

Developing a national measure for predictable public transport: bus, rail and ferry

April 2018

S Rashidi, L Schmitt, P Ranjitkar, T Rabel, S Sood, L Baker, V Ivory and H Rezaie
Opus International Consultants, Wellington and Auckland

NZ Transport Agency research report 641

Contracted research organisation – Opus International Consultants Ltd

ISBN 978-1-98-851298-3 (electronic)
ISSN 1173-3764 (electronic)

NZ Transport Agency
Private Bag 6995, Wellington 6141, New Zealand
Telephone 64 4 894 5400; facsimile 64 4 894 6100
research@nzta.govt.nz
www.nzta.govt.nz

Rashidi, S, L Schmitt, P Ranjitkar, T Rabel, S Sood, L Baker, V Ivory and H Rezaie (2018) Developing a national measure for predictable public transport: bus, rail and ferry. *NZ Transport Agency research report 641*. 81pp.

Opus International Consultants was contracted by the NZ Transport Agency in 2016 to carry out this research.



This publication is copyright © NZ Transport Agency. This copyright work is licensed under the Creative Commons Attribution 4.0 International licence. You are free to copy, distribute and adapt this work, as long as you attribute the work to the NZ Transport Agency and abide by the other licence terms. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. While you are free to copy, distribute and adapt this work, we would appreciate you notifying us that you have done so. Notifications and enquiries about this work should be made to the Manager Research and Evaluation Programme Team, Research and Analytics Unit, NZ Transport Agency, at NZTAresearch@nzta.govt.nz.

Keywords: measures, performance, predictability, public transport, reliability

An important note for the reader

The NZ Transport Agency is a Crown entity established under the Land Transport Management Act 2003. The objective of the Agency is to undertake its functions in a way that contributes to an efficient, effective and safe land transport system in the public interest. Each year, the NZ Transport Agency funds innovative and relevant research that contributes to this objective.

The views expressed in research reports are the outcomes of the independent research, and should not be regarded as being the opinion or responsibility of the NZ Transport Agency. The material contained in the reports should not be construed in any way as policy adopted by the NZ Transport Agency or indeed any agency of the NZ Government. The reports may, however, be used by NZ Government agencies as a reference in the development of policy.

While research reports are believed to be correct at the time of their preparation, the NZ Transport Agency and agents involved in their preparation and publication do not accept any liability for use of the research. People using the research, whether directly or indirectly, should apply and rely on their own skill and judgement. They should not rely on the contents of the research reports in isolation from other sources of advice and information. If necessary, they should seek appropriate legal or other expert advice.

Acknowledgements

The project team would like to acknowledge the input of the following people for their valuable contribution to this work:

- Steering Group Chairman Paul Clark, then Ernest Albuquerque
- Steering Group members Ellen Cox (NZTA) Wayne Hastie (GWRC) Shannon Boorer (ECan) then Edward Wright (ECan)
- External peer reviewers: Professor Graham Currie (Monash University); Professor Hesham A. Rakha (Virginia Polytechnic & State University)

We would also like to thank all of the people who gave their time to attend the industry workshops in Auckland, Wellington and Christchurch

Abbreviations and acronyms

AT	Auckland Transport
AVL	automatic vehicle location
AWT	actual wait time
CTOC	Christchurch Transport Operations Centre
DOT	Department of Transportation, US
ECan	Environment Canterbury
FWHA	Federal Highway Administration (US)
GIS	geographic information system
GPS	global positioning system
GWRC	Greater Wellington Regional Council
KPI	key performance indicator
LCH	lost customer hours
MBTA	Massachusetts Bay Transportation Authority
NCHRP	National Cooperative Highway Research Program
PT	public transport
PTOM	Public Transport Operating Model
PTV	Public Transport Victoria
RTI	real-time information
Transport Agency	NZ Transport Agency

Contents

- Executive summary** 7
- Abstract** 10
- 1 Introduction** 11
 - 1.1 Key research questions/project objectives 11
 - 1.2 Focus on ‘in-vehicle travel time’ 11
- 2 Methodology** 12
 - 2.1 Key project stages 12
 - 2.2 Literature review 12
 - 2.3 Assess modification potential using New Zealand PT data 13
 - 2.4 Validation workshops 14
 - 2.5 Conclusions and recommendations 15
- 3 Literature review** 16
 - 3.1 Definition of ‘predictability’ and related measures 16
 - 3.1.1 NZ Transport Agency road traffic measures 16
 - 3.2 New Zealand public transport measures 17
 - 3.2.1 Auckland Transport 17
 - 3.2.2 Greater Wellington Regional Council 20
 - 3.2.3 Environment Canterbury/Canterbury Transport Operation Centre 21
 - 3.3 International measures (industry practice) 24
 - 3.3.1 Public Transport Victoria, Australia 24
 - 3.3.2 Massachusetts Bay Transportation Authority, US 25
 - 3.3.3 Transport for London, UK 27
 - 3.4 International measures (academic) 29
 - 3.4.1 Travel time reliability measures 29
 - 3.4.2 Reliability measures from operators’ point of view 30
 - 3.4.3 Reliability measures from the customer’s point of view 31
 - 3.4.4 Travel time reliability indices 31
 - 3.5 Review of measures 39
 - 3.5.1 Aggregation and ‘fitting’ to the Transport Agency road index 42
 - 3.5.2 Shortlisted measures 43
- 4 Assess modification potential of measures using New Zealand PT data** 45
 - 4.1 Preparing data: analytics, error detection and cleansing 45
 - 4.1.1 Missing data treatment 46
 - 4.1.2 Outliers and extreme values identification: 47
 - 4.2 Applying the measures 47
 - 4.2.1 Shortlisted predictability measures 47
 - 4.2.2 Aggregation technique 48
 - 4.2.3 Aggregated results - bus network 49
 - 4.2.4 Aggregated results - train network 49
 - 4.2.5 Observations 50
 - 4.3 Threshold testing 50
 - 4.4 Overall lessons from data testing 51

4.4.1	Data issues – obtaining data	51
4.4.2	Preparing data	51
4.4.3	Conclusions	52
5	Validation testing – workshops	54
5.1	Participants	54
5.2	Structure of the workshops	55
5.3	Outcomes from the workshops	55
5.3.1	How PT predictability is relevant to stakeholders.....	55
5.3.2	Data sources for measuring predictability	59
5.3.3	Overarching objectives of the research	60
5.3.4	Assessing the measures	62
5.3.5	Exploring thresholds	66
5.4	Overall conclusions from the workshops	67
6	Conclusions and recommendations	70
6.1	Summary of outcomes	70
6.2	Evaluation of measures	70
6.3	Implications.....	72
6.3.1	Data preparation	72
6.3.2	Service frequency and predictability	73
6.3.3	Advanced technology and predictability	73
6.4	Recommendations and next steps.....	73
7	References	75
	Appendix A: Multimodal dashboards from Auckland Transport.....	79

Executive summary

The ability to reliably predict public transport (PT) journey times is critical for network operators and transport authorities to measure, monitor and target improvements to the PT network, with flow-on effects for customers. In fact, customers have identified PT reliability as a key issue; however, it is not measured particularly well in New Zealand. Currently, a range of predictability-related measures for PT is used in different parts of New Zealand, without one uniform measure deployed. Likewise, the NZ Transport Agency (the Transport Agency) has a measure of travel-time predictability for road but does not have a comparable measure for PT. This road-based measure may not be easily transferable to PT. The present research sought to address these issues by developing a measure for predictability performance that can be used across different transport modes and in different parts of the country. This report documents the research related to investigating and fulfilling this aim.

Primary aims of the research were:

- Identify and develop the best measure for PT travel time predictability based on a review of literature, a practice review, ascertaining the benefits and limitations of measures, and comparison with the current Transport Agency road-based measure and validation testing.
- Develop a nationally aggregated set of travel time predictability information across regions, travel modes and different times of day to 'trial' shortlisted measures.

It should be noted that the scope of finding an 'optimal' measure was limited to a measure that could be used to compare in-vehicle travel time rather than one that encompassed various legs of journeys.

Method

In order to achieve these aims, the following research was conducted in New Zealand between August 2016 and August 2017:

- A review of literature involved undertaking a local and international review of predictability and reliability measures used for PT or private vehicle travel, and also included evaluation of measures. Measures were evaluated in terms of their ability to meet the overarching aims of the research and consideration was also given to the fact that, ideally, effective performance measures should be: applicable to users, easy to calculate, accurate and able to be clearly and consistently interpreted. From this review, and with consideration of the potential for inter-modal and inter-regional aggregation, a shortlist of three preferred measures was developed including the buffer index, modified buffer index and planning index.
- Assessing modification potential of measures using New Zealand PT data: The shortlisted measures were applied to a nationally aggregated set of PT travel data from across regions and PT modes. This research stage involved obtaining New Zealand PT datasets, preparing data, applying shortlisted measures of data alongside the Transport Agency's road index, undertaking threshold sensitivity testing and interpreting the outcomes.
- Validation testing – workshops: Research was then presented to key stakeholders (PT management authorities and operators) through interactive workshops in Auckland, Wellington and Christchurch. Attendees were asked to provide feedback on their understanding of shortlisted measures, whether they preferred any of the measures or thought they were useful, and what they thought the implications of adopting the measures might be.

Results

The measures found in industry practice and literature fell into four main categories: 1) schedule adherence measures, 2) statistical ranges, 3) buffer time and 4) tardy trip indicators. Schedule adherence measures are the most common type of predictability measure used by PT authorities in New Zealand and abroad. These measures are particularly useful for assessing operators as key performance indicators. Unfortunately, this type of reliability measure is unsuitable for comparison with existing road-based reliability because the road-based measures do not have a prescriptive schedule. The buffer time measures, in general, are popular in the United States among network planners, in operations, and are customer focused metrics that utilise the worst-case percentiles travel time to highlight the expected delay on top of 'normal' travel conditions. Three buffer-time measures were shortlisted and applied to aggregated PT data alongside the Transport Agency's road index to evaluate the 'closeness of fit'. The literature review did not reveal any reliability or predictability measures that were being applied to both PT services and private cars by one agency. However, subsequent stages of the project revealed that Auckland Transport (AT) has begun using the modified buffer index and a variant of the planning index for analysis across modes.

In order to assess the modification potential of measures, PT management agencies in Wellington, Auckland and Christchurch were contacted to obtain data. It took time to receive data and the data received required significant 'cleansing' work to prepare it for aggregation and analysis. Once the data was ready, it was aggregated and the three shortlisted measures were then applied to the AM peak period (7am to 9am) March 2016 aggregate datasets alongside the Transport Agency's road index. The outcomes of the data testing indicated the shortlisted measures were comparable to the Transport Agency's road index but were sensitive to the thresholds adopted. The road-based thresholds initially applied did not provide much differentiation in performance outcomes. The data testing revealed that the shortlisted measures are all linearly related – that is they are closely related and the results of the performance analysis produced comparable results across different measures. This suggested there was not a compelling case for one particular measure to be used, although mean-based measures were more sensitive to extreme observations in the data than median-based measures.

