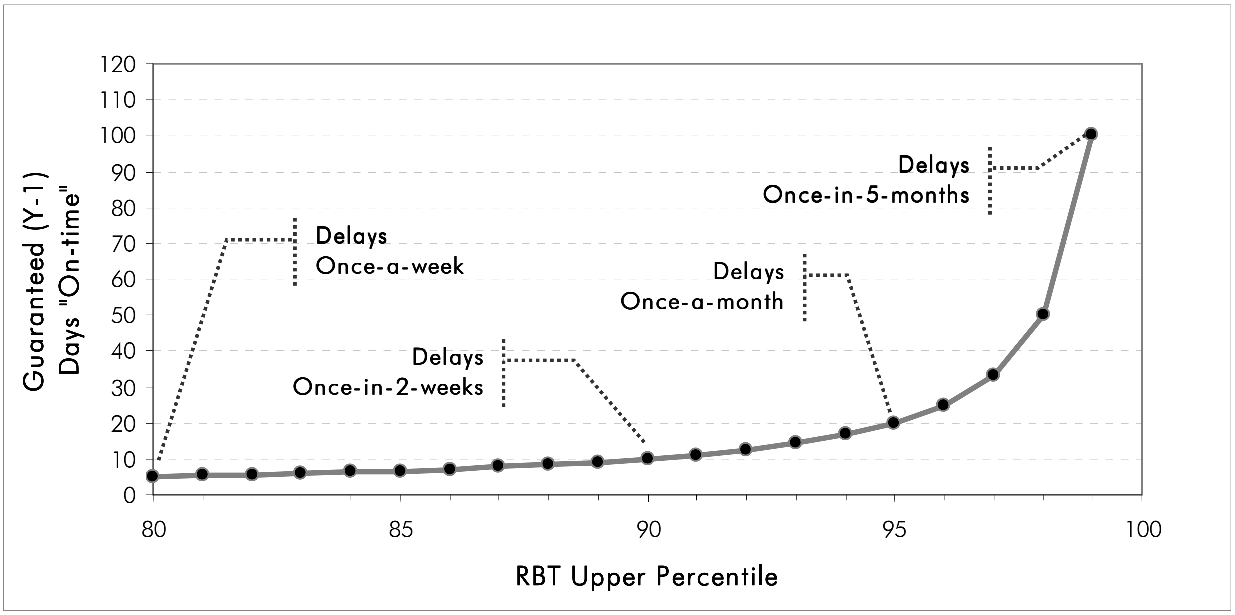


Figure 4-1: The odds of a late arrival as a function of the RBT upper percentile value

Figure 4-1 shows how the 90th percentile upper percentile value guarantees a late arrival once every ten journeys, the 95th percentile about once every 20, and 99th percentile once every 100 trips. Selection of a value that is too low so that it loses its relevance to passengers should also be avoided.

The second consideration for setting this parameter takes into account the degree to which the travel consistency being sought is realistic to achieve from an operational perspective. Henderson et al. (1991b) also point out that just as the odds of late arrival decrease non-linearly with higher percentile values, so does the amount of effort required to guarantee those odds. For example, a policy that evaluates the service (hence sets performance targets) based on the 99th percentile travel time implies that it must guarantee that only one out of every 100 journeys will exceed that travel time, making the measure highly sensitive to small delays (i.e. those that affect more than 1% of the total pool of journeys). However, operators would also have to balance this with a situation where the percentile value is so low that it is no longer sensitive to reasonably large delays (e.g. for the 80th percentile, those delays that affect over 20% of total trips).

Third, the characteristics of the data must be considered, with a focus on the effects of individual behaviour and minimum sample size on estimates of performance. With respect to the former, Chan (2007) found that a small proportion of journeys had unreasonably high journey times for various O-D pairs. This was explained by individual behaviour, such as passengers waiting behind the gate-line for friends, and it was found that the proportion of these trips could be safely capped at 1% of the total number of journeys. This translated to an upper percentile value of N equal to the 99th percentile travel time in order to control for the effects of outlier journey times. Second, as the percentile travel time being studied increases, it is necessary to have larger sample sizes to reduce the margin of error due to non-consecutive travel time observations (see section 3.4.2). In the case of N = 95th percentile, a minimum sample size of 20 observations would be required to avoid measuring the maximum travel time observation as the upper boundary for the RBT. The minimum sample size rapidly increases to 100 observations when N = 99th percentile, having broader implications on the applicability of the reliability measures through reduced system coverage.

In this study, N is set to the 95th percentile travel time. This value was successfully used by Chan (2007) for the London Underground. In addition, a large proportion of studies on reliability measures found in the literature (see Chapter 2) tended to focus on the 90th or the 95th percentile, supporting the validity of the value as a reasonable upper limit for the RBT. Lastly, a once-a-month chance of late arrival serves as a reasonable initial value that could be updated through market research studies. Ultimately, however, this is a policy decision by the transit agency which must balance the level of sensitivity with which to evaluate the operation with the service quality demands by customers.

Reliability Buffer Time: RBT = Difference between the 95th and the 50th Percentile Travel Times

This value represents the amount of "buffer" time that needs to be budgeted into one's schedule *above* the typical travel time, in order to be sure of on-time arrival at one's destination with probability 0.95. It is a measure of the compactness of the distribution, as recommended by Abkowitz et al. (1978), yet conceptually represents the effects of travel time variability on scheduling considerations. It focuses on the impact of variability on "late" arrivals as recommended by the criteria from the previous section, and is supported as a better measure of reliability than the standard deviation by previous empirical studies (Lam & Small, 2001). In addition, the positive skew in the travel time distributions observed from Oyster data make the use of traditional measures of compactness such as the standard deviation inappropriate (Abkowitz et al., 1978). Finally, the measure is straightforward to estimate and interpret. Moreover, its unit of time (i.e. minutes) is compatible with performance monitoring systems like the Underground's JTM.

Temporal Aggregation of Performance

The level of aggregation of the travel time distribution for a particular O-D pair used to measure the RBT can be readily modified to meet the needs of the transit agency implementing the measure. At the most disaggregate level, the RBT can be measured at the shortest time interval that satisfies minimum sample size requirements (e.g. a 5 minute interval for high volume journeys), and within a single day (sample period). The RBT can also be applied at higher temporal levels of aggregation in order to be compatible with estimation periods required by a particular transit agency, or simply to obtain a different perspective on the performance of the system. For example, the application of Equation 4-1 to an O-D pair for departures during the AM Peak over 20 weekdays (i.e. a month) using the parameter values recommended in this section is given by:

$$\textbf{Reliability Buffer Time} = (\textbf{95}^{\textbf{th}}\ \textbf{percentile travel time} - \textbf{50}^{\textbf{th}}\ \textbf{percentile travel time})_{\textit{O-D, AM Peak, 20 Weekdays}} \qquad [4\text{-}2]$$

The corresponding travel time distribution and RBT are graphically depicted in Figure 4-2 as an illustration.

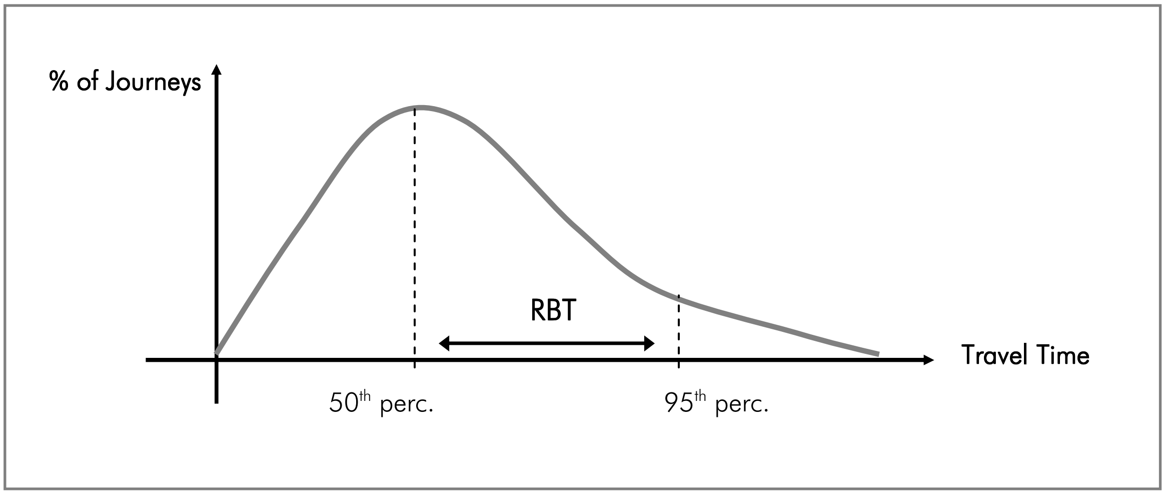


Figure 4-2: RBT metric applied to a total travel time distribution for an O-D pair

Spatial Aggregation of Performance

The RBT metric is estimated based on the distribution of total passenger travel times for a particular origin-destination pair. Under certain applications, however, it might be desirable to estimate performance at higher units of spatial aggregation, such as for an entire line. At the line level, the measure could be used by management concerned with improving the overall passenger experience, or as an input into global indicators of network performance.

Though it would be possible to aggregate all same-line journeys to estimate a travel time distribution at the line level, the results from the application of the RBT metric would be difficult to interpret as they relate to the passenger experience. One approach to measure reliability at the line level would be to estimate the performance of each O-D pair and find a weighted average using the volume of passengers completing each journey, as given by:

$$RBT_{Line} = \frac{\sum_{OD\in Line} Vol_{OD} \cdot RBT_{OD}}{\sum_{OD\in Line} Vol_{OD}} \qquad [4\text{-}3]$$

Where Vol_{OD} = total passenger journeys for origin-destination pair "OD" within "Line", and
RBT_{OD} = reliability buffer time for origin-destination pair "OD" within "Line".

By weighing the performance of each O-D pair by the volume of trips that experienced it, higher weight is given at the aggregate level to the reliability of journeys that affect the highest proportion of passengers. Only same-line journeys are considered in this initial approach. However, the level of reliability experienced by passengers transferring onto the line could be estimated through the use of a network assignment.

4.1.3: Advantages of the Reliability Buffer Time Metric

The RBT metric satisfies several of the conceptual and mathematical criteria specified in section 4.1.1. Regarding the first conceptual guideline, the measure is representative of the passenger's perspective as it quantifies the effects of service unreliability on the scheduling considerations of travelers. This is an improvement with respect to supply-side based measures of reliability (see section 2.1.4).

In addition, the metric is straightforward to both estimate and interpret, as called for by the second conceptual guideline. Its estimation is facilitated not only by the low amount of computational complexity required, but also by the large volume of data available from AFC systems used to find travel time distributions for each O-D pair. Also, the way changes in performance translate to benefits and costs to passengers is easily interpreted from the RBT metric results, both because of the conceptual representation of travelers' definition of reliability and because of the use of units of time (i.e. minutes).

Moreover, the RBT metric satisfies the majority of the mathematical recommendations obtained from previous studies. This includes the metric's ability to quantify the compactness of the travel time distribution and its independence from scheduled performance values, as well as its focus on "late" arrivals caused by unreliability. Lastly, the use of percentiles by the RBT metric is appropriate when travel time distributions are skewed, as is empirically observed in subsequent chapters.

4.2: Performance Classification – Recurrent vs. Incident-related Delays

The second layer of the reliability framework proposes a way to break down performance using AFC data that is useful for gaining insight into the causes of unreliability and enhancing the initial set of reliability measures presented in the previous section. The motivation for this addition to the framework is presented in section 4.2.1, which considers some of the areas where the RBT metric could be made more robust. Section 4.2.2 provides some of the intuition behind the approach used to classify performance using empirical observations from the performance of the Underground to illustrate. Section 4.2.3 defines two categories of performance and explains how they can be used to understand the changes in reliability observed in the previous section. Lastly, a methodology for classifying performance is developed in section 4.2.4.

4.2.1: Considerations Regarding the Application of the RBT Metric

Despite the initial set of advantages of the RBT metric, there are three general areas where the measure's usefulness to transit agencies could be enhanced. These relate to the requirements raised by conceptual guidelines 3, 4, and 5 from section 4.1.1, and serve as motivation for the development of the remainder of the framework.

1. **Comparability and Aggregation of Results** – the study by Chan (2007) found that service characteristics such as journey length and the scheduled headway were positively correlated with the level of travel time variability experienced by passengers. This made it difficult to compare performance across O-D pairs with different characteristics, and to aggregate their performance to a line level measure that could be used to compare lines with one another. The aggregation approach proposed in Equation 4-3 to measure reliability at the line level also suffered from these effects, being applicable mainly for comparison of performance for a single line over time. This issue was partially addressed by Chan, which proposed to measure reliability at the line level by only including O-D pairs on the trunk portions to control for the effect of scheduled headways (see section 2.3.2). Abkowitz et al. (1978) also tried to address the issue by proposing to normalize the measure of compactness by the typical travel time (i.e. controlling for the effects of journey length). Both studies, however, concluded that further work was still required to deal with these effects properly.

2. **Useful as input to the Evaluation of Performance** – there are two considerations that reduce the usefulness of the RBT metric in the context of performance evaluation. First, the existence of an "irreducible" variability caused by the discrete nature of transit service (i.e. even with null headway variability, random passenger arrivals would lead to variable waiting times) would lead to a certain level of unreliability that would be impossible to reduce. Second, empirical observations in this research indicated that the long-run reliability (e.g. over a period of four weeks) was highly sensitive to the performance of a handful of days with high levels of delays attributable to incident-related disruptions (see section 4.2.2). This raised the question of whether operators should be evaluated for the performance of the system including disruptions, or only on the delivery of service during typical conditions. This last point relating strongly to the recommendation by Abkowitz et al. (1978) of having a separate measure for the "Likelihood of Extremely Long Delays," which the RBT metric does not distinguish from its measure of compactness. Both these issues could be addressed through the estimation of a baseline measure of reliability to control for the minimum variability of travel times, and the proper evaluation of operator performance.
3. **Provides Insight into Causes of Unreliability** –the RBT metric has the potential to help operators identify the causes of unreliability and inform strategies to address them. However, due to the use of total travel time variability estimates obtained from demand-side data, the measure does not yield information on the sources behind observed changes in reliability (e.g. changes due to train crowding, service irregularity, etc.). This limitation can be addressed through the use of additional information on the operations of the system.

4.2.2: Intuition for Performance Classification

One of the desired applications of the framework is in the area of performance monitoring, which requires the ability to quantify and compare performance over time. This would allow operators to detect deterioration in performance requiring remedial action, or an improvement in service quality that can, hopefully, be attributed to specific strategies. When comparing the level of reliability experienced by passengers across several weeks in the sample of Oyster data, interesting and important patterns of variation were observed. Figure 4-3 shows the median travel time and the RBT across four weeks during February 2007, for four of the highest-volume O-D pairs in the Underground. It shows two overall patterns with regards to the distribution of travel times for each week. First, the median travel time (i.e. the typical travel time to complete a journey) is relatively constant across over time for each O-D pair (accounting for the expected variation of +/- 1-minute discussed in section 3.4.2). Second, the level of reliability, as captured by the RBT, varies appreciably from week to week and across O-D pairs. For example, the journey from Waterloo to Canary Wharf station on the Jubilee line (shown in blue) went from a buffer time of 74% of the typical journey time in the first week, to a buffer time of only 22% of the median travel time two weeks later. Conversely, these same values for journey from Brixton to Oxford Circus on the Victoria line ranged only from 21% to 35% over the same four weeks.

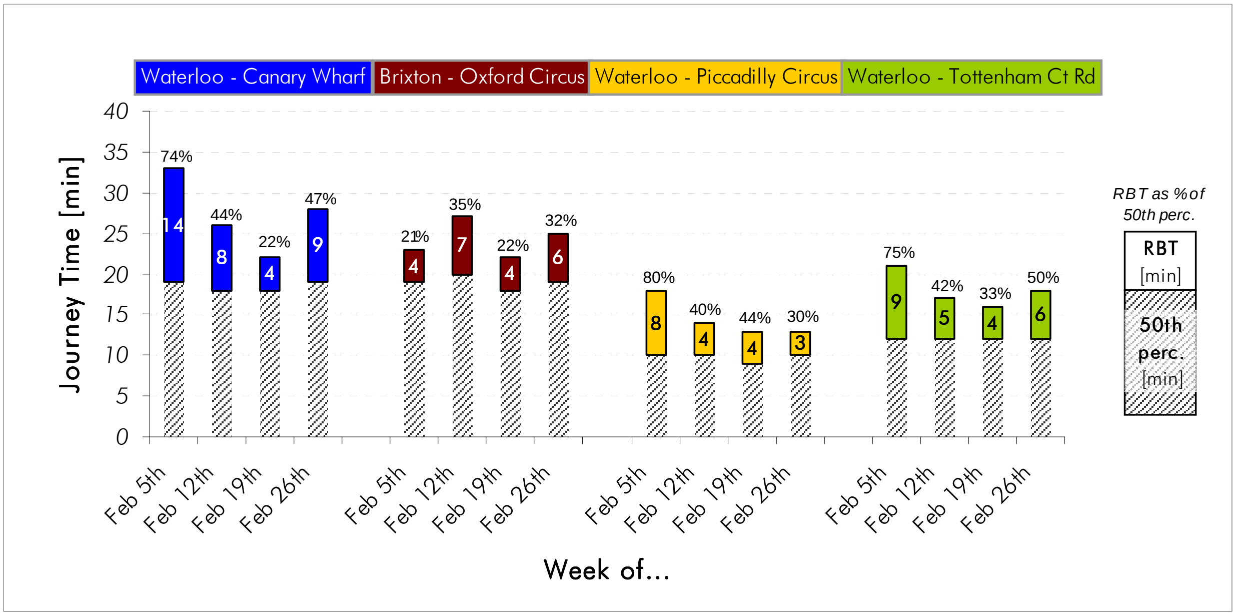


Figure 4-3: Comparison of the weekly median journey time and the RBT for four Underground O-D pairs – AM Peak, Feb. 2007

Figure 4-3 can also be interpreted as saying that when an individual enters the Underground to make a journey, his particular total travel time is likely to come from a different probabilistic distribution, depending on the week and O-D pair being traveled on. That is to say, not only is there variability in journey times, but also variability in the *distribution* of journey times itself.

Additional insight was gained by comparing reliability across various days for a single O-D pair. First, it was observed that a few of the days exhibited travel time distributions with fat tails as compared to the majority of the days, representing a relatively large proportion of journeys that exhibited very high travel times. This is illustrated in Figure 4-4 for journeys from Waterloo to Canary Wharf over five weekdays in February 2007 during the morning peak.

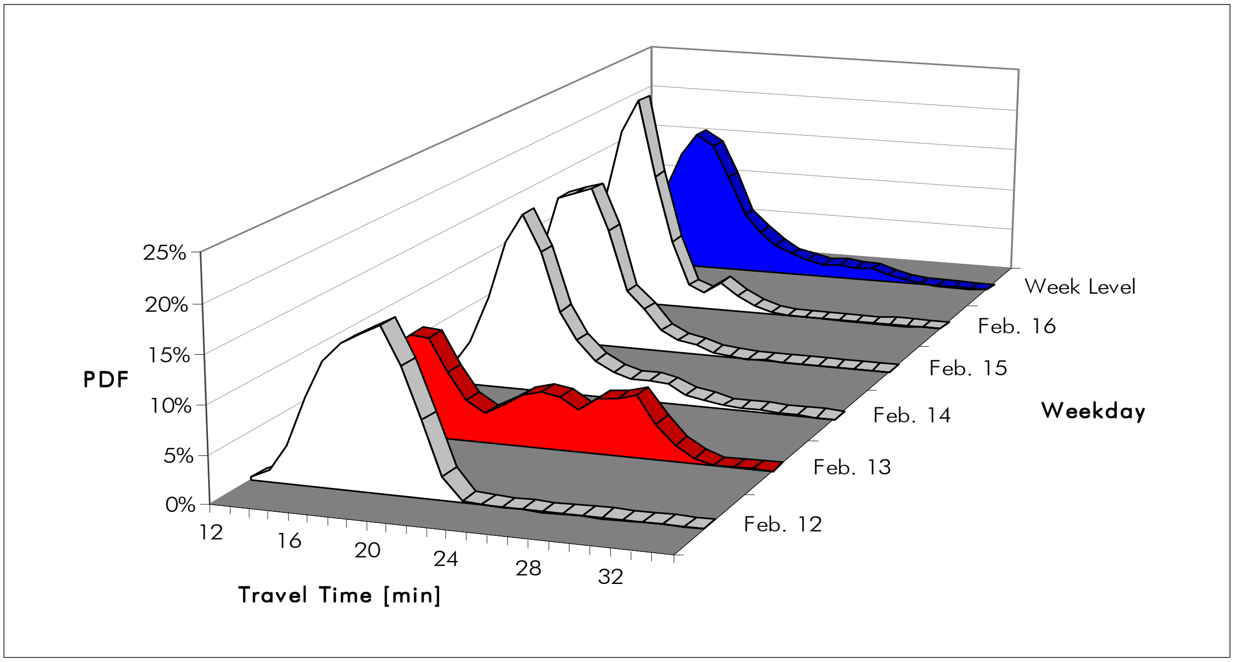


Figure 4-4: Comparison of Oyster AM Peak travel time distributions for journeys from Waterloo to Canary Wharf

The travel time distribution on February 13th shown in Figure 4-4 clearly deviated from the shape of the distribution shown by the remaining four weekdays, with a non-decreasing probabilistic function (i.e. bi-modal distribution). In addition, the remaining days had similar travel time distributions that were at the same time more compact and typically exhibited a "smoother" drop towards the higher end of journey times. This pattern implies that the long-term variability of travel times, experienced by passengers is sensitive to the performance of a few days with extremely high levels of delays. The travel time distribution at the week level shown in Figure 4-4 (in blue) reflects the effect of the performance on February 13th on the long-run variability of travel times through its tail that is fatter than the remaining weekdays, leading to a higher 95th percentile journey time (i.e. reliability buffer time). Figure 4-5 quantifies the nature of this effect by comparing the RBT at the weekly level with the RBT required had performance on February 13th followed the pattern of the remaining weekdays.

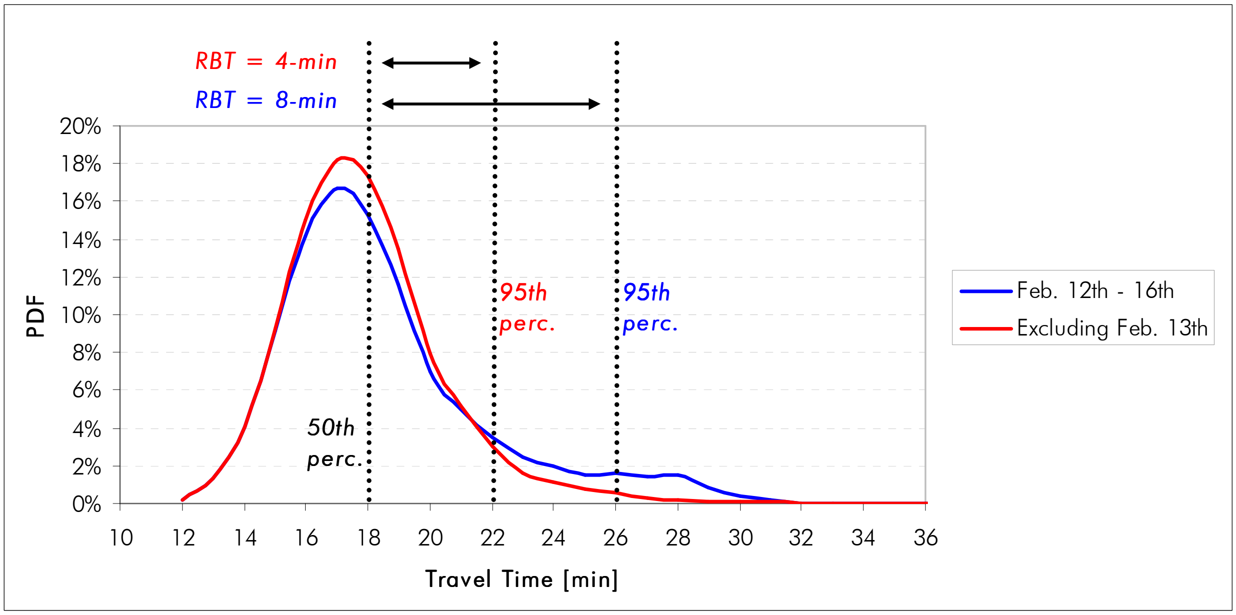


Figure 4-5: Contribution of the performance on Feb. 13th to the RBT for the week of Feb. 12th – 16th, 2007 for journeys from Waterloo to Canary Wharf

As suggested by empirical observations of the Underground's performance, the apparent irregularity of the travel time distribution across O-D pairs and over time and the sensitivity of this distribution to the performance of a few days with high levels of delays, can be explained by the occurrence of large disruptions in the system. These effects have important implications in terms of the sensitivity of the reliability measures to the various causes of unreliability, and their use as inputs into performance evaluation efforts. The next section builds upon this insight to make the proposed measures of reliability, and the framework, more robust.

4.2.3: Performance Categories – Recurrent vs. Incident-related Reliability

The changes in reliability described in the previous section can be understood by thinking about service quality as the product of various inputs that go into the execution of the operation. These inputs or factors of unreliability can be grouped along two dimensions first introduced in section 2.1.3: the degree of predictability of their effects, and whether they are endogenous or exogenous to the provision of service (i.e. controllable or uncontrollable by the transit agency).

At one extreme are those factors that occur in a recurring or systematic way and that have predictable impact on the performance of the system. These factors include both fixed components of the service like station spacing and scheduled frequencies, as well as recurring levels of passenger demand. These inputs are expected to produce, above the "irreducible" variability of travel times attributable to the discrete nature of transit service (see section 4.2.1), a similar outcome in terms of passenger travel times. That is, the "irreducible" variability of travel times, which is also expected to repeat itself over time, contributes to (i.e. is a subset of) the recurrent reliability of the system.

On the other end of the spectrum are those factors that affect performance in an unpredictable way, both in terms of how often and when they occur (i.e. randomly), and the magnitude of their impact. These factors include operations control interventions and other day-specific conditions, as well as incidents and other unforeseen disruptions, with their contribution to performance occurs in addition to the service produced by recurring factors. In the case of incidents, performance is expected to worsen from what would have been observed under typical conditions.

Through the lens of these two overall categories of factors, the changes in reliability discussed in the previous section for the Underground can be better understood. First of all, passenger travel times resulting from a typical day of operation (i.e. undisrupted conditions) can be expected to follow a distribution that would be similar over time. Secondly, when this routine performance is disrupted by the occurrence of day-specific factors like incidents, two visible effects on passengers could be expected: (a) journey times would increase from their usual value, leading to a higher proportion of passengers experiencing higher delays than usual, and (b) the level of delays under each disruption in the service would be unlikely to replicate due to the random/unpredictable nature of the effects of incidents. From this, the large and varying delays observed in the tail of the travel time distribution for a handful of days can be attributed in part to the occurrence of incidents, and the similarity of performance across the remaining days to the contribution of repeating factors as discussed above.

Through a separation of performance along these two dimensions, the contribution of incidents to unreliability above and beyond the travel time variability observed under recurrent conditions could be quantified. This would help improve the proposed measure of performance as an input into evaluation efforts by providing insight into the causes of unreliability.

Classification of Performance – Recurrent vs. Incident-related Travel Times

Based on the discussion above, it is proposed that the overall reliability of the system be decomposed into the performance observed under recurrent conditions, and the performance measured when the service is affected by incident-related disruptions. This reflects a mixture model intuition where the overall passenger travel time distribution is simply a probabilistic combination of individual distributions for the recurrent and incident-related performance of the system. For example, assuming a probability of suffering a large disruption on any given day, P[Incident], an individual's long-run distribution of travel times can be expressed as:

$$\mathbf{T.T.}_{\text{Overall}} = \mathbf{P[}\text{Incident}\mathbf{]} * \mathbf{T.T.}_{\text{Recurrent}} + \mathbf{P[}\text{No Incident}\mathbf{]} * \mathbf{T.T.}_{\text{Incident-related}} \qquad [4\text{-}4]$$

Where $\mathbf{T.T.}_{\text{Overall}}$ = distribution of travel times experienced by passengers in the long-run,
$\mathbf{T.T.}_{\text{Recurrent}}$ = distribution of travel times under recurrent conditions,
$\mathbf{T.T.}_{\text{Incident-related}}$ = distribution of travel times under incident-related conditions, and
$\mathbf{P[}\text{No Incident}\mathbf{]} = 1 - \mathbf{P[}\text{Incident}\mathbf{]}$.

The expected shape of each of the three distributions is illustrated in Figure 4-6, as well as the proportion of journeys experiencing each type of service as captured through the number of journeys in the y-axis.

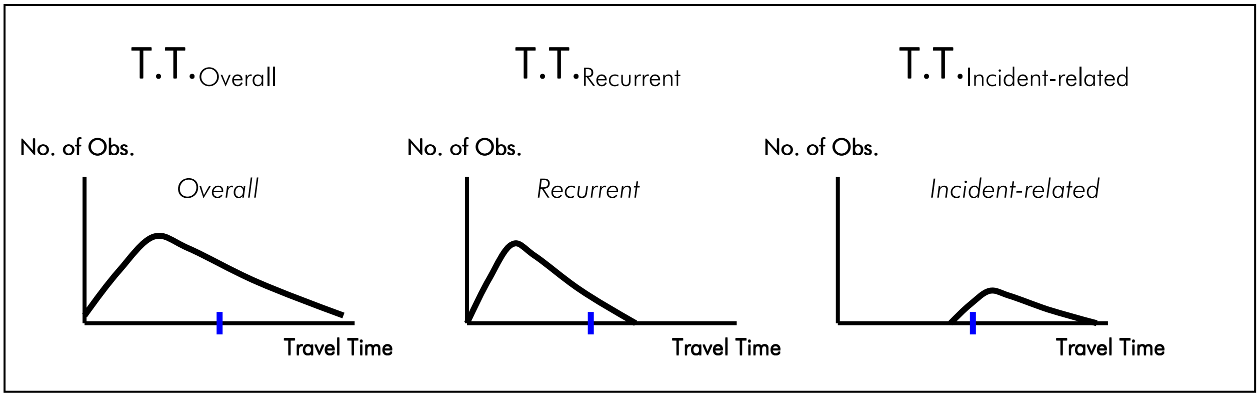


Figure 4-6: Illustration of mixture model intuition

The left-most histogram in Figure 4-6 represents the overall travel time distribution for a particular journey, and has the largest number of total journeys because it is the result of probabilistically combining the journeys categorized under the remaining two distributions of performance: recurrent and incident-related. The expectation that passenger travel times are lower under recurrent conditions than under disrupted conditions is also illustrated through the position of the histogram relative to the mark on the travel time (horizontal) axis.

Based on this discussion, the framework proposes to categorize the overall performance for individual days into two categories that reflect the unreliability caused by incidents and recurrent factors. One approach is to find natural "break points" or clusters by comparing performance across days without prior assumptions regarding the categories and their definitions. The other approach, preferred in this case, is to statistically classify the performance of each day into one of the two a-priori performance classes as already discussed. The next section presents the set of definitions and methodology for doing so.

4.2.4: Methodology for Performance Classification

The two categories of performance presented in section 4.2.2 are used to develop a method for classification. These categories can be defined as:

- **Recurrent Performance**: a day when performance, as shown by the passenger travel time distribution, was similar to that on other days (repeatable performance), and when the level of delays was near the non-incident related level of delays (i.e. low level of delays).

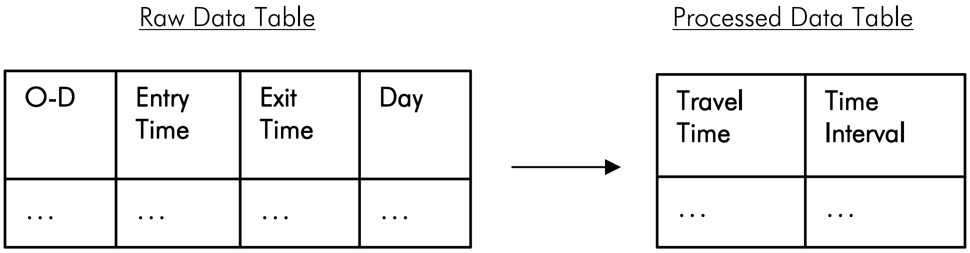
- **Incident-related Performance:** a day when performance, as shown by the passenger travel time distribution, was statistically different from the performance shown by other days (non-repeatable performance), and when the level of delays was above that from the remaining, majority, of incident-free days.

The approach is based on the identification of those days whose performance is statistically likely *not* to have occurred under recurrent conditions. That is, it is assumed that all days are not affected by incidents, unless proven otherwise. This leads to a focus on the identification of a subset of days whose performance is statistically significantly different from the pattern of performance shown by the remaining days in the sample. This is different from finding the subset of days that were "most likely to have occurred under recurrent conditions." The approach for classifying performance into these two categories is described next.

Outline of Classification Approach

The following is a graphical outline of the steps followed to classify performance based on the travel time distributions obtained from Smart Card data. Specified in *red* (italicized) are the parameters used in each step, and in green (underlined), their recommended values. The four steps required in the classification process are illustrated in Figures 4-7-a through d.

1. **Estimation of Travel Time Distribution(s) from Oyster Data –** Estimate passenger travel time distribution from Smart Card entry-exit data for a particular sample period (e.g. 1 day). Then, assign each journey to a *time interval* (of a certain duration) based on the time of entry into the system. For example, the morning peak could be defined as a single 3-hr time interval or as six 30-minute time intervals. This step makes it possible to (a) measure performance at more disaggregate levels, and (b) change the sensitivity of the classification approach to delays of different magnitudes (aggregation at smaller time intervals being more sensitive to smaller delays), and (c) to modify the sample size of a travel time distribution for a particular O-D pair to satisfy minimum sample size requirements (see section 3.4.2). Figure 4-7-a illustrates the estimation of two additional fields required for this step from the pre-processed or "raw" Smart Card data: travel time and the time interval when the journey began.



Raw Data Table

O-D	Entry Time	Exit Time	Day
...	...	...	...

Processed Data Table

Travel Time	Time Interval
...	...

Figure 4-7-a: Classification Approach – Step 1

2. **Selection of appropriate Indicator of Delays –** Using the travel time distribution for each origin-destination pair, a day (sample period is set to this value), and time interval (e.g. O-D-Day-AM Peak), an *indicator for the level of delays* must be selected. This indicator should represent the number of passengers experiencing long travel times (i.e. delays) during travel on that particular day. In this case, the indicator of delays was set to the 95th percentile travel time because of its sensitivity to smaller disruptions (only requires that 5% or more of all journeys within an O-D-Day be delayed), and to reflect the parameters recommended for the RBT metric (see section

4.1.2). Figure 4-7-b illustrates the indicator of delays for an O-D pair's travel time distribution during each of the time intervals, for a single day.

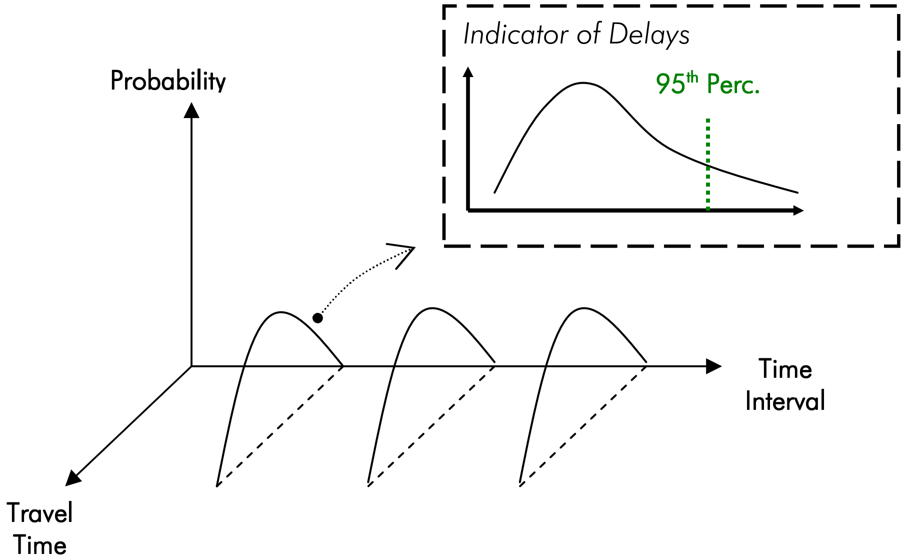


Figure 4-7-b: Classification Approach – Step 2

3. **Classification of Performance based on Indicator of Delays –** For the travel time distributions for one O-D pair within one time interval (e.g. 3-hr morning peak) over multiple days, identify the day-specific intervals with delays that are statistically significantly different from the delays found in the remainder (majority) of the days. A *statistical classification* method is used to identify these days (i.e. incident-related performance). Stepwise Regression is applied for this classification (see Appendix B for a detailed presentation of this method). Figure 4-7-c illustrates the comparison of the 95th percentile travel time across days during a time interval for journeys on a single O-D pair, highlighting those that were identified as incident-related as compared to the remaining observations, identified as belonging to the recurrent performance of the system.

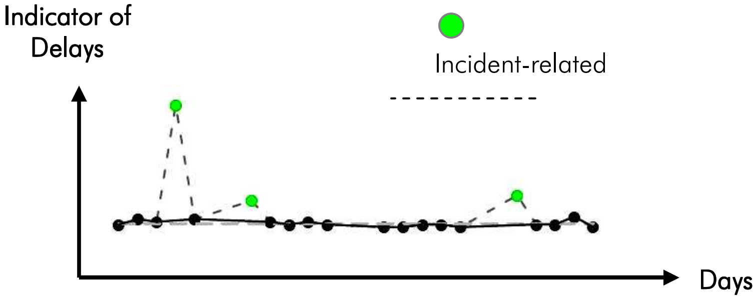


Figure 4-7-c: Classification Approach – Step 3

4. **Aggregation of Performance –** The recurrent and incident-related travel time distributions for each O-D pair within one time interval are estimated by *aggregating the journeys for each of the days* identified under each of the two performance categories is applied. The journeys within each period of interest of each day under each category are directly pooled together to estimate the recurrent and incident-related travel time distributions. The RBT metric can now be used to estimate the level of reliability experienced by passengers under each of the two performance conditions. Figure 4-7-d illustrates the aggregation of journeys from days identified under each category, showing how those journeys during days identified as being incident-related were not used to estimate the recurrent performance of the system.

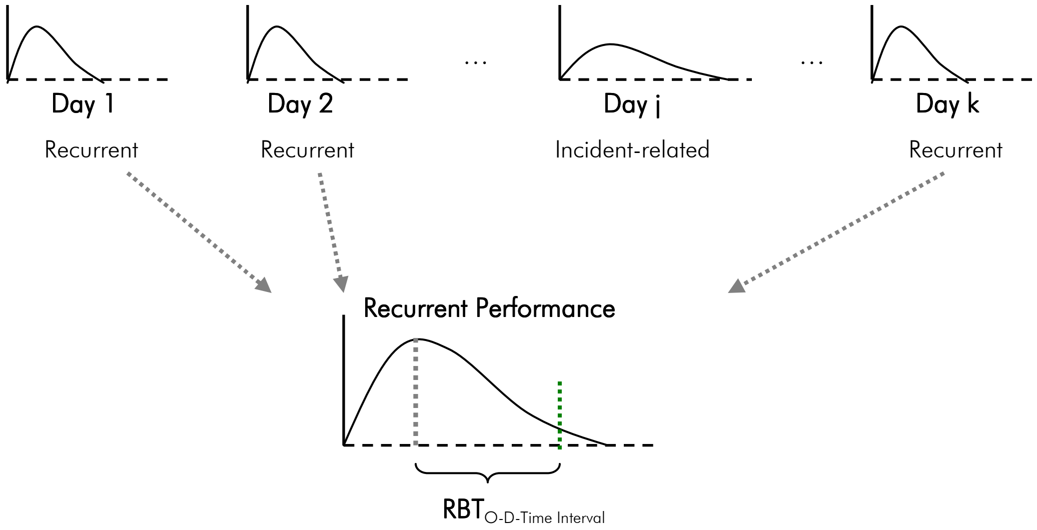


Figure 4-7-d: Classification Approach – Step 4

The reliability obtained from the recurrent travel time distribution can then be interpreted as the performance experienced by passengers under typical conditions, attributable to the characteristics of the journey being made as well as the discrete nature of transit service. The difference between this performance and the reliability measured from the overall travel time distribution is the contribution of incident-related disruptions to unreliability. The ability to distinguish between these two performance categories is useful not only for quantifying what the average contribution of incidents is to unreliability, but also for enhancing the proposed set of reliability measures.

4.3: Excess Reliability Buffer Time and Framework Extensions

The ability to decompose the performance of the Underground based on the framework described above provides useful insight into the contribution of different factors to unreliability. This section makes use of this insight to develop more robust measures of reliability that can be used as part of the routine monitoring of the service. In sections 4.3.1 and 4.3.2, two additional measures of reliability are proposed, extending the RBT metric to address some of the considerations raised earlier (see section 4.2.1). Section 4.3.3 ends by discussing the effectiveness of the proposed measures as an input into efforts towards evaluation of performance.

4.3.1: Definition of the Excess Reliability Buffer Time Metric

One of the most important characteristics of any measure of performance is the definition of a baseline that represents a "benchmark" performance. This baseline allows the analyst to define quantitatively a desired or expected performance, and use it to separate the causes and levels of unreliability that are tolerated and penalized. As discussed in section 4.2.1, the absence of a baseline made it difficult to use the RBT metric for comparative analyses across different parts of the network. With the ability to decompose performance, the framework makes it possible to consider extending the RBT metric to include a baseline performance reflective of the typical conditions of the system.

The general structure of the measure is defined by comparing the actual level of reliability experienced by passengers, with the reliability that they *should* have experienced had everything gone according to plan, or free from large disruptions. More concretely, the Excess Reliability Buffer Time (ERBT) metric is defined as "the buffer time needed by passengers to be 95% sure of on-time arrival at their destination, *in addition* to the buffer time required had the service gone according to plan." This time is found by measuring the difference between the overall RBT for a particular O-D pair, time interval, and sample period, and the baseline measure for the RBT. That is, the ERBT is given by:

$$\textbf{Excess Reliability Buffer Time} = (\textbf{RBT}_{\textbf{Overall}} - \textbf{RBT}_{\textbf{Baseline}})_{\textit{O-D, Time Interval, Sample Period}} \quad [4\text{-}5]$$

The baseline measure of the RBT in Equation 4-5 is ultimately a policy decision, where the transit agency using the measure must determine what level of performance should be considered as an acceptable reference point. In the case where the repeated performance of the system that can be expected given a set of fixed inputs into the service (i.e. free from the effects of large disruptions) is taken as the baseline, the baseline can be set to be the RBT of the recurrent travel time distribution found through the developed framework. This would be advantageous for transit agencies as it would offer a realistic view on the performance that could be achieved on a repeated basis given the existing service characteristics.

Setting the recurrent RBT as the baseline measure of reliability, the ERBT measure could be used to capture the additional unreliability added by large non-recurrent incidents above that caused by recurrent factors and smaller events. As an example, for an O-D pair during the AM Peak considering a span of 20 weekdays, Equation 4-5 would be given by:

$$\textbf{Excess Reliability Buffer Time} = (\textbf{RBT}_{\textbf{Overall}} - \textbf{RBT}_{\textbf{Recurrent}})_{\textit{O-D, AM Peak, 20 Weekdays}} \quad [4\text{-}6]$$

A natural question that arises when applying the ERBT is the following: How often should the baseline performance of the system be re-estimated? The answer to this depends on several considerations, including the presence of seasonality effects in service reliability and changes in the characteristics of the service (e.g. new scheduled headways) over time that must be accounted for.

In the case where the recurrent performance is used as a baseline measure, the ability to estimate a long-run level of reliability would improve proportionally with the number of days that should be compared with each other and classified. This study explored two alternative sample periods for measuring the baseline RBT: using a 20-weekday and a 40-weekday definition (i.e. the available data set). This simply meant that in the former case, a baseline performance was estimated for each four-week period in February and November 2007, and in the latter case both four-week periods shared a common baseline performance. This is separate from the definition of the sample period used to monitor and report changes in the overall performance, which in the case of JTM would correspond to a single four-week period. Analysis comparing the two approaches using data for the Underground did not show significant differences in estimates of the recurrent performance, suggesting both that differences due to seasonality between these months appeared to be minor, and that a truly recurrent performance could be estimated over the long-run. The 40-weekday definition of the sample period when estimating the baseline RBT was adopted in the application of the framework to the London Underground in section 6.1.

Spatial Aggregation of Performance

A line level measure of the ERBT can be estimated similarly to the approach described for the overall RBT metric, weighing the performance of each O-D pair by the volume of passengers experiencing that level of reliability, as given by:

$$ERBT_{Line} = \frac{\sum_{OD \in Line} Vol_{OD} \cdot ERBT_{OD}}{\sum_{OD \in Line} Vol_{OD}} \qquad [4\text{-}7]$$

Where Vol_{OD} = total passenger journeys for origin-destination pair "OD" within "Line", and
ERBT_{OD} = excess reliability buffer time for origin-destination pair "OD" within "Line".

Similar to the line level measure of the RBT, only same-line journeys are considered in this aggregation procedure. However, the approach can be easily extended to capture the performance experienced by journeys involving one or more transfers through the use of a network assignment. Also, from the approach used to estimate the overall RBT and ERBT at the line level as represented in Equations 4-3 and 4-7, it becomes straightforward to derive the relationship:

$$RBT_{Line} = RBT_{Line,Baseline} + ERBT_{Line} \qquad [4\text{-}8]$$

Where $RBT_{Line,Baseline}$ = the line level measure of the baseline RBT.

Equation 4-8 simply states that the measure of the overall RBT at the line level is the sum of the baseline RBT at the line level, using a similar aggregation approach at shown in Equation 4-

3, and the ERBT at the line level as defined above. This simple relationship makes it possible to decompose the overall reliability at the line level into a baseline and excess level of performance, parallel to that possible at the O-D pair level.

In addition to using the ERBT metric to provide an estimate of the level of unreliability for a particular O-D pair associated with incident-related conditions, it can also be extended to estimate the proportion of passengers receiving unreliable service based on an established baseline performance.

4.3.2: Definition of the Percentage of Unreliable Journeys Metric

Following the recommendations by Abkowitz et al. (1978), the ERBT metric is extended to obtain a measure for the likelihood of extremely long delays, or similarly, the proportion of passengers experiencing unreliable service. First, the framework is used to define a reliable journey as being shorter than or equal to the 95th percentile travel time under recurrent conditions. By construction, this definition implies that it is acceptable that 5% of journeys made under recurrent conditions experience unreliable service. However, when all journeys, (including those made under incident-related performance) are compared to this definition, the Percentage of Unreliable Journeys (PUJ) is likely to increase, and mathematically is given by:

$$\textbf{Percentage of Unreliable Journeys} = \left(\textbf{Percentage of Overall Journeys with T.T.} > \textbf{RBT}_{\textbf{Recurrent}}\right)_{\textit{O-D, Time Interval, Sample Period}} \quad [4\text{-}9]$$

Naturally the results from this measure will include 5% of all journeys made under recurrent conditions. Because by construction the duration of these journeys is accepted as being higher than the threshold for unreliable service, they can be subtracted from the overall number of unreliable journeys. The difference would produce the percentage of unreliable journeys *in excess* of the value the proportion that would have been observed in the absence of incidents. This is given by a simple extension of Equation 4-9, where 5% of the journeys under recurrent conditions is subtracted from the percentage of overall journeys (total) that were considered unreliable:

$$\begin{aligned}&\textbf{Percentage of Excess Unreliable Journeys} = \\ &\quad \left(\textbf{Percentage of Overall Journeys with T.T.} > \textbf{RBT}_{\textbf{Recurrent}}\right) \\ &\qquad - \left(\textbf{Percentage of Jrnys under recurrent perf. with T.T.} > \textbf{RBT}_{\textbf{Recurrent}}\right)_{\textit{O-D, Time Interval, Sample Period}} \quad [4\text{-}10]\end{aligned}$$

Equations 4-9 and 4-10 are illustrated in Figure 4-8, where the theoretical histograms for the recurrent and the overall travel time distributions are compared.

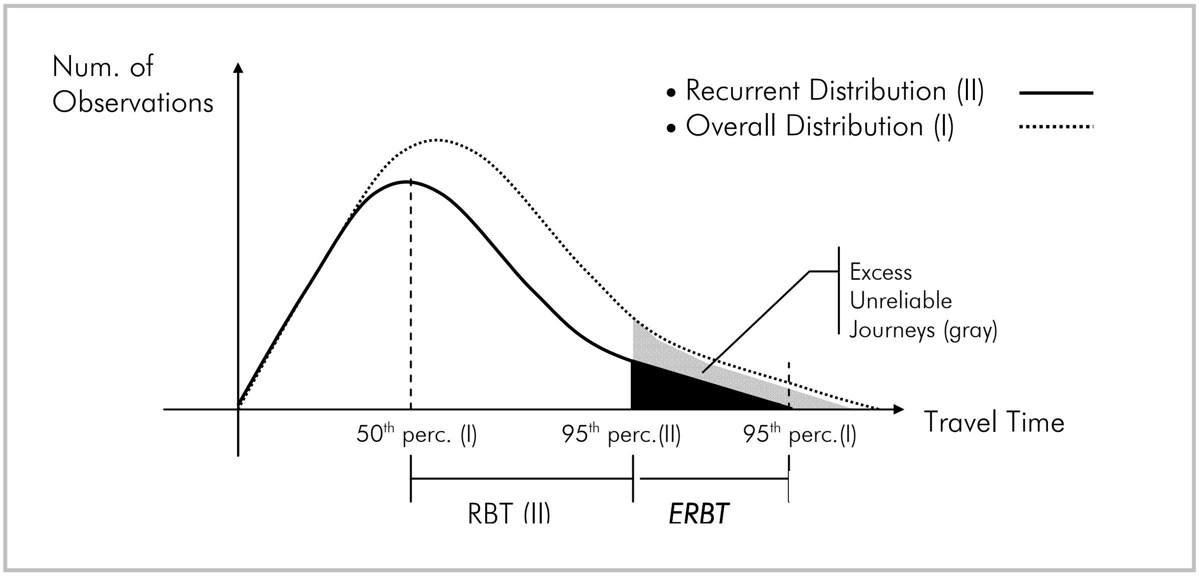

Num. of Observations
• Recurrent Distribution (II)
• Overall Distribution (I)
Excess Unreliable Journeys (gray)
Travel Time
50th perc. (I)
95th perc.(II)
95th perc.(I)
RBT (II)
ERBT

Figure 4-8: Illustration of the ERBT and PUJ

The area shaded in black represents those journeys that by construction experienced unreliable service and were made on days classified to be part of the recurrent performance of the system (i.e. 5% of all journeys on recurrent days). The area shaded in gray is the number of unreliable journeys that occurred in excess of those journeys considered unreliable by construction, and it represents the impact of incidents on the proportion of all trips receiving unreliable service.

4.3.3: Assessment of the Excess Reliability Buffer Time Metric

The proposed ERBT metric addresses some of the main considerations regarding the RBT metric introduced in section 4.2.1. Some of these advantages include:

- **Useful as input to the Evaluation of Performance** – First, by using the performance of the system under undisrupted conditions as the baseline value, the ERBT metric addresses the existence of an "irreducible variability" for every journey by not penalizing the performance for it. Second, the RBT Metric was identified as being highly sensitive to the delays caused by incident-related disruptions. The ERBT metric separates the effects of these disruptions on unreliability from the unreliability associated with recurrent variability (which includes the "irreducible" variability as discussed in section 4.2.3), enabling the estimation of the amount of buffer time added by incidents. In addition, the PUJ measure provides an estimate for the likelihood that a passenger would experience extremely long delays (i.e. unreliable service).

- **Insights into the Causes of Unreliability –** From the outset, the separation of performance as observed through Oyster data into two categories relates directly to the different types of factors that can produce changes in the reliability of the system. Because of the use of demand-side data, it is difficult to establish a one-to-one relationship between the passenger experience and what occurs in the operations of the system. Through the framework, however, it becomes possible to distinguish between the contribution of incidents and the characteristics of the service, making it possible to gain insight into the different causes of unreliability (see section 5.2). The way these factors are separated is also advantageous due to the way they relate to the two categories of strategies introduced in section 2.1.3, useful for improving performance. Namely, factors whose performance occurs on a repeated, predictable basis are more suitable for being addressed by preventive strategies that can be applied over the long-term (i.e. planning), as opposed to incidents and other non-recurring factors which must be dealt with on a more reactive basis (i.e. real-time).

- **Comparability and Aggregation –** The influence of the characteristics of each O-D pair on the variability of travel times could theoretically be partly controlled for through the $RBT_{Recurrent}$ component of the ERBT metric. This could make it possible to compare performance across O-D pairs with different characteristics more fairly, if in fact the ERBT is empirically not sensitive to O-D pair characteristics. In this case, aggregating O-D pair performance would address some of the issues presented in Chan (2007) in her proposal for a line level measure of reliability (see section 2.3.2). However, preliminary analysis for one line in the system indicated the presence of a non-negligible amount of correlation between estimates of ERBT and the length of a journey (as given by the median travel time). This suggests that further research is required to ascertain the degree to which ERBT controls for the effects of the characteristics of a particular journey (see section 7.3).

More broadly, the framework can be used to develop applications directed at improving service reliability. Some applications are developed in Chapter 6 for the particular case of the London Underground. However, the framework is first used to characterize the reliability of the Underground at both the micro and broader levels of analysis in Chapter 5, as well as provide further insight into the causes of unreliability.

Chapter 5: Characterization of AM Peak Performance in the Underground

This chapter demonstrates how the proposed framework can be used to both quantify reliability using AFC data, and to gain insight into the contribution of different factors to unreliability. This is achieved starting in section 5.1 with a quantitative description of reliability as it is viewed by passengers. Here reliability is characterized at a detailed level of analysis by using three O-D pairs to illustrate how the reliability framework is applied. In section 5.2, an analysis is presented to uncover several general causes of unreliability. This latter analysis is applied at a broader level, studying 800 of the highest-volume O-D pairs in the system.

5.1: Framework Application to Three O-D Pairs

This section provides a disaggregate view of reliability by characterizing the performance of three O-D pairs, distinct in their characteristics, over two periods of four weeks. Specifically, section 5.1.1 describes the variability of travel times as directly experienced by passengers. Section 5.1.2 applies the framework to classify performance and shows how, through validation, incident-related performance can be attributed to the occurrence of incidents impacting Underground performance. Finally, section 5.1.3 applies the proposed reliability measures to the three O-D pairs under study in order to illustrate the use of the framework as an input into the evaluation of performance.

5.1.1: Variability of Passenger Travel Times and the Reliability of Service

At the heart of the concept of reliability is the level of consistency or variability of travel times experienced by passengers over time. That is, the lower the variation of the typical time needed to complete a journey, the easier it is for passengers to both develop accurate expectations of the system and modify their behaviour to reduce the disutility of travel (e.g. changes in their time-of-departure), due to the smaller "buffer" time that must be budgeted into one's travel.

The variability of travel times experienced by passengers (i.e. observed from Oyster data) is characterized for three O-D pairs along four dimensions of variability, based on earlier work by Recker et al. (2005) in their study of travel time reliability. These four dimensions are:

- **Individual Traveler Variability:** This type of variability represents the variation in journey times across individual travelers, caused by both differences in their personal characteristics (e.g. walking speed, willingness to board a crowded train), and their interaction with the service (e.g. different arrival times at the platform relative to the next vehicle departure). It can also be thought of the variability of journey times measured across passengers traveling on the same day within a very short time frame, where the service conditions can be assumed to be relatively constant (e.g. 5–15 minutes). Identifying this type of variability is important when evaluating performance because embedded in it is the "irreducible" variability inherent in any O-D pair's travel time distribution that results from the discrete nature of transit service.

- **Within-day Variability:** Variation in the journey time distribution *within* a particular day is also of concern, as it recognizes that either the schedule service is not uniform throughout the day, or that time-of-day specific conditions (e.g. crowding) could have an impact on the distribution of travel times experienced by passengers. The larger the variation of the passenger experience within a day, the more important it is to measure reliability over shorter periods of time in order to accurately represent the concerns of passengers traveling at different times during a day.

- **Day-to-day Variability:** The day-to-day variability relaxes the idea of constant service conditions and instead examines the variation in passenger travel times that arise because of both the stochastic nature of the operation, as well as the influence of day-specific factors like the weather. This type of variability is important because it helps determine passengers' long-term expectations of the service, and therefore their ability to make optimal travel choices (e.g. decisions on the departure time, route, and mode). It is also important from an operational perspective because it reflects the ability to consistently carry out the planned service.

- **Seasonal Variability:** Differences in travel times across longer periods of time (e.g. month-to-month) are important to take into account when measuring reliability first because they provide a sense for what the appropriate level of aggregation over such periods should be, and second because controlling for them helps account for adjustments made by passengers when developing their expectations of the service (e.g. a passenger might adjust their view of the service during rainy season).

Within these types of variability, it is important to distinguish between the variability of "typical" or average journey times, and the variability of the second moment (e.g. variance) of the distributions. Previous studies like that of Recker et al. (2005) have focused on looking at the variability of average journey times. This study examines the variability of the median travel times for a journey, as well as the way the variance of the travel time distribution changes over the four dimensions of variability identified above.

Description of Three Origin-Destination Pairs

The three origin-destination pairs selected for this initial characterization of the reliability of the Underground represent some of the highest volumes of Oyster journeys (i.e. journeys made on Oyster cards) in the system during the morning peak period. These journeys were also selected based on their varying characteristics, in order to gain insight into the different causes of unreliability (more formally explored in section 5.2).

Tables 5-1-a,b,c describe the characteristics of each O-D pair, situating it within the broader Underground network. The five characteristics shown for each journey include the line and direction for the trip, the average scheduled headway during the AM Peak[13], the number of stops as a proxy for journey length, the rank of the O-D in terms of the total volume of passengers carried during the morning peak[14], and the percentage of all journeys for that O-D pair using an Oyster card[15]. The location of each O-D pair within the Underground network is indicated within the system map (see Appendix A).

Table 5-1-a: O-D pair # 1 – Waterloo to Canary Wharf

• **Line-Direction**	Jubilee line – Eastbound
• **AM Peak Average Scheduled Headway [min]**	2-5
• **No. of Stops**	5
• **Rank (Daily AM Peak Journeys)**	2 (7961)
• **Oyster Use [% trips]**	28%

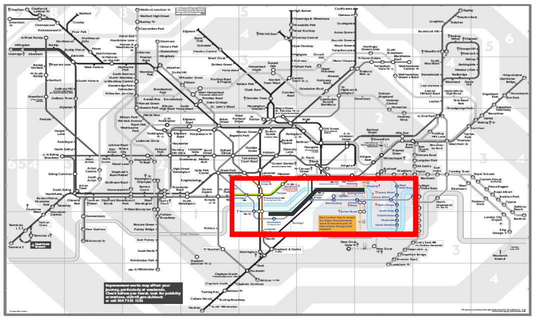


Table 5-1-b: O-D pair # 2 – Victoria to S. Kensington

• **Line-Direction**	District and Circle lines – Westbound
• **AM Peak Average Scheduled Headway [min]**	2-3
• **No. of Stops**	1
• **Rank (Daily AM Peak Journeys)**	22 (1316)
• **Oyster Use [% trips]**	43%

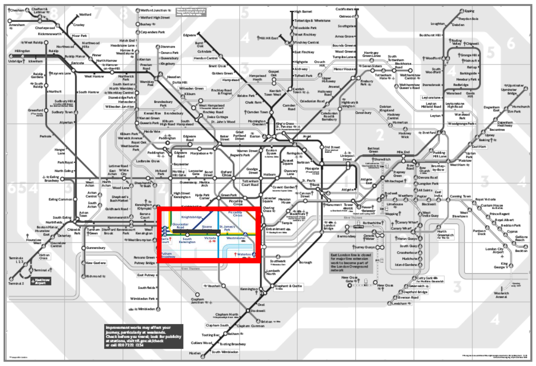
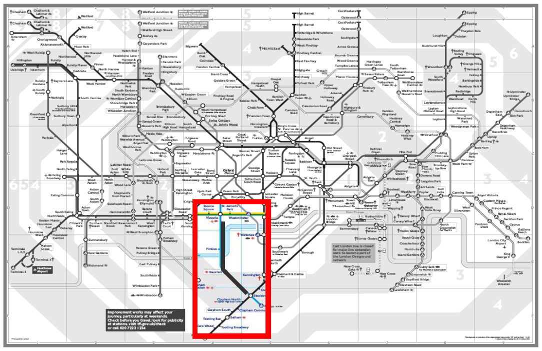


Table 5-1-c: O-D pair # 3 – Tooting Broadway to Victoria Station (Interchange @ Stockwell)

• **Line-Direction**	Northern & Victoria lines - Northbound
• **AM Peak Average Scheduled Headway [min]**	2-5
• **No. of Stops**	9
• **Rank (Daily AM Peak Journeys)**	517 (232)
• **Oyster Use [% trips]**	81%



Each O-D pair represents a different combination of the five characteristics described earlier. For example, the first two journeys occur within a single line, whereas the third requires an interchange[16]. These three journeys also range in length, from a 1-stop journey with a typical

[13] Data source: TfL Published Timetables (http:// tfl.gov.uk) – October 2008.

[14] Data from AM Peak O-D Matrix 2006 (Chan, 2007).

[15] Percentage of Oyster journeys approximated using the Oyster volumes from Feb. 16th, 2007.

[16] Due to the layout of the journey, virtually all passengers traveling on this O-D pair are expected to use this single path.

travel time of 7 minutes (O-D pair # 2) to a 9-stop journey with a typical travel time of 30 minutes (O-D pair # 3). Finally, four different lines are represented through these O-D pair examples.

Passenger Travel Time Variability – 50thPercentile

Quantifying the consistency of the "typical" travel time across the different levels of temporal aggregation is important for two reasons. First, this part of the reliability measure represents the performance that passengers can come to regularly expect over time. Therefore, it is necessary to understand its variability (or lack thereof) in order to have a sense for how "typical" it really is. Second, the median travel time is at the heart of the proposed measures of reliability, and its stability over time is of concern so that conceptually the reliability buffer time remains appropriate (i.e. the buffer time *above* the typical time required to complete a journey).

The median travel time for all journeys during the AM Peak for the three O-D pairs is compared over 20 weekdays in February 2007 and summarized in Table 5-2 to reflect the day-to-day variability experienced by passengers.

Table 5-2: Day-to-day variability of the median travel time – AM Peak, Feb. 2007

	50th Percentile Travel Time [min]		
Date	OD # 1	OD # 2	OD # 3
Mon, 5-Feb	20	7	27
Tue, 6-Feb	20	7	26
Wed, 7-Feb	18	7	27
Thu, 8-Feb	24	7	29
Fri, 9-Feb	18	7	26
Mon, 12-Feb	18	7	28
Tue, 13-Feb	20	7	26
Wed, 14-Feb	18	7	27
Thu, 15-Feb	17	7	25
Fri, 16-Feb	17	7	32
Mon, 19-Feb	18	8	25
Tue, 20-Feb	17	7	25
Wed, 21-Feb	19	7	26
Thu, 22-Feb	18	7	25
Fri, 23-Feb	17	7	26
Mon, 26-Feb	18	7	27
Tue, 27-Feb	18	8	27
Wed, 28-Feb	18	8	25
Thu, 29-Feb	22	7	25
Fri, 1-Mar	22	8	25
Median T.T.	**18**	**7**	**26**
% Days w/in 1 min	**70%**	**100%**	**85%**

The median journey time for each O-D pair for the entire four-week period is shown in the second to last row in Table 5-2, representing the long-run "typical" travel time that could be expected by passengers. The last row is the percentage of days that fall within +/- 1 minute of the long-run median travel time. As expected from the discussion in the previous chapter, the daily median travel time is relatively stable on a day-to-day basis, with the percentage of observations falling within 1-minute of the long-run median travel time ranging from 70% to 100%. This suggests that the indicator for the "typical" travel time used in the proposed reliability metrics is in fact typical for the origin-destination pairs examined, including one requiring an interchange.

The consistency of the median travel time within a day is also explored by comparing its variation both across the hours of the morning peak period, as well as the day-to-day consistency within each of these hours. Table 5-3 summarizes this for all weekdays in February 2007.

Table 5-3: Within-day variability of the median travel time – AM Peak, Feb. 2007

	50th Percentile Travel Time [min]					
	OD # 1		OD # 2		OD # 3	
Hour within AM Peak	Median T.T.	% Days w/in 1 min	Median T.T.	% Days w/in 1 min	Median T.T.	% Days w/in 1 min
7:00-8:00am	17	80%	7	95%	25	95%
8:00-9:00am	19	55%	7	100%	27	80%
9:00-10:00am	18	75%	7	100%	25	70%

Table 5-3 shows how for the journey between Victoria station to South Kensington, the typical travel time remained unchanged throughout the morning peak. This was different from the performance of the remaining journeys, where there the median travel time increased during the peak hour of service, possibly due to the effects of demand peaking. Another important pattern visible for these O-D pair is the relationship between the day-to-day consistency of the median journey time (as shown by the % of Days w/in 1-min of the median for the period), and the within-day consistency. Specifically, higher variation across days for the typical travel time was accompanied by a higher variation across the three hours of the morning peak, indicating a less reliable service overall.

The presence of seasonal variation in the median journey time was also explored by comparing the performance between two four-week periods in February and November 2007 for each of the three O-D pairs. Table 5-4 shows this, and also indicates the day-to-day variability of the median travel time within each four-week period.

Table 5-4: Seasonal variability of the median travel time – AM Peak, Feb. & Nov. 2007

	50th Percentile Travel Time [min]			
	February		November	
O-D Pair	Median T.T.	% Days w/in 1 min	Median T.T.	% Days w/in 1 min
O-D # 1	18	70%	19	90%
O-D # 2	7	100%	7	100%
O-D # 3	26	85%	27	80%

The values in Table 5-4 do not suggest strong seasonal variability, with long-run median journey times being similar between the two four-week periods examined, accompanied by minor changes in the day-to-day variability for each of the periods (as shown by the percentage of days within 1 minute). This can be interpreted as saying that even though there is some variability across days, which itself is similar across different periods, the long-term median travel time is stable and therefore serves as a good indicator for the "typical" journey time.

In addition to the median journey time, another key aspect of performance is the shape and spread of the travel time distribution, as well as its consistency over time. The degree to which the spread of the distribution varies over the four dimensions of variability is characterized next.