When presented with the results of the research in the validation workshops, key stakeholders had a range of reactions and feedback. Some workshop attendees were more confident in understanding and interpreting the measures than others. Generally, there was a sense the measures might be appropriate for network planning but would require some translation for the outcomes to be meaningful for customers and other parties. Some attendees expressed concern about trying to 'fit' a predictability measure to the Transport Agency's road index measure as many found the existing measure to be complicated and non-intuitive. The workshops further revealed that stakeholders felt selection of any shortlisted measures depended on what aspect of reliability one wanted to examine and that care needed to be taken in comparing modes and developing thresholds. There was not an overwhelmingly strong preference for any of the shortlisted measures but when pressed, some respondents from the Wellington workshop seemed to prefer an absolute travel time measure (with a range to show variability). Everyone seemed to understand and like punctuality measures, with which they were familiar, but given the non-applicability to the objective of comparability with private vehicle measures, their next preference of the shortlisted measures presented seemed to be the planning index and the modified buffer index. The Wellington contingent thought the planning index seemed flexible in its applicability to convert into useful measures for both network planners and customers. In the Auckland workshop, we also learned that since the literature review for this project was conducted, AT had begun using the modified buffer index, referring to this as 'reliability', and a measure similar to the planning index which AT refers to as 'delay'.

Conclusions and recommendations

Overall, the research project was largely able to achieve the overarching aims. Any of the three shortlisted measures would be appropriate to use for aggregate comparisons of reliability across modes. However, it is recommended that different predictability measures be used depending on which aspect of PT reliability needs to be examined. Different measures can show variability from slightly different perspectives:

- With their reliance on average rather than median, the Transport Agency road index and the buffer index are more sensitive to extreme observations than median-based measures like MBI.
- The buffer index can be useful for looking at how much fluctuation occurs on average along a route.
- The planning index can be relatively easily converted to total journey times and offers an absolute minimum and maximum (considered most understandable for customers).
- Punctuality is a common and easily understood measure used to evaluate reliability for PT but does not meet the objective of being comparable to private vehicle travel.

The fact that AT is already using two of the shortlisted measures (or variants of) across multiple modes suggests that both the modified buffer index and the planning index are appropriate for comparing reliability aspects across modes (and regions). However, it may be useful to present performance outcomes under more easily understood terms and using colour-coding. AT's dashboard example of performance reporting provides a useful example. The appropriate thresholds to use for different measures require more consideration and may vary by region or mode.

Other lessons were also learned. One, aggregating data across regions and modes is time-intensive and complex. Another issue is that while the focus of the research was on in-vehicle travel time, in the context of 'real-world' perceptions of PT reliability, waiting time is very important to customers and their perceptions of PT reliability. Some reliability measures may therefore be less relevant to customers when there is high service frequency. Thirdly, advanced technology may be changing the opportunities for PT predictability analysis and also customer expectations.

The report concludes with some recommendations and suggestions for the direction of future research:

- For network planning, use either (or both) the modified buffer index and planning index for analysis, as they are both statistically buffered from data outliers.
- Undertake further research applying each of the shortlisted measures to private vehicle travel; ideally this should be done for the same reporting period as for PT data.
- Undertake further research to determine ideal thresholds for the shortlisted measures.
- Provide customers with simple travel times or a range of absolute travel times. Alternatively, the modified buffer index and planning index outcomes can be presented using more easily understandable names and outputs, modelled on what AT currently produces.

Finally, in considering the future of aggregating to compare modes, there is potential to weight modes based on the number of people being moved which may be useful to consider. However, this matter requires further consideration and research.

Abstract

The ability to reliably predict PT journey times is critical for network operators and transport authorities to measure, monitor and target improvements to the PT network, with flow-on effects for customers. Research conducted in New Zealand between August 2016 and August 2017 aimed to identify and develop an optimal measure for PT predictability. This involved undertaking a local and international review of predictability/reliability measures used for PT or private vehicle travel, and included evaluation of measures. From this review, and consideration of the potential for inter-modal and inter-regional aggregation, a shortlist of three preferred measures was developed including: buffer index, modified buffer index and planning index. Shortlisted measures were applied to a nationally aggregated set of PT travel data from across regions and PT modes. This data testing helped assess 'fit' to the NZ Transport Agency's road index, modification potential, and revealed that the shortlisted measures are all linearly related, with comparable results across different measures. This suggested there was not a compelling case for one particular measure to be used. Validation workshops further revealed that stakeholders felt selection of any shortlisted measures depended on what aspect of reliability one wanted to examine and that care needed to be taken in comparing modes and developing thresholds.

1 Introduction

In 2016 Opus was commissioned to undertake research to investigate an appropriate performance measure of 'predictability' for public transport (PT) in New Zealand. The ability to reliably predict PT journey times is critical for network operators and transport authorities to measure, monitor and target improvements to the PT network. Similarly, predictable PT is important for existing and potential PT customers. For example, a person working a fixed shift with a mandatory start time *depends* on their bus getting in at the scheduled time. Unpredictable journey times can mean schedules are artificially lengthened to slow down services to better meet timetables. This can result in passengers having frustrating waits at timing points. While customers have identified PT reliability as a key issue, it is not measured consistently or particularly robustly in New Zealand.

Currently a range of predictability-related measures for PT are used in different parts of New Zealand, without one uniform measure deployed. The NZ Transport Agency (the Transport Agency) has a measure of travel time predictability for road but does not have a comparable measure for PT. This road-based measure assesses deviations from mean travel times for 15-minute travel timeslots within a pre-defined 5% time buffer. This structure may not be easily transferable to PT. The present research sought to develop a measure for predictability performance that can be used across different transport modes and in different parts of the country. This report documents the research related to investigating and fulfilling this aim.

1.1 Key research questions/project objectives

The primary aims of the research were:

- Identify and develop the best measure for public transport travel time predictability based on a review of literature, a practice review, ascertaining the benefits and limitations of measures, comparison with the current Transport Agency road-based measure, and validation testing.
- Develop a nationally aggregated set of travel time predictability information across regions, travel modes and different times of day to 'trial' shortlisted measures.

1.2 Focus on 'in-vehicle travel time'

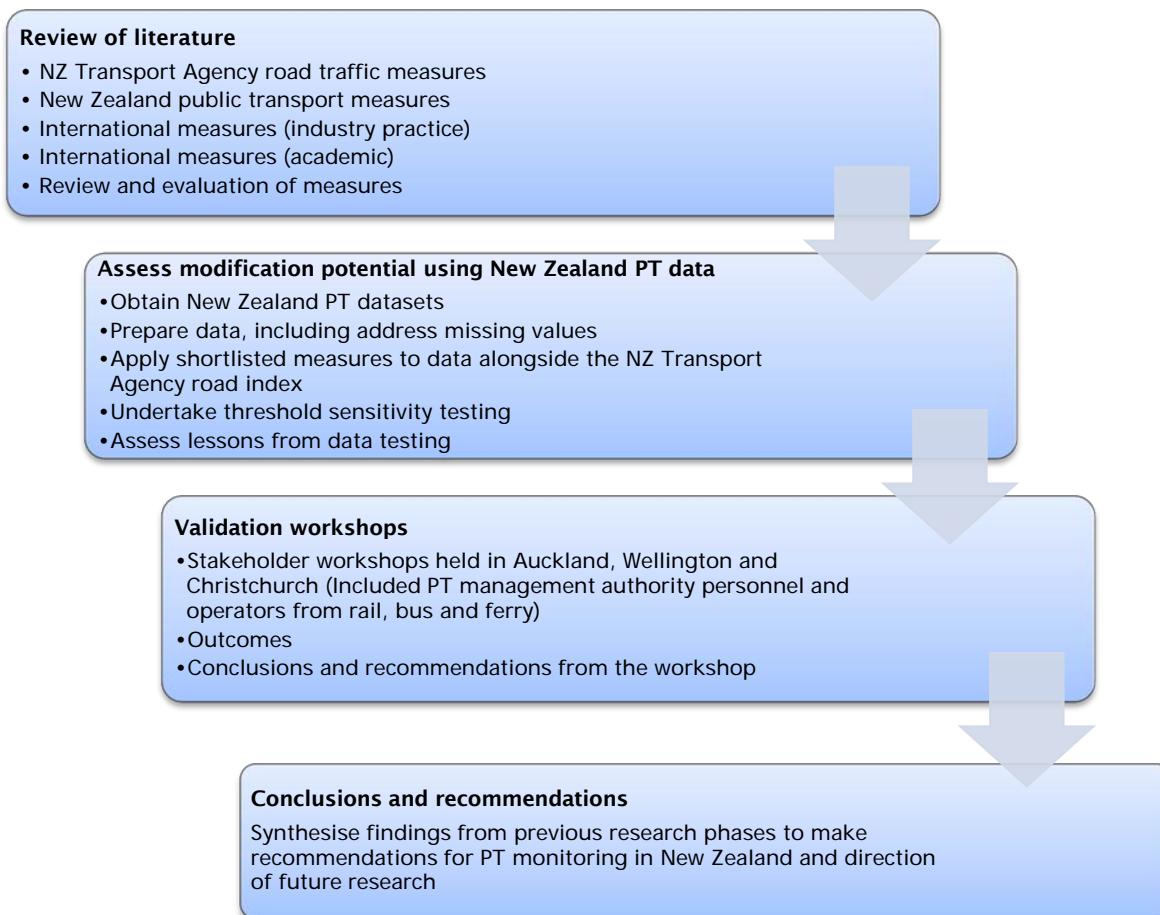
It is important to note the aim of this research was limited to a focus on 'in-vehicle' travel time. It is acknowledged that public transport journeys include multiple stages of travel including: 'first mile' travel to a stop or station, sometimes purchasing of fares, wait times for services, sometimes transfers (and additional wait times), in-vehicle travel, and finally 'last mile' travel to one's final destination. The emphasis to research solely 'in-vehicle' travel times is because of the aim to develop a measure that can be used to compare travel across modes, including private vehicle travel.

2 Methodology

2.1 Key project stages

The research was undertaken between August 2016 and August 2017. An overview of the key research stages is provided in figure 2.1 and shows that the research methods included a review of literature, data testing and validation workshops, which led to the considered development of conclusions and recommendations.

Figure 2.1 Summary of research methodology



The research methods were agreed between the researchers and steering group using an iterative approach – that is, findings from earlier stages of the research informed the direction of the later stages of research. A more detailed description of each of the key research methods follows.