Passenger Travel Time Variability – Reliability Buffer Time

It is important to understand the variability of the 95th percentile travel time from day-to-day in order to determine the stability of the spread of the distribution over time. A higher variation of the 95th percentile across days would lead to a higher RBT required of passengers in the long-run to account for those days with high levels of delays. To get a sense for the how changes in the travel time distribution across days would affect the long-run RBT, the difference in the 95th percentile for each day and the long-run (i.e. 20-weekday) median travel time is estimated and compared to the RBT for the four-week period. Table 5-5 summarizes these values for the three selected O-D pairs, and estimates the coefficient of variation (CV) for the difference between the daily 95th percentile and the long-run 50th percentile travel times. In addition, the last row in the table normalizes the RBT for the period by the median travel time, which could be thought of as a proxy for the length of the journey. This would help account for differences in perception where passengers completing a long journey might view a certain buffer time as less severe than that same level of variability being experienced for shorter journeys. In addition, initial insight into the relationship between variability and trip distance can be gained.

Table 5-5: Day-to-day and individual traveler variability for the RBT (absolute and normalized by distance) – AM Peak, Feb. 2007

	95th Percentile Day - 50th Percentile Period [min]		
Date	OD # 1	OD # 2	OD # 3
5-Feb	7	4	6
6-Feb	13	3	7
7-Feb	3	5	4
8-Feb	36	5	6
9-Feb	7	5	3
12-Feb	3	5	6
13-Feb	11	5	3
14-Feb	7	4	12
15-Feb	4	5	7
16-Feb	4	4	12
19-Feb	3	5	5
20-Feb	3	5	5
21-Feb	6	4	4
22-Feb	3	5	7
23-Feb	3	4	7
26-Feb	4	2	8
27-Feb	6	5	6
28-Feb	9	6	4
29-Feb	12	6	5
1-Mar	10	10	7
Period RBT	9	5	8
CV(95th – 50th)	0.96	0.32	0.39
Period RBT/50th	50%	71%	31%

From Table 5-5 it is possible to capture the way the spread of the travel time distribution varies from day to day. In the case of O-D pair # 1, the standard deviation represents around 96% of the average daily difference in the 95th percentile and the long-run median travel time. This variation over time was less severe for the remaining O-D pairs, which tended to have a more stable travel time distribution, implying that in the short-run it is easier for passengers to predict the necessary buffer time in order to arrive on-time at their destination with 95% certainty. The last row in Table 5-5 captures another aspect of the story, representing the buffer time as a proportion of the median travel time for each O-D pair. It shows how in the case of O-D pair # 1, a passenger would need to increase the median travel time by a factor of 1.5 in order to be 95% certain of on-time arrival while traveling on any given day. This is lower than the factor of 1.71 required of passengers on O-D pair # 2, where because of the short duration of the journey (i.e. median travel time of 7 minutes), a 5 minute buffer time would seem more onerous than a 9 minute buffer time for an 18-minute journey.

Referring back to the idea of individual traveler variability, even when the characteristics of the service remain unchanged (e.g. during a very small time-interval such as 30 minutes during one day), differences in travel times across individual travelers will still be observed. Using the journeys from O-D pair # 1 made on February 19th, a day with the minimum 95th percentile travel time (see Table 5-5), during the 30-minute interval after 8:00am, the variability across individual passengers can be captured. Figure 5-1 graphically represents this distribution.

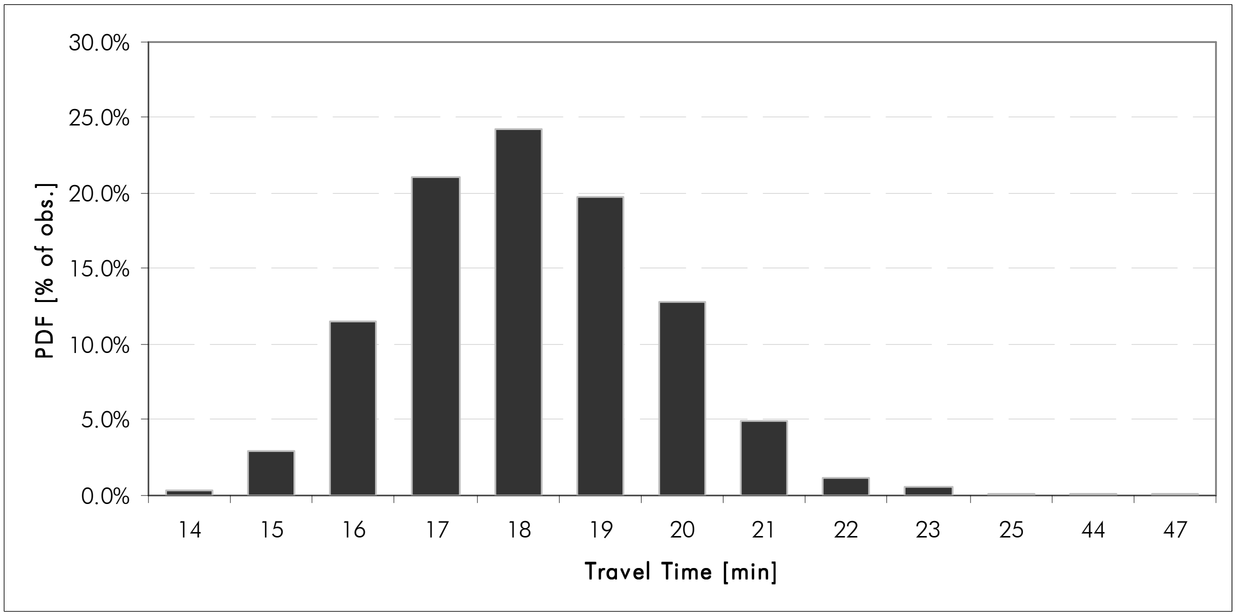


Figure 5-1: Travel time distribution for Waterloo to Canary Wharf – 8:00-8:30am, Feb. 19th, 2007

The distribution in Figure 5-1 alludes not only to the variability of travel times under recurrent performance, but also to the irreducible level of travel time variability caused by the discrete nature of transit service. In this particular example, the difference between the 95th percentile and the median travel time of 3 minutes reflects the scheduled headway for this journey of 2-3 minutes (i.e. passengers arriving randomly would sometimes wait the full headway).

In addition, the variation of the RBT *within* a day over a 20-weekday period is also quantified. Table 5-6 captures this for each hour within the AM Peak, and the proportion of the median travel time that each RBT represents.

Table 5-6: Within-day variability of the RBT – AM Peak hourly, Feb. 2007

	Reliability Buffer Time [min]					
	OD # 1		OD # 2		OD # 3	
Hour within AM Peak	RBT	RBT/50th	RBT	RBT/50th	RBT	RBT/50th
7:00-8:00am	8	*47%*	4	*57%*	6	*24%*
8:00-9:00am	9	*47%*	5	*71%*	9	*33%*
9:00-10:00am	7	*39%*	5	*71%*	7	*28%*

The changes in the Reliability Buffer Time across the hours of the morning peak suggest that there could be a non-negligible effect of demand congestion on performance, as observed for O-D pairs # 1 and 3. This difference between the RBT over the three morning period hours also underlines the need to measure reliability at finer levels in order to capture the actual experience of passengers departing at different times of the day.

Lastly, the consistency of the Reliability Buffer Time during the two four-week periods in February and November, 2007 are compared to gain insight into the long-term consistency of the variability of travel times, and any possible changes due to seasonality. Table 5-7 summarizes, for each of the three O-D pairs, the magnitude of the RBT during each four-week period, as well as the variability of the difference in the 95th percentile for each day and the long-run 50th percentile travel time.

Table 5-7: Seasonal variability of the RBT – AM Peak, Feb. & Nov. 2007

	RBT (PD 95th - PD 50th Perc.) [min]					
	February			November		
Day-to-Day Variability	RBT	RBT/50th	CV (95th – 50th)	RBT	RBT/50th	CV (95th – 50th)
O-D # 1	9	50%	*0.96*	6	32%	*0.79*
O-D # 2	5	71%	0.32	5	71%	0.32
O-D # 3	8	31%	0.56	9	33%	0.85

As expected, the performance of each O-D pair differed when comparing the RBT of each four-week period. For journeys between Waterloo and Canary Wharf (O-D pair # 1) the buffer time differed across both periods, reflecting the high level of variability of the spread of the distribution over time, also observed earlier in Table 5-5 and Figure 4-3. The opposite was true for the performance of O-D pair # 2, with similar levels of reliability in February and November.

More importantly, however, is the positive correlation that can be observed in Table 5-7 between the RBT for a period and the level of day-to-day variability of the 95th percentile travel time. This relationship could be explained in part by the occurrence of incident-related disruptions, which would not only increase the variability of the spread of the distribution across days, but also increase the overall buffer time for passengers. The unpredictable nature of incidents could also help explain the differences in the buffer time from month-to-month, leading one to conclude that in order to ascertain the existence (or absence) of any seasonal variability in the spread of the travel time distribution, the recurrent reliability should be compared.

Through this preliminary characterization of reliability, the importance of focusing not only on the predictability of the typical travel time for a journey, but of the consistency of the spread of the travel time distribution becomes clear. The performance of the selected O-D pairs showed that even though the median journey time was consistent over the four dimensions of variability looked at, the spread of the distribution was not, highlighting the weakness of existing reliability measures that focus solely on average performance.

5.1.2: Classification of Performance & Validation of Incident-related Reliability

This section applies the reliability framework to the three selected origin-destination pairs in order to illustrate how their reliability can be broken down into recurrent and incident-related performance. Additionally, incident log data from the Underground is used to validate the incident-related performance category.

Classification of Performance – Recurrent and Incident-Related Delays

The framework is first use to classify the performance of journeys on O-D pair # 1 by identifying those days that exhibited a high level of non-recurrent delays relative to the performance observed during the remaining weekdays in February 2007. These days were treated as candidates for having suffered disruptions, and were classified as belonging to the incident-related category.

The second step of the application of the framework (see section 4.2.2 – "*Outline of Classification Approach"*) estimates the travel time distribution for each weekday's AM Peak has along with the 95th percentile, which is used to represent the level of delays experienced by passengers on that particular day. Each day is then classified using the stepwise regression approach (see Appendix B), which uses the 95th percentile travel time for each day to assign each level of performance into the recurrent and incident-related categories. Figure 5-2 illustrates the results of this classification approach for the first O-D pair.

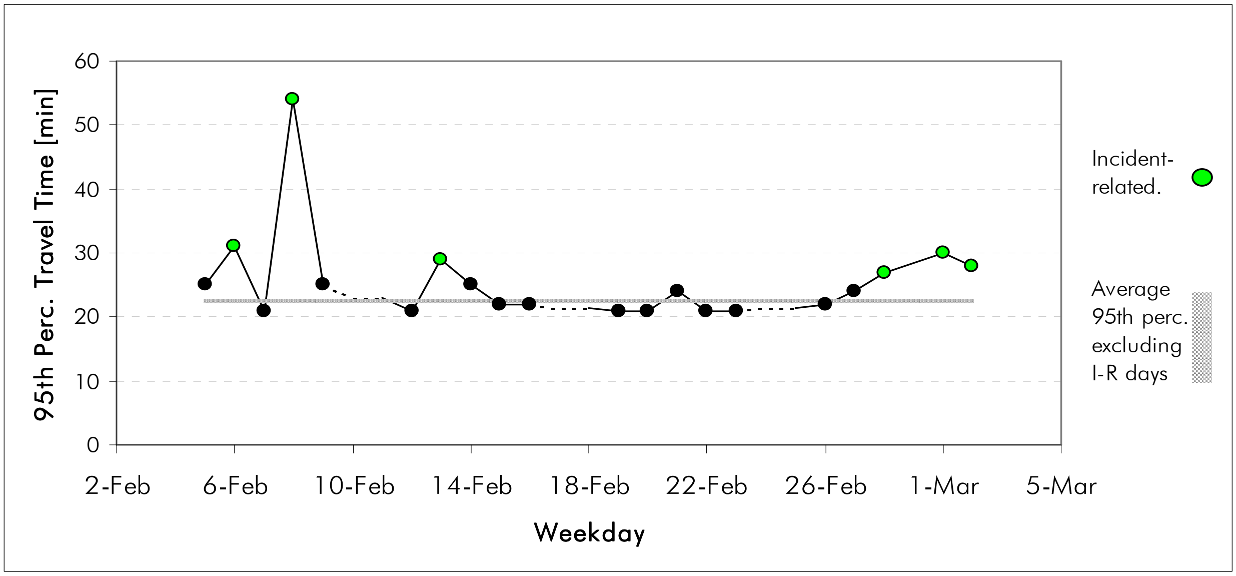


Figure 5-2: Classification into recurrent and incident-related performance – O-D pair # 1, AM Peak, Feb. 2007

The peak periods indicated in (lighter) **green** represent those days exhibiting a 95th percentile travel time that was classified as incident-related performance at the 5% significance level using the stepwise regression approach. This can simply be interpreted as saying that there is a very small chance (less than 5%) that the particular observation classified as being incident-related was not due to that observation coming from a completely different underlying travel time distribution (i.e. different performance conditions) when compared to the performance exhibited by the remaining days (majority).

Throughout this study, the 5% significance level was used to classify performance as it provided a reasonable level of sensitivity towards identifying candidate days whose service suffered disruptions caused by incidents. Figures 5-3 and 5-4 show the classification of each weekday-AM Peak during February 2007 for the other two O-D pairs. For the second and third cases, there were fewer days classified as belonging to non-recurrent performance conditions.

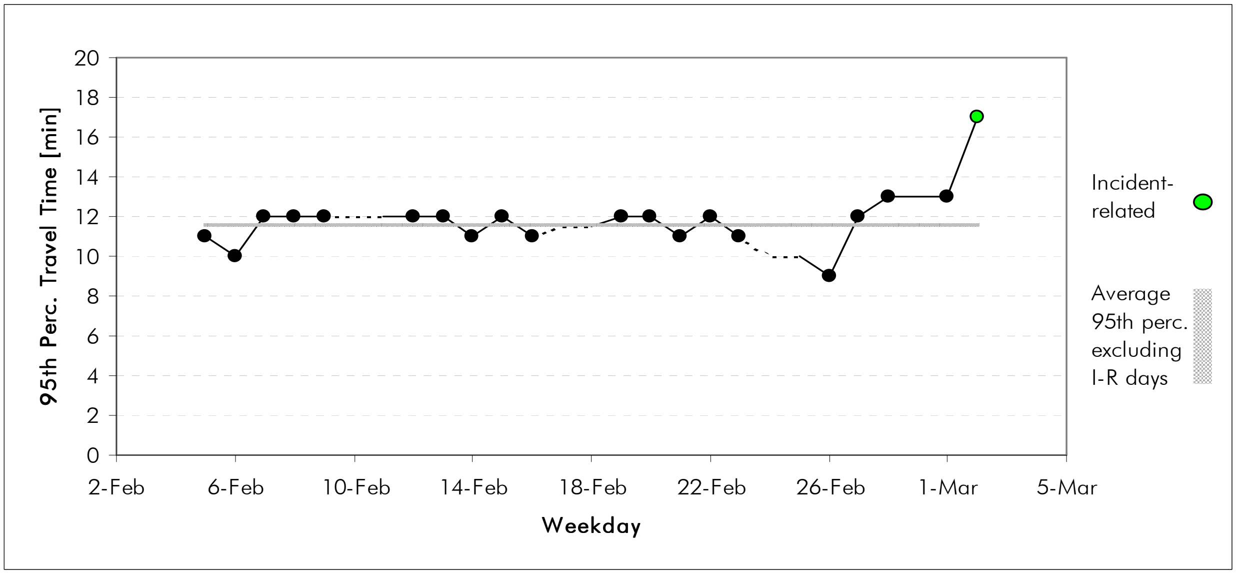


Figure 5-3: Classification into recurrent and incident-related performance – O-D pair # 2, AM Peak, Feb. 2007

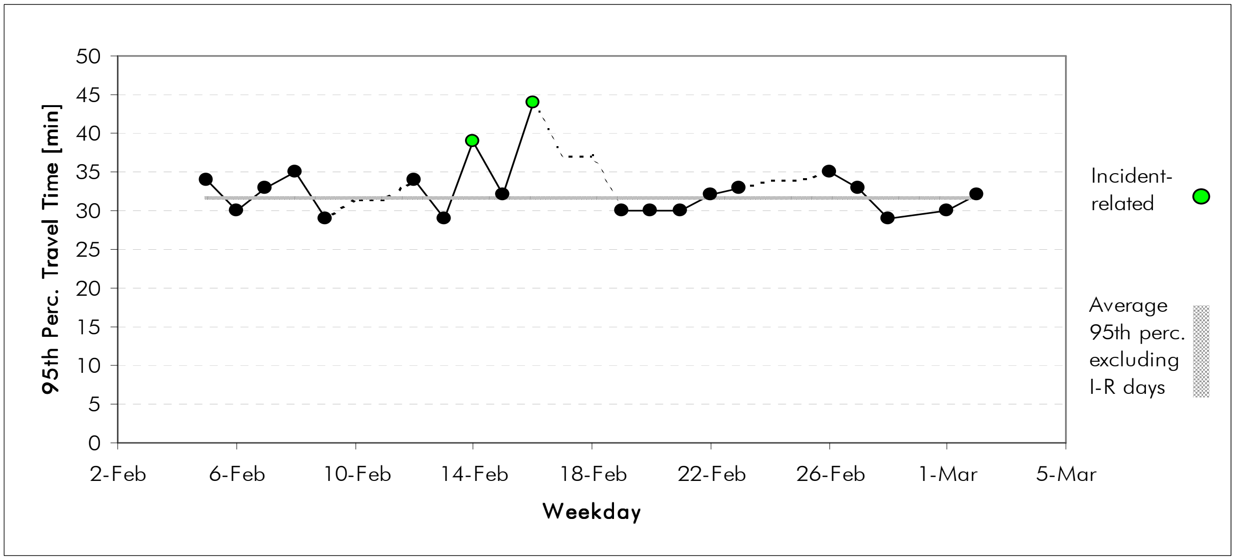


Figure 5-4: Classification into recurrent and incident-related performance – O-D pair # 3, AM Peak, Feb. 2007

Given that the classification of peak periods into recurrent and non-recurrent performance was determined based solely on Oyster passenger data, it is impossible to know for certain whether the changes in performance were in fact caused by incidents. The following section attempts to validate this approach by examining the Underground's incident log data for the same days.

Non-recurrent Performance and Incident Log Data Validation

Underlying the classification approach used as part of the reliability framework was the hypothesis that incident-related disruptions in the system were the primary cause of large and non-recurrent delays. Because the approach thus far has relied solely on Oyster passenger data, those days classified as incident-related were treated as *candidates* for having suffered disruptions. Using data from incident logs, the relationship between incident-related (i.e. non-recurring) performance and the actual occurrence of disruptions in the system is validated.

The incident log data are obtained from the NACHs system, which is used by the Underground to estimate the level of passenger delays caused by incidents (see section 3.3.1). This data source has multiple fields of detailed information for each incident, including the date (Date), time-of-day the incident first started (Start Time), the line it occurred on (Line), the stop or segment where it happened (Location), the cause of the disruption (Cause), the effect on service it had (Result), and finally the resulting estimated level of delays incurred by passengers in units of hundreds of passenger-hours (Indicative NAX). This source of data is used alongside the information obtained from Oyster to understand what led to the large delays observed on some of the non-recurrent performance periods for each of the three selected O-D pairs.

The incident log data are first compared with the observed Oyster journeys for O-D pair # 1, in order to illustrate how incident-related performance can be the result of disruptions in the system caused by incidents. Figure 5-2 identified 6 days as being incident-related at the 5% significance level (from the highest to lowest level of delays observed through the 95th percentile): February 8th, 6th, 13th, March 2nd and 1st, and February 28th.

Figure 5-5 shows the duration for each individual trip made throughout the morning peak by time of entry into the system on the day with the highest level of delays, February 8th. On this particular day, passengers experienced up to an hour of travel time, or around 40 minutes above the median travel time under recurrent conditions. Also, very erratic journey durations seemed to take place before 9:00am, both in terms of no entries into the system occurring for short time intervals (as can be seen by the "gaps" in Oyster journeys), and their high degree of variability of travel times once they entered the system. These effects are summarized in the plot on the upper-right corner, showing the travel time probability density function for that peak period.

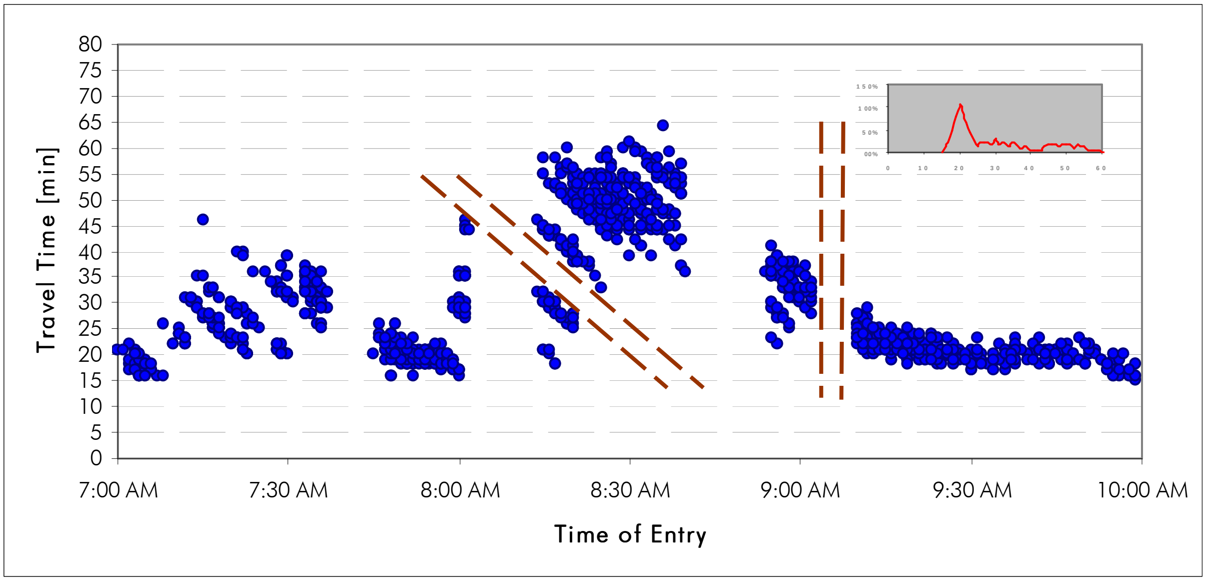


Figure 5-5: Individual Oyster journey times during the morning peak on O-D pair # 1 – February 8th, 2007 & the corresponding travel time distribution

The passenger travel times shown in Figure 5-5 can also be read by paying attention to two types of breaks or discontinuities, examples of which are indicated in the figure. First there are the diagonal gaps between the different passenger journeys. These can be explained as the result of long headways, where for passengers that entered the station at the same time (horizontal axis value held constant), one discrete group of people experienced a lower travel time than the remaining group. The group that experienced the lower travel time was plausibly able to board the service available when they arrived at the platform (either because of faster walk times or because of their ability to board a crowded train), whereas the remaining group had to wait a long headway (represented by the size of the gap) for the next service. This type of discontinuity is also observed on some of the other incident-related days shown next. The second type of discontinuity has to do with passengers not entering the station for a certain period of time (vertical gaps). This can be explained by the temporary closure of the station gatelines (e.g. to avoid crowding on the platform), and is less common on the other days with incident-related performance.

Listed in the incident log for the Jubilee line, there were numerous entries referring to extreme weather conditions that took place earlier that morning. As a result, several parts of the system were severely affected, leading to high levels of passenger delays. Table 5-8 summarizes the entries found in the incident log obtained from the NACHs system for February 8th.

Table 5-8: Summary of the incident log for the Jubilee line – AM Peak, February 8th, 2007

Date	Start Time	Line	Location	Cause	Result	Indicative NAX
8/2/2007	8:32 AM	JUBILEE	Bond St	Fleet – Defective in Service	Train Delays	84.9
8/2/2007	9:00 AM	JUBILEE	Neasden Depot	Extreme Weather - Snow & Ice	Train Cancellations	94.7*
**10 Similar entries throughout the morning peak - Ind. NAX shows total weather-related delays*						

The "Indicative NAX" field in this case can be used to get a sense for the impact of the disruption(s) on passenger travel times relative to the impact of other incidents. For the month of February 2007, the average total Indicative NAX for each day during the AM Peak was 44.1, showing that relative to the typical level of incidents, February 8th was one of the worst. Table 5-8 also shows that under certain circumstances one cannot attribute the abnormal performance experienced by passengers to a single particular incident. In this case, a primary cause of the observed delays was the general weather conditions of that particular day rather than the occurrence of any specific event.

Other delays on incident-related days can also be attributed to the information on the state of the operation from incident log data. Figure 5-6 and Table 5-9 show these two sources of information together for February 13th.

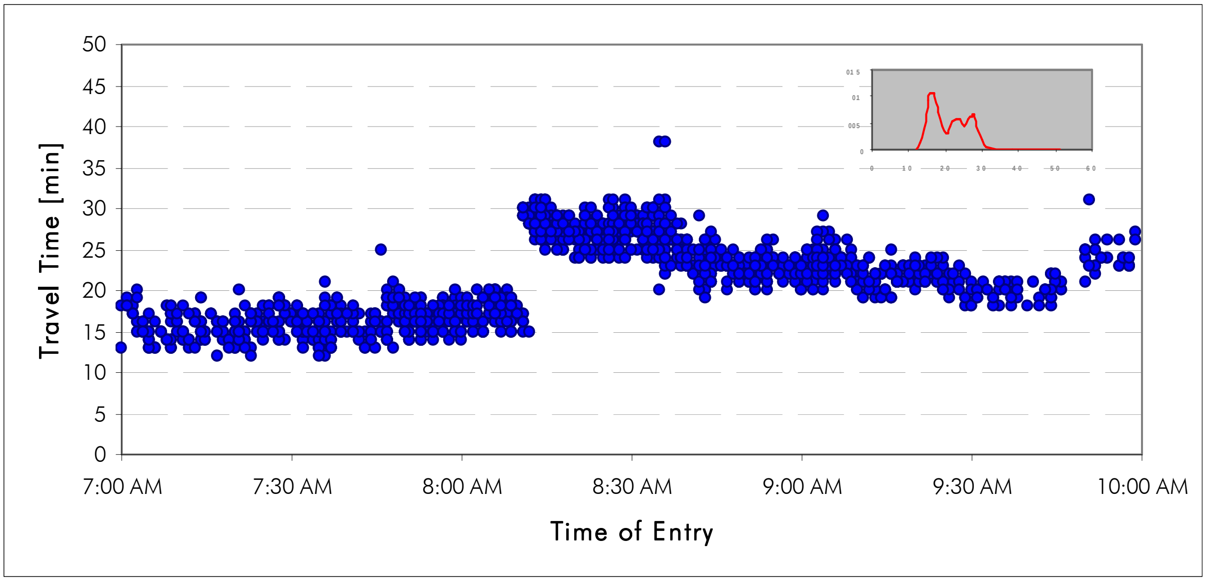


Figure 5-6: Individual Oyster journey times during the morning peak on O-D pair # 1 – February 13th, 2007 & the corresponding travel time distribution

Table 5-9: Summary of incident log for the Jubilee line – AM Peak, February 13th, 2007

Date	Start Time	Line	Location	Cause	Result	Indicative NAX
13/2/2007	7:01 AM	JUBILEE	Wembley Park	Fleet - Defective in Service	Train Delays	2.4
13/2/2007	8:06 AM	JUBILEE	Canary Wharf	Customers - Crowding	Train Delays	3.5
13/2/2007	8:26 AM	JUBILEE	North Greenwich	Track Power Failure	Train Delays	25.2
13/2/2007	9:32 AM	JUBILEE	London Bridge	Customers - Disruption	Train Delays	2.1

From the Oyster journey times plotted in Figure 5-6, the major disruption on the Jubilee line affecting passengers travelling from Waterloo to Canary Wharf at approximately 8:10am becomes apparent, given the sudden increase in journey times experienced by all passengers.

These changes in the passenger experience can be attributed to some of the incidents logged for this particular morning peak. For example, the third entry in Table 5-9 indicates a loss of traction current near Waterloo and Canary Wharf resulting in a relatively large delay as seen by the Indicative NAX column (the average delay size for each incident during this period for the Jubilee line was 8.2 NAX units). The abnormal journey times during the remainder of the morning peak could be related to this large incident, as well as smaller ones like those shown in the other entries in Table 5-9.

It is also important to note from the comparison for trips on February 13th that there is not a one-to-one relationship between the events logged throughout the day and the changes in travel time observed through Oyster data. That is, there are incidents occurring on a daily basis in the Underground with some having a larger impact than others. In the case above it is more difficult to attribute specific changes in Oyster travel times to smaller incidents, because they could easily be masked by other sources of uncertainty in travel time, despite still having a disruptive effect on performance. Nevertheless, by looking at individual journeys through Oyster data, it is still possible to visually identify passengers delays that are not consistent with the performance expected under normal conditions, and incident log data provides insight into the causes of these changes.

A similar comparison for February 28th illustrates the case when a small number of large incidents took place, making it easier to relate them to changes in performance observed through Oyster. Figure 5-7 shows how passengers experienced relatively stable travel times during the morning peak until around 8:30am, where all journey times gradually increased until about 9:45am, at which time the operation began to recover.

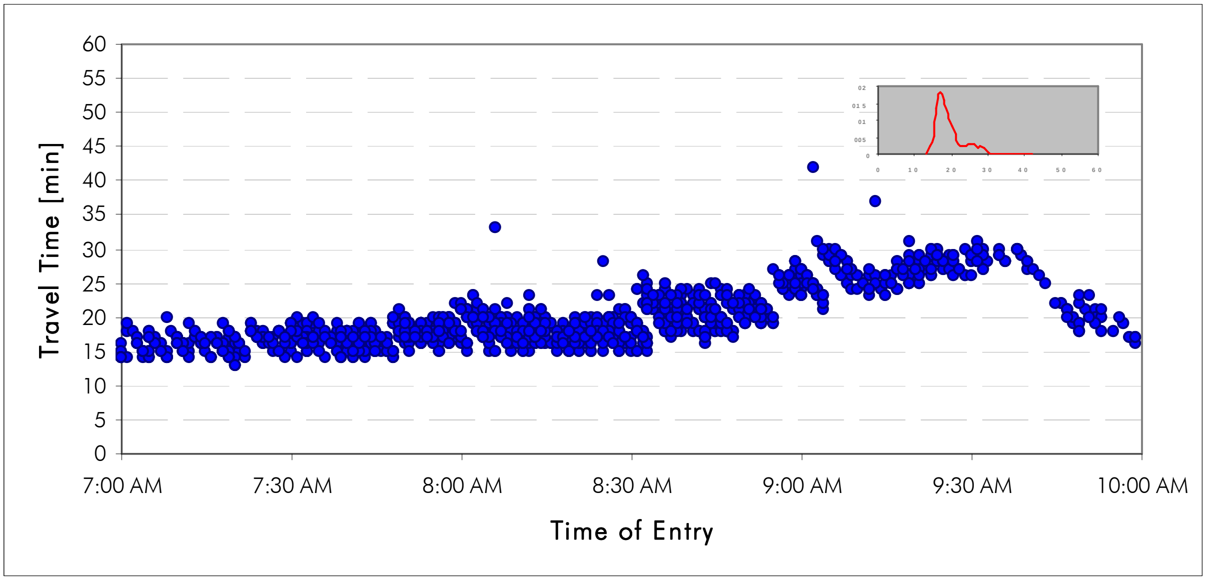


Figure 5-7: Individual Oyster journey times during the morning peak on O-D pair # 1 – February 28th, 2007 & the corresponding travel time distribution

Table 5-10 shows the only entry recorded in the incident log for that day during the morning peak. There is a strong correspondence between that entry and what was observed in Figure 5-

7, with the time of occurrence of the incident (8:46am) being similar to the time when passengers began to experience delays.

Table 5-10: Summary of incident log for the Jubilee line – AM Peak, February 28th, 2007

Date	Start Time	Line	Location	Cause	Result	Indicative NAX
28/02/2007	8:46 AM	JUBILEE	Baker Street	Signal Failure	Signal Failure(s)	156.2
This was the only recorded Incident for this day - AM Peak						

In addition to performing validation of incident-related peak periods for a same-line O-D pairs, the framework is also effective when applied to journeys involving more than one line. The performance observed from Oyster for O-D pair # 3, which involves a transfer from the Northern line to the Victoria line at Stockwell station, is validated with incident log data. Figure 5-8 shows the travel times experienced by passengers completing the entire journey on February 16th, an incident-related peak period, from Tooting Broadway to Victoria station, the trips starting at Tooting Broadway and ending at the transfer station (Stockwell), and the trips for passengers starting at the transfer station and ending at Victoria station.

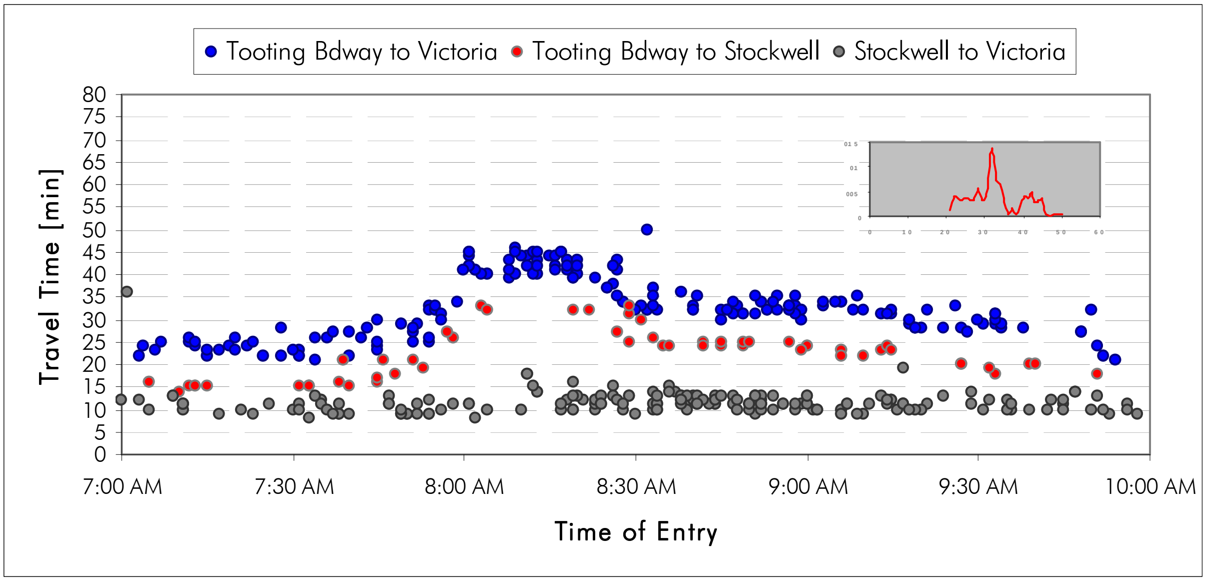


Figure 5-8: Individual Oyster journey times during the morning peak on O-D pair # 3 and journey legs – February 16th, 2007 & the corresponding travel time distribution for the full journey

Figure 5-8 shows how journeys travelling from Stockwell to Victoria station experienced nearly constant travel times throughout the morning peak, indicative of disruption-free service during that day for the Victoria line. However, journeys travelling on Northern line from Tooting Broadway to Stockwell experienced an increase in travel times around 8:00am, suggesting the occurrence of incident-related disruptions during that day. Entries from the incident log for the Northern line, shown in Table 5-11, supports this claim, also helping to explain the classification of journeys on O-D pair # 3 as belonging to incident-related performance.

Table 5-11: Summary of incident log for the Northern line – AM Peak, February 16th, 2007

Date	Start Time	Line	Location	Cause	Result	Indicative NAX
16/02/2007	7:15 AM	NORTHERN	Morden	Staff Uncovered	Train Delay	0.6
16/02/2008	7:18 AM	NORTHERN	Kennington	Customer - Emergency Alarm	Train Delay	6.3
16/02/2009	7:18 AM	NORTHERN	Morden	Staff Uncovered	Train Cancellation	0.2
16/02/2010	8:16 AM	NORTHERN	Oval	Fleet - Defective	Train Delay	81.5
16/02/2011	8:36 AM	NORTHERN	Borough	Fleet - Defective	Train Delay	60.6

The fourth entry in Table 5-11 reveals that a train was detained at Oval station (one stop beyond the transfer station, Stockwell) for over 13 minutes, leading to severe train delays also reflect in passenger travel times from Oyster. This validation raises additional questions regarding the effect of interchanges on reliability, including whether journeys on two or more lines are more susceptible to severe disruptions, given that only one line needs to be disrupted in order to affect the entire journey. Conversely, one could ask whether there is a dampening effect on the impact of incidents on delays because only a portion of the trip is affected. An initial exploration of the effect of transfers on reliability is presented in section 5.2.

Despite the difficulty in establishing a one-to-one relationship between incident log data and passenger travel time data from Oyster, it is still possible to see how the incident-related performance category identified through the framework corresponds to the actual occurrence of incidents. This direct comparison, however, also raises three types of issues related to the classification of performance based on the observed travel times of passengers. First is the issue of erroneously classifying a day that experienced one or more incidents as being part of the recurrent category. This could happen if the impact of the incidents were small enough so as to not cause the level of delays to vary significantly from that of the remainder of the days. Preliminary validation of this error shows that incidents occurring on recurrent days were minor in terms of passenger delays. This is partially a consequence of using the 95th percentile as our indicator for delays as part of the classification process (i.e. a highly sensitive indicator of delays). The second issue has to do with where one draws the boundaries when quantifying the delays related to particular incidents. The estimates of passenger delays from NACHs take into account any "ripple" effects in the system spatially and over time. In this case passenger delays are estimated using Oyster data for a specific part of the system (i.e. O-D pair) and during a specific time interval (i.e. AM Peak), making it difficult to directly connect the level of delays as shown by the Indicative NAX column in the incident log data, and the level of delays for an O-D pair as shown by Oyster. Thirdly, there is the issue of the accuracy of NACHs *estimates* of delays, and how they correspond to an actual *measurement* of delays, with the expectation that by the nature of the data, the latter would represent the passenger experience more accurately. Nevertheless, this initial validation with incident log data strongly supports the incident-related performance category estimated through the proposed Oyster-based reliability framework. The next section draws a connection between the categories of performance used in the framework and the mixture-model intuition described earlier in section 4.2.3.

As was shown in the previous chapter (see Figure 4-6), the overall distribution over several days can be taken as the probabilistic combination of those journeys made on peak periods where the performance of the system followed routine conditions, and those peak periods when passengers experienced the effects of severe disruptions. Each distribution can be interpreted differently, starting with the overall travel time distribution as representing the actual passenger experience over the long-run, regardless of the causes of unreliability in the system. The shape and nature of the overall reliability was amply described in section 5.1.1, noting the different ways that one can measure travel time variability, and how this translates into the actual passenger experience.

The second distribution, relating to recurrent performance, can be understood as the long-term reliability that would have been experienced by passengers had the service been absent of severe disruptions. This type of performance is expected to have a lower reliability buffer time than both the performance during incident-related peak periods, and the performance aggregated over all days (i.e. overall performance). Also, the reliability buffer time for peak periods with recurrent performance is, by definition, expected to be more predictable across days than the buffer time during incident-related peak periods and the overall observed performance. Table 5-12 compares the RBT for the recurrent performance days, and its day-to-day variability through the coefficient of variation, with the overall performance; the latter simply being drawn from the results shown in Table 5-5.

Table 5-12: Day-to-day variability of the RBT for the overall and recurrent performance categories for three O-D pairs – AM Peak, February 2007

	RBT (overall and recurrent) [min]			
O-D Pair	RBT overall	RBT recurrent	CV(95th – 50th) overall	CV(95th – 50th) recurrent
O-D # 1	9	5	0.96	0.38
O-D # 2	5	5	0.32	0.21
O-D # 3	8	7	0.56	0.37

The two right-most columns in Table 5-12 show that the consistency of the RBT across days increased for all three O-D pairs once the incident-affected days were excluded from the sample. This means that incidents contribute not only to higher RBT over the long-run, as shown by the first two columns, but also to more variability in the spread of the daily travel time distribution over time. From the passengers' perspective, one possible interpretation is that if someone was making a one time journey (e.g. visit to the doctor), it would be easier to predict the buffer time required by this person for that particular day to allow for in order to arrive on-time, if one could be certain that there would be no disruptions that day. For passengers making the journey on a repeated basis, this uncertainty across days is captured by the additional buffer time required over and above what is required under recurrent performance.

Finally, the non-recurrent travel time distribution can be understood as the experience of passengers during those particular days where large incidents took place. Figures 5-9 through 5-11 illustrate the relationship between these performance categories for each of the three cases.

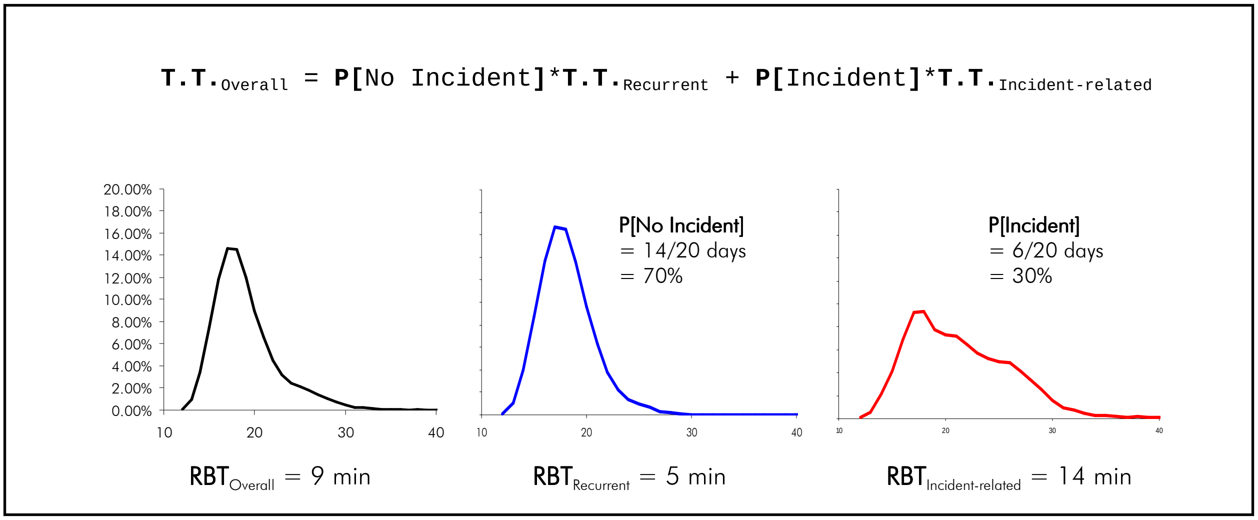


Figure 5-9: Break down of overall performance into recurrent and incident-related conditions and their probabilistic combination for O-D pair # 1

The overall travel time distribution for O-D pair # 1 is broken down into the performance categories used in the framework in Figure 5-9. The majority of weekdays during February 2007 (14 of 20) experienced more reliable service than did the remaining six days, which were most likely affected by large disruptions. The figure also shows how around 70% of the time, when a passenger travelled from Waterloo to Canary Wharf during the morning peak, he or she required only about 5 minutes of buffer time in order to reliably complete his or her journey on-time. Also, with a probability of 0.3, these same passengers were likely to experience a disruption, whose impact on those peak periods meant that most passengers (i.e. 95%) needed about 14 minutes of buffer time to reach their destination; 9 minutes more than on an undisturbed day. Similar interpretations can be made for the other two cases, shown in Figures 5-10 and 5-11.

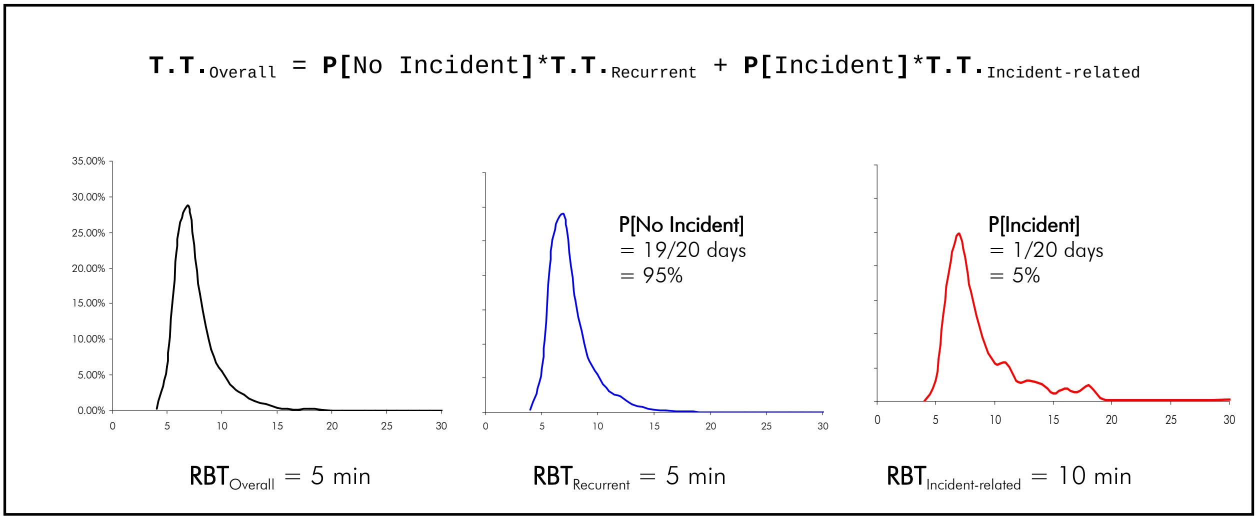


Figure 5-10: Break down of overall performance into recurrent and incident-related conditions and their probabilistic combination for O-D pair # 2

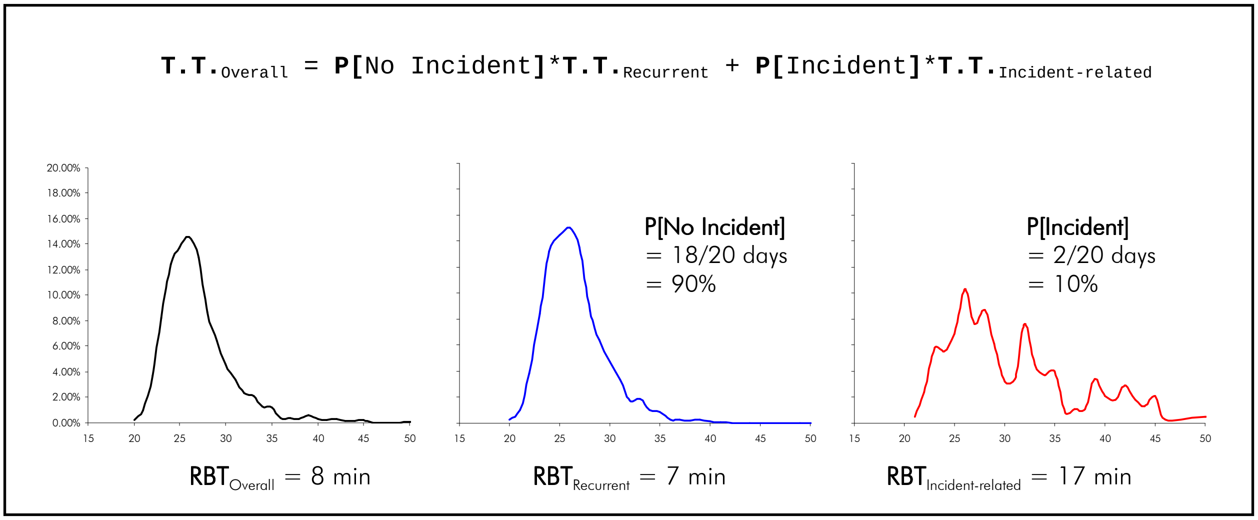


Figure 5-11: Break down of overall performance into recurrent and incident-related conditions and their probabilistic combination for O-D pair # 3

For O-D pairs # 2 and 3 there was a lower number of days classified as incident-related performance than that experienced by passengers of O-D pair # 1. Figures 5-10 and 5-11 show that the difference between what the buffer time "could have been" (i.e. recurrent performance), versus what it actually was (i.e. overall performance), was not large at only 1 minute difference. In the case of O-D pair # 2 (i.e. passengers travelling from Victoria station to South Kensington), there appeared to be only one day which was affected by large incidents, where 95% of all passengers required twice as much buffer time than that required on the remaining 19 days in order to complete their journey on time. Despite the poor performance observed on this day, the monthly reliability buffer time remained at the 5-minute level. The decomposition of performance for the third O-D pair shows a slightly larger impact of incidents, with their occurrence having severe effects on 2 of the 20 days of travel during February, leading to an increase in the monthly reliability buffer time of 1 minute over what it could have been without disruptions.

As seen above, the proposed framework can be used to gain useful insight into the reliability of the Underground and the potential for incident-related disruptions to worsen the passenger experience. In addition, the framework proposed two new measures of reliability which are applied to the three selected O-D pairs in the next section.

5.1.3: Application of Reliability Measures

In this section the Excess Reliability Buffer Time and the Percentage of Unreliable Journeys metrics proposed in Chapter 4, are applied to the three selected O-D pairs. Through these two measures, both the "depth" and "breadth" of the impact of unreliability on service quality can be captured, respectively.

The ERBT metric captures the impact of incidents on the travel time reliability experienced by passengers in units of time. That is, it provides a sense for the "depth" of unreliability (including the contribution of non-recurring disruptions) in terms of "how bad" performance was for a particular passenger making a journey (as opposed to the number of passengers affected), compared to what it could have been had there not been any major incidents.

To estimate the ERBT, the RBT under recurrent conditions must be subtracted from the RBT actually experienced by the entire passenger population (overall performance). The difference represents the buffer time that was experienced by passengers *in excess* of the buffer time that would have been required in the absence of incidents. Figure 5-12 illustrates this comparison for journeys on O-D pair # 1 during the four weeks in February 2007.

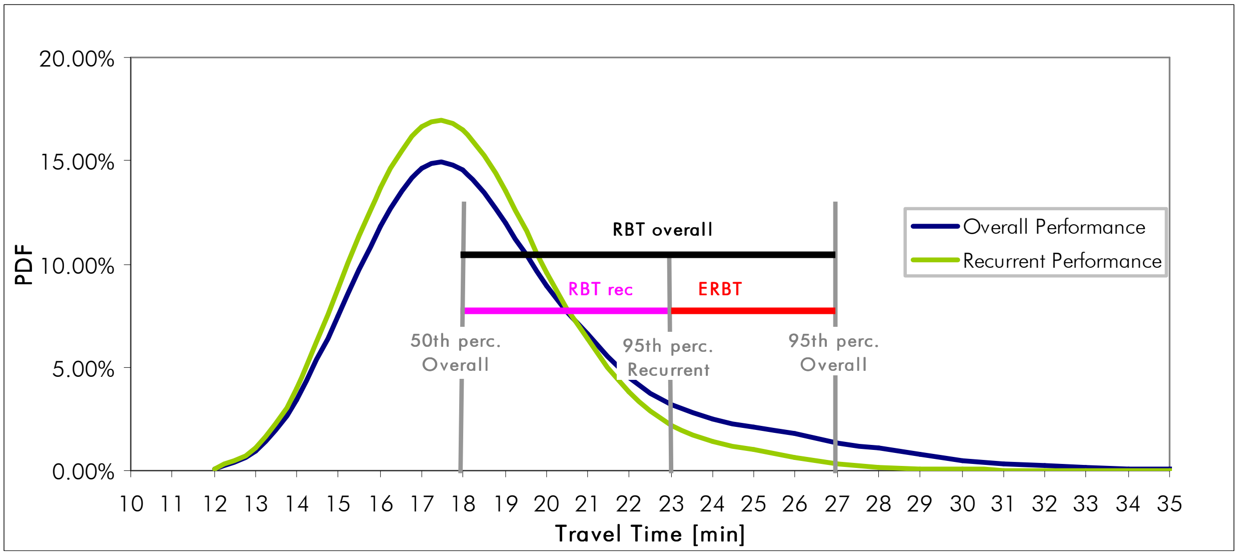


Figure 5-12: Illustration of the ERBT estimated for O-D pair # 1 – AM Peak, Feb. 2007

Figure 5-12 shows the value of the RBT from the overall travel time distribution to be 9 minutes, meaning that given the conditions during the month of February, a passenger travelling on a repeated basis would require this amount of time in addition to the typical travel time (here estimated to be 18 minutes) to guarantee an on-time arrival 95% of the time. In addition, the RBT from the recurrent performance for this O-D pair was only 5 minutes, meaning that this was the buffer time that passengers would have had to budget into their schedules had the system not been subjected to non-recurring incidents. Based on these two values, it is clear that the Excess RBT measure represents the additional buffer time that passengers had to leave, because of the disruptive effects of incidents on performance. Here the ERBT is naturally 4 minutes, or just over 20% of this journey's median travel time.

The ERBT is also applied over the four-week periods in February and November 2007 for each of the selected O-D pairs. In this case, however, the recurrent RBT was estimated using data from both periods, in order to capture the long-run baseline performance (see section 4.3.1), as well as have a fixed reference point to compare the overall reliability across the two months. Table 5-13 summarizes the ERBT using this baseline value for the morning peak during

February and November 2007. The long-run median travel time was used, explaining the difference in 1 minute of the overall RBT for O-D pair # 1 for February shown in Figure 5-12 (9 minutes), and that estimated in Table 5-13 (8 minutes).

Table 5-13: ERBT metric estimates for O-D pairs # 1, 2, and 3 – AM Peak, Feb. & Nov. 2007

	O-D pair # 1		O-D pair # 2		O-D pair # 3	
RBT recurrent [min]	5		5		7	
Period	February	November	February	November	February	November
RBT overall [min]	8	6	5	5	8	10
ERBT [min]	4	1	0	0	1	3
(ERBT/50th perc.)	16%	5%	0%	0%	4%	12%

For the first O-D pair in Table 5-13, the passenger experience was better for the month of November, where the overall RBT was closer to the recurrent reliability buffer time (i.e. ERBT = 1 minute, which is only 5% of the typical journey time for that O-D pair). This meant that from the perspective of passengers, if they were accustomed to incident-free performance, the observed travel time variability for November would not have been a severe increase over this, as a proportion of their journey length. The second O-D pair also provides some interesting insight into the reliability achieved for travel on this O-D pair. For both four-week periods, the actual level of reliability experienced by passengers was the recurrent performance. This could be attributed to either the characteristics of this O-D pair such as its short journey length (1-stop), or to the performance of the District line for this segment. The results for the third O-D pair reveal a slightly higher level of unreliability during November, with a 2 minute increase in RBT over the overall performance during February. However, relative to the median travel time for that journey, RBT constituted 4 and 12%, which might not be a significant level of additional unreliability from the point of view of passengers.

The "Breadth" of Unreliability – Application of the Percentage of Unreliable Journeys Metric

The second metric is focused less on "how bad" the performance was for a typical journey, but more on the number of passengers affected by that level of performance. More precisely, the Percentage of Unreliable Journeys measures the proportion of passengers that experienced journey times greater than a travel time limit which represents a "reliable" journey. This latter value is defined here as the sum of the typical travel time for a journey (50th percentile) and the reliability buffer time that would have been needed under recurrent conditions. This threshold for defining a journey as "reliable" or "unreliable" is appropriate from the transit operator's and passenger's perspectives since it represents the journey time that can be reasonably expected from the system on a repeated basis (given that service is not disrupted). Any journey durations above this threshold would be "out-of-the-ordinary" and unexpected, and would be thought of as an "unreliable" trip. Figure 5-13 illustrates this value for O-D pair # 1.

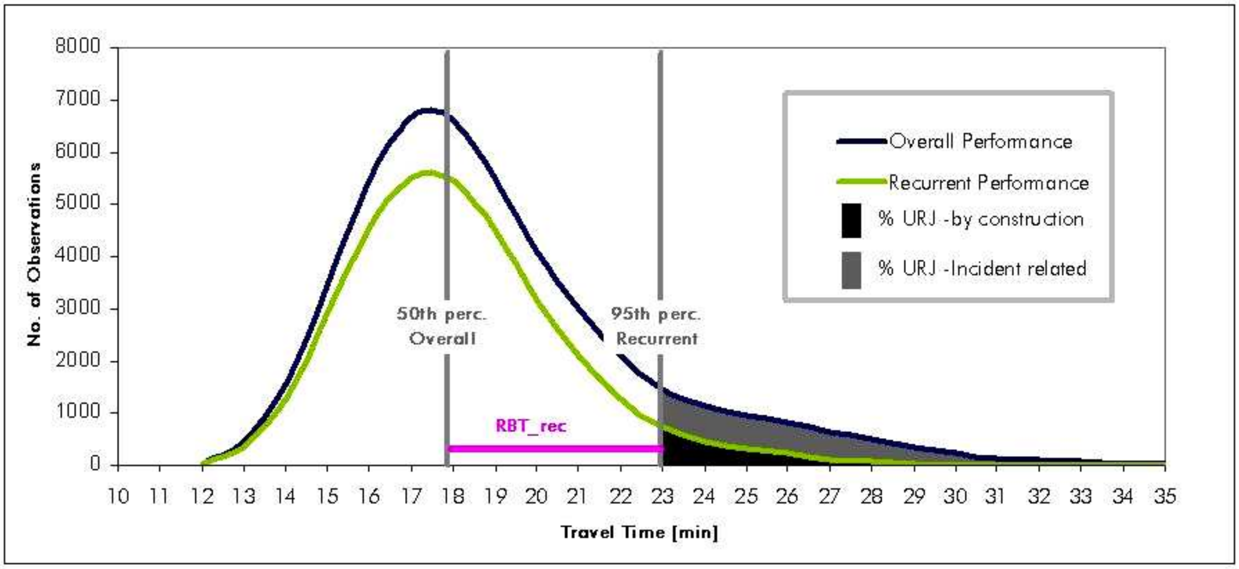


Figure 5-13: Illustration of the PUJ metric for O-D pair # 1 – AM Peak, Feb. 2007

The combined areas shaded in black and gray in Figure 5-13 represent the total number of journeys for O-D pair # 1 that experienced "unreliable" service (i.e. travel times greater than the recurrent 95th percentile travel time). By construction, this area includes 5% of all journeys that took place under recurrent performance, captured by the black shaded area, which means that even under recurrent conditions, 1 in 20 journeys are still expected to experience a journey time beyond the threshold for "reliable" service. The area shaded in gray captures the remaining number of unreliable journeys and can be considered to be the additional unreliable journeys attributable to the occurrence of non-recurring incidents. Table 5-14 summarizes the number of unreliable journeys as a proportion of the total number of journeys, for the three origin-destination pairs during the months of February and November.

Table 5-14: PUJ metric for O-D pairs # 1, 2, and 3 – AM Peak, Feb. & Nov 2007

	O-D pair # 1		O-D pair # 2		O-D pair # 3	
	February	November	February	November	February	November
PUJ (Gray + Black)	9%	7%	5%	5%	6%	8%
PUJ incident-related (Gray)	6%	4%	2%	2%	3%	5%
ERBT [min]	3	1	0	0	1	3

As expected, the proportion of passengers experiencing poor performance is highly related to the "depth" of unreliability. For higher levels of ERBT there were a higher percentage of passengers experiencing unreliable service overall, with every minute of excess reliability buffer time leading to an increase of around 1-4% in the proportion of passengers experiencing unreliable service due to incidents. For example, for O-D pair # 1, the values in Figure 5-13 and in Table 5-14 show that during the month of February, a 3 minute excess buffer time was experienced by 6% of all passengers travelling during the morning peak, compared to the performance for the month of November where a lower impact of incidents was observed

(ERBT of 1 minute), with only about 4% of passengers being affected. In the case of O-D pair # 2, the impact of non-recurring incidents led to a total of 5% of all journeys having a travel time greater than the recurrent 95th percentile travel time. This implies that even though incident-related performance contributed to around 2% of these journeys, they were still not enough to go beyond the 5% threshold (due to the use of the 95th percentile value), and increase the ERBT for both February and November.

Contribution of O-D Pair Performance to System Unreliability – Weighted-ERBT Metric

Another way to measure the performance of the same three origin-destination pairs is in terms of their contribution to overall system unreliability. A variation of the ERBT metric can be used to quantify this, where the original metric is weighed by the total passenger volume for a particular O-D pair. The result is in units of passenger-minutes, and reflects both how unreliable a journey is in terms of pure performance (i.e. "depth"), and how many passengers were affected by it in absolute numbers (as opposed to a percentage). Using absolute passenger volumes is appropriate for comparing the contribution to unreliability by different O-D pairs across the system. Table 5-15 summarizes the results from applying this measure, termed the Weighted-ERBT metric, to each of the three origin-destination pairs.

Table 5-15: Weighted ERBT metric for O-D pairs # 1, 2, and 3 – AM Peak, Feb. & Nov. 2007

	O-D pair # 1		O-D pair # 2		O-D pair # 3	
	February	November	February	November	February	November
ERBT [min]	3	1	0	0	1	3
Total Passenger Trips [17]	7,961	7,961	1,316	1,316	232	232
W-ERBT [pax-min]	23,883	7,961	0	0	232	696

Figure 5-15 shows the contribution of each O-D pair to the overall unreliability of the system and compares it to their ERBT (i.e. "depth" of unreliability). For example, even though the level of unreliability during February for O-D pair # 1 and during November for O-D pair # 3 had both 3 minutes of ERBT and a similar percentage of passengers experiencing unreliable service (9% and 8% respectively), because of the significantly higher volume of passengers making the former journey, its contribution to total unreliability at the system level was greater by a factor of 30.

This extension to the ERBT metric could be used to determine the parts of the system where the largest improvements in overall unreliability could be attained through improvements in the performance of the service.

[17] Total Passenger Trips estimated from AM Peak O-D Matrix 2006 (Chan, 2007).

5.2: Reliability Factors and their Impact on Service Quality

Apart from using the reliability framework to monitor performance on a regular basis, it can also be used to quantify the contribution of different factors influencing reliability. The contribution of some of these factors is explored in this section in order to (a) quantify the contribution of incidents and recurrent factors to unreliability as part of the proposed framework, and (b) gain a more in-depth understanding of reliability and show the potential of this source of data for policy analysis and future research.

The section begins in 5.2.1 with a review of the factors that could affect reliability and some of the evidence found in previous studies for their impact. Section 5.2.2 attempts to quantify the contribution of some of these factors through a linear regression analysis using Oyster Smart Card data.

5.2.1: Overview of Factors and Preliminary Findings

In section 2.1.3, some of the factors responsible for changes in the reliability of passenger journey times were highlighted. These included the characteristics of the service, or those factors that have a recurrent effect on performance, as well as the more unpredictable causes that may be mitigated through real-time strategies, as opposed to more long-term planning interventions. Understanding the effects of these factors helps not only to design more robust measures of performance, but also for informing policies that aim to improve service reliability.

Recurrent Reliability Factors

Among the various recurrent factors that are expected to influence the level of travel time variability for a particular journey, the following three are considered: journey length, scheduled vehicle headways, and whether a journey involves one or more interchanges.

The **journey length** is expected to be positively correlated with the variability of travel times because as a passenger spends more time on a vehicle, his or her exposure to random variation caused by dwell times at intermediate stations, or train speeds due to line traffic, increases. This additional exposure means that there are more chances to have the journey increase in duration, with the cumulative effect being that the extreme travel times would tend to be longer in absolute terms than for shorter journeys. Also, in the particular case of Oyster data where several journey travel time components are measured together, longer journeys would make the in-vehicle travel time a larger proportion of the overall journey time (because access/egress and wait times are independent of journey length), which could make its variability the dominant variability in determining the total journey time variability (e.g. as opposed to being driven by the variability of walk time).

The **scheduled headway** is also expected to be positively correlated with a higher level of total travel time variability. This is due to the increasing range between the longest and the shortest possible waiting time for passengers as the time interval between vehicles increases (assuming random passenger arrivals at platform). As will be discussed later on in this section, this effect can be observed when comparing trunk and non-trunk portions of the service on the same line. The actual service headways and their variability is clearly also important to take into account when explaining journey time variability.

A third intrinsic characteristic of a journey likely to influence reliability is whether a journey **involves an interchange**. This is mainly because of the presence of additional platform walking and waiting times during the additional leg(s) of the journey, which are expected to contribute to the variability of the total journey time above what it would have been, had the journey been within stations on the same line.

Due to the repeating nature of their impacts, these factors help explain the travel time variability observed during recurrent conditions identified as part of the framework in the previous chapter.

Non-recurrent Reliability Factors

In addition to recurrent factors, non-recurring events are also expected to have a large impact on reliability. Two types of events are **operations control interventions** and **incident-related disruptions**. In the case of the former, it is difficult to predict how a particular action will affect the variability of travel times experienced by passengers and future research is recommended in this area (see section 7.3). For the latter, section 5.1 illustrated the way service disruptions caused by incidents would lead to a higher variability of travel times both during the particular peak period that the disruption took place, as well as over time (i.e. long-run overall variability).

Finally, a variable to account for any possible **seasonality effects** in performance was included, where certain times of the year might show consistent levels of travel time delays and variability that differ from other periods, caused by factors such as natural fluctuations in passenger demand and changes in weather. In the case of the Underground, any significant effects of this kind were not expected. The JTM does not take them into account for the two four-week periods studied here, suggesting that they might not be dominant. In addition, no major schedule changes were made between February and November 2007 that would be expected to change the recurrent performance of the system.

Evidence for the Impact of Reliability Factors

There is a limited amount of work on quantifying the relationship between the factors mentioned above and the variability of travel times experienced by passengers. The previous study by Chan (2007) found evidence for the effects of some of these factors through a cross-tabulation approach, where changes in the spread of the travel time distribution were compared side-by-side with differences in the journey attributes. Some of the findings from the study are presented in Table 5-16.

Table 5-16: Comparison of RBT for Piccadilly line O-D pairs (adapted from Chan, 2007)

	Trunk O-D pairs			Non-trunk O-D pairs		
Median Journey Time Range [min]	Mean RBT [min]	Standard Deviation of RBT [min]	Number of O-D pairs	Mean RBT [min]	Standard Deviation of RBT [min]	Number of O-D pairs
5-10	9.14	9.26	7	8.06	2.44	10
10-15	6.19	3.52	92	7.52	3.60	89
15-20	6.15	3.52	96	7.91	1.89	86
20-25	5.92	2.04	92	8.00	1.78	91
25-30	7.04	2.67	68	8.82	2.19	93
30-35	7.65	2.85	61	10.23	3.13	107
35-40	9.05	4.30	47	10.45	3.72	114
over 40	9.31	5.26	29	11.81	4.59	115

In particular, the study found that, except for the shortest journeys, as the median travel time for a journey increased, the level of unreliability as measured by the RBT, also increased. Table 5-16 also shows that journeys taking place on trunk portions of the Piccadilly line, with average scheduled headways of around 3 minutes, tended to have a variability of travel times than trips starting at non-trunk portions with longer scheduled headways. Additional analyses in the study by Chan (2007) found similar effects for another line in the Underground, as well as evidence for the effects of train load levels on travel time variability.

The evidence for the effects of various factors on unreliability provided by this earlier study goes in accordance with the hypotheses presented in this section. The following section attempts to quantify the impact of several factors of unreliability at a more general level of analysis, by considering 800 of the highest-volume O-D pairs in the system.