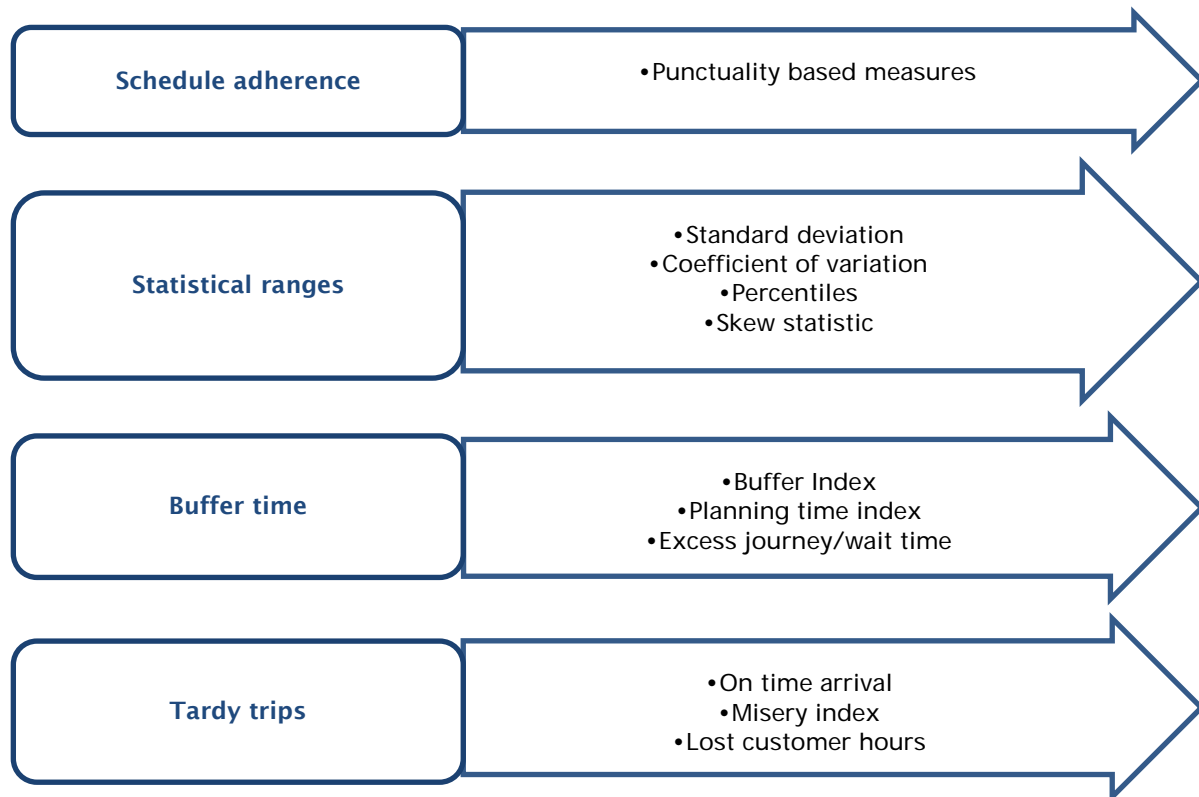
2.2 Literature review

The research began with a literature review which included a review of both academic papers and a practice review to determine what measures were being used in both New Zealand and abroad. As an integral component of the research method, the key approach to this review was to ‘cast the net wide’ to provide a great understanding of the potential measures available and their respective advantages and

disadvantages in the New Zealand PT context. This assisted in developing a more confident **shortlist** of preferred measures to be 'tested' using actual PT data.

Four main categories of measures were examined in the literature and are outlined in figure 2.2. These are discussed in detail in the literature review (chapter 3).

Figure 2.2 Overview of measures examined in the literature review



Once the literature review was complete, a presentation of its key findings was given to the steering group to obtain feedback and confirm the direction of the remainder of the research project. In addition, the literature review report was peer-reviewed by a public transport expert based in Australia who provided helpful comments and identified some further resources to include. Some informed adjustments were then made to the literature review.

2.3 Assess modification potential using New Zealand PT data

In the next stage of research, PT datasets were obtained from the three biggest cities in New Zealand: Auckland, Wellington and Christchurch, and the shortlisted predictability performance measures were applied to the data.

Data testing involved four main stages:

- 1 Obtaining New Zealand public transport datasets
- 2 Preparation of data, including addressing data abnormalities
- 3 Applying shortlisted measures to the data
- 4 Examination of results and undertaking threshold sensitivity testing.

A brief overview of each of these research stages follows below. More detail on the research method is provided in chapter 4 of this report.

Obtaining access to PT datasets required careful negotiation with key PT authorities in New Zealand (AT, Greater Wellington Regional Council (GWRC) and Environment Canterbury). The authorities requested a detailed scope of the nature of the data required, including specific routes and time periods. Four key datasets were obtained: bus data for routes in Auckland, Wellington and Christchurch; and rail information for Wellington. Unfortunately, for having full inter-modal PT coverage, no ferry data was obtained.

As is typically the case with data, the data was not in a format that was completely ready for analysis so some work was undertaken to prepare the data. This involved data analytics, error detection and cleansing. Data preparation and dealing with abnormality in a dataset are important issues that must be considered appropriately. Previous research by Rashidi (2014) revealed that some research conducted for modelling and estimation of bus travel time had not treated and reported the presence of missing values in their analysis. This undermines the robustness of the analytics and the efficacy of the outcomes. Some of the causes of missing data issues with automatic vehicle location (AVL) and automatic passenger counting data included:

- missing or zero recordings of arrival and departure time
- duplicate recordings of bus identification (ID)
- unreasonable arrival or departure records (eg departure times are less than arrival times)
- wrong stop ID recorded
- outliers (unreasonably small or large records).

Once the data was 'cleaned', the shortlisted measures selected following the literature review were applied to the data. The results were then critically examined and it became apparent that directly applying road based measures to the PT data worked, but was not overly informative at explaining and highlighting variations in PT predictability. This was deemed to be, in large part, because the shortlisted measures that were in use for PT had largely originated from private car-based measures. As such, and to better understand this issue, some sensitivity analysis, or exploration of adjusting thresholds, was then undertaken as part of this research stage. This is explained in more detail in section 4.2.

2.4 Validation workshops

The third major research stage was to conduct three validation workshops where the research findings from the literature review, data testing and application of the shortlisted measures were presented to a number of relevant practitioners to obtain their feedback. Industry contracts from PT authorities and PT operators (across all PT modes) were invited. Workshops were held in Auckland, Wellington and Christchurch. The workshops were interactive and involved learning how PT predictability was relevant to the attendees and then presented the research including a synopsis of the literature review and the data testing. The workshops sought to address the following key questions with stakeholders:

- whether stakeholders understood the measure(s) being proposed
- if stakeholders thought the measure(s) would be useful
- what they thought the implications of adopting the measure(s) would be.

2.5 Conclusions and recommendations

Finally, a considered set of conclusions and recommendations was developed. First the outcomes of the three main research phases were synthesised to provide a number of key conclusions emerging from the research. A discussion was held about the implications of these conclusions and in the wider context of transport, including with the rapid 'disruptive' changes happening in transport associated with advanced digital technology. Finally, a number of recommendations are provided for PT performance monitoring into the future and for the future direction of research.

3 Literature review

The literature review sought to identify lessons and insights to develop a PT (bus, rail and ferry) travel time predictability measure that can be compared to the Transport Agency predictability measure currently applied to road traffic. To achieve this, an extensive international literature review was undertaken, drawing from industry practice in New Zealand and overseas, and from academic literature. Existing predictability (and reliability) measures for both PT and private vehicles were examined. The latter was necessary as many of the existing PT reliability measures used are 'schedule adherence' measures, which are, by nature, incompatible with 'fitting' to private vehicle travel. This issue is explained more throughout this section. The literature review not only identified different measures available but also documented the advantages and disadvantages of various approaches as part of the evaluation of available measures. From this evaluation, three shortlisted measures were then identified for further evaluation alongside the Transport Agency index in the next stage of the research: 'assessing the modification potential using New Zealand data'. The literature review also included a short review to guide that stage of research with a review of closeness of fit and aggregation techniques and the care required in comparing PT with private vehicles.

3.1 Definition of 'predictability' and related measures

3.1.1 NZ Transport Agency road traffic measures

Since there are many concepts and terminologies used to define travel time performance of road traffic and, to a certain extent, PT, it is useful to allocate definitions to each of these concepts. Of particular significance is the term 'predictability', which the Transport Agency has adopted as a key measure of customer expectation (CTOC 2015). In a memo directed to the National Transport Operation Centre, the Transport Agency defines 'predictability' as a:

'threshold' and measures whether a time period exceeds this threshold or is under this threshold. The threshold is referred to as the "buffer", and this approach may be called the 'buffer measure'. Currently, the threshold is defined as the rolling 12-month average (average over last 12 months from current month) plus 5%. (CTOC 2015)

To calculate predictability, each peak hour per day during a monthly period (in 15-minute intervals) is assessed against this buffer and then allocated a value of zero if the observed travel time exceeds this buffer or 1 (one) if it is below. Predictability is then expressed as a percentage and is equal to the average of the sum of the ones and zeros in the peak hour period of that month. The percentage is directly proportional to predictability, ie a lower percentage indicates a lower predictability. The measure is then compared against the previous months for performance reporting purposes. For the remainder of this report, this Transport Agency predictability measure is referred to as the 'Transport Agency road index'.

The CTOC memo (2015) refers to a 'customer focus', which suggests that the measure is meaningful and useful from a customer point of view. While there may be potential to develop a measure, or insights from it that are meaningful to the customer, the measure in its present use is solely intended for performance reporting from an operator perspective.

In addition to the 'predictability' measure, the Transport Agency also measures travel time performance in several ways. These have been summarised in table 3.1.

Table 3.1 Summary of common travel time performance measures (NZ Transport Agency)

Term	Definition
Predictability (CTOC 2015)	Each 15-minute time interval is compared against a 12 month rolling average (+5% buffer). The interval is recorded as a pass or fail (1 or 0). The predictability for a time period, ie am peak (730–830) is the average (1 and 0's) across each 15-minute interval in that time period. This is the key measure of customer expectation. The measure sets an expectation threshold and determines which days/periods exceed this threshold.
Average travel time (CTOC 2015)	The mean (average) daily travel time from the start of the route to the end of the route expressed in minutes and calculated at 15-minute intervals.
Reliability (CTOC 2015)	Standard deviation of travel time divided by average minutes travel time (as per Austroads) – coefficient of variation. This provides a measure of change on a route. It is used more extensively by technical staff to flag issues and then drill-down and identify reasons.
Delay (CTOC 2015)	Delay is defined as the difference between two travel times. Two delay measures are calculated, the first based on recurrent (or typical) delay and the second based on delay due to unusual conditions (events, incidents, or atypical days). The calculation for typical/recurrent congestion involves calculating the mean travel time for the month as the first step (measure no. 1), estimating the free-flow travel time, and calculating the difference between these two travel times through the day. Used to describe the magnitude of opportunity that exists on each route and across each region – the overall total system delay.
Punctuality (NZ Transport Agency 2016)	Percentage of scheduled service trips leaving between 59 seconds before and 4 minutes 59 seconds after the scheduled departure time of selected points.
Congestion (CTOC 2015)	Calculated by dividing the average peak period speed by the route speed limit to yield a percentage, eg if the am peak average speed was 40 km/h, the result for the example presented above would be $40/50 = 80\%$. A lower percentage indicates more congestion.