5.2.2: Linear Regression Analysis – Contribution of Factors to Unreliability

Building on the prior work discussed above, as well as the insights presented in the first section of this chapter, this section attempts to quantify the reliability effects of some of the factors presented above. Can recurrent performance in fact be explained by fixed journey attributes (i.e. inherent characteristics), and if so, how much variability is added above this underlying level of performance by large disruptions?

This question is addressed using a linear regression analysis because it captures the *simultaneous* contribution of the different factors to the observed level of unreliability, in an attempt to overcome the limitations of a two-dimensional comparison of the data, as conducted in the prior study referred to above. Given these considerations, the objective of this analysis is not to provide a comprehensive explanation for all of the underlying variability observed in travel times, but rather to:

- Expand the exploratory work begun by Chan (2007) by quantifying the effects of the different reliability factors, and show the potential for future studies in this area;

- Validate the framework's concepts of differential contributions to unreliability from recurrent factors and incident-related disruptions;

- Generalize the characterizations in section 5.1 for the three O-D pairs by quantifying the level of unreliability observed in the system for a larger sample of O-D pairs over multiple days at the peak period level.

For this initial analysis, a simple model specification is proposed that relates two recurrent factors and the presence of incident-related disruptions to the Reliability Buffer Time metric.

Model Specification and Variable Definitions

The regression analysis performed here is framed temporally for each AM Peak, and spatially at the O-D pair level. In this way, one can study the interaction between the characteristics of each journey and its level of unreliability, while still capturing the influence of broader system-wide factors like incidents on the passenger experience. Drawing from some of the reliability factors identified previously and the framing of the analysis just explained, and taking into account the fact that the data involve both a cross-section of O-D pairs as well as daily observations over 8 weeks (i.e. panel data structure), the model specification consists of:

$$RBT_{i,j} = \alpha + \beta_1 \cdot JrnyLength_{i,j} + \beta_2 \cdot XFER_i + \beta_3 \cdot INC_{i,j} + \beta_4 \cdot FEB_{i,j} + \varepsilon_{i,j} \qquad [5\text{-}1]$$

Where the variables are defined as follows:

- *$RBT_{i,j}$ = (95th percentile of peak period j – 50th percentile of the four-week period containing j for O-D pair i)*. This variable represents the Reliability Buffer Time for passengers departing during peak period j, on a particular O-D i. The use of the four-week median travel time provides a fixed reference point for evaluating a particular peak period's level of travel time variability, representing the long-run travel time that passengers might come to expect. In addition, the period median travel time helps to address the presence of correlation between the daily median travel time and the daily 95th percentile travel time.

- *$JrnyLength_{i,j}$ = 50th percentile for four-week period containing peak period j for O-D pair i.* The length of the journey is represented by the median travel time for that particular O-D pair during that four-week period (February or November). This is to capture the inherent contribution of journey length on unreliability as opposed to the peak-period specific performance. The journey length is measured for each four-week period in order to account for the possibility of seasonality (i.e. overall differences in performance), captured by the last independent variable.

- *$XFER_i$ = 1 when journey i involves one or more transfers, 0 otherwise (i.e. binary variable).* Because a higher level of travel time variability is expected for journeys involving an interchange, due to the additional platform wait time, this dummy variable is set equal to 0 in the case of a same-line journey. There are two aspects of this variable definition that could slightly reduce the accuracy of the analysis. First, there is a small proportion of O-D pairs lying on the branched portions of a line that were identified as being "same-line", even though in actuality they would be expected to use multiple lines. The effect of this small proportion of "transfer" O-D pairs would be to slightly underestimate the value for the dummy variable, assuming that the presented hypothesis is correct. Second, this definition groups together single interchange journeys with trips requiring more than one transfer point. This could lead to a slight mis-interpretation for the coefficient of this variable as solely capturing the effects of the additional wait time at the transfer station. In reality, it could also capture multiple travel paths or more than one platform wait time. However, because the O-D pairs selected for the analysis involve high passenger volumes and relatively short to medium distances, the vast majority of the journeys would be reasonably expected to be completed with only 1 interchange.

- *$INC_{i,j}$ = 1 if the performance of peak period j was identified to be incident-related using the classification approach at the 5% significance level, 0 otherwise (i.e. a binary variable).* This dummy variable represents those peak periods whose performance differed significantly from the travel time variability of the remaining peak periods. Section 5.1.2 showed how these peak periods could be attributed to incident-related disruptions in the system. By definition, this binary variable only captures the effects of the largest disruptions, and as discussed in the previous section, clearly some incidents of small impact may have occurred on days identified as part of the recurrent performance set.

- *$FEB_{i,j}$ = 1 if the performance of peak period j for O-D pair i took place during the four-weeks in February 2007, 0 otherwise (i.e. November 2007).* This dummy variable attempts to capture any seasonal differences in performance between the two four-week periods studied.

- *$\varepsilon_{i,j}$ = random error term for observation during peak period j for O-D pair i.*

Summary of Data Set used in the Analysis

A total of the 800 largest O-D pairs were selected from the Underground system, including 632 same-line journeys with the highest Oyster volumes in the system, and the remaining 168 O-D pairs with the highest-volume O-D pairs that involved one or more transfers. The AM Peak travel time distribution for each O-D pair across 40 weekdays was determined, leading to a total of 32,000 observations (800*40) evenly split between four weeks in February and November 2007. Table 5-17 summarizes the data used in the analysis.

Table 5-17: Summary of data characteristics –800 O-D pairs

Variable	Average	Min	Max
RBT [min]	7.7	0	99
JrnyLength [min]	22.3	4	53
XFER [1,0]	0.21	0	1
INC [1,0]	0.18	0	1
FEB [1,0]	.5	0	1

On average, the Reliability Buffer Time was just under 8 minutes for the selected O-D pairs, meaning that for the average journey, that much additional time above the median travel time had to be allowed for to have 95% of on-time at the destination. Figure 5-14 below shows the distribution of the Reliability Buffer Time for each O-D pair (800) and peak period (40) in the sample, with the majority of O-D pairs having a buffer time between 4 and 10 minutes, and a minority of observations exhibiting extremely high buffer times (e.g. 99 minutes) attributable to severe disruptions during those particular days.

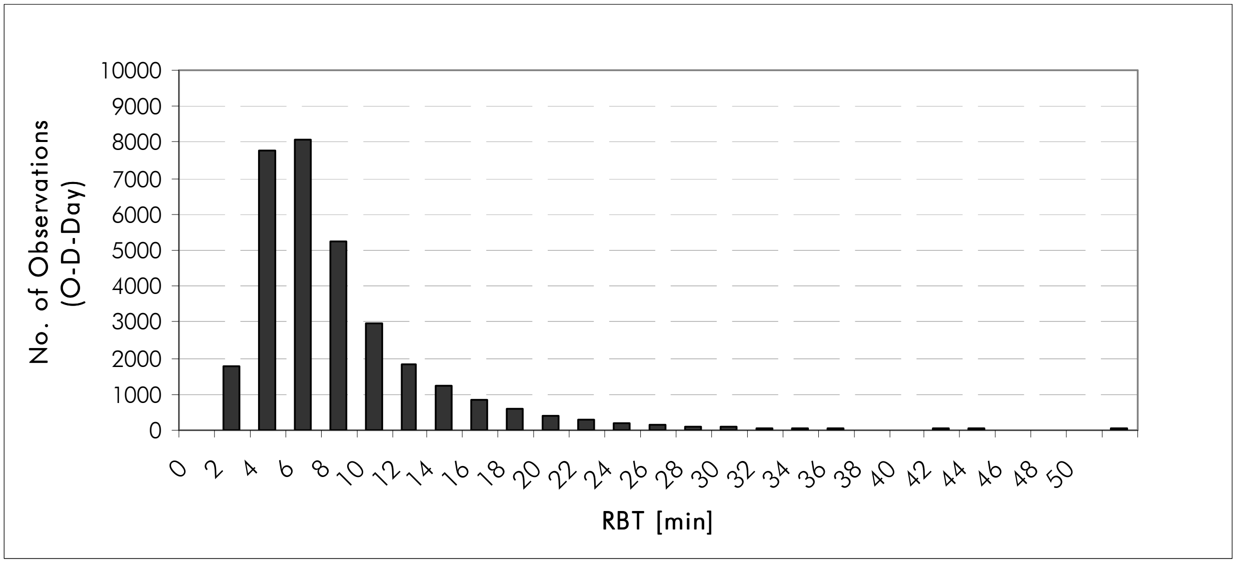


Figure 5-14: Distribution of the RBT for 800 O-D pairs – AM Peak, Feb. & Nov. 2007

The selected journeys had an average travel time of 22.3 minutes, or a median travel time of 22 minutes. This last figure is slightly less than the system-wide median travel time of 26 minutes found by Chan (2007), showing that the O-D pairs used in the analysis were overly representative of the short to medium range journeys in the system, with a smaller proportion of longer distance journeys[18]. Figure 5-15 shows the distribution of the median travel time for the O-D pairs in the sample.

[18] This can be explained by the decreasing proportion of high-volume journeys as one moves away from Central London, caused by the lower penetration rate of Oyster for users of National Rail. Also, the higher population and employment density of Central London would lead to a concentration of shorter high-volume trips.

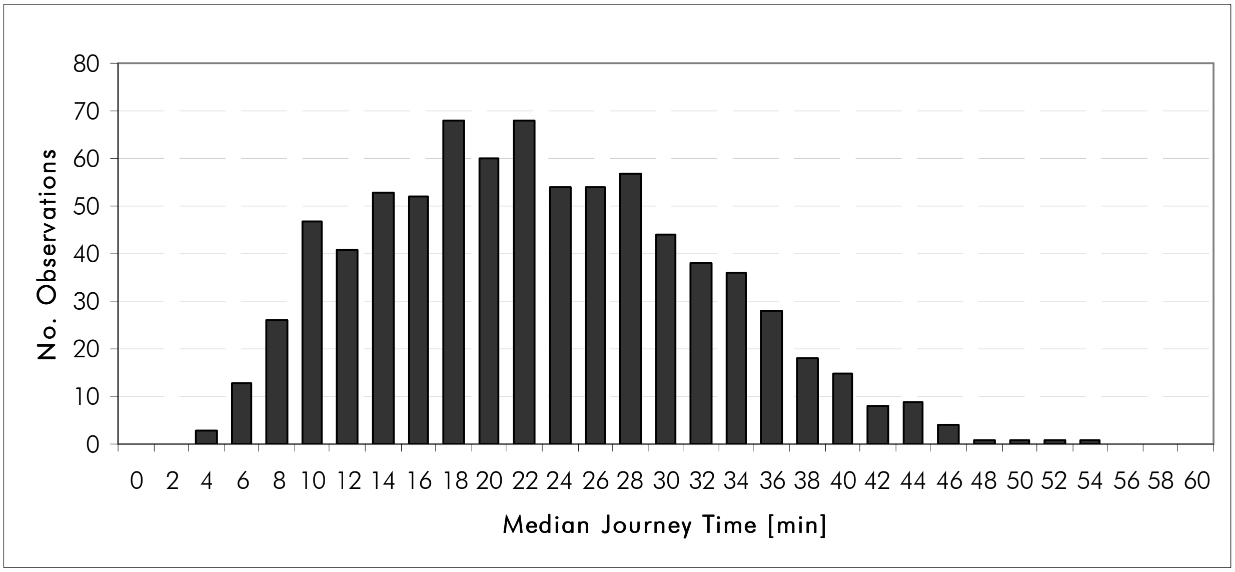


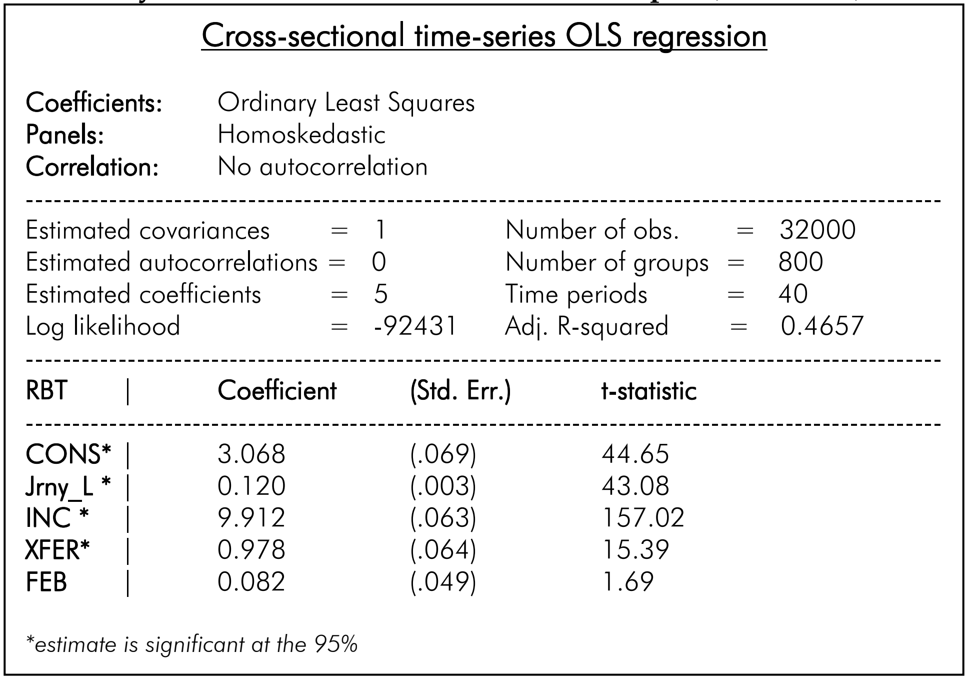
Figure 5-15: Distribution of median journey times for 800 O-D pairs

Finally, Table 5-17 shows that, on average, 21% of the O-D pairs in the sample involved one or more transfers, and that 18% of the days, or around 3.6 weekdays per month, were subject to incident-related disruptions.

Linear Regression Estimation Process and Results

The model specified in Equation 5-1 is first estimated using the Ordinary Least Squares (OLS) procedure to get a sense for the magnitude of the relationship between the explanatory variables and the level of reliability observed for each of the O-D pairs. This "naïve" approach is based on several assumptions, some of which are revisited later in this section. Specifically, it assumes that the errors are independently and normally distributed with an expected value of 0 and with a constant variance. The results from the initial OLS estimation are summarized in Table 5-18.

Table 5-18: Summary of OLS estimation results – 800 O-D pairs, AM Peak, Feb. & Nov. 2007



```
                 Cross-sectional time-series OLS regression

Coefficients:    Ordinary Least Squares
Panels:          Homoskedastic
Correlation:     No autocorrelation
------------------------------------------------------------------------------------------
Estimated covariances      =  1          Number of obs.      =  32000
Estimated autocorrelations =  0          Number of groups    =  800
Estimated coefficients     =  5          Time periods        =  40
Log likelihood             =  -92431     Adj. R-squared      =  0.4657
------------------------------------------------------------------------------------------
RBT      |      Coefficient      (Std. Err.)        t-statistic
------------------------------------------------------------------------------------------
CONS*    |      3.068            (.069)             44.65
Jrny_L * |      0.120            (.003)             43.08
INC *    |      9.912            (.063)             157.02
XFER*    |      0.978            (.064)             15.39
FEB      |      0.082            (.049)             1.69

*estimate is significant at the 95%
```

The OLS estimation results show that both the signs and magnitudes of the regression coefficients are consistent with the stated a priori expectations. The constant term in this first estimation of the model reveals that around 3 minutes of buffer time were needed for same-line journeys absent of incidents and after controlling for the effects of distance, transfers, and seasonality. This value is likely to capture the variability contributed by the platform wait time and in-station access time, and is within the order of magnitude that would be expected as seen from the discussion below for the contribution of an interchange.

The effect of distance is captured by the median travel time variable (Jrny_L), which shows that, assuming a linear relationship, for every additional minute that a journey typically takes, the Reliability Buffer Time is increased by around 0.12 minutes (7 seconds). This means that for the average journey in the system, which takes around 22 minutes on a typical day, the basic journey distance effect would lead to an increase in the buffer time of about 2.64 minutes. This is an important contribution to unreliability if one takes into account the fact that the average observed RBT was around 8 minutes for the sample, meaning that almost 1/3 of the observed average variability can be attributed to the effects of this fixed attribute.

Another important result is the contribution that the other recurrent factor, XFER, has to unreliability. When a journey involves an interchange, the variability of that journey increases by approximately 56 seconds. This value is within the order of magnitude one would expect if our hypothesis is correct, where the additional wait time caused by the second leg of the journey would be the main cause of the increase in the total travel time variability. For example, if a uniform wait time distribution for 3 minute headway service is assumed (i.e. service is perfectly regular and passenger arrivals at station is random), its variance would be around 0.75 minutes squared. If the wait time of the initial journey is a random variable that is independently distributed from the second wait time caused by the transfer, then it follows that the variance of the total wait time distribution is a linear sum of the variances from the individual wait time distributions. Therefore, the second wait time distribution could be expected to contribute to the overall travel time distribution by an order of magnitude around the derived value, or more precisely its square root (in units of minutes) of 0.87. This is similar to the parameter value obtained from estimation of the specified model.

The third explanatory variable, INC, captures the effect that incidents have on the variability of travel times for the particular peak period in which it took place. It reveals that, on average, when a large disruption occurs due to one or more incidents, the Reliability Buffer Time is increased by about 10 minutes above what it would normally be for that peak period. When the probability that a large disruption takes place (P[Incident] = 0.18) is taken into account, the occurrence of incidents explains on average about 1.85 minutes of the observed RBT, or about 24% at this rate of disruption[19].

Finally, the dummy variable for February (FEB) was insignificant at the 95% level. This means that there were no significant differences in performance between February and November 2007. This result is important because it validates the selected approach for estimating a reliability baseline for the ERBT metric (see section 4.3.1) where the recurrent performance of the system is estimated over various periods combined under the assumption of no seasonality effects. This finding was specific to the two four-week periods used in this analysis, and estimation of the recurrent performance for different periods would have to verify the absence of seasonal changes in performance.

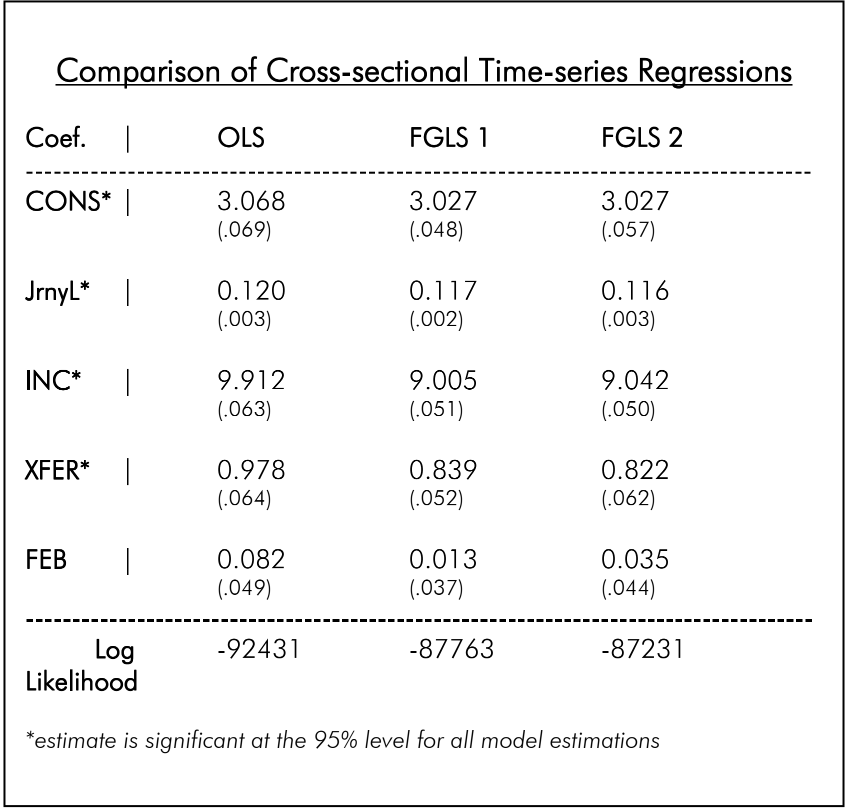
In addition to the values of the coefficients being as expected, Table 5-18 above shows that all except that of seasonality were statistically significant at the 95% level. This aspect, however, requires further investigation mainly because, for panel data, the assumptions on which the OLS procedure is based often do not hold to be true. Specifically, the assumption constant variance for the error term would not be expected when the existence of individual effects for each O-D in the panel (i.e. heteroskedasticity) is considered. Also, because in addition to cross-sectional data there are time-series observations, it is possible to have autocorrelation between the error terms within a particular group (i.e. O-D pair). However, the OLS coefficient estimates are still unbiased and consistent in the presence of heteroskedasticity and autocorrelation, implying that the values discussed previously are still valid. Rather, the overall effect of these OLS assumptions being violated is to bias the variance of the estimates where they are no longer the most efficient estimators. This means that in order to correctly interpret the significance of the coefficients, the possible presence of these two characteristics in the data must be corrected for.

Two additional estimation attempts were performed to illustrate how further work could improve the initial estimation results and possibly gain more reliable insight into the causes of unreliability. The second approach estimates the model using a Feasible Generalized Least Squares (FGLS) approach in order to correct for heteroskedasticity. The third estimation assumes a first-order serial correlation to correct for the possibility of autocorrelation among the observations within a group, in addition to the presence of heteroskedasticity.

[19] The actual effect of incidents on the RBT over 20 days depends on other factors such as the frequency of disruptions, the size of the disruptions and the level of recurrent travel time variability already present.

The first application of FGLS ("FGLS 1") assumes that the regression residuals are uncorrelated but have unequal variances across the O-D pairs, which is sensible given the degree to which the performance of each O-D is expected to be independent of the other O-D pairs. This might not be necessarily true in cases where O-D pairs overlap, or there are other effects which would make their performance related. However, as an initial correction and given the wide array of journeys being considered in the panel, this residual structure was adopted. The correction for serial correlation is labeled "FGLS 2", where a first-order autoregressive (AR1) process that is specific for each O-D pair was assumed, while also assuming heteroskedasticity as in the FGLS1 estimation. Table 5-19 compares the outcomes of the three estimation procedures.

Table 5-19: Comparison of estimation results – 800 O-D pairs, AM Peak, Feb. & Nov. 2007



Comparison of Cross-sectional Time-series Regressions

Coef.	OLS	FGLS 1	FGLS 2
CONS*	3.068 (.069)	3.027 (.048)	3.027 (.057)
JrnyL*	0.120 (.003)	0.117 (.002)	0.116 (.003)
INC*	9.912 (.063)	9.005 (.051)	9.042 (.050)
XFER*	0.978 (.064)	0.839 (.052)	0.822 (.062)
FEB	0.082 (.049)	0.013 (.037)	0.035 (.044)
Log Likelihood	-92431	-87763	-87231

**estimate is significant at the 95% level for all model estimations*

After estimating the linear regression by controlling for heteroskedasticity and specifying a panel-specific autoregressive process to account for serial correlation, the estimates of the coefficients remained similar to the initial unbiased and consistent OLS estimates. However, more efficient estimators are now found using the second and third approaches (as evidenced by the lower standard errors than under OLS), allowing one to claim from these estimation results that the coefficient estimates are in fact significant at the 95% level, with the exception of FEB which remained insignificant.

It is clear that additional research is required to fully understand the contribution of other causes of unreliability besides those tested here. Possible improvements include estimating the relationship of travel time variability and the actual (as opposed to scheduled) service frequency for a particular O-D pair, the inclusion of line and direction specific dummy variables, and performing the analysis at more disaggregate levels of time in order to account for the effects of demand congestion within the morning peak. In addition, non-linear interactions between the different factors could be explored, as well as segmentation of the data to find more specific results (e.g. do incidents have a larger impact on transfer journeys?). In addition, improvements in the estimation procedure can be done, including testing for the degree to which heteroskedasticity and autocorrelation actually have an effect on the model results, and the most appropriate form for controlling for their effects (e.g. the appropriateness of the AR1 assumption).

However, the similarity of the results of the three estimation procedures, as well as the improvements in the model fit (as revealed by the increase in the log likelihood), indicated that this type of analysis using Smart Card data and the reliability framework has appreciable potential for providing useful insights into the nature of reliability.

Chapter 6: Framework Applications for the London Underground

In this chapter, two applications of the reliability framework for the London Underground are developed. The first application, presented in section 6.1, shows how the framework can be used as part of on-going service quality monitoring efforts by proposing an extension to the Journey Time Metric. The reliability extension is then used to quantify the performance of the Victoria line during the morning peak. Section 6.2 presents the second application, which uses the framework to mitigate the impacts of unreliability by proposing improvements in the information provided to passengers through the Journey Planner trip planning software currently used by TfL.

6.1: Reliability Extension of the Journey Time Metric – Victoria Line Application

One of the primary applications of the framework is in the area of performance monitoring and subsequent evaluation. This section proposes an extension to the Journey Time Metric to quantify the reliability experienced by Underground passengers and applies it to one line in the system. Section 6.1.1 describes the reliability extension and its estimation methodology. Section 6.1.2 introduces the Victoria line with a discussion of the level of coverage attained through the proposed reliability measures. Section 6.1.3 applies the reliability extension by estimating the recurrent and excess levels of reliability buffer time for passengers of this line, followed by a discussion of the contribution of reliability to the overall travel time experienced by passengers in section 6.1.4.

6.1.1: Proposed Extension – Reliability Buffer Time Component

The structure of the RBT and ERBT metrics (see Equations 4-1 and 4-5, respectively), including their estimation in units of time and across various levels of spatial and temporal aggregation, makes them feasible to estimate within the existing framework of the Journey Time Metric. Given that the five existing components of a journey capture the average trip time, the Reliability Buffer Time is proposed as a sixth journey component in JTM, capturing the effects of total travel time variability on passengers (see section 3.3.1).

The estimation process for the reliability extension would be similar to that already used by other components of JTM with regards to the aggregation of results. Spatially, the estimation process would begin by measuring reliability at the O-D pair level, and using the total passenger volumes for each journey to find a weighted average performance at the line level, as described by Equation 4-3. The line level estimates could then be used to estimate reliability at the network level where line volumes are again used to find a weighted average performance. Temporally, the reliability extension would set the sample period of the RBT to match the four-week reporting cycle used by JTM (i.e. 20 weekdays), and set the time interval to match the desired operating time period of the day (e.g. morning peak).

In addition, the RBT metric and its variations can be easily applied to represent the four existing layers of estimation found in JTM: the actual, baseline, excess, and weighted performance for each component of a journey. First, a direct measurement of the RBT metric can be used to quantify the overall or actual level of reliability observed during a particular four-week period. Second, the recurrent RBT presented in Chapter 4 as part of the reliability framework parallels the existing baseline performance measured by JTM for each of its existing components of a journey. The difference between the recurrent and the overall RBT leads to the excess level of unreliability (i.e. ERBT), which is compatible with measures of excess journey time in JTM. Lastly, the fourth layer of estimation in JTM weighs the performance of each journey component by the value of time that passengers place on it relative to the value for one minute of in-vehicle travel time (see section 3.3.1). The RBT can also be weighed by the value of time passengers place on it, with possible estimates for this value discussed in section 6.1.4. In addition, the weighting of the individual components by their VOT also makes the contribution of unreliability to service quality, as captured by the reliability extension, comparable to the contribution of the average travel times to the passenger experience. This makes it possible to bring together the performance for the various components of a journey, including the reliability extension, and find a single performance value, as currently practiced by JTM.

6.1.2: Victoria Line Description and O-D Pair Coverage

The reliability extension is applied to one line in the system in order to demonstrate its potential use within the Journey Time Metric, and to obtain initial estimates of the reliability currently experienced by passengers of the Underground, and the contribution of incidents to service quality.

The Victoria line runs north-south through Central London and is a useful case for the application of the reliability extension to JTM because of its relatively uniform service configuration throughout (i.e. no branches, all sections have high frequency during morning peak), and with only 16 stations, it is one of the shortest in the system (see Appendix A). Also, it is one of the most heavily used lines in the Underground network, providing large sample sizes for the analysis.

Regarding the latter point, it is important to establish the level of coverage that can be achieved when applying the RBT metric to the Victoria line using Oyster data. The proportion of all O-D pairs represented in the line will depend on the number of journeys completing them on Oyster. If a minimum sample size of 20 passenger journeys is required to measure the **overall RBT** over the four-week sample period, for the AM Peak time interval, it is possible to measure reliability over all possible O-D pair combinations on the Victoria line.

However, when estimating the **baseline RBT** (i.e. the recurrent RBT), a larger minimum sample size is required. This is due to the classification methodology used to separate recurrent from incident-related performance (see section 4.2.4), where each *peak period* is required to have at least 20 passenger journeys (as opposed to the pooled four-week sample of journeys). If the two four-week periods during February and November 2007 are pooled to estimate the long-run recurrent RBT (see section 4.3.1), 150 of the 240 (16*15 stations) O-D pairs possible in the Victoria line are covered. These are divided between 60 northbound and 90 southbound journeys, reflecting the directional imbalance in demand that exists for the line (i.e. southbound travel has higher overall volumes during the morning peak). These 150 O-D pairs represent

95.9% of all Oyster journeys on the Victoria line during the AM Peak for February and November 2007. The high coverage possible through Oyster data indicates that the reliability extension is feasible to use within JTM for routinely monitoring service quality, as it provides an accurate snapshot of the performance of the line as a whole.

6.1.3: Estimation of Victoria Line Reliability Buffer Time

The reliability extension is applied to the 150 O-D pairs in the Victoria line during the AM Peak for a four-week period in February and November 2007. First, the baseline reliability buffer time is estimated at the line level, followed by an estimation of the overall reliability as experienced by passengers. These two performance estimates are then used to find the excess level of unreliability for the line.

Estimation of Victoria Line Baseline (Recurrent) Reliability Buffer Time

The first part of the estimation process is to determine the amount of unreliability that passengers would be exposed to under typical conditions, or the baseline buffer time. The recurrent performance is estimated at the O-D pair level for the 150 journeys identified in the previous section using a pooled sample period of 40 weekdays. Figure 6-1 provides an example of the recurrent reliability for a sample of journeys starting from the endpoints of the trunk portion and ending at each successive station down the line in each direction.

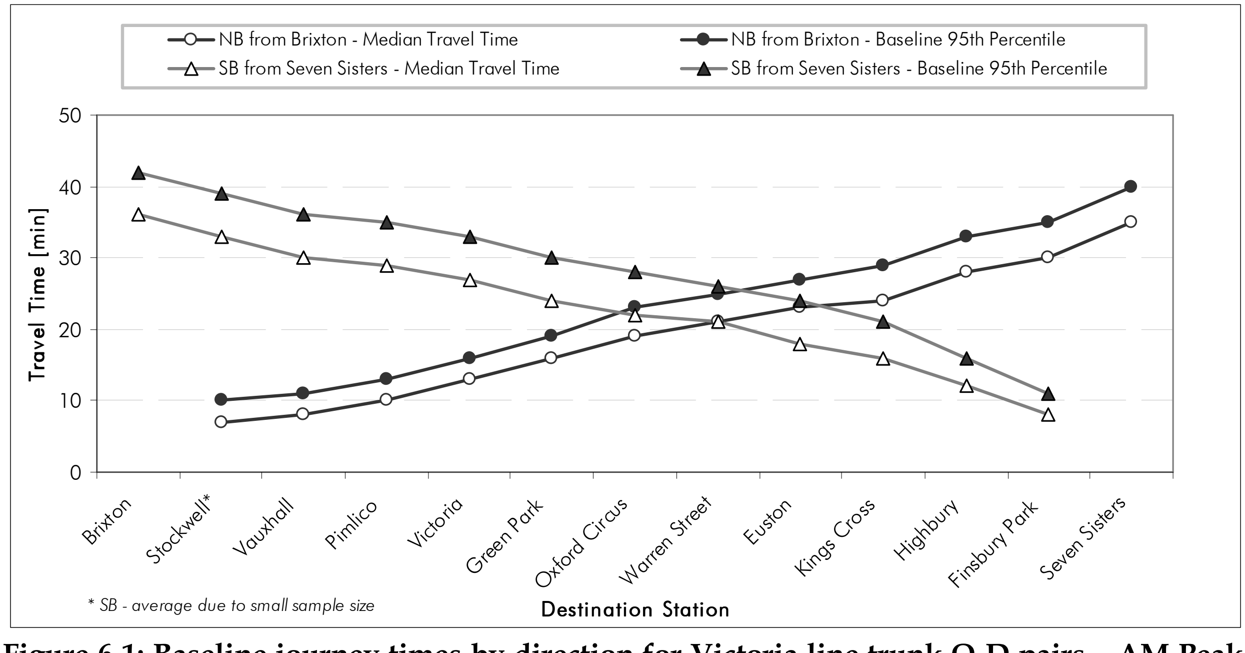


Figure 6-1: Baseline journey times by direction for Victoria line trunk O-D pairs – AM Peak, Feb. & Nov. 2007

The baseline reliability buffer time for each O-D pair is captured by the difference between the 95th and the 50th percentile travel time shown in Figure 6-1. The effect of journey length on the variability of travel times for Victoria line journeys is evidenced by the fact that there is an increase in the RBT from 3 minutes for a one-stop journey in both directions, to 5 and 6 minutes when reaching the end of the trunk portion in the north- and southbound directions, respectively. This represents an increase in buffer time of between 2-3 minutes due in part to journey length, which closely corresponds to estimates from section 5.2. Specifically, if a rate of increase in the reliability buffer time of 0.12 minutes for every minute of additional median travel time (i.e. journey length) is used, the difference in the buffer time for a one-stop journey and for a journey traversing the full trunk portion would be approximately 3.4 minutes (28 minutes of additional median travel time * 0.12 reliability buffer time increase rate). The performance for the remaining O-D pairs on the line was similarly estimated.

The recurrent reliability buffer time for the entire line is then estimated by aggregating the performance at the origin-destination pair level using the approach given by Equation 4-3. The recurrent performance at the line level by direction is estimated both using actual and uniform (i.e. unweighted by demand) passenger volumes in order to estimate the contribution of demand to the final aggregated performance. Figure 6-2 compares the performance of the Victoria line by direction for both passenger volume profiles.

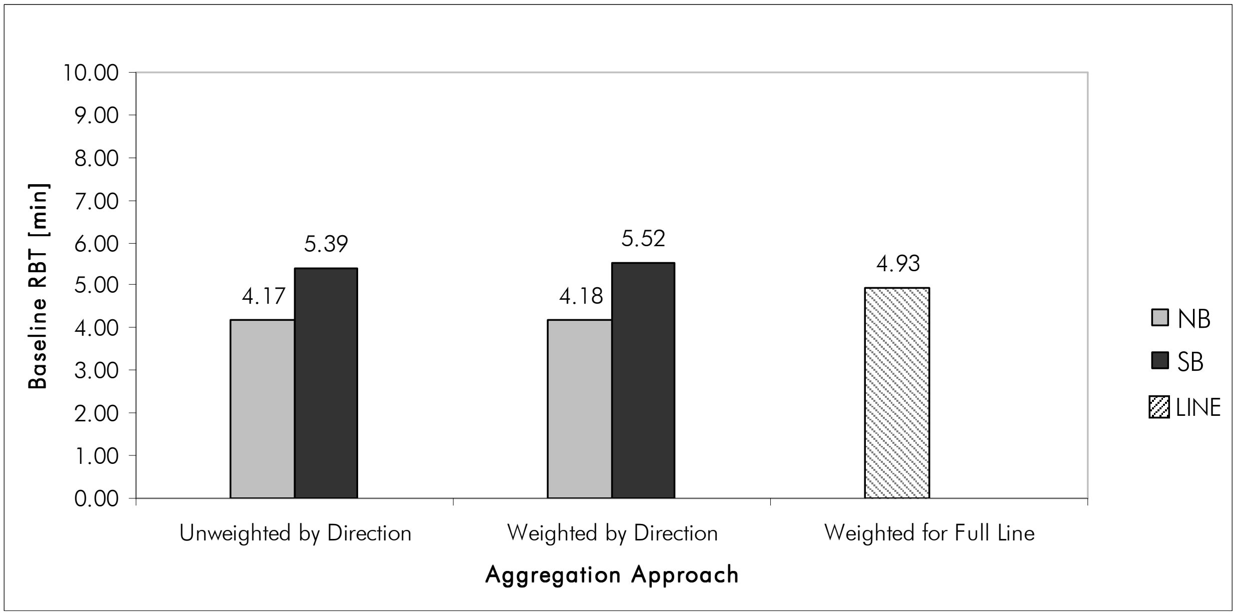


Figure 6-2: Baseline RBT – by direction and entire line, unweighted and weighted by passenger volumes – AM Peak, Feb. & Nov. 2007

Two important patterns emerge from Figure 6-2. First, when weighing performance by passenger demand, the northbound reliability buffer time had a negligible increase compared to the increase observed for the southbound direction. This could be because journeys on the highest-volume northbound O-D pairs were equally likely to experience unreliability above and below the average performance. Also, it could be interpreted as saying that southbound O-D pairs with higher-than-average passenger volumes also had a tendency to experience worse-than-average levels of reliability (i.e. higher reliability buffer times).

The second, more important pattern, has to do with the differences in the reliability buffer time for each direction of the line, both when weighted and unweighted by passenger demand.

These directional imbalances were first found by Chan (2007) when estimating the reliability of the Victoria line, and were explained by differences in performance and demand. Similar to Chan's findings, the southbound direction exhibited a much higher buffer time than did the northbound direction in this analysis. This is expected due to the different distribution of journey lengths and demand profile during the AM Peak for each direction. More specifically, the Victoria line has both a higher proportion of journeys making longer trips in the southbound direction than it does in the northbound direction, as shown by Figure 6-3.

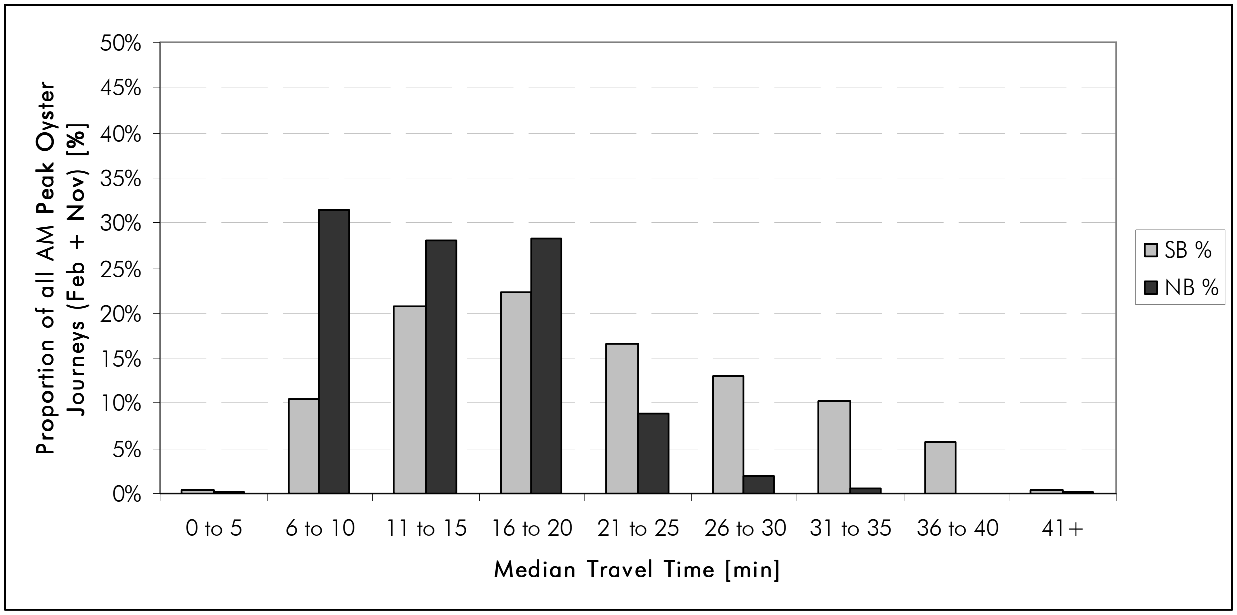


Figure 6-3: Proportion of journeys by median travel time for Victoria line by direction –AM Peak, Feb. & Nov. 2007

The higher proportion of longer journeys being made in the southbound direction for the Victoria line, as shown in Figure 6-3, helps explain the higher recurrent reliability buffer time observed in Figure 6-2 for the southbound direction, given the positive relationship between journey distance and travel time variability previously discussed. Finally, the right-most bar in Figure 6-2 represents the performance for the full line (i.e. both directions) during the AM Peak, which is the level of aggregation used within the context of JTM, and is adopted as the baseline performance on which to compare the overall level of reliability for the line, estimated next.

Estimation of Victoria Line Overall & Excess Reliability Buffer Time

In order to estimate the excess reliability buffer time at the line level, the overall RBT must be estimated first. Using Equation 4-2, the overall RBT for each of the 150 O-D pairs is estimated using a sample period of four weeks, in order to match the reporting frequency of JTM. After using Equation 4-3 to aggregate the performance of each origin-destination pair to the line level, an overall reliability buffer time is obtained for each individual four-week period (i.e. February and November). It then follows from Equation 4-8 that the ERBT at the line level can be easily estimated by comparing the overall RBT with the recurrent RBT estimated previously. Figure 6-4 compares the reliability of the Victoria line for each four-week period.

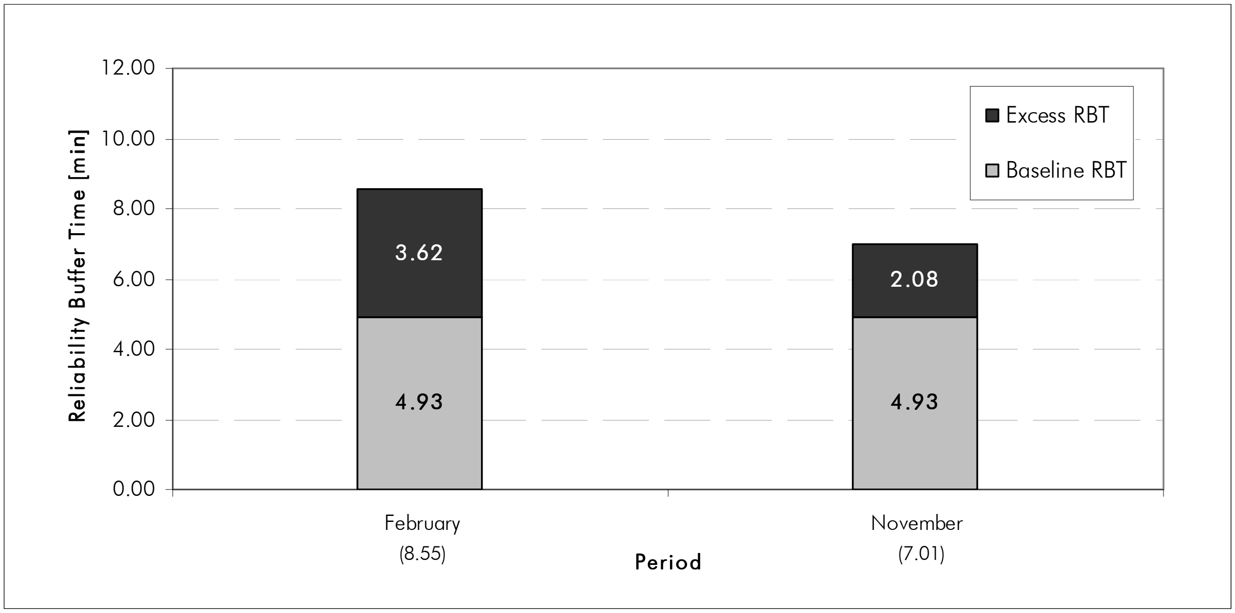


Figure 6-4: Victoria line Baseline and Excess RBT– AM Peak, Feb. & Nov. 2007

The line level results show that incident-related disruptions had a larger effect on reliability during the month of February, adding 3.62 minutes to the baseline buffer time. This is an increase of 73% in the amount of time passengers would be required to allocate to their travel to be sure of on-time arrival above that required under recurrent conditions. The corresponding value for the period in November is somewhat lower, showing a 42% increase in the baseline buffer time. Nevertheless, the contribution of disruptions to the total unreliability of the line is appreciable, being comparable in magnitude to the total contribution of recurrent factors, including the schedule and actual vehicle frequencies and passenger demand, as captured by the baseline performance. In addition, Figure 6-4 highlights the sensitivity of the line level reliability buffer time to changes in performance over time, which existing measures of average performance do not exhibit, making it a useful addition to JTM.

The JTM extension can also be used to obtain a more detailed view of performance by estimating the RBT for each direction. Figure 6-5 shows the baseline and excess RBT for the two observed periods by line-direction, and shows that the higher excess buffer time during February was largely a result of poor reliability in southbound performance, with a nearly doubling of the baseline RBT to a total of 10.74 minutes. In contrast, the northbound performance was more similar across the two periods, adding between 37-56% above the baseline buffer time.

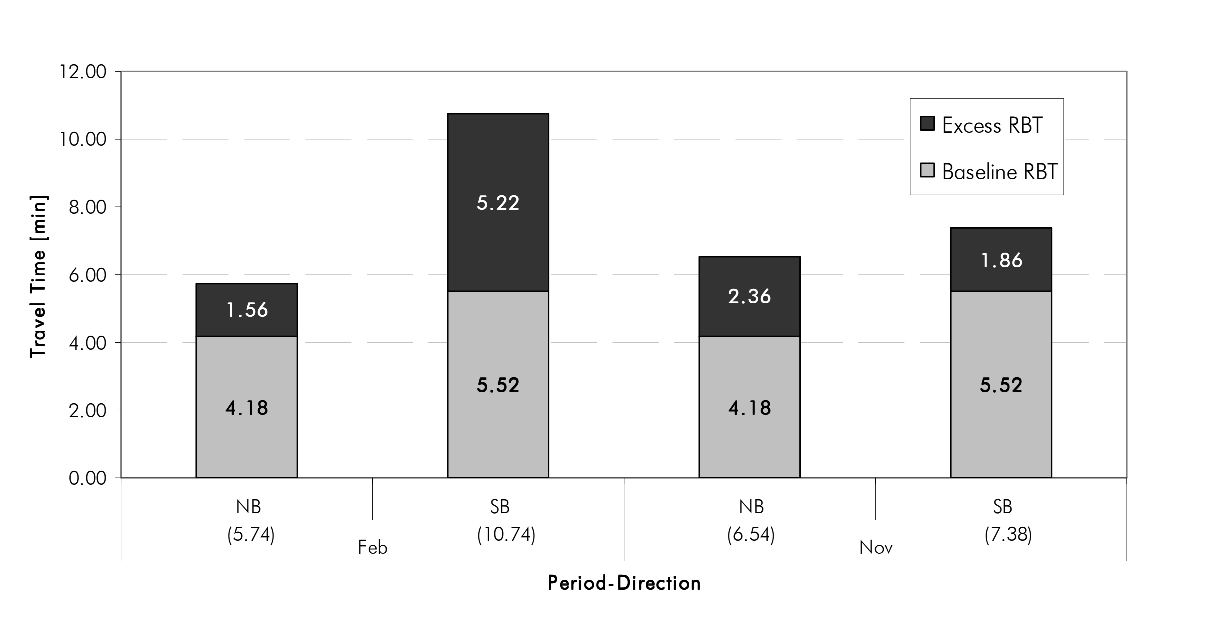
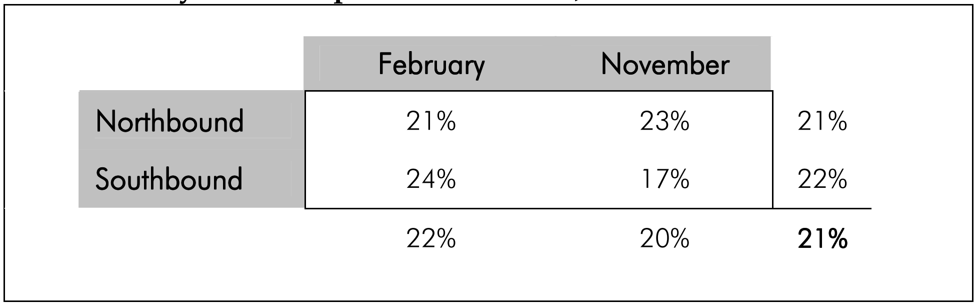


Figure 6-5: Victoria line Baseline and Excess RBT by direction – AM Peak, Feb. & Nov. 2007

The difference in the level of excess RBT between each direction of the line can be explained either through the severity of the individual disruptions, or through their rate of occurrence, or both. Even though a more detailed analysis is required to determine whether and why incident-related disruptions tended to lead to greater delays on the southbound direction of the Victoria line, it is possible to gain some insight by observing the frequency at which large disruptions took place. Taking the ratio of the number of days classified as incident-related to the total number of days during a period to be an indicator for the frequency of incident-related disruptions, it is possible to better understand the differences performance shown in Figure 6-5. For example, the ratio of incident-related days to the 20 weekdays in November shown on average across southbound O-D pairs was 17%, whereas this same value for the northbound direction was 23%, helping to understand the overall worse performance of the latter direction during this period. Table 6-1 summarizes the ratio of the average number of incident-days across O-D pairs by direction and four-week period.

Table 6-1: Proportion of days identified as incident-related for the Victoria line by direction-period – AM Peak, Feb. & Nov. 2007



	February	November	
Northbound	21%	23%	21%
Southbound	24%	17%	22%
	22%	20%	**21%**

As would be expected, the frequency at which incident-related days were identified corresponds to the changes in overall performance observed at the line-direction level for each period. In particular, the proportion of days identified as being affected by disruptions for the

Victoria line (21%) was similar to the rate found in section 5.2 (18%) for a broader sample of O-D pairs during these same periods. This highlights the importance of further understanding the underlying probability of experiencing a disruption, which could be similar throughout various parts of the system, motivates further analysis in this area (see section 7.3).

6.1.4: Contribution of Reliability to Perceived Travel Time

The final step in the estimation of JTM consists of weighing each travel time component by the value-of-time (VOT) passengers place on it relative to the value of the in-vehicle travel time. Though the actual value passengers place on reliability as represented by the buffer time is beyond the scope of this research, there are several empirical and theoretical estimates from previous studies that can be used to get a sense for the contribution of reliability to overall passenger perceptions of service quality, particularly as compared to the contribution of average performance. This section reviews the range of possible values of reliability, and applies them to estimate the contribution of reliability to the total travel time as perceived by passengers of the Victoria line.

Estimates of the Value of Time for Reliability

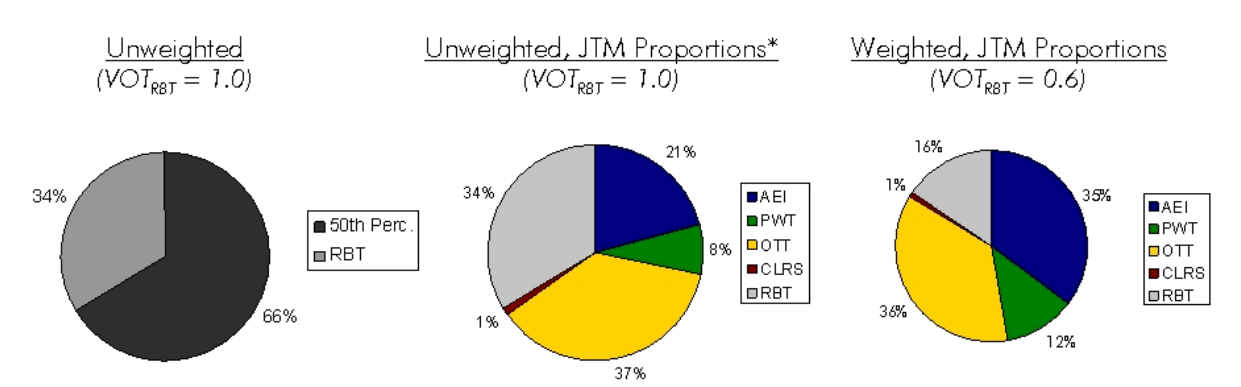
There are a range of estimates for the value of reliability found both through empirical and analytical methods. Two studies in particular propose values of reliability which may be useful. The study by Furth et al. (2006a) proposes a value for a reliability buffer wait time measure similar to the reliability buffer travel time measure of this study. It argues that the value for a minute of reliability should be 0.75 that of one minute of average travel time, with the criterion that the reliability value should be less than 1 because of the fact that during the majority of trips, the full buffer time is never actually realized (e.g. if spent at the destination after an early arrival). The second study, by Lam & Small (2001), determines several estimates for this value, whereby men valued reliability (as captured by the difference between the 90th and the 50th percentile travel time) at 0.66 of the median journey time, while women valued reliability at 1.40 of the median journey time.

Ultimately, the value passengers place on reliability will depend on many personal factors, including risk-averseness and personal preferences, but also trip characteristics like time of day and trip purpose. The next section explores the contribution of unreliability to total perceived journey time on the Victoria line, first with a direct comparison with the contribution of average journey time components as measured by JTM, and then by exploring the impact of unreliability over a range of possible values of time.

Contribution of Reliability to Victoria Line Service Quality

The overall reliability buffer times experienced by passengers during February and November are compared to the median travel time for both periods for the entire line in order to compare the contribution of unreliability to service quality relative to the contribution of average travel time components. In the analysis, the median Oyster travel time is substituted for the aggregate measure of travel time currently found by JTM, to ensure a consistent comparison as the latter is derived from other data sources.

First, a direct aggregation of the estimated median travel time value of 16.71 minutes, with the overall RBT estimated for February of 8.55 minutes leads to a total line level journey time of 25.26 minutes, unweighted by VOT. Then, in order to compare the contribution of reliability to total perceived travel time relative to the contribution of individual journey components, the median travel time for the line is broken down. To do this, the proportion of the total journey time for each component in JTM is used, observed during the same four-week period and normalizing out the Ticket Purchase Time component (TPT) because it is by default not captured by Oyster-based estimates of total travel time. The first two pie charts in Figure 6-6 reflect these calculations.



* total 101% due to rounding

Figure 6-6: Contribution of Victoria line JTM journey components and RBT to total journey time, unweighted and weighted by value of time – AM Peak, Feb. 2007

The leftmost pie chart in Figure 6-6 shows that the RBT for the Victoria line represents around 34% of total journey time when unweighted by VOT, with the median journey time composing the remaining 66%. The middle pie chart decomposes the median travel time into five individual journey components. Of these components, on-train time (OTT) provides the single largest contribution to unweighted journey time (37%), followed by Access and Egress time (AEI) with 21%. Platform Wait Time (PWT) has less of an impact, with 8% of total travel time, followed by the contribution of Closures & Incidents (CLRS) of around 1%. Even though these values represent the breakdown of each component to the total travel time experienced by passengers, they do not reflect the differential values with which they are perceived. In order to do so, a VOT for the reliability buffer time is introduced.

There are a range of empirical and theoretical values for the VOT for reliability. In order to obtain a lower bound on the contribution of reliability to total perceived travel time, a conservative value of time is selected (i.e. undervaluing reliability). Using a reasoning similar to that presented by Furth et al. (2006a), and representing the lower end of empirical estimates of the worth of one minute of buffer time relative to the value of one minute of in-vehicle travel time, $VOT_{RBT} = 0.6$ is selected.

Weighting the RBT by this value, and weighting the remaining components by the VOT used in JTM, the rightmost pie graph in Figure 6-6 is obtained. This initial estimate shows that the overall RBT contributed to approximately 16% of the total perceived journey time for trips on the Victoria line during the morning peak in February 2007. This contribution is comparable to that of other components of the journey such as the wait time (PWT), which contributed to 12% of total perceived journey time in this particular case. This indicates that by not explicitly measuring reliability, an important part of service quality is ignored.

Another important observation from this analysis has to do with the contribution of incident-related disruptions to unreliability compared to their contribution to average performance. For the Victoria line during this period, of the 16% of weighted (by value of time) reliability buffer time, 7% is due to incident-related disruptions (i.e. excess RBT). This contribution to total perceived travel time is large relative to the estimated contribution of Closures and Incidents to average travel time (CLRS) of around 1%, despite the higher value of time placed on the latter (VOT_{CLRS} = 2.01). This suggests that incident-related disruptions in the system have a large impact on service quality through their effect on unreliability, which would go unmeasured when focusing only on the impact of disruptions on average performance.

Finally, given that a value of time for the RBT was somewhat arbitrarily assumed, and that the estimated contribution of this component to total perceived journey time depends on this value, a sensitivity of the journey time component contributions (including that of RBT) to the value of time of the RBT is estimated and shown in Figure 6-7.

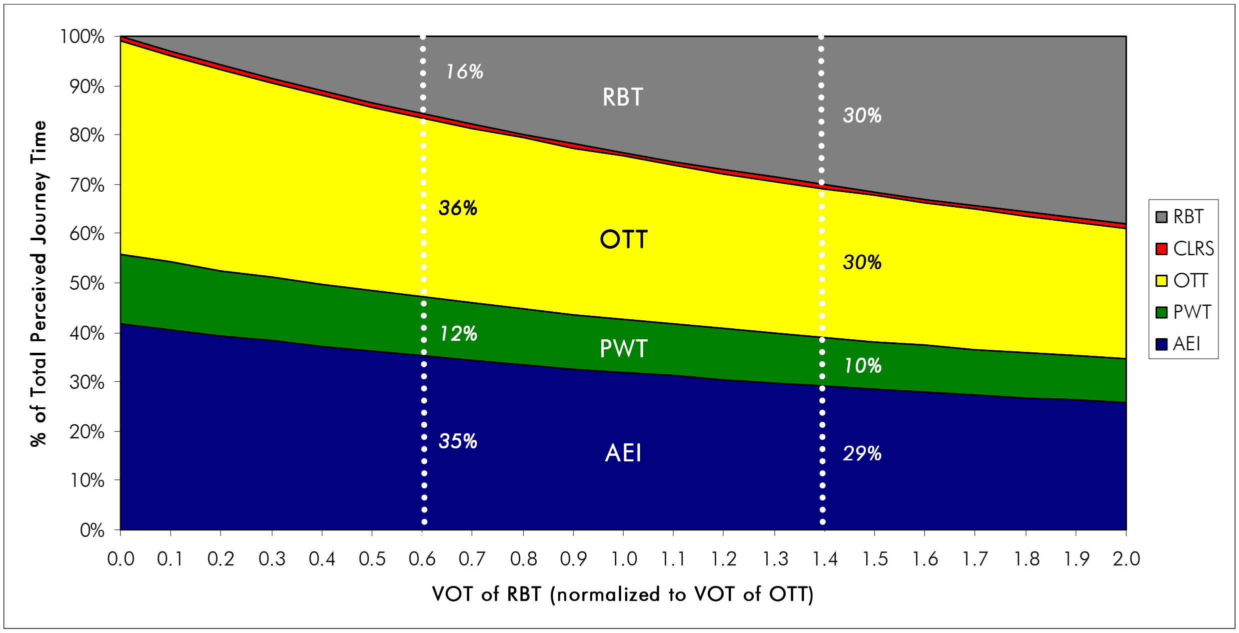


Figure 6-7: Sensitivity analysis of Victoria line perceived journey time across values of time for RBT – AM Peak, Feb. 2007

In Figure 6-7, when a (conservative) value of time for the buffer time of 0.6 was assumed, the contribution of the RBT to total perceived travel time was 16% as shown in Figure 6-6. As this value is incremented, the contribution of RBT to total perceived journey time increases appreciably. For example, when the VOT_{RBT} = 1.4, where a minute of buffer time is 1.4 times more important than a minute of average in-vehicle travel time, the RBT contributes nearly a third of the total perceived journey time. Such high values of time placed on reliability are plausible under certain circumstances (e.g. a large penalty for late arrival), and illustrate not only the potentially large contribution of reliability to the service quality of the Underground, but also the importance of explicitly measuring it in on-going performance monitoring efforts.

6.2: Reliability and Travel Information – TfL Journey Planner Application

The focus of the research to this point has been on quantifying the actual reliability of the system. Regarding the impact of reliability on passengers, as perceptions of travel time variability deviate from the actual variability of the system, the ability of passengers to minimize their disutility of travel through optimal choices diminishes. It is in ameliorating the gap between expectations and reality that passenger information becomes useful, as it allows passengers to make optimal travel choices in the face of uncertain travel times.

This section explores the potential use of the Oyster-based framework for supplementing the existing information provided to passengers through London's trip planning software, the Journey Planner, and providing passengers with reliability information. Section 6.2.1 reviews the way passenger misperceptions can worsen their travel experience, and what types of travel information can help mitigate the impacts of unreliability. Section 6.2.2 then quantifies the potential improvement in reliability that users of the Journey Planner would accrue by including travel time reliability information for the Underground. Finally, alternative ways to present reliability information to passengers through Journey Planner is explored in section 6.2.3.

6.2.1: Service Quality and Reliability Information

The variability of travel times increases the disutility of travel for passengers by forcing them to tradeoff between the time they allocate for completing a journey and the certainty of arriving at the desired time. This tradeoff, however, is subjected to the gap between passenger perceptions and the actual performance of the system. These misperceptions make it more difficult for passengers to make optimal travel choices, thereby increasing the disutility they experience above and beyond that already caused by the service itself. In the case of reliability, an over or underestimation of the level of variability would lead passengers to allow for too much or too little buffer time, respectively, than they would have had to actually budget in order to reach their destination on-time (see section 2.2.3).

The work by Ettema & Timmermans (2006) covers these ideas in-depth, and it postulates four types of travel information, summarized in Table 6-2, that can be used to ameliorate gaps between perception and reality: retrospective, descriptive (qualitative and quantitative), and predictive.

Table 6-2: Types of travel information (adapted from Ettema & Timmermans, 2006)

Type of Information		Contents
Retrospective		Historical conditions
Descriptive	- Qualitative	Current specific condition in qualitative terms
	- Quantitative	Current specific condition in quantitative terms
Predictive		Predicted specific conditions for given departure time

The four categories of travel information can be thought of differing along the degree of completeness of information that is presented to the user. Retrospective information, concerned with performance conditions that have already been realized, is useful for giving passengers a sense for the average performance (i.e. travel times) they can expect, with some typical variation (e.g. the historical travel time distribution). Descriptive information is concerned with the current state of the transport system, providing information either through qualitative (e.g. disrupted service on the Jubilee line) or quantitative means. It allows passengers to update their decisions at the time of travel, as opposed to only relying on knowledge of average historical performance, and can add value when provided above other kinds of information. Finally, predictive information provides, like descriptive information, a point value for where in the historical travel time distribution (observed through retrospective information) future conditions will lie. As with all forecasts, this will involve some level of uncertainty which the passenger must adjust for.

The authors also discuss how these different types of information can be useful for addressing different kinds misperceptions, including those regarding the accuracy or **correctness** of the travel information and its **completeness**. In the former case, misperceptions due to incorrectness might occur when passengers are making a journey for the first time and have no a priori information (i.e. from personal experience) to base their travel decisions on. The provision of the average journey time through a trip planning tool, for example, would help these passengers align their expectations of the service with its actual state, even if only providing a snapshot of the average conditions of the system (e.g. through retrospective information). The other type of misperception arises when passengers do not have complete information when making a journey. This forces them to account for a high degree of uncertainty in their choices, despite the accuracy of any information they may already have. This can be the case when passengers have a good understanding of the average conditions of the system but not any information on conditions at their time of travel, or the degree to which conditions may vary about the average time to complete a journey. For example, more complete historical information could include the expected variability of journey times. As another example, descriptive information could help passengers update their knowledge of the current state of the system.

Reducing both types of misperceptions will improve passengers' ability to make better travel decisions. In addition, under the assumption that information already provided to passengers is correct, the impacts of unreliability can be mitigated by improving the completeness of travel information as it reduces the uncertainty faced by passengers.

6.2.2: Analysis of Reliability and TfL Journey Planner Trip Information

This section explores the potential for using the Oyster-based framework to improve the travel information provided through Journey Planner with the objective of mitigating the effects of unreliability. An analysis is done to determine the extent to which current Journey Planner information is both correct and complete, and the impact improvements in this information could have on passengers. First, the characteristics of this tool are presented.

Description of Journey Planner

The Journey Planner[20] is Transport for London's trip planning tool made available to passengers over the internet and mobile devices. It presents different travel options to users for completing their journeys within the TfL network, including bus, rail and non-motorized forms of travel, as well as the expected journey duration for each trip.

For travel information regarding the London Underground, expected travel time estimates were obtained from operating schedules for the service. Users of the tool specify an origin and destination, among other travel preferences, and a departure time (by the minute) for the present or upcoming days. In return, they obtain each segment of the journey by mode, each with an expected travel time duration, scheduled time of departure, and advisories (i.e. qualitative descriptive information) on any planned engineering works and closures taking place at the time of travel.

The ability to measure individual passenger travel times through Oyster data and to quantify the reliability of different journeys provides an opportunity to improve upon the information currently provided by Journey Planner in two respects. First, it can help calibrate the currently provided scheduled travel times with the typical or median travel times experienced by passengers and as revealed by Oyster. This would help address the problem of the potential accuracy of the provided information. Second, by providing information on the variability of travel times, in addition to the already presented expected value, information on the performance for a given trip would be more complete, helping passengers make more informed decisions. A quantitative analysis is performed next for the 50 highest-volume origin-destination pairs in the Underground during the morning peak to determine the potential benefits from these two types of improvements.

Journey Planner Travel Time Information - Correctness

The correctness or accuracy of the expected travel time information currently provided by Journey Planner is compared to the typical passenger travel times as revealed by Oyster data. For the 50 highest-volume O-D pairs in the system, the expected travel time information provided by Journey Planner was found to consistently underestimate the actual typical travel time, as revealed by Oyster. Figure 6-11 shows the difference between the Oyster median journey time and the Journey Planner times for the 50 journeys observed.