3.2 New Zealand public transport measures

Various PT performance measures are used by the regional authorities managing the bus, rail and ferry systems in the main centres of New Zealand. The review of New Zealand public transport measures was limited to Auckland, Wellington and Christchurch, as these are the main centres with sizeable public transport systems.

3.2.1 Auckland Transport

3.2.1.1 Public transport punctuality

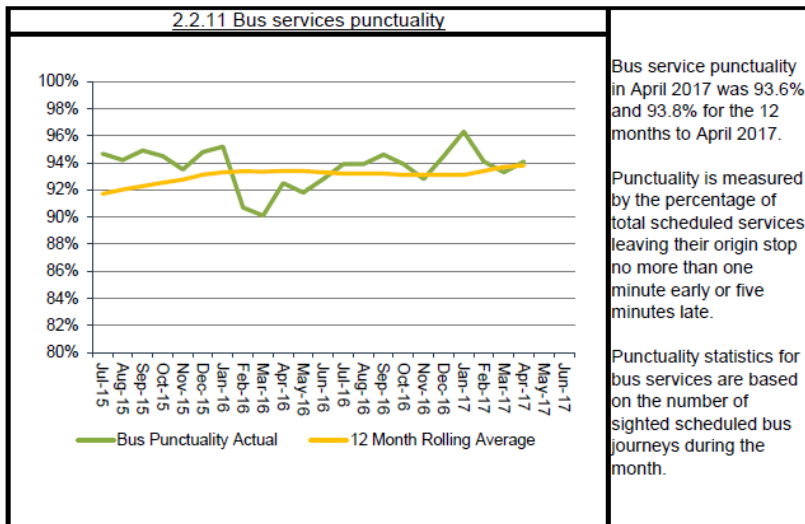
Similar to the Transport Agency punctuality measure, AT measures 'punctuality' for all its PT modes (ie buses, trains and ferries) and defines punctuality as the:

percentage of trips between 59 seconds before and 4 minutes 59 seconds after the scheduled departure time of selected points¹.

¹ Note that at the time of writing this report AT only looked at trips *starting* within between 59 seconds before and 4 minutes 59 seconds after the scheduled departure time. There is an intention however to expand this to intermediate points and the destination (ref: <https://at.govt.nz/media/679083/Item-11ii-Bus-reliability-and-punctuality-performance.pdf>)

At the time of writing, AT's 'selected points' referred only to the origin stations/stops. This means that the actual running time of the PT service is ignored. 'Punctuality' is reported to the board monthly in the monthly performance indicator report (AT 2016). Figure 3.1 shows a typical chart from this monthly report.

Figure 3.1 Punctuality reporting for public transport (AT 2017)



AVL data is used to calculate these measures and is now quite comprehensively collected from the majority of PT services in Auckland.

Table 3.2 provides a summary of the advantages and disadvantages of the AT punctuality measure, along with its applicability to the Transport Agency road index measure.

Table 3.2 Review of AT punctuality measure

Advantages	Disadvantages
<ul style="list-style-type: none"> Simple measure, easy to aggregate and express as percentage. Can aggregate across different routes. Measure is designed to ensure operators are maintaining key performance indicators and is independent from infrastructure inadequacies (ie control point from origin station). Consistent measure across all public transport modes and therefore can be aggregated across PT modes. Data is readily attainable to the operator through the AVL system. Possible to disaggregate into line or route level and to also investigate individual trips. 	<ul style="list-style-type: none"> Measure may be difficult for some customers to understand and is based solely on performance reporting. Difficult to identify which routes/time periods are having issues as they are hidden inside an aggregated percentage. It is an 'average' measure and therefore may disregard variable travel times. Does not capture the relative distribution of 'lateness' or 'earliness', ie a service could be 8 minutes late but under this methodology it is given the same weighting as a service which is 5 minutes late. Also, treats early and late arrivals with equal weight. Users are affected in different ways by late services (a delay) compared with early running (if a customer misses a low-frequency service due to the bus leaving early they would experience a greater delay). There is not a clear explanation of how the thresholds have been selected. These measures are sensitive to threshold determination. Uses departures from the origin station as the control point. Thus, this does not capture travel time reliability through the entire route and intermediate stations. There is no relative weighting in terms of how many people are affected, ie delay per passenger.

In addition to the above punctuality measure which has been publicly reported for some time, as described in chapter 5 during the validation workshops, AT reported that it now also calculates and publicly reports (via board reports issued 10 times a year²) other predictability-related measures which are referred to as 'reliability' and 'delay'. A brief description of each of these measures follows.

3.2.1.2 Reliability

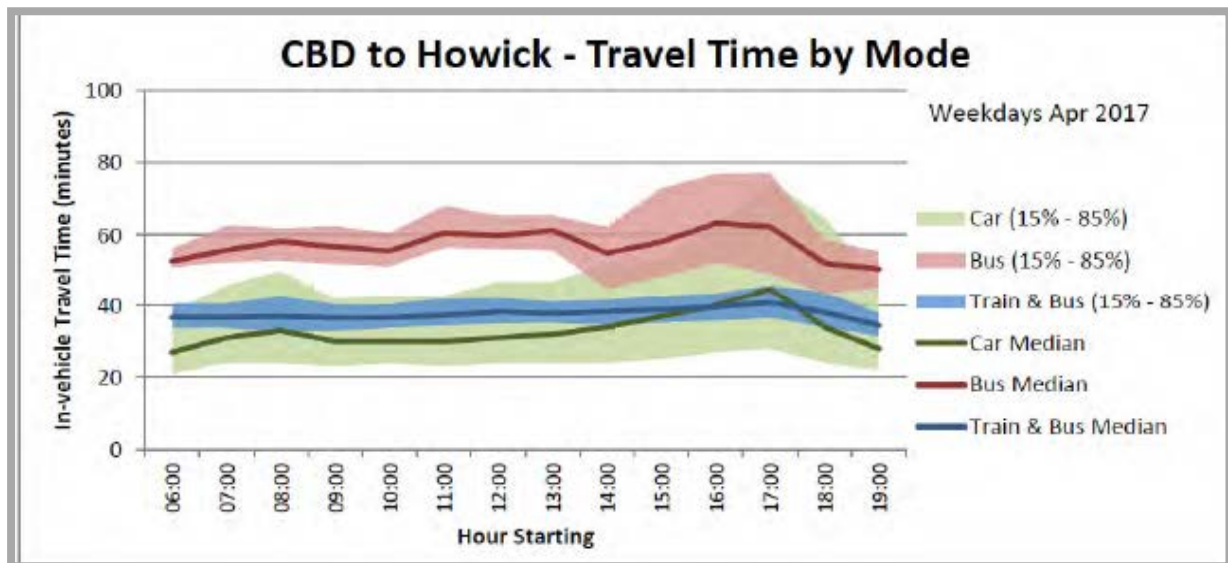
AT (2017) defines its 'reliability' measure as 'the additional travel time needed relative to typical travel time'. AT uses median for 'typical' travel time, and examines the 85th percentile travel time for the time period of interest. A recent board report provides the following example:

During the April 17 AM peak, the 85th percentile was 64% longer than the typical travel time. Therefore, if a typical AM peak journey took 20 minutes, a motorist would need to allow an additional 12.8 minutes, for a total of 32.8 minutes, to be 85% certain of arriving on time (AT 2017).

This measure is also used in practice by members of the Auckland Motorways Alliance who refer to it as a modified buffer index which is the name used to describe this measure in much of this report.

AT later advised this reliability measure was in use across multiple modes of transport including car, bus, train and bicycle. Figure 3.2 depicts an example of AT comparing travel time reliability across modes for an area of Auckland and shows that the measure is only comparing *in-vehicle* travel time.

Figure 3.2 'Reliability' reporting for public transport (AT 2017)



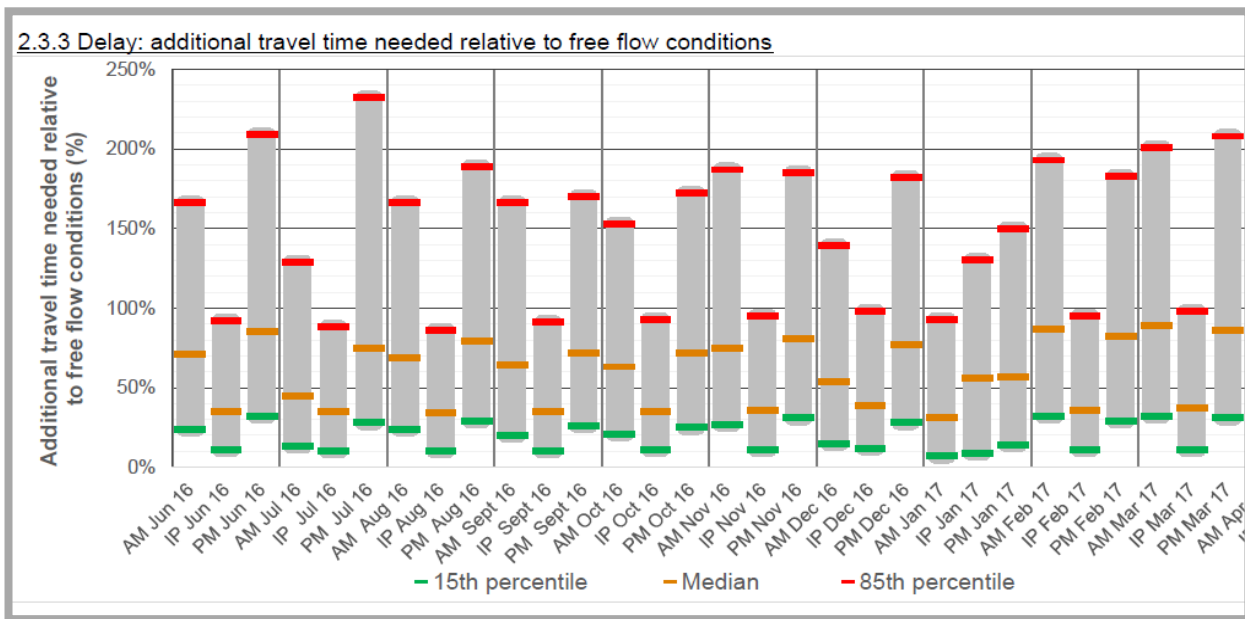
² For an example of a full board report visit: <https://at.govt.nz/media/1973778/item-122-monthly-indicators-report-april-2017-covering-paper.pdf>

3.2.1.3 Delay

AT also uses a measure called ‘delay’ which is related to the Planning Time Index documented in academic literature (as discussed in section 3.4). In a recent board report, AT (2017, p17) defines ‘delay’ as the ‘additional travel time needed relative to free flow conditions’.