[20] http://tfl.gov.uk/journeyplanner

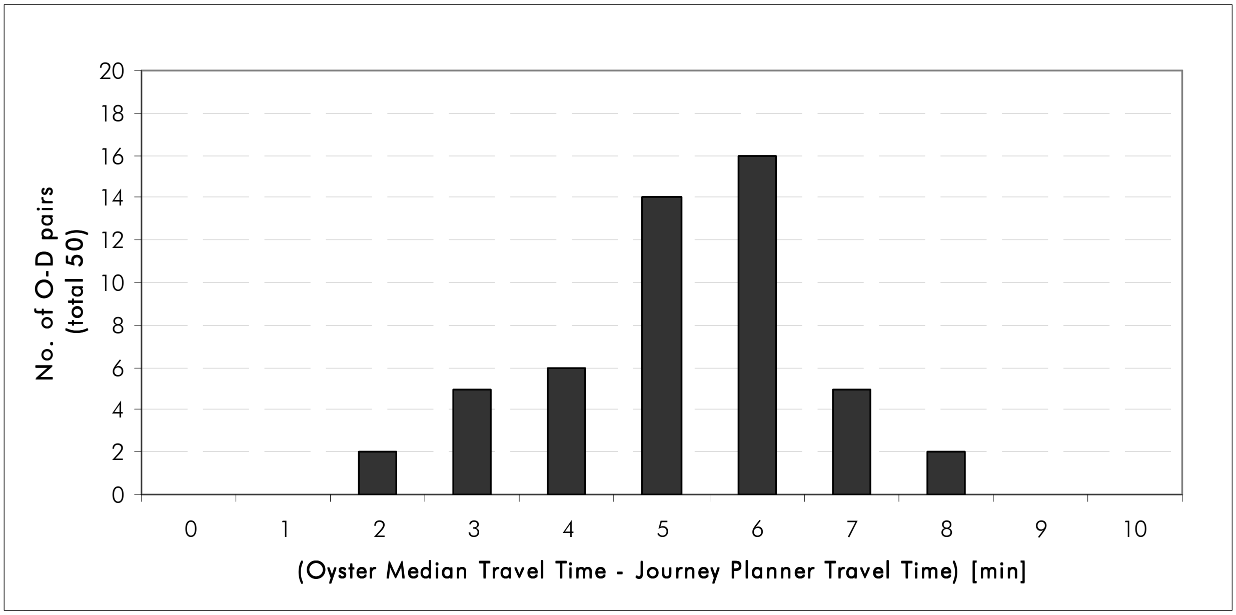


Figure 6-8: Histogram of the difference between Oyster median travel times and the Journey Planner travel times for 50 highest-volume O-D pairs – AM Peak, Nov. 2007[21]

Approximately two thirds of the 50 O-D pairs had an underestimation of around 4 to 6 minutes for the typical travel time. When normalized by the Oyster median travel time, this difference was found to represent between 30-50% for 43 of the 50 observed journeys. That is, not only was there a gap between the Journey Planner and Oyster travel time values, the gap also represented a large proportion of the total journey time that passengers had to account for. In addition, because of the shape of the travel time distribution (i.e. skewed to the right with the average greater than the median), a comparison of the Journey Planner times with the average Oyster travel times would only exacerbate the degree of underestimation.

This inaccuracy would also lead to greater uncertainty not only in terms of the expected travel time for a journey, but also with regards to the likelihood of experiencing a very large delay compared to the published information. In the particular case of the Journey Planner tool, as the expected travel time is underestimated, the probability that an individual will arrive at his or her destination before the desired arrival time (i.e. on-time) decreases, assuming reliance on the published schedule. Figure 6-9 shows the probability of on-time arrival using the travel time published in Journey Planner for the selected 50 O-D pairs. The large majority of passengers traveling on these O-D pairs had a probability of on-time arrival of less than 10% if they used the travel times published in Journey Planner.

[21] Journey Planner travel times for Figures 6-8 through 6-10 obtained December 2008.

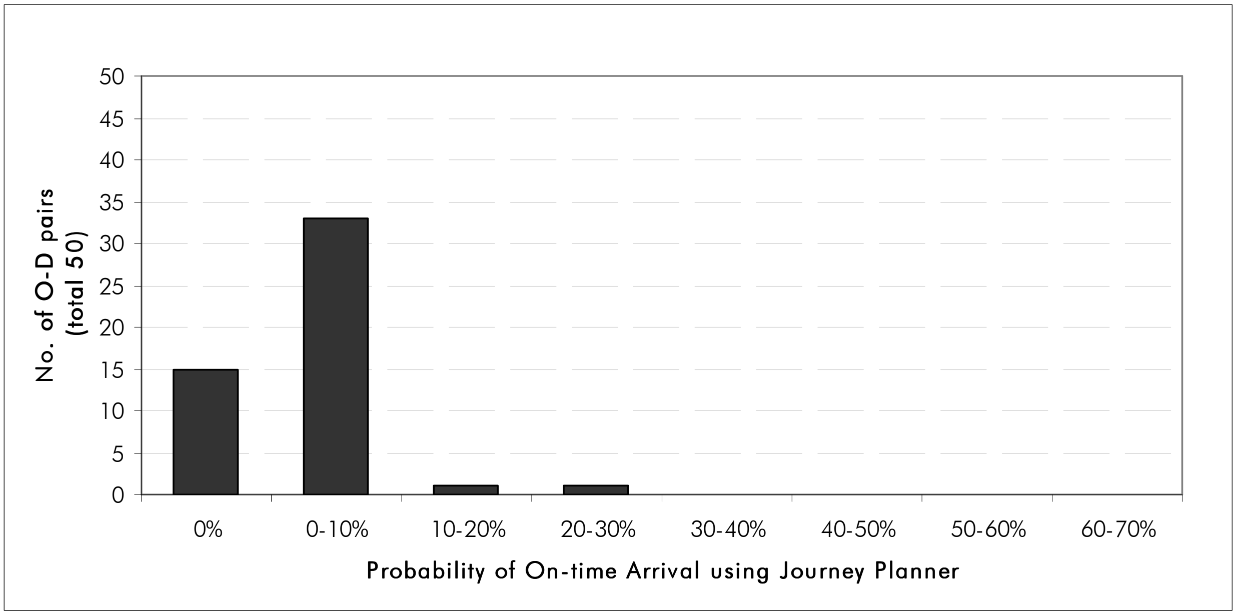


Figure 6-9: Probability of on-time arrival using Journey Planner travel times for 50 highest-volume O-D Pairs – AM Peak, Nov. 2007

One possible explanation for these results is that Journey Planner travel times are derived from the operating schedule, which includes only vehicle departure and arrival times at particular stations (i.e. excluding the time spent *within* a station). A straightforward improvement could come from using the median travel times historically observed from Oyster data, with the immediate improvement of guaranteeing that at least 50% of all journeys would arrive on-time at their destination, if they relied solely on the information published in Journey Planner. Regardless of the accuracy of the published expected travel times, information on the variability of travel times would still be required to help passengers increase their chance of on-time arrival at their destination.

Journey Planner Travel Time Information – Completeness

In addition to having information on the expected travel time for a journey, passengers would benefit from information on the distribution of travel times. As discussed in section 2.2.2, as the variability of travel times increases, the probability of late arrival increases for the same amount of time allocated to complete a journey. Providing passengers with the reliability buffer time for a particular journey would help them arrive on-time with a greater degree of certainty, than if they had relied solely on expected travel time information.

Two types of information could help reduce this kind of uncertainty for passengers. First, the historical variability of travel times could be provided. This would allow passengers to select an optimal departure time based on previously observed performance. Second, historical information could be supplemented by real-time advisories on current travel conditions, which could warn passengers of abnormal situations (e.g. incident-related performance). This second source of information could help passengers make a judgment as to the probable location within the travel time distribution a journey on that particular day would fall on (i.e. the tail of the distribution). Also, if the historical performance focuses on recurrent conditions, real-time travel advisories could help passengers determine the degree to which the historical travel time distribution would hold to be true, and whether any additional countermeasures would be needed to complete the journey on-time.

The importance of providing passengers with information not only on the expected journey time but also on its variability is illustrated by estimating the ratio of the Oyster 95th percentile travel time to the median travel time for each of the 50 highest-volume O-D pairs in the Underground. Figure 6-10 summarizes the results, showing that for 43 O-D pairs, passengers would be required to allow for a buffer of *at least* 40% the median travel time in order to be 95% certain of on-time arrival.

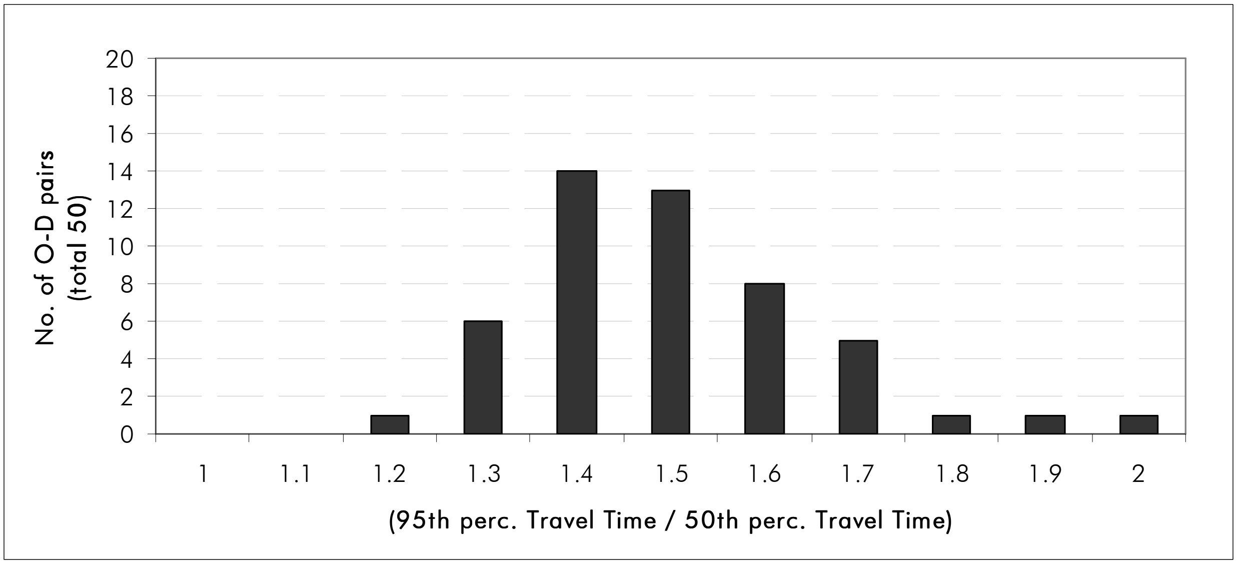


Figure 6-10: Histogram for the ratio of the Oyster 95th and the 50th percentile travel times for the 50 highest-volume O-D pairs – AM Peak, Nov. 2007

The results from Figure 6-10 indicate that even if passengers were provided with accurate information on the expected journey time (as captured by the Oyster median travel time), they would still have to leave a significant margin in order to counter the effects of travel time variability. In particular, by providing passengers with objective information on the variability of travel times, their reliance on perceived levels of reliability is reduced resulting in better travel decisions.

The next section discusses how reliability information might be presented to passengers in a way that is compatible with the Journey Planner, assuming that the benefits of using Oyster as a data source for travel times will by default increase the accuracy of the information.

6.2.3: Reliability Addition to TfL Journey Planner

Two important aspects to consider when presenting reliability information through the Journey Planner are the type of information given to passengers and the way it should be displayed.

The first consideration includes deciding between presenting passengers with the historically observed level of travel time variability, or whether the travel time variability under typical conditions (i.e. recurrent performance) should be provided. As observed in previous chapters, because of the unpredictable nature of incident-related disruptions, if the overall travel time variability was provided to passengers, it could oscillate from period to period, reducing the effectiveness of the information. That is, knowledge of the travel time variability of the previous month might not be helpful to passengers trying to be sure of on-time arrival during the *current* month. One way to address this issue is to provide passengers with information on the recurrent performance of the system, which is much more stable over time, and supplement it with real-time travel advisories warning of conditions at the time of travel (i.e. qualitative – descriptive information). This would allow passengers to understand the degree to which any reliability information is accurate (i.e. in the absence of severe disruptions), reducing any chances for providing misleading information about the system.

However, for a system like the Underground, the decision to provide information on the recurrent reliability to passengers is not straightforward as it must tradeoff between providing more accurate (i.e. better forecast) information to passengers, at the cost of being less comprehensive. This is important given the high rate of occurrence of incident-related disruptions (around 20% chance), where if only recurrent performance information is taken into account by Journey Planner, passengers would not be protected from a late arrival at their destination around 1 in 5 journeys. Conversely, providing the overall buffer time (i.e. including incident-related effects) would likely lead passengers to over-budget the amount of time to complete a journey, thus leading to excessively early arrivals at the destination. The way passengers trade off these two considerations should be studied further by market research using estimates of reliability under both types of conditions from the framework developed here. This research recommends providing passengers with the recurrent level of travel time variability, supplemented by descriptive – qualitative information at the time of travel, warning of typical or disrupted service conditions.

The second consideration on the most appropriate way to present reliability buffer time information to passengers also deserves some discussion. Ultimately, travel time information attempts to reduce uncertainty, making it important to consider whether passengers should be provided with exact numerical values for the reliability buffer time for a particular journey, or to provide a range of values; the main tradeoff considering the degree to which passengers might misinterpret a precise numerical value as being a "guarantee" of on-time arrival, when in fact it is simply a probabilistic figure aimed at helping them improve their chances of completing a journey on-time in all but the most extreme cases. On the other hand, providing reliability buffer time information in a less exact format (e.g. range of values, qualitative description) might be difficult for passengers to interpret, or maybe too imprecise for those wishing to rely on it anyway.

One simple way to present buffer time values while avoiding an overly precise prescription might be to present a Reliability Buffer Time Index, where buffer times are grouped into discrete 5 minute intervals. This would give the passenger a range of values for how much time to allocate in addition to the expected travel time already being provided. This approach has the weakness that grouping reliability buffer time values into discrete ranges (e.g. of 5 minutes in size) might lead to such a low level of resolution that the information is less useful to passengers. In the case of the Underground, a large proportion of all journeys have a median travel time of less than 20 minutes. Using a range of values that is too blunt might represent a high proportion of the expected time for the journey, reducing the effectiveness of the measure as compared to passengers simply adding their own margins using their intuition (e.g. add 10 minutes of buffer time to most journeys). On the other hand, if the size of the discrete buffer time range becomes too small, it is simply better to return to the provision of a point estimate. On balance, it is recommended that passengers be provided with a precise recurrent buffer time value for their journey, supplementing it with descriptive – qualitative information on the conditions of the system at the time of travel. A simple example for how this information might be presented as a part of Journey Planner is included next.

A Simple Example – Illustration of a Journey Planner Reliability Addition

A simple journey within the Underground is used to illustrate not only the benefit to passengers in terms of providing more accurate information through Oyster data and additional information on reliability, but also as a way to propose a visual layout for the information.

One could imagine a passenger making a journey for the first time from her home in Bow Road, a residential area of London in zone 2, to her 9:00am business appointment in St. James' Park station in zone 1. Figure 6-14 maps the journey, showing that it is a one-seat journey eastbound on the District line.

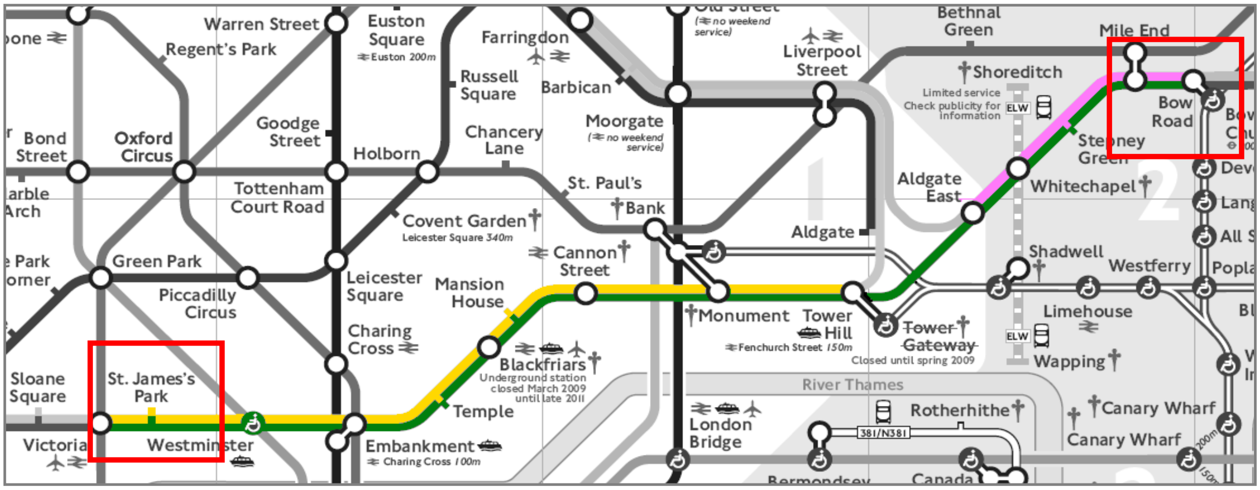
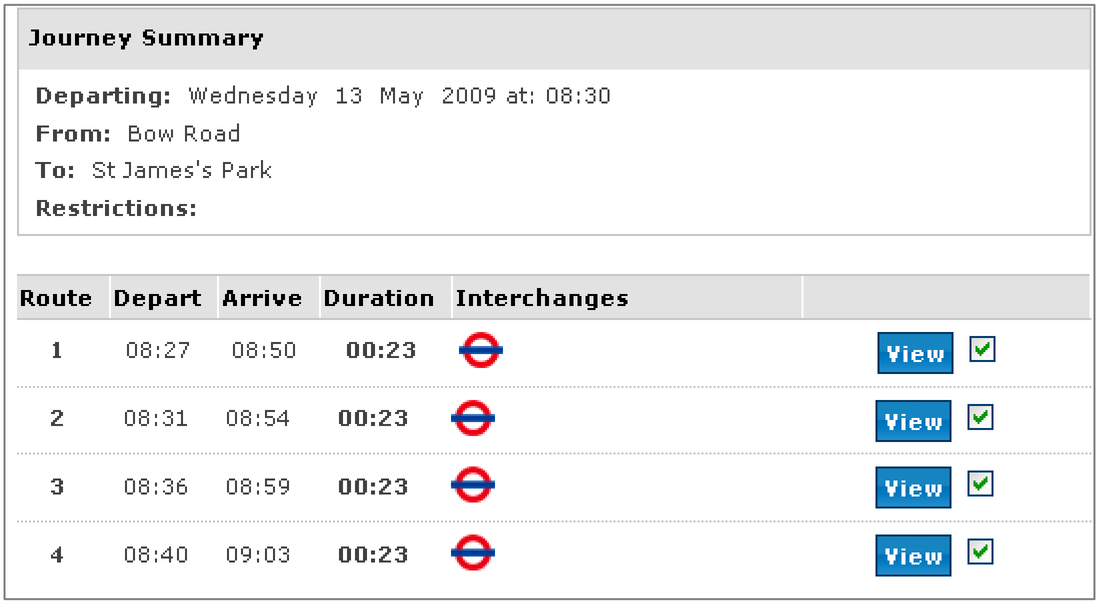


Figure 6-11: Illustration of the journey between Bow Road station and St. James' Park on the District line

Due to the purpose of the trip, and the fact that it is the passenger's first time completing it, the Journey Planner tool is consulted to select an appropriate departure time to ensure an on-time arrival the destination (i.e. business appointment). This will produce several different departure time choices for the Underground, each presenting a journey time of 23 minutes, and therefore a recommended departure time of 8:36am[22] for an arrival 9:00am on-time arrival. Figure 6-12 is a snapshot of the information for this trip from Journey Planner.



Journey Summary

Departing: Wednesday 13 May 2009 at: 08:30

From: Bow Road

To: St James's Park

Restrictions:

Route	Depart	Arrive	Duration	Interchanges	
1	08:27	08:50	**00:23**	⊖	View ☑
2	08:31	08:54	**00:23**	⊖	View ☑
3	08:36	08:59	**00:23**	⊖	View ☑
4	08:40	09:03	**00:23**	⊖	View ☑

Figure 6-12: Snapshot of travel information for Bow Road to St. James' Park provided by Journey Planner

Based on the inaccuracies of Journey Planner information discussed in the previous section, this passenger can expect to have a very low chance of arriving on-time if she departs at 8:36am. To obtain a sense for the likelihood of late arrival, the predicted travel times from Journey Planner are compared against the travel time distribution for that particular journey as revealed by Oyster. Figure 6-13 plots the cumulative travel time distribution for the journey using data from February 2007, and the 23 minute travel time recommended by Journey Planner. From the figure it is possible to see how using the Journey Planner travel time would lead to a probability of on-time arrival of less than 1% (i.e. the minimum recorded journey time).

[22] Journey Planner travel times for Figures 6-12 and 6-13 obtained May 2009.

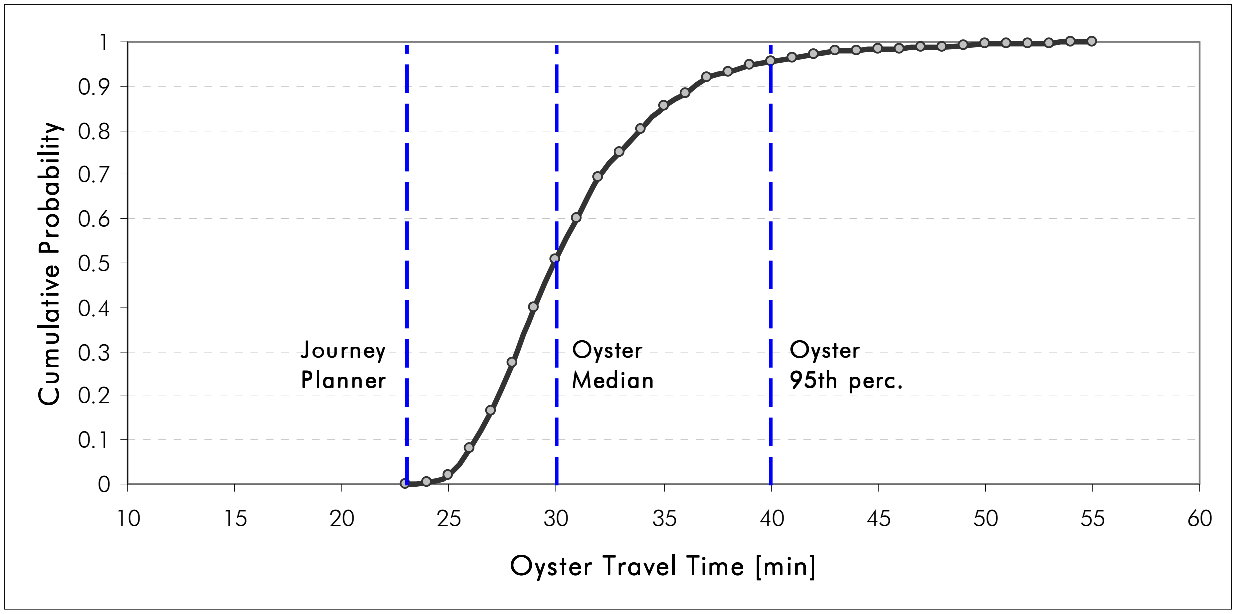
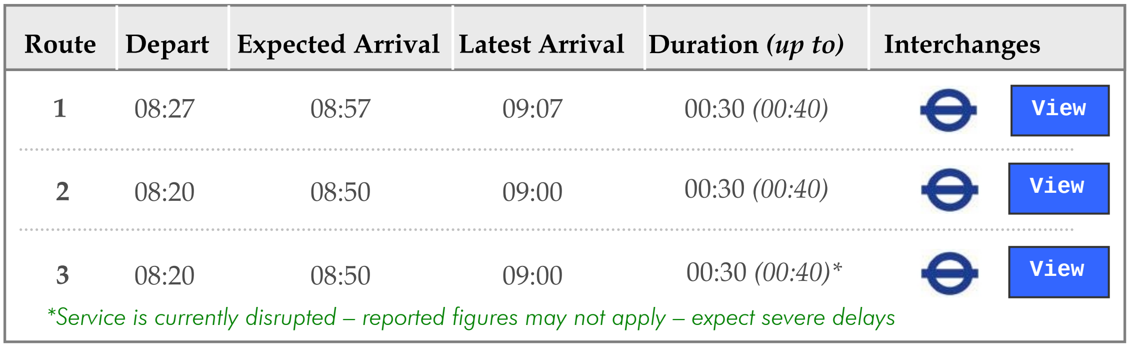


Figure 6-13: Cumulative probability distribution for the recurrent performance from Bow Road to St. James' Park station – AM Peak, Feb. 2007

In this example, providing passengers with the median journey time from Oyster would have recommended a departure time of 8:30am with an expected arrival time at 9:00am. This, however, would not address the fact that the probability of on-time arrival would remain low at only 50% when using the median. Therefore, the reliability buffer time should be presented to passengers, in addition to the median travel time, with an illustration of a possible presentation format shown in Figure 6-14.



Route	Depart	Expected Arrival	Latest Arrival	Duration *(up to)*	Interchanges
1	08:27	08:57	09:07	00:30 *(00:40)*	View
2	08:20	08:50	09:00	00:30 *(00:40)*	View
3	08:20	08:50	09:00	00:30 *(00:40)**	View

**Service is currently disrupted – reported figures may not apply – expect severe delays*

Figure 6-14: Proposed presentation of reliability information in Journey Planner

Three possible journey options that passengers could follow are presented in Figure 6-14. The first option (Route 1) is designed for users that select the "Depart at" feature of the Journey Planner (i.e. not concerned about on-time arrival). It includes the 95th percentile travel time simply as additional information next to the expected journey duration. This is shown by the value in parentheses under the "Duration" field, which is the 95th percentile travel time for a journey departing at 8:27am.

The second travel option (Route 2) is intended for passengers who select the "Arrive by" feature of the Journey Planner, hence expressing a desire to minimize the chances of late arrival at their destination. In this case, the software could recommend the departure time that would guarantee an arrival by 9:00am with 95% certainty (i.e. using the 95th percentile travel time). This option would then simply indicate to passengers that by departing earlier, they could also expect to arrive a few minutes before their desired arrival time. This is captured by the field "Latest Arrival", which simply indicates that the passenger will arrive at her destination at 9:00am with 95% certainty using that earlier departure time, and can expect to arrive at 8:50am around half the time.

The third travel option shown in Figure 6-14 shows the same recommended travel times as before for this same journey, however with a qualitative warning about disrupted conditions on the line. This feature, which currently exists in Journey Planner and is updated in real-time, informs passengers that the travel times presented need not apply, and that they should factor in an additional amount of time to be safe. The exact amount of additional time to be allowed for under disrupted conditions is not included in this example but is possible to estimate roughly using the Oyster-based framework. In the general case, however, it is not recommended.

A final advantage from using Oyster data comes not only from its ability to provide passengers with more accurate (i.e. including all components of a journey) and more complete (i.e. reliability buffer time) information through Journey Planner, but also in terms of its ease of implementation. This source of data makes it possible to estimate, at low marginal cost of collection, the travel times for each individual trip in the network for a season or year at a time based on the previous historical performance. As the relationship between service schedules and passenger travel times as revealed by Oyster is better understood, passenger information could depend less on the historically observed travel times, and begin to make use of forecasts in the event of a change in the service.

Finally, providing reliability information on tools like London's Journey Planner could complement efforts from the operations side to reduce uncertainty for passengers, thereby increasing their ability to make more informed choices that in the end serve to improve the service quality of the Underground. The set of reliability measures proposed in this framework could be of immediate use to TfL towards this effort given their ease of interpretation to passengers and their cost-effectiveness to estimate on behalf of the transit agency.

Chapter 7: Research Summary & Conclusions

This chapter concludes the thesis with a summary of the research in section 7.1, including a description of the objectives and set of questions this work set out to answer and what was achieved in addressing them. Section 7.2 summarizes the main findings of the research and their broader implications, drawing from specific conclusions to support them. Lastly, section 7.3 discusses a set of future research directions, distinguishing between more immediate extensions and the broader challenges to be approached in the medium to long-term.

7.1: Research Summary

For over four decades, the concept of reliability has been of great concern to transit agencies and researchers not only because of its effects on the cost-effectiveness of the operation, but also because of its impact on the quality of the service as perceived by passengers. The ability to quantify and understand reliability in the latter context has slowly increased over the years, particularly as additional theoretical and empirical work developed and applications supported by new sources of data such as AVL began to appear (Furth et al., 2006b).

This research has the overarching objective of bringing together much of the work to date on reliability in order to develop a set of measures for quantifying and understanding this dimension of performance from the perspective of passengers. The proposed reliability measures take advantage of the availability of individual journey time information obtained from Smart Card data in systems like the London Underground, where the required validation upon exit from the system made this type of analysis possible.

A framework was developed consisting of set of reliability measures, based on empirical observations from this and previous work regarding the way travel time variability changed across days and across O-D pairs in the Underground with different characteristics. It made use of these insights to break down performance into different a priori categories (i.e. classification approach), themselves related to the different types of factors that affect reliability, including incident-related disruptions and the characteristics of the service. Through the proposed framework, insights into the causes of unreliability were obtained, as well as a practical approach for quantifying reliability that could be used as an input into the evaluation of performance.

Finally, this work developed a set of initial applications based on the proposed framework to help transit agencies improve reliability in the near-term with a focus on enhancing the existing service quality monitoring regime for the Underground and expanding on Transport for London's primary trip planning tool for passengers.

7.2: Summary of Findings

There are four overarching findings from this research. While they are empirically developed in the context of the Underground, they reflect the broader contribution of this work to the on-going body of research on the topic of service reliability. These findings are:

1. **Reliability is an important part of service quality in the Underground, relative to average performance, and should be accounted for explicitly.** This research quantified the level of unreliability that passengers experienced at the individual origin-destination pair and line levels. Initial estimates showed that when reliability is factored by the value of time passengers place on it relative to how much they value average performance, its contribution to total perceived journey times was comparable to the contribution of other important trip components like platform wait time. For the specific case of the Victoria line in London, a lower bound on the contribution of reliability to total perceived journey time was estimated at 16%, which was comparable to the perceived contribution of waiting time (12%). This suggests that the current focus on average performance ignores an important "hidden" part of service quality, and foregoes capturing any benefits to passengers that might accrue due to upgrades in the infrastructure, tighter service controls, and other actions leading to more consistent service overall. The framework proposed and developed here can be used to capture the impact of unreliability on passengers in the context of the routine monitoring of the system.

2. **Incidents have a large impact on service quality through reliability, which may be underestimated through a sole focus on average performance.** The framework made it possible to quantify the contribution of incident-related disruptions to unreliability above the contribution of service characteristics and performance under typical conditions. In the particular case of the London Underground, incident-related disruptions took place often and contributed to a significant portion of the overall level of unreliability experienced by passengers. Initial estimates of this contribution for the Victoria line indicated that around 40% of the overall unreliability could be attributed to incident-related disruptions, or approximately 7% of total perceived journey time. This was estimated to be greater than the effect of closures and incidents on average delays, which for this same line were found to contribute around 1% of total perceived journey time. The ability to quantify the reliability of the Underground with a specific consideration of non-recurrent incidents allows for the development of strategies aimed at cost-effectively improving service quality. Towards this effort, the reliability framework can be used to enhance existing measurement systems to measure delays produced by incidents (see Section 7.3).

3. **The occurrence of severe disruptions in the service can be detected through individual passenger travel times as captured by AFC Smart Card data in systems with entry/exit-validation.** This research also demonstrates that the occurrence of incident-related disruptions in the service can be detected by observing changes in passenger travel times. One particular case of this was developed in this research, showing how one can reasonably detect the occurrence of disruptions in the system

using passenger data from Oyster, where these detected disruptions were validated using actual incident log reports.

4. **The framework can be used to provide reliability information to passengers through existing information systems, helping to mitigate the effects of uncertainty on service quality.** Misperceptions of the actual level of travel time variability in the Underground can lead passengers to make sub-optimal decisions when trying to deal with unreliability. Providing passengers with information on the actual level of reliability reduces their uncertainty about the performance of the system and mitigates the effects of unreliable service. An application of the framework was developed for one of TfL's passenger information systems – the Journey Planner tool. It was found that the information currently provided by the tool led to a low probability of on-time arrival, and passengers were required to allow for a high proportion of the currently presented travel time as a margin.

7.3: Future Research Directions

There are several threads of research stemming from this work that transit agencies and analysts could find useful to explore. In some cases, these threads are simple extensions to this research, while in others, they present new directions to be explored by future work. This section highlights the most important additional work that can be developed by future researchers.

- **Perform reliability analysis at higher levels of resolution.** One of the benefits of the AFC-based reliability framework is that because of the high level of disaggregation of the data, it is possible to measure performance at very high levels of resolution. An important application of this could involve studying how performance changes within the AM Peak (i.e. at the 15-minute level), allowing analysts to identify the contribution of changes in demand to unreliability, in addition to the factors already studied in this thesis. Preliminary analysis shows that the methodology developed here is scalable to smaller time intervals, sample size notwithstanding, at the origin-destination pair level.

- **Extend performance categories through a clustering approach.** The methodology developed here as part of the framework assumed a priori categories for classifying performance (e.g. recurrent and incident-related). However, it is conceivable to attempt to develop a new set of categories a posteriori, based on the results from the data through statistical cluster analysis. This could provide room for new insights into the performance of the system, and possibly a more refined breakdown of performance. Some of these additional performance categories could be used as a further input into the evaluation of the service, providing operators with a higher level of resolution for establishing a baseline performance (e.g. "best case" performance conditions).

- **Revisit existing incident management tools and improve their estimates through AFC-based measurement of performance.** One of the inherent advantages of AFC-data is that they provide researchers with *measurements* of passenger's experienced travel times, as opposed to *estimates* from models and analytical approaches. These direct measurements could be useful in terms of the study and estimation of the effect of incidents on passenger delays. In the case of the Underground, the reliability framework could be used to revisit some of the estimates from the Nominally Accumulated Customer Hours (NACHs) system and provide additional insight into the way passengers are really affected by, and even respond to, the occurrence of disruptions in the system.