An example of the reporting of this from a recent board report is provided in figure 3.3. It is worth noting that this example is for road networks generally; however as is discussed in section 5.3, it became apparent in the validation workshops that this measure is used across modes, including PT.

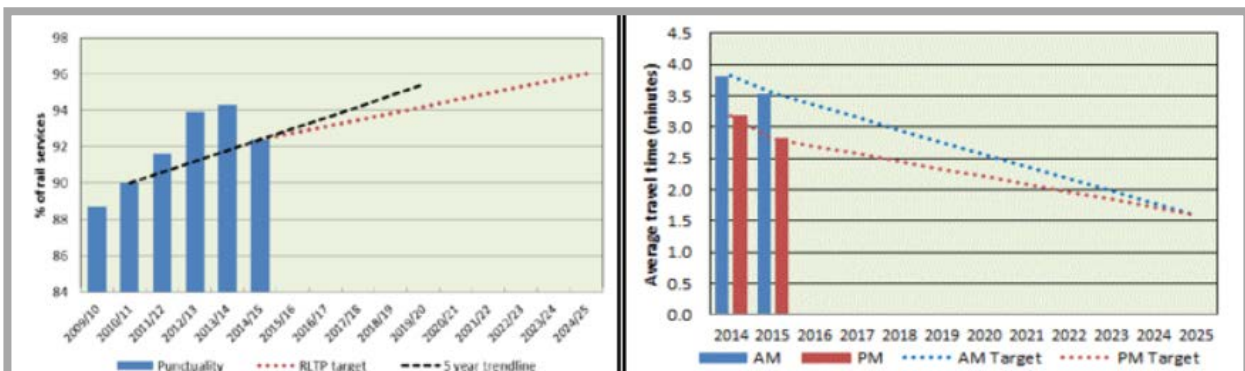
Figure 3.3 ‘Delay’ reporting for roads (AT 2017)



3.2.2 Greater Wellington Regional Council

GWRC uses a schedule adherence ‘punctuality’ measure for the Wellington train network. Punctuality in this instance relates to ‘percentage of train services which arrive and depart Wellington Station within five minutes of scheduled time’. For buses, GWRC measures performance using ‘average lateness’. Figure 3.4 presents charts from GWRC (2015) that illustrate how the measure is reported.

Figure 3.4 GWRC rail ‘punctuality’ (left) and ‘average lateness’ for buses (right) (GWRC 2015)



There is no published data for ferries in Wellington. GWRC reports performance by aggregating their measures to one year (core bus and rail routes only) and does not report at the same frequency as AT.

Currently data is accessed through the PT real-time information (RTI) system, which operates on the majority of Wellington's core rail and bus services (but there was no RTI on ferries at the time of writing).

During the first steering group meeting for this research project, GWRC advised they had adopted a new measure superseding the above. The new method aligns with the expected changes to the Public Transport Operating Model (PTOM) and uses the same definition as above but now considers selected key stops/stations through an entire route. The new measure better meets customer expectations and is much stricter than the previous single point measurement. However, the changes now mean that historic monitoring results cannot be compared as the two methodologies are incomparable. It is understood the new measure has been adopted for both bus and trains.

The measure has similar advantages and disadvantages to those described for the AT punctuality metric in table 3.2. One exception is that GWRC considers the measurements from the destination station as well as arrival station (Wellington Railway Station), which means the PT running time through the entire route is considered for many journeys.

3.2.3 Environment Canterbury/Canterbury Transport Operation Centre

Two separate measures for Christchurch are currently being used by Environment Canterbury (ECan) and the Christchurch Transport Operation Centre (CTOC).

ECan uses a measure termed 'reliability' (but very similar to punctuality), which is the percentage of trips having at least one late time-point departure along the bus route (figures 3.5 and 3.6). This is classed as anything later than six minutes or earlier than four minutes.

This is undertaken for all bus routes and the departure time is checked at a bus stop that is selected as a 'timing point'. A timing point is usually a key location such as a major interchange or shopping mall. Each route has approximately five to ten timing points.

Figure 3.5 ECan reliability performance for June 2016

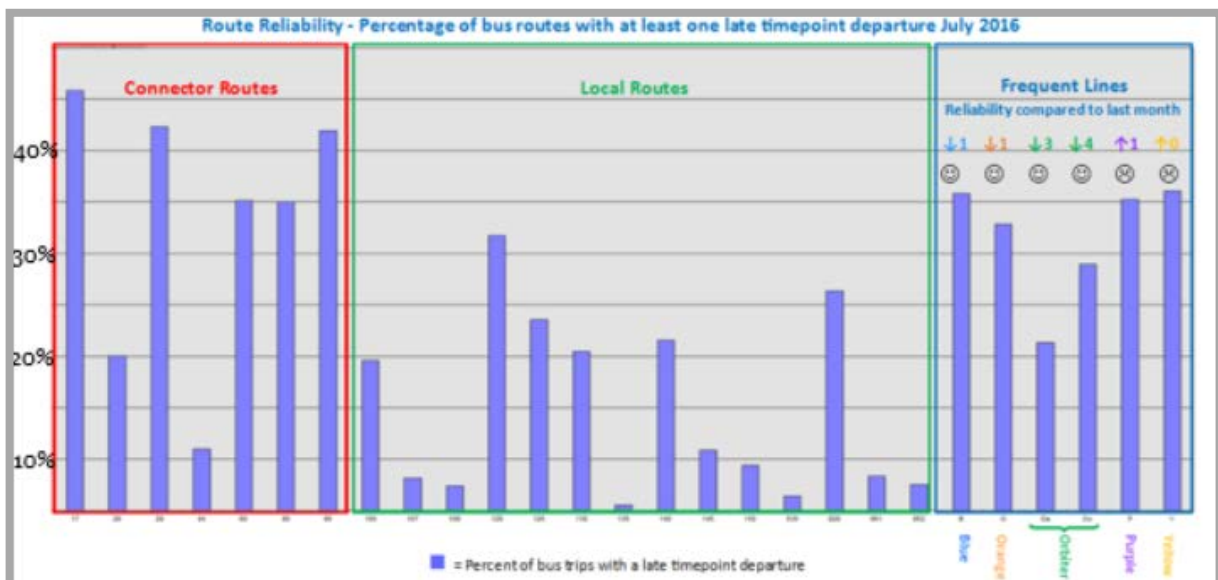
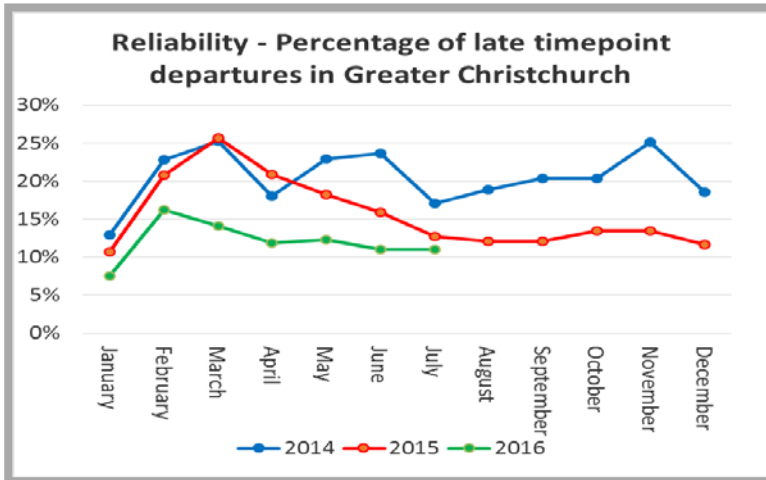


Figure 3.6 ECan reliability performance long term

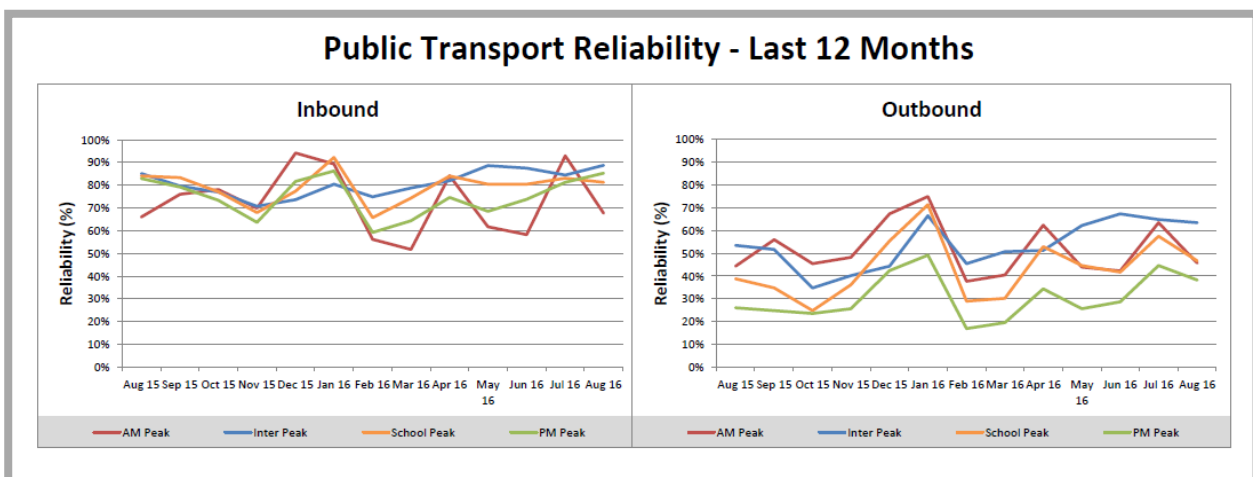


It is noted ECan does not run train services and only has one ferry service between Lyttelton and Diamond Harbour.

Meanwhile CTOC has adopted the concept of 'reliability', which in this instance relates to total percentage of scheduled bus services that arrive two minutes after or two minutes before the scheduled time at key time points.

In principle, this is the same concept as 'punctuality' as used by AT and GWRC. The only difference is the time threshold, with AT and GWRC both adopting one minute under and five minutes over the scheduled time. This is reported monthly along with other CTOC travel time performance measures for general road traffic. An example is shown below in figure 3.7. The other difference is that ECan uses the actual running time of the service compared with the scheduled running time (instead of using arrival/departure time compared with timetabled arrival/departure time). The scheduled running time is defined as the time difference (from timetable) between two stops of interest for a particular service.

Figure 3.7 Example of CTOC monthly reporting for PT reliability (CTOC 2016)



The ECan measure has very similar advantages and disadvantages as the punctuality measures used by AT and GWRC. However, in addition, it considers the actual running time of the journey which means the travel time through the entire route is considered and this is somewhat more customer oriented.

A summary of the metrics used by AT, GWRC and ECan is provided in table 3.3.

Table 3.3 Summary of New Zealand PT performance measures

Term	Definition	Region
Punctuality	Percentage of scheduled service trips leaving no more than 59 seconds before and 59 seconds after.	Auckland (trains, ferries and buses) and GWRC (trains only) – (AT 2016; GWRC 2015)
Reliability	Whether service is cancelled or not. Presumably binary measure – either '1' or '0'. Usually shown as a percentage between the actual number of services run for a given time period (month) and scheduled number of services.	Wellington – all public transport modes (GWRC 2015)
	Percentage of scheduled services which arrive 2 minutes before or 2 minutes after the scheduled time.	Christchurch (buses only) – (CTOC 2016)
	Percentage of trips with at least one late time-point departure along the bus route – this is classed as anything later than 6 minutes or earlier than 4 minutes.	Christchurch (ECan)
	Additional travel time needed relative to typical (median) travel time (usually examined for the 85th percentile travel time).	Auckland (AT 2017)
Average lateness	Used for buses by GWRC. Averages the aggregated lateness (from schedule) across all services.	Wellington (buses only) – (GWRC 2014)
Delay	Additional travel time needed relative to free flow conditions (can examine for various percentiles).	Auckland (AT 2017)

Overall the review of PT predictability measures being used in New Zealand revealed some consistency in the approach of measuring reliability via schedule-adherence measures but there was inconsistency in the thresholds adopted. Punctuality-like measures were typically defined as some variation of a deviation of the actual travel time from scheduled time, with a pre-determined threshold for late and early services. For Wellington and Auckland this threshold is five minutes late and one minute early; in Christchurch, for CTOC, it is two minutes for both earliness and lateness, and for ECan, it is any late time-point departure along a bus route later than six minutes or earlier than four minutes. This is then expressed as a percentage, usually aggregated monthly for a set of core routes.

Although in principle the concept is very similar amongst the three organisations, there are some inconsistencies in the terminology used. 'Punctuality' and 'reliability' are used interchangeably, which can lead to confusion. 'Reliability' in the sense of the GWRC measure relates to the percentage of PT services that were cancelled, ie they did not run at all or did not reach their final destination.

One of the big differences in punctuality reporting is that ECan uses the stop-to-stop running time against a scheduled running time (which can be worked out from the timetable) and they also disaggregate out the analysis into different time periods and directions. This is advantageous in that it does not skew results, for example when peak and non-peak directions are combined. A major disadvantage with the AT punctuality measure is that it compares the departure time at the origin. With this methodology, the running time of the actual service is completely excluded and as such the measure does not provide a useful means of understanding the overall performance of the service. It is utilised instead to determine the punctuality of the operator starting the service on time.

The punctuality measures are not particularly meaningful to customers (especially those using higher frequency services where schedule adherence becomes less relevant), rather they are geared towards measuring contractual performance by operators and to report the PT network's performance at a high level, ie board meetings and annual reporting. Knowing that AT reported 90% punctuality for all bus routes

gives no indication to customers on how their personal routes performed and no information about actual delays against which to plan their future trips. The travel conditions of the remaining 10% are not mentioned either.

All of the punctuality-related measures from the three centres are used to ensure that operators are achieving their KPIs and hence designed to measure travel time variability as a result of operator non-performance (rather than variability resulting from infrastructure and adverse traffic conditions). With these in mind, the measures provide little information on problems occurring on specific services/corridors, time periods and how many customers are affected. Nor do they allow for opportunities to be identified, eg to increase frequency on a route without investing in more vehicles by removing traffic (or even artificial schedule) delays. However, its simplistic nature provides a way in which significant deviations can be identified easily and interrogated further if required. All methods make use of the AVL/RTI data sources which enable consistent data formats and have flexibility in selecting an analysis to cater for specific needs.

3.3 International measures (industry practice)

3.3.1 Public Transport Victoria, Australia

Public Transport Victoria (PTV) in Australia reports on its network operational performance for metropolitan trains and trams, regional trains as well as metropolitan and regional route bus services. It uses both 'reliability', also termed 'service delivery', and 'punctuality' measures as defined below:

- Punctuality – measured as a percentage of on time arrivals at specified monitoring points. Range as outlined below:
 - metropolitan trains – no later than four minutes and 59 seconds after the timetabled arrival time
 - regional trains – no later than five minutes and 59 seconds after the timetabled arrival time for short distance services and 10 minutes and 59 seconds after the timetabled arrival time for long distance services
 - buses – no more than five minutes late at any point on the timetable.
- Reliability – measured as a proportion of the timetable that is actually delivered by the operator. Performance results are made available daily (for trains and trams) on PTV's website and reported monthly, quarterly and annually summarised quarterly in its 'Track record' reporting series³. PTV performance reporting is also linked to customer compensation when targets for punctuality and reliability are not met. An example of the output can be seen in figure 3.8. Performance measures of 'delivery' and 'punctuality' are posted on the train operator's website and at railway stations.

Data is collected through telemetry, global positioning system (GPS) and manually recorded by individual providers. This process is supported by independent sample surveys undertaken by PTV to ensure the information supplied is accurate.

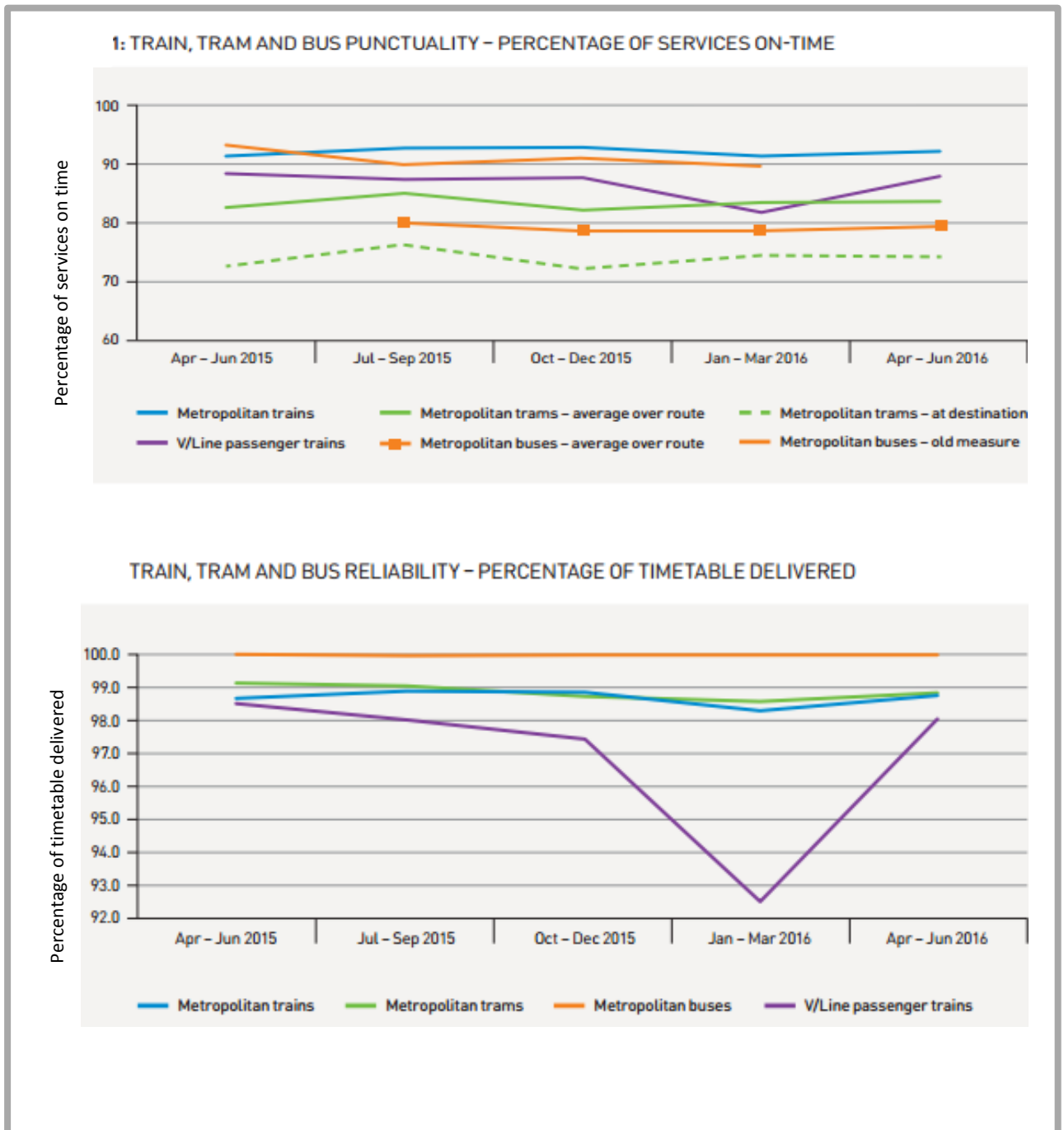
3.3.1.1 Advantages and disadvantages of the Public Transport Victoria measure

The PTV measure has similar advantages and disadvantages to the New Zealand punctuality metrics but with the additional benefit in that, although it is largely an operator focused metric, its measurement

³ www.ptv.vic.gov.au/about-ptv/data-and-reports/operational-performance/

carries a benefit for customers because it is linked to the compensation scheme. Since it is a schedule adherence measure it cannot be used to compare reliability with road traffic.