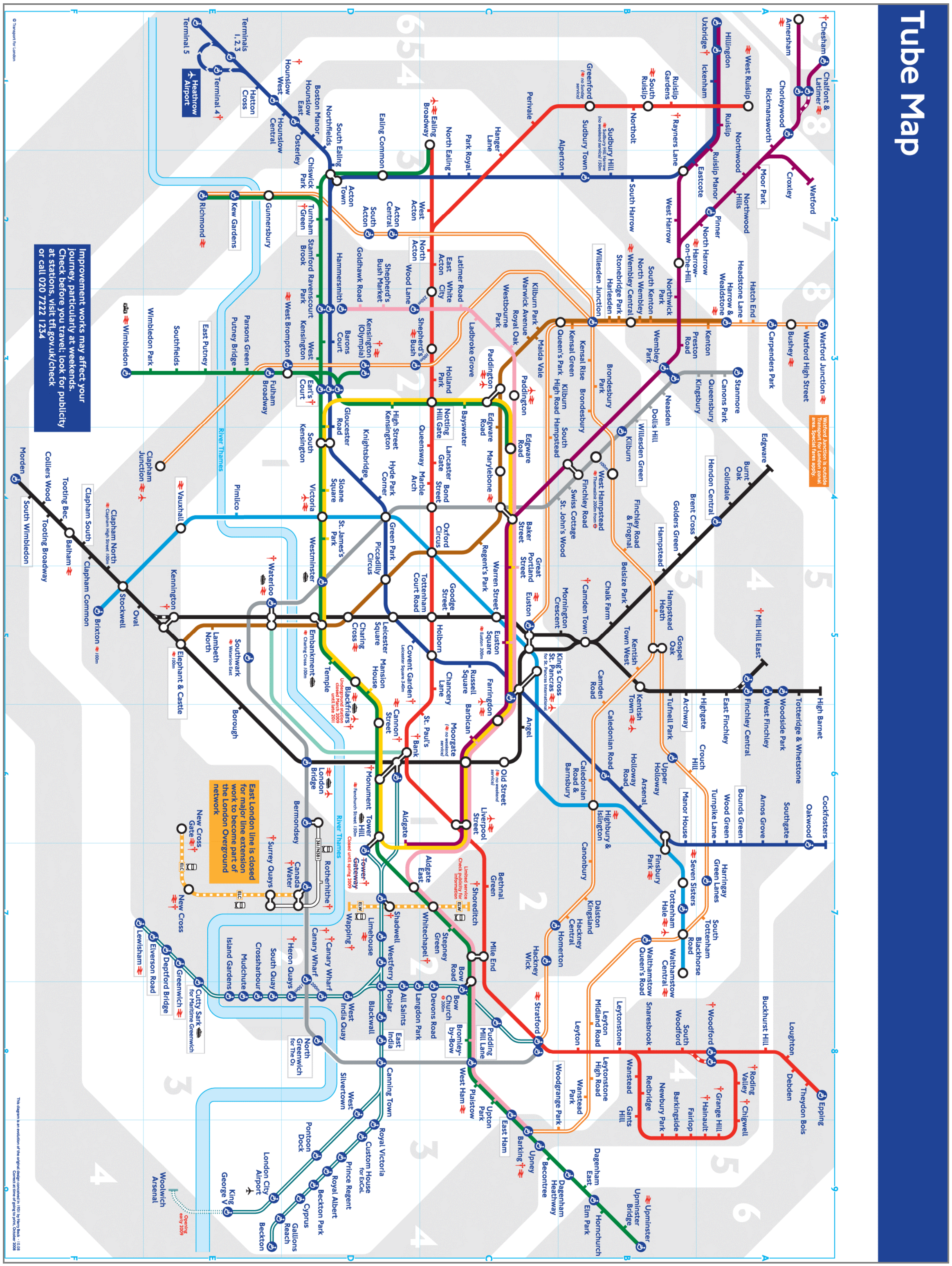
- **Explore the use of AFC-based reliability measurements as inputs to the evaluation of service quality.** One of the immediate uses of the reliability measures proposed through the framework is in the context of service evaluation. The ability to capture a recurrent level of performance could, in theory, help control for the contribution of service characteristics to unreliability. This would be useful when evaluating the service, as a baseline performance that largely captures the effects of the execution of the operation on reliability could be estimated. A preliminary analysis to determine the degree to which the ERBT is independent of the effects of service characteristics (i.e. the recurrent RBT controls for the contribution to unreliability of service characteristics), did not provide strong evidence to support this claim. Specifically, there was a non-negligible degree of correlation between the ERBT and the median travel time (representing journey length) for two periods on the Victoria line. Further research could be conducted not only to determine the strength of this relationship, but also to develop extensions to the proposed set of reliability measures that better separate between the unreliability caused by the characteristics of transit service and the delivery of the operating plan.

- **Study passenger behaviour and preferences, including how reliability affects departure time and path choice, and how passengers value this attribute of service.** The ability to assess the reliability of the system could be used for more detailed analyses of passenger behaviour and preferences. That is, path and departure time choice models could, through additional data (e.g. path choice surveys), be refined by taking into account the reliability of the different travel alternatives. From this type of analysis, the value of reliability relative to average journey time could also be determined empirically, serving as an input into the evaluation of service quality.

- **Estimate the contribution of additional factors to unreliability, including operations control interventions and additional service characteristics.** In this study a first attempt to quantify the contribution of some of the factors to reliability was made using a regression analysis. The evidence from this analysis showed that there is a large potential to further explore the contribution of other factors of unreliability (e.g. scheduled frequency, headway regularity, etc.). More interestingly, measurements of changes in reliability could be studied alongside operations control interventions to understand how passengers are affected by decisions made regarding the delivery of the service.

- **Develop more accurate analytical or simulation-based approaches for estimating the reliability experienced by passengers using the AFC-based framework as a reference point for calibration.** The data used for this study depended on the existence of both entry- and exit-validation at the system in question in order to estimate total passenger journey times. However, for heavy rail systems with only entry-validation (many current AFC systems fall within this category), as well as bus modes, estimating reliability using AFC Smart Card data requires inferences regarding journey destinations thus posing more challenges. Therefore it is important to use the ability to directly measure travel times to develop other approaches for *estimating* reliability without relying on exit-validation information. These additional approaches could be analytically- or simulation-based, with the expectation that systems lacking exit-validation information can benefit as well from an improved ability to quantify reliability.

Appendix A: Map of System and Fare Zones



Source: *Transport for London website (http://tfl.gov.uk), downloaded October, 2008*

Appendix B: Data Classification Method & Software Code

<u>Data Classification Method:</u> Stepwise Regression

This procedure classifies observations into two categories by finding the subset of observations within a sample that differed significantly from the remainder of the observations, with the expectation that the latter subset have similar values and occur a majority of times, and the former subset have very different values that do not repeat.

It is based on a sequence of regression estimates, where each model specification represents a different subset of observations being classified as a separate category. During each iteration the procedure compares the estimation results from the latest specification with the results from the previous specification through an F-test, in order to determine which observations should be included or excluded for an improved fit around the observed data. This approach follows the general structure of a stepwise regression procedure, explained in detail in Draper & Smith (1998), and more generally in Wolberg (2006). The classification approach applied in this research is described next.

Description

For a set of observations, this method answers the question:

> *If the i^{th} observation is removed from the sample, would the fit of the* **remaining** *observations around their mean be better than the fit of the* **previous** *set of observations (remaining + i^{th} observation) around their original mean?*

If the improvement in fit is large enough at a certain level of statistical significance, the removed observation(s) are labeled as a belonging to incident-related conditions, and the remaining observations as belonging to recurrent conditions. The process is applied sequentially until removing an observation does not lead to a statistically significant improvement in fit of the remaining sample.

Assumptions

1. The performance of the Underground during the time interval (TI) in question, e.g. 6:00-9:00am, is similar across days. The performance of each day-specific period is measured by the selected indicator of delays; in the case of this study the 95th percentile of the travel time distribution for that O-D-TI.

2. The performance of the Underground will be dissimilar to the other days only when a disruption, such as an incident, occurs. These disruptions will only increase travel times (i.e. delays); therefore, any deviations of the performance from typical conditions should only increase the indicator of delays.

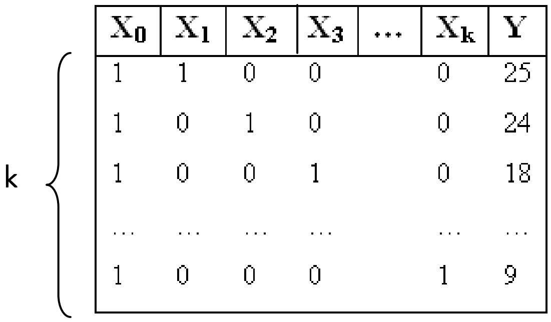
Estimation Procedure

1. Sort in decreasing order all the observations "Y" for that O-D-TI. In this case, Y = the 95th percentile travel time of each day, for an O-D-TI.

2. Estimate the following linear regression models, using the structure of example data shown in Figure A-1.

model 1: $Y = \beta_o$
model 2: $Y = \beta_o + \beta_1 X_1$
model 3: $Y = \beta_o + \beta_1 X_1 + \beta_2 X_2$
…
model m: $Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m$

Where Y = Indicator of Delay values (sorted in decreasing order),
X_i =*i*th dummy variable (taking the value 1 for the *i*th observation and 0 otherwise),
k = number of observations (day specific periods), and
m = number of models to be estimated = the number of dummy variables (m < k-1).



X_0	X_1	X_2	X_3	…	X_k	Y
1	1	0	0		0	25
1	0	1	0		0	24
1	0	0	1		0	18
…	…	…	…		…	…
1	0	0	0		1	9

Independent variable matrix

Figure A-1: Structure of stepwise regression model specifications

Note that because of the way the dummy variables, X, are defined (one for each period in the sample, i.e. one for each observation), the *i*th model estimates the mean of the (k-i+1) remaining observations. Therefore, the values for the estimated coefficients would be:

model 1: $\hat{\beta}_o = \overline{Y}_{1,k}$

model 2: $\hat{\beta}_o = \overline{Y}_{2,k}$, $\hat{\beta}_o + \hat{\beta}_1 = Y_1$

model 3: $\hat{\beta}_o = \overline{Y}_{3,k}$, $\hat{\beta}_o + \hat{\beta}_1 = Y_1$, $\hat{\beta}_o + \hat{\beta}_2 = Y_2$

…

model m: $\hat{\beta}_o = \overline{Y}_{m,k}$, $\hat{\beta}_o + \hat{\beta}_1 = Y_1$, … , $\hat{\beta}_o + \hat{\beta}_{m-1} = Y_{m-1}$

Where $\overline{Y}_{i,j}$ = average across observations *i* to *j*, and
Y_i = indicator of delay value for observation *i*.

It is important to be aware that the number of dummy variables that can be added must be less than the number of observations in the sample less one, in order for the least-squares method of estimating the coefficients in the regression to be valid (i.e. m < k-1).

3. The structure of the models estimated in the previous step has two characteristics that make it possible to identify the subset of the data that is significantly different from the majority (i.e. classified as incident-related performance):

(i) Models 2 through "m" are restricted forms of each other, as can be seen through the setting of (β_1, β_2, …, $\beta_m = 0$) to zeros. In this way, every model *j* can be restricted to become model *i*, where *i* < *j*.

For example:

- Unrestricted model (model 3): $Y = \beta_o + \beta_1X_1 + \beta_2X_2$
- restriction: (β_1=0, β_2=0)
- Restricted model (model 1): $Y = \beta_o$.

(ii) The contribution of each observation to the Error Sum of Squares (ESS) is zero if there is a dummy variable present in the model for each of those observations.

The implication of the first characteristic is that because every model *i* is a restricted version of model *j* (where *i* < *j*), they can be compared using an F-test to test the null hypothesis that the retricted model is true (i.e. that the regression on the restricted model explains the observed data better than the unrestricted model). The F-statistic is estimated as follows (Pindyck and Rubinfeld, 1998):

$$F\text{-}stat = \frac{(ESS_R - ESS_{UR})/q}{ESS_{UR}/(k-(m+1))} \qquad \text{[A-1]}$$

Where $ESS_R = \Sigma(Y_i - \hat{Y}_i)^2$ for each observation Y_i in the restricted model,
$ESS_{UR} = \Sigma(Y_i - \hat{Y}_i)^2$ for each observation Y_i in the unrestricted model,
q = number of restrictions = 1 (due to incremental comparisons),
k = number of observations, and
m = number of dummy variables.

From Equation A-1, one would expect that as the model is restricted, a decline in the fit would be observed with the Error Sum of Squares for the restricted model increasing ($ESS_{UR} < ESS_R$). This is similar to the notion that as variables are added to the model, the fit around the data can only improve (and is related to why an Adjusted R^2 is used to estimate goodness-of-fit when adding variables to a regression, as opposed to the R^2 directly). The F-test takes this into account by comparing the change in the ESS when restricting the model with the number of restrictions used, *q*. If the increase in ESS$_{R}$ is not large compared to the number of restrictions placed, then the null hypothesis (β_{j-1}=0) is not rejected. However, if the increase in the ESS$_{R}$ is large enough despite the number of restricted coefficients, then the null hypothesis can be rejected and one can conclude that the variables omitted through the restriction were significantly different as a group from the remaining observations.

The ability to compare the restricted to the unrestricted models in terms of detecting outliers, however, comes from the second characteristic. It states that because of the structure of the models, with a dummy variable added for each observation Y_i (in descending order), the ESS$_{UR}$ will only measure the fit of the remaining observations around their mean. For example, if model 2 is compared to its restricted version, model 1, the following estimates for the coefficients would be obtained:

model 1: $\hat{\beta}_o = \overline{Y}_{1,k}$

model 2: $\hat{\beta}_o = \overline{Y}_{2,k}$, $\hat{\beta}_o + \hat{\beta}_1 = Y_1$

In terms of the ESS$_{UR}$, the contribution of the first observation, Y_1, would be 0. This is seen by noting that: ESS$_{UR}$ = $\Sigma(Y_i - \hat{Y}_i)^2$ = $(Y_1 - \hat{Y}_1)^2$ + $\Sigma(Y_{2\text{-}k} - \hat{Y}_{2,m})^2$. And, since $\hat{Y}_1 = \overline{Y}_1 = Y_1$, the ESS$_{UR}$ = $\Sigma(Y_{2,k} - \hat{Y}_{2,m})^2$, and the "fit" of model 2 will depend only on observations 2 through k. What this implies in terms of detecting the subset of statistically different observations (i.e. incident-related performance), is that what the unrestricted models 2 through *m* are doing is effectively controlling for a set of observations (1 to m-1), and comparing the fit of the model of the remaining ones to the first model where all observations are taken into account. Similarly, it is possible to iterate through the sample of observations, where the contribution of each observation to changes in the fit of the model can be controlled for and used to compare between models.

The p-value for the F-test comparing the restricted and the unrestricted models is estimated and compared to a certain level of significance determined as the threshold with which a certain subset of observations is considered to be sufficiently different from the remaining majority to be classified as a separate group (i.e. incident-related performance). The p-value then refers to the following:

> *The* ***p-value*** *is the probability that the decrease in ESS resulting from unrestricting the model (while giving up degrees of freedom for the variables added) was due to normal variation in the data, and not a fundamental improvement in the model.*

In this case, a low p-value would indicate that it was unlikely that after unrestricting the model, its fit would improve due to normal variation in the data. Conversely, it would mean that if the p-value is less than a certain level of significance, the null hypothesis can be rejected and the improvement in the model fit can be attributed to an actual improvement in the model specification's ability to explain the observed data (i.e. the observations are better explained by their mean when certain observations are excluded by treating them as statistically different from the remaining observations and are therefore considered to be incident-related).

4. The procedure then continues to iterate by testing each time whether the next observation should belong within the subset identified as incident-related (i.e. statistically different) or whether separating the next observation does not meet the threshold of significance established previously.

Specifically, the model specification is updated to compare the fit of the model with the *i*th dummy variable and the (*i*+1)st dummy variable, capturing the likelihood that the classification of that *additional* observation into the separate subset of observations made the fit of the new specification sufficiently better, despite having lost a degree of freedom due to the additional dummy variable. The process iterates through the different specifications, starting with the comparison between model 1 vs. model 2, model 2 vs. model 3, and so on, until the p-value for that particular comparison is greater than the desired level of statistical significance. Once the p-value exceeds the significance threshold, the null hypothesis of the F-test cannot be rejected and the iteration stops by concluding that "the marginal benefit in terms of statistical fit of removing the (*i*+1)st observation did not outweigh its cost in terms of an additional variable added to the specification." As a result, all the values until the *i*th observation are classified under the incident-related category, and from the (*i*+1)st on, as belonging to recurrent conditions. Figure A-2 illustrates graphically this iterative procedure.

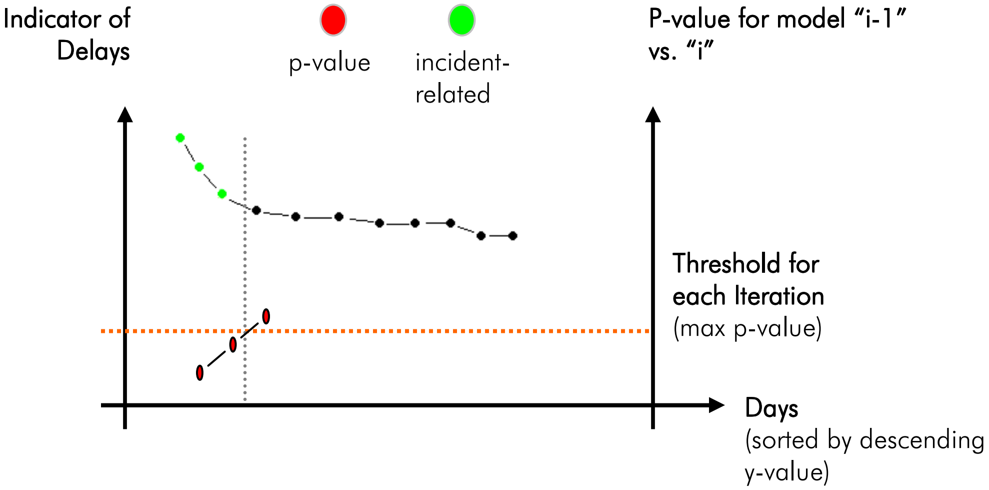

Indicator of Delays
🟥 p-value
🟩 incident-related
P-value for model "i-1" vs. "i"
Threshold for each Iteration (max p-value)
Days (sorted by descending y-value)

Figure A-2: Illustration of iteration in classification procedure

The vertical axis on the left of Figure A-2 represents the indicator of delays, in this case the 95th percentile of the travel time distribution of each day (represented by the horizontal axis). The corresponding values for 11 days are shown sorted in descending order of magnitude (i.e. not chronologically). The vertical axis on the right represents the p-value for each comparison between each model specification, with only the first three values shown in red. Finally, the horizontal red dashed line represents the threshold determined for accepting or rejecting classifying a particular observation as belonging to the incident-related subset of observations.

This simple illustration shows how the first three observations (in green) were classified as incident-related because during the first iteration of the procedure, the p-value was lower than the threshold for accepting the given classification. This iteration continues until the third comparison, which examined whether to include the fourth observation into the subset of incident-related observations. In this case, because the p-value for this comparison was higher than the determined threshold, the iteration stops and the most recent observation is classified as belonging to the recurrent performance for this journey.

Data Classification Software Code:
"R" Programming Language for Classification Procedure

The code for the classification procedure previously described was written in the "R" statistical programming language, both because of its versatility, as well as accessibility (i.e. free access). There are four basic steps using the "R" programming language that are required to apply the classification procedure here described. They are:

1. Import data into "R" server using the following command:

```
DATA1 <- read.table("input1.txt", header = TRUE)
```

This command simply reads the data stored in "input1.txt" and stores it into a data structure named "DATA1". The format of "input1.txt" should include the following fields in order, with headers included (can be saved from MS Excel as .txt directly): "Station Origin Code", "Station Destination Code", "Time Interval", "Day Code", and "95th percentile Travel Time". All fields must be integer values, with the first two representing Oyster station codes. The third field is available in case one wants to analyze various time intervals (e.g. 15-minute intervals within AM Peak) simultaneously. "Day Code" uses a numeric code to represent each individual day, where in this case the Underground's calendar system was used (e.g. February 5th, 2007 = 9897). Finally, the discontinuous (i.e. integer) 95th percentile journey time for that day's travel time distribution is entered.

2. Sort the information in DATA1 in terms of decreasing 95th percentiles, keeping the start and end locations as well as time interval (if used) in the correct order. That is, when running the classification approach on multiple O-D pairs simultaneously, it does only one O-D pair (and Time Interval) at a time. The following command is used for this:

```
DATA2 <- DATA1[order(DATA1$STARTLOC,DATA1$ENDLOC,DATA1$TIMEINT,-DATA1$P95),]
```

In this case DATA2 is the new data structure where the sorted table is stored, and STARTLOC and ENDLOC are simply the field names for the start and end stations, and TIMEINT represents the field containing the time interval. Notice the negative sign in front of the P95 field, indicating a decreasing sort.

3. Upload the program that contains the classification procedure code into "R". This can be done simply by copy-pasting the following code into the "R" command prompt (note: this step only needs to be done once for every session in "R"):

```
# CLASSIFICATION PROCEDURE output_function4.txt - by David L. Uniman - 2008
#------------------------------------------------------------------
#Description: Code of a function called "output4" which classifies data based on stepwise regression to identify
#whether the classification of one more observation (the next one) into a separate subset (i.e. incident-related
#performance) is statistically significant enough for three threshold levels: 10%, 5%, and 1%.

#Variable Description
#q - number of variables being restricted in each F-test comparison (set to 1 b/c stepwise)
#N - number of observations
#ESS - estimated sum of squared errors (yi-y_hat)^2
#Keepit_10% - equal to 0 if observation [j] had a p-value < 10%, or that the F-test
#(restriction) was significant, thus rejecting null hypothesis and claiming that
#the observation was classified into the subset and the unrestricted model is preferred
#(which accounts for this variable)
#daynum = number of days per period to process at once. Can modify when estimating
#cumulative baseline, where for example 40 days were processed simultaneously (set daynum = 40).
#data = data input for program to run. Has to have specified order of fields.


output4 <- function(daynum = 20, data = DATA2) {

        y <- data[,5]

        #loop to read each 20 days from larger trip table and define placeholder arrays     #used later
        #------------------------------------------------
        for(i in 1:(length(y)/daynum)) {
                ytemp <- y[(daynum*(i-1)+1):(daynum*i)]


                N <- length(ytemp)
                ESS_col <- array(0,N)
                pval_col <- array(0,N)
                F_stat_col <- array(0,N)
                keepit_10 <- array(0,N)
                keepit_05 <- array(0,N)
                keepit_01 <- array(0,N)

                #loop to estimate the ESS for each model specification
                #-----------------------------------------
                for(j in 1:(N-2))                          {
                q <- 1
                ESS_col[j] <- sum( (ytemp[j:N] - mean(ytemp[j:N]))^2)


                                                                    }

                ESS_col[N-1] <- 12345
                ESS_col[N] <- 12345

                #loop to estimate the F-stat and p-value for each model comparison
                #----------------------------------------
                for(h in 1:(N-3))                          {
                k <- (h+1)

                F_stat_col[h] <- ( (ESS_col[h] - ESS_col[h+1])/q )/(ESS_col[h+1]/(N-k))

                pval_col[h] <- 1-pf(F_stat_col[h],q,(N-k))
```

```
                                                    }

F_stat_col[N-2] <- 12345
F_stat_col[N-1] <- 12345
F_stat_col[N] <- 12345


pval_col[N-2] <- 12345
pval_col[N-1] <- 12345
pval_col[N] <- 12345



#loop to estimate the Keepit arrays at alpha = 10%
#----------------------------------------
k <- 1
while((k<N) & (pval_col[k] <= 0.1))
{
        k <- k+1
}
keepit_10[1:(k-1)] <- 0
keepit_10[k:N] <- 1




#for alpha = 5% significance
#----------------------------------------
k <- 1
while((k<N) & (pval_col[k] <= 0.05))
{
        k <- k+1
}
keepit_05[1:(k-1)] <- 0
keepit_05[k:N] <- 1




#for alpha = 1% significance
#----------------------------------------
k <- 1
while((k<N) & (pval_col[k] <= 0.01))
{
        k <- k+1
}
keepit_01[1:(k-1)] <- 0
keepit_01[k:N] <- 1




#assign pval_col[array] and keepit_a%[array]'s to final output columns
#------------------------------------------
if(i==1)
{
```

```
                        pfinal <- pval_col
                        kfinal_10 <- keepit_10
                        kfinal_05 <- keepit_05
                        kfinal_01 <- keepit_01
                }
                else
                {
                        pfinal <- c(pfinal,pval_col)
                        kfinal_10 <- c(kfinal_10,keepit_10)
                        kfinal_05 <- c(kfinal_05,keepit_05)
                        kfinal_01 <- c(kfinal_01,keepit_01)
                }


                                        }

                datafinal <-
        cbind(data,p_val=pfinal,keep_10=kfinal_10,keep_05=kfinal_05,keep_01=kfinal_01)

                                                                        }
```

4. Run the program that contains the classification procedure on the data stored in DATA2, and store the output (classified data) into DATA3. Do so by entering this command, noting that the name of the program is "output4":

```
DATA3 <- output4()
```

The form of the output is simply the original first five fields of DATA2 (which in turn is simply DATA1 sorted), plus four additional fields added on at the end. These fields are "p_val", which is the p-value for each iteration of the model (next to the value that was being tested as a part of the subset of incident-related performance), and "keep_10", "keep_05", and "keep_01", which are simply binary indicators with 1 meaning that the value should be retained into the larger group of observations, and 0 meaning that it should be classified as incident-related performance, for a threshold level of 10%, 5%, and 1% (with the latter being the most conservative – i.e. requires largest difference of values to classify observations into incident-related subset). To increase the sample period to 40 weekdays, the argument inside the function is set to "daynum = 40".

5. Finally, the classification results are exported so they can be read by other software and used for analysis. The following command is used to export the results into a .txt file once again, in this case named "outputname.txt":

```
write.table(x=DATA3, file="outputname.txt", quote=FALSE, row.names=FALSE,
col.names=FALSE, sep=",")
```

Note that the output specifications are such that the data can easily be imported into MS Excel for data analysis. In the output, the field names are not included.

Bibliography

Abkowitz, M., Slavin, H., Waksman, Englisher, L., and Wilson, N. (1978). *Transit Service* Reliability, Report UMTA-MA-06-0049-78-1. USDOT Transportation Systems Center. Cambridge, MA.

Abkowitz, M. (1983) *The Transit Service Reliability Problem and Potential Solutions. Proceedings of the August 1982 Transit Reliability Workshop*. USDOT Urban Mass Transportation Administration

Bates, J., Polak, J., Jones, P., Cook, A. (2001) *The Valuation of Reliability for Personal Travel*. Transportation Research (Part E), Vol. 37, No. 2, pp. 191-229.

Batley, R., (2007) *Marginal Valuations of Travel Time and Scheduling, and the Reliability Premium*. Transportation Research (Part E), Vol. 43, pp. 387-408.

Bertini, R., El-Geneidy, A. (2004) *Generating Transit Performance Measures with Archived Data*. Transportation Research Record 1841, pp. 109-119.

Bowman, L., Turnquist, M. (1981*) Service Frequency, Schedule Reliability and Passenger Wait Times at Transit Stops*. Transportation Research, Vol. 15A, No. 6, pp. 465-471.

Chan, J. (2007). *Rail OD Estimation and Journey Time Reliability Metrics Using Automated Fare Data*. Thesis, Master of Science in Transportation, MIT.
`

Cham, L. (2006) *Understanding Bus Service Reliability: A Practical Framework Using AVL/APC Data*. Thesis, Master of Science in Transportation, MIT.

De Jong, G., Kroes E., Plasmeijer, R., Sanders, P., Warffemius, P. (2004) *The value of reliability*. Proceedings of the European Transport Conference – Strasbourg, 4-6 October, 2004.

Draper, N., Smith, H. (1998) *Applied Regression Analysis*. 3rd Edition, John Wiley & Sons.

Ettema, D., Timmermans, H. (2006) *Costs of Travel Time Uncertainty and Benefits of Travel Time Information: Conceptual Model and Numerical Examples*. Transportation Research (Part C) Vol. 14, pp. 335-350.

Evans, J., et al., (2004) *Traveler Response to Transportation System Changes*. TCRP Report 95-c. Transportation Research Board, Washington D.C.

Fattouche, G. (2007) *Improving high-frequency bus service reliability through better scheduling*. Thesis, Master of Science in Transportation, MIT.

Furth, P., Muller, T. (2006a) *Service Reliability and Hidden Waiting Time*. Transportation Research Record 1955, pp. 79-87.

Furth, P., Hemilly, B., Muller, T., Strathman, J. (2006b) *Using Archived AVL-APC Data to Improve Transit Performance and Management*. TCRP Report 113. Transportation Research Board. Washington D.C.

Greater London Authority Transport Home Page (1999) Mayor of London, London Assembly, and Greater London Authority, accessed Jan. 2009 <http://www.london.gov.uk/gla/transport.jsp>

Henderson, G., Adkins, H., Kwong, P. (1990) *Toward a Passenger-Oriented Model of Subway Performance*. Transportation Research Record 1266, pp. 221-228.

Henderson, G., Adkins, H., Kwong, P. (1991a) *Regularity Indices for Evaluating Transit Performance*. Transportation Research Record 1297, pp. 3-9.

Henderson, G., Adkins, H., Kwong, P. (1991b) *Subway Reliability and the Odds of Getting There on Time*. Transportation Research Record 1297, pp. 10-13.

Hollander, Y. (2006) *Direct Versus Indirect Models for the Effects of Unreliability*. Transportation Research (Part A), Vol. 40, pp. 699-711.

Kittelson & Associates, et al. (2003a) *A Guidebook for Developing a Transit Performance-Measurement System*. TCRP Report 88. Transportation Research Board. Washington D.C.

Kittelson & Associates, et al. (2003b) *Transit Capacity and Quality of Service Manual*. TCRP Report 100. Transportation Research Board. Washington D.C.

Lam, T., Small, K. (2001) *The Value of Time and Reliability: Measurement from a Value Pricing Experiment*. Transportation Research (Part E) Vol. 37, pp. 231-251.

Lomax, T., Schrank, D., Turner, S., Margiotta, R. (2003) *Selecting travel time reliability measures*. Texas Transportation Institute and Cambridge Systematics, Inc.

Love, A., Jackson, P. (2000) *The Meaning of Reliability*. Qualitative Research Commentary Report S.00756, London Buses, UK.

Martland, C.D. (1972) *Rail Trip Time Reliability: Evaluation of Performance Measures and Analysis of Trip Time Data*. Studies in Railroad Operations and Economics, Vol. 2, MIT Report No. 72-37.

Multisystems, Inc. (2003) *Fare Policies, Structures and Technologies: Update*. TCRP Report 94. Transportation Research Board. Washington D.C.

Noland, R.B. (1997) *Commuter Responses to Travel Time Uncertainty Under Congested Conditions: Expected Costs and the Provision of Information*. Journal of Urban Economics, Vol. 41, pp. 377-406.

Noland, R.B., Small, K. (1995) *Travel Time Uncertainty, Departure Time Choice, and the Cost of Morning Commutes*. Transportation Research Record 1493, pp. 150-158.

Noland, R.B., Polak, J.W. (2002) *Travel time variability: a review of theoretical and empirical issues*. Transport Reviews, Vol. 22 (1), pp. 39-54.

Osuna, E. E., Newell, G. F. (1972) *Control Strategies for an Idealized Public Transportation System*. Transportation Science, Vol. 6, pp. 52-57

Paine, F. T., Nash A. N., Hille S. T., Brunner, G. A. (1976) *Consumer Attitudes Toward Auto vs. Public Transport Alternatives*. Journal of Applied Psychology, Vol. 53 (6), pp. 472-480

Pindyck, R.S., Rubinfeld, D.L. (1998) *Econometric Models and Economic Forecasts*. 4th edition, Irwin McGraw-Hill.

Polak, J.W., Oladeinde, F. (2000) *An empirical model of travellers' day-to-day learning in the presence of uncertain travel times*. In M.G.H. Bell and C. Cassir (eds.), Reliability in Transport Networks. Hertfordshire: Research Studies Press.

Price, E. (2005) *Customer Insight Report: Train Service*. London Underground, UK.

Recker, W., et al. (2005) *Considering Risk-Taking Behavior in Travel Time Reliability*. Research Report UCB-ITS-PRR-2005-3, Institute of Transportation Studies, University of California, Berkeley.

Small, K. (1982) *The scheduling of consumer activities: work trips*. American Economic Review, Vol. 72 (3), pp. 467-479.

Strathman, J., Kimpel, T., Dueker, K. (2000) *Time Point-Level Analysis of Passenger Demand and Transit Service Reliability*. Final Technical Report TNW2000-03, Portland State University.

TfL – Group Transport Planning and Policy (2006) *Transport 2025: Transport vision for a growing world city*. Transport for London, U.K.

TfL – Group Business Planning and Performance (2007a) *TfL Business Plan 2005/6 to 2009/10*. Transport for London, UK.

TfL – Transport Planning Business Operations (2007b) *London Travel Report 2007*. Transport for London, UK.

TfL (2008), *TfL Guide to Fares and Tickets*, on http://tfl.gov.uk, Transport for London, UK.

TfL (2009), *TfL Journey Planner Homepage*, on http://http://journeyplanner.tfl.gov.uk, Transport for London, UK

Wainberg, S., Powell, G. (2008) *Performance Indicator Record*. Personal Correspondence – 20 March 2008. Unpublished Documents, Transport for London, U.K.

Wilson, N., Nelson, D., Palmere, A., Grayson, T., Cederquist, C. (1992) *Service-Quality Monitoring for High-Frequency Transit Lines*. Transportation Research Record 1349, pp. 3-11.

Wolberg, J. (2006) *Data Analysis Using the Method of Least Squares: Extracting the Most Information from Experiments*. New York, Springer.

Wood, P. (2008). *LU Station Ticket Sales Trends*. Personal Correspondence – 21 October 2008. Unpublished Documents, London Underground, U.K.