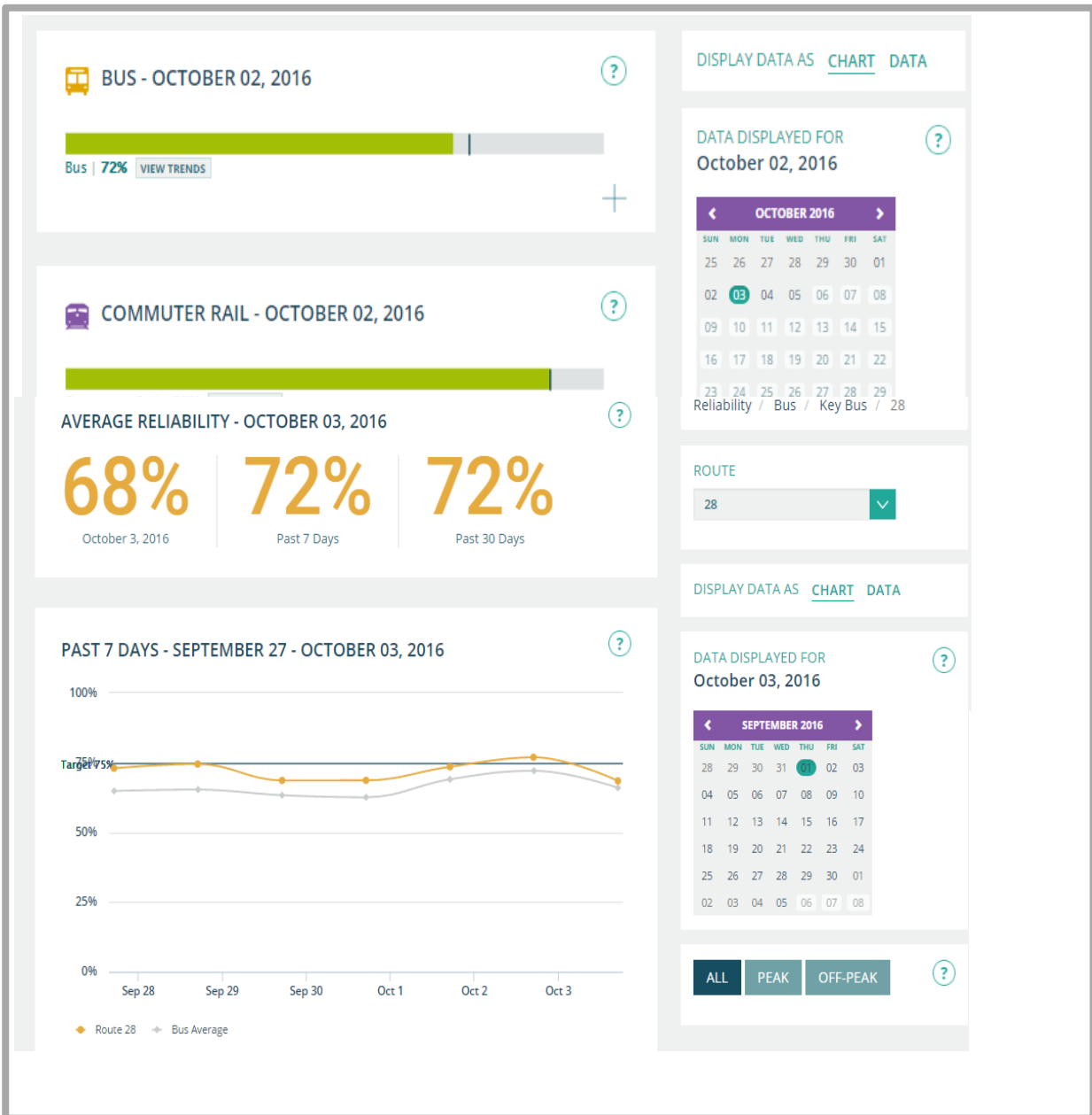
Figure 3.8 Example of PTV quarterly reporting for PT punctuality and reliability (Public Transport Victoria nd)



3.3.2 Massachusetts Bay Transportation Authority, US

Massachusetts Bay Transportation Authority (MBTA 2016) reports 'reliability' for buses, commuter rail and the subway network, and this is available on its dashboard. An example of the output can be seen in figure 3.9.

Figure 3.9 Example of reliability reporting on the Massachusetts Bay Transportation Authority dashboard (Source: MBTA 2016)



The data is available for each day and can be downloaded for a specified period. It can also be viewed for peak and off-peak services and is measured as follows:

- Buses – measured at each end of the route and at key stops in the journey. For services that come every 15 minutes, or less, reliability is calculated as the percentage of buses that are no more than three minutes later than the scheduled interval since the last bus (three-minute buffer). For other bus services, MBTA measures the percentage of stops the bus leaves between one minute before and six minutes after the scheduled time.
- Trains – reliability is measured by the percentage of trains that arrive at the final stop within five minutes of the schedule.

Thus, both bus and train measures are primarily related to schedule adherence punctuality; however, for more frequent bus services (with headways of 15 minutes or less), the reliability performance measure is a little different and is based on 'schedule intervals' associated with the previous scheduled service. This may be a more customer-centric approach as it better accounts for waiting periods but is also a little less intuitive.

Reliability data for buses is collected via GPS and a broader vehicle technology suite called TransitMaster. For rail, a signalling system and GPS are used to collect data simultaneously.

A summary of the advantages and disadvantages of the MBTA's performance reporting is provided in table 3.4.

Table 3.4 Summary of MBTA characteristics, advantages and disadvantages

Advantages	Disadvantages
<ul style="list-style-type: none"> • Data is readily attainable via the dashboard on the MBTA website. • Data is updated daily with trends displayed for the past 7 days, and 30 days. Further data can be downloaded. • Information can be separated into peak/off peak period and different types of services. • Simple measure, easy to aggregate and express as percentage. Can aggregate across different routes. • Measure considers the infrastructure delays as intermediate locations are monitored (rather than just the origin point). • Customer focus comes from the website reporting platform where users can readily access past service information. • Ridership data is also made available on the dashboard. 	<ul style="list-style-type: none"> • Does not capture the relative distribution of 'lateness' or 'earliness'. • Inherently still an operator-based metric with some improvements to the customer aspect through the website platform. The metric itself means very little to the customer and their personal journeys. • Schedule adherence measure means that it cannot be used to examine reliability across PT and roads.

3.3.3 Transport for London, UK

Transport for London applies several different measures for their array of PT services. These are summarised below:

3.3.3.1 Bus (Reed 2013)

- *Low frequency services*: where a service operates with four scheduled buses an hour or fewer, London treats this as low frequency and the contract with the bus operator and the report of performance is based on a percentage 'on time' calculation where on-time is within -2.5 minutes (early) or +5 minutes (late) of the scheduled arrival time.
- *High frequency services*: where a service operates more than four buses an hour, London treats this as high frequency and the calculation method is based on the calculation of 'excess waiting time', this being the difference between the scheduled waiting time and the actual waiting time as recorded at points along the route. The actual waiting time is weighted in accordance with the headways which are timetabled. An example of this type of reporting is provided in table 3.5.

Table 3.5 Example of London high frequency bus service reporting (adapted from Transport for London 2016a)

Measure	This quarter	Same quarter a year ago
Average scheduled waiting (minutes)	4.97	4.93
Average excess waiting (minutes)	1.12	1.10
Average actual waiting (minutes)	6.09	6.02
% Chance of waiting <10 minutes	81.3%	81.7%
% Chance of waiting 10–20 minutes	16.8%	16.5%
% Chance of waiting 20–30 minutes	1.7%	1.6%
% Chance of waiting >30 minutes	0.3%	0.3%

3.3.3.2 Underground (Transport for London 2016b)

The travel time reliability measure for the London Underground system is particularly customer focused. Travel time reliability on the underground system is expressed as the excess journey time. This is the time in minutes to complete an average journey on the network over and above the scheduled journey time, weighted by customer time values.

Transport for London breaks journeys down into stages and gives each one:

- a scheduled time length, ie how long a given journey should take if everything goes as planned
- a value of time based on how customers feel about that leg of the journey, eg going up an escalator has a value of time of 1.5, whereas walking upstairs has a value of time of 4, because it makes the perceived journey time longer.

Under this model the stages of the journey accounted for include:

- time from station entrance to platform
- ticket queuing and purchase time
- platform wait time
- on-train time
- platform to platform interchange
- time from platform to station exit.

In each period, actual journey times are measured and then compared with the schedule. The difference between the two is the measure of lateness, referred to as excess journey time. This is therefore a measure of how efficiently London Underground is providing its scheduled or 'stated' service: the more reliable the service the lower the excess journey time. The calculation includes the impact of planned closures. An advantage of this measure is that it is customer-centric in taking account of not just in-vehicle time but other components of PT trips and also value of time for customers. The major disadvantage of this measure is that it has been designed around the specific characteristics of the London Underground and is not easily transferrable to other modes of transport. In particular, its reliance on PT schedules means that it cannot be used for private vehicle travel on the road network.

The underground system also uses 'lost customer hours' which is the total extra journey time, measured in hours, experienced by Underground customers (based on patronage) as a result of all service disruptions with durations of two minutes or more. For example, an incident at Oxford Circus during a

Monday to Friday peak gives rise to a much higher number of lost customer hours than an incident of the same length in zone 6 on a Sunday morning. In the actual London Underground performance reports, lost customer hours are shown for each subway line for the time period and for a (rolling) 'moving annual average'. In addition, for each line, another graph is provided indicating the causes of lost customer hours (eg 'fleet', 'staff').

A summary of the advantages and disadvantages of Transport for London's reliability performance measures is provided in table 3.6.

Table 3.6 Summary of advantages and disadvantages of TfL's reliability reporting measures

Advantages	Disadvantages
<ul style="list-style-type: none"> • Data is separated by each service line in the Underground. • Bus service shows distribution in wait time rather than just the arbitrary 5 or 3 minutes used by other operators. • Highly customer oriented with the underground using total journey time from origin station entrance to destination station exit and waited by passengers' perceptions/values of time. • The lost customer hours measure ensures unreliable services are weighted to reflect patronage of each service. • Also provide detailed reporting of causes of lost customer hours. 	<ul style="list-style-type: none"> • Unclear at what time periods the analysis is conducted for (ie peak hours or all day). • Inconsistency in the measures for each PT mode. • Excess journey time relies on schedule adherence so is not appropriate for comparing private vehicle and PT predictability. • Requires highly sophisticated data platform and analysis, ie Automatic Fare Collection Oyster Card and gated entry at station. • High cost in collecting and monitoring data. • Lost customer hours is an average metric and does not measure variation.

3.4 International measures (academic)

Internationally, 'reliability' was the most commonly used term to describe trip inconsistencies, variability and predictability. For this reason, this section describes how reliability has been used.

3.4.1 Travel time reliability measures

Travel time reliability in general terms is defined as a measure of trip consistency during a specific time period in a specific location. It takes into account more than daily congestion and is attributed to route inconsistencies due to unexpected delay (Kimley Horn and Associates 2011). Commuters are faced with traffic congestion on a daily basis and plan their trip based on their experience of the network. However, unexpected congestion may impose a delay and lead them to arrive late at their destination. As a result, travel time reliability has been identified by some as the most important factor affecting ridership satisfaction (Gaffney 2006).

In PT, however, travel time reliability definitions vary. From one point of view, it is defined as the consistency of PT travel time and ability of the PT system to maintain regular headway and adhere to a schedule (Chen et al 2009). Mazloumi et al (2008) have distinguished between travel time reliability and variability definitions. Travel time variability is variation in travel time while travel time reliability is the level of trip consistency to expected arrival time.

As summarised in table 3.7, using Sydney as a case study, Currie et al (2012) reviewed 10 indicators of urban bus service reliability (and variability). The indicators included the percentage of services cancelled; the percentage departing 'on time'; the percentage arriving 'on time'; excess waiting time; the average

lateness of services; variability measures (standard deviation etc.); a reliability buffer index; passenger rating of reliability, customer complaints, and a customer journey time delay measure (at stop and on bus). The latter measure was added to an initial 2006 review of indicators by the research team (Douglas et al 2006) and was enabled by the development of AVL data which provides more easily available travel time data. Indicators were rated for five attributes: ease of understanding, customer focus, fidelity and objectivity and cost/effort efficiency. Of the 10 indicators ranked in Currie et al (2012), the two highest performing measures were 'excess waiting time' and 'customer delay'. Although 'excess waiting time', is very customer-centric, the measure does not reflect an 'in-vehicle' travel time which is outside the scope of the present research. Customer delay performed well under the authors' assessment but relies on stop-to-stop bus times which can be onerous to collect. The authors found bus frequency to affect the evaluation of measures. For instance, excess waiting time is easier to calculate on higher frequency routes and, the percentage of buses departing on time, a schedule adherence measure that was highly rated in the 2006 study, is more difficult to calculate on higher frequency routes where it can be harder to link to the schedules.

Table 3.7 Summary of reliability measures ranking (Currie et al 2012)

Reliability indicator	Ease of understanding	Customer focus	Fidelity and Objectivity	Cost/effort efficiency	Overall rating
% Buses cancelled	High	Low	Low	Low	Medium
% Departing on-time	High	Low	Medium	Medium	Medium
% Arriving on-time	High	Low	Medium	Medium	Medium
Excess waiting time	High	High	Medium	Medium	High
Average lateness	High	Low	Medium	Medium	Medium
Variability measures	Low	Low	Low	Medium	Low
Reliability buffer	Low	Medium	Low	Low	Low
Passenger rating	Medium	High	Medium	Low	Medium
Customer complaints	High	High	Low	Medium	Medium
Customer delay	High	High	High	Low	High

As evidenced by the evaluation criteria in the table by Currie et al (2012), PT reliability can be measured from different perspectives, eg from the point of view of an operator or a passenger and some indicators serve one vantage or another better.

3.4.2 Reliability measures from operators' point of view

Usually, on-time performance/punctuality (and or schedule adherence) and headway regularity are considered as effective measures to report journey reliability from an operator's point of view. On-time performance is the percentage of trip departure time deviation from schedule time point (Nakanishi 1997). The Dutch Ministry of Transport states a train is punctual if it is not more than three minutes late (Ruben

et al 2011). However, the three-minute threshold has been criticised because using multiple punctuality thresholds makes for better analysis (Ruben et al 2011).

Although there are wide variations in the definition of on-time performance, it is usually used to report the proportion of buses that are not more than one minute early and no more than five minute late (Chen et al 2009). Late (percentage of buses more than 10 minutes late) and not observed buses can also be reported and this measure can highlight routes with serious problems (Nakanishi 1997).

'Head way regularity' measures how evenly distributed PT services are compared with scheduled services (Cramer et al 2008). 'Coefficient of variation of head way' is another common method to investigate headway regularity (Furth et al 2006). Typically, headway regularity measures are used for high-frequency services (usually defined as headways of 12 minutes or less) (Schil 2012).

3.4.3 Reliability measures from the customer's point of view

Transport for London (2011) conducted an extensive study on how customers view PT reliability. A significant finding from this study was that customers value personalised information and metrics. The study confirmed that the aggregated route and network-wide performance measures (usually expressed as percentages or averages) are meaningless to customers as these metrics do not relate to their own experiences. Customers also wanted travel time reliability to be considered at the whole journey level (ie service wait time, running time, exit time).

From a customer point of view, PT reliability is mainly concerned with waiting-time related measures including; excess waiting time and potential waiting time. Thus, for PT reliability from a passenger's vantage, Transport for London proposes using excess waiting time as a key indicator of reliability for high frequency routes (5+ services per hour) (Reed 2013). The excess waiting time is the difference between average 'schedule waiting time' and 'actual waiting time'. The 'average schedule waiting time' for high-frequency routes can be calculated as half the service frequency (as the passengers randomly arrive at bus stops) if the service ran exactly as scheduled, assuming no 'rare' events occur that hold up vehicles for long periods of time. 'Potential waiting time' is the difference between budget time (budget time is the 95th percentile waiting time) and mean waiting time (Furth et al 2006). For long-headway service, excess waiting time is defined as the difference between the mean and 2-percentile departure deviation (it is assumed experienced travellers will arrive early to decrease the probability of missing a bus to 2% or less) (Furth et al 2006). For example, if the passenger budgets 10 minutes for waiting, but the bus arrives after three minutes, the seven-minute difference is the potential waiting time. In this case the traveller will arrive seven minutes earlier at the destination. While this sounds beneficial, it might not be appreciated by someone who would rather have spent that time differently.

3.4.4 Travel time reliability indices

The literature review revealed many measures for travel time reliability. This section explains them in more detail. It should be noted that some of the measures below are road-based measures of reliability rather than strictly PT measures.

3.4.4.1 Standard deviation

Standard deviation is a classic statistical technique that measures the spread of the data. The more concentrated the data is around the mean, the smaller the standard deviation will be. It is generally represented by sigma (σ).

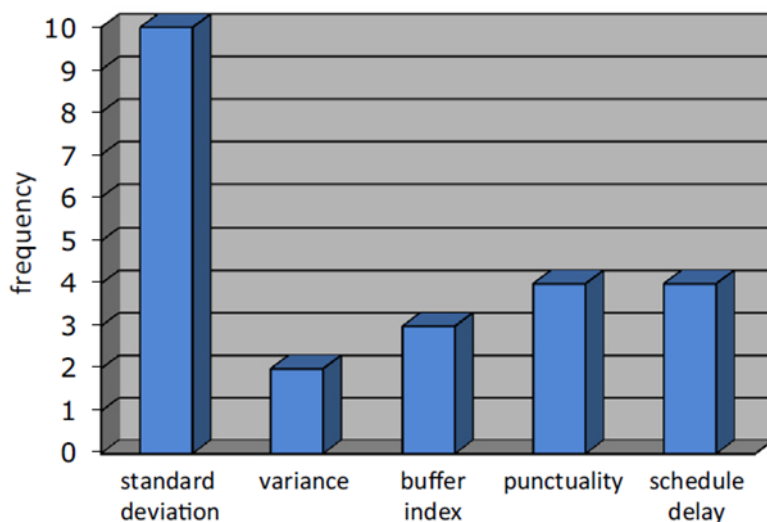
(Equation 3.1)

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X - \bar{X})^2}{N - 1}}$$

where X is individual measurement of travel time, \bar{X} is mean travel time and N is sample size.

Standard deviation is sometimes used as a proxy for other reliability measures (Dowling et al 2009) and is a convenient measure when the reliability of travel time is calculated with classic mathematical or statistical models (Dong and Mahmassani 2009). The US Federal Highway Administration (FHWA 2006) discourages its use as a reliability performance measure because it is not easily understood by non-technical audiences nor easily related to everyday commuting experiences, and it treats early and late arrivals with equal weight, whereas they should not be treated with equal weight. More recently, (de Jong and Bliemer 2015) interviewed international experts on travel and transport time reliability. One of the questions asked was ‘which operational definition of reliability would you recommend for including reliability in the cost-benefit analysis?’ The results are presented in figure 3.10 and show that standard deviation had the most support among the experts as a measure of reliability.

Figure 3.10 Most appropriate definition of reliability for use in cost- benefit analysis: frequency distribution of answers of the experts (Source: de Jong and Bliemer 2015)



Arguments for using the standard deviation were:

- It can be empirically measured.
- It has an indirect base in theory, since Fosgerau and Karlström (2010) showed the formal equivalence with the scheduling model (at least for modes without timetables, such as the car).
- It is relatively easy to include in standard transport models because it does not require schedules to be imported into the model.
- Related to the previous, since it requires no formal scheduling model, it also does not require preferred arrival times, for which specific survey interviews would be needed or reverse engineering (Kristoffersson 2013).
- It often provides a good fit to stated-preference data (choices between alternatives that differ in terms of reliability are often well explained by a model that includes the standard deviation).

- It can capture a residual (non-scheduling-related) value (eg anxiety).

Arguments against using the standard deviation are:

- It is sensitive to outliers.
- It is dependent on other aspects (such as other moments) of the travel time distribution, though it does not properly pick up the form of the tail and skew.
- It cannot be used to compare variability between routes of different route lengths since standard deviation increases with route length.
- It is not additive over links. Even when link travel times are independent of each other, simple summation of standard deviations (unlike the variance) over links will not give the standard deviation of the route that travels through these links. Aggregating variance as opposed to standard deviation may resolve the last argument, however, in a congested network, congestion spreads backwards from the original bottleneck, creating dependence among the travel times of adjacent links. In this situation, the variance is not additive over links either.

3.4.4.2 Coefficient of variation

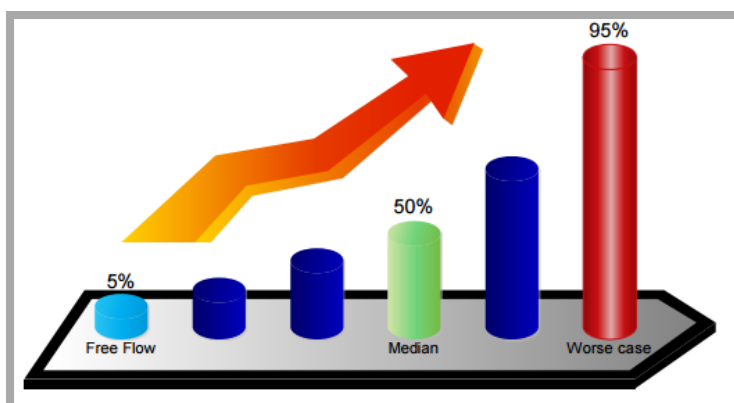
The coefficient of variation is the ratio of the standard deviation to the mean. It experiences the same disadvantages as standard deviation and as a result, use of the coefficient of variation is also discouraged (FHWA 2009 and Cambridge Systematics et al 2008). However, Mazloumi et al (2008) recommend using the coefficient of variation when reliability is compared between routes with different lengths. Mazloumi et al (2010) demonstrate the coefficient of variation has an inverse to constant relationship with the average travel time. The relationship is strong in sections with low average travel time, whereas it tends to be relatively constant as the average travel time increases (Mazloumi et al 2010).

Meanwhile, *Percent variation = coefficient of variation × 100%* (Cambridge Systematics 2008) mathematically has the same characteristics as the coefficient of variation. However, because the percent variation is expressed as a percentage of average travel time, it is more easily understood by the public. This measure was adopted by the 1998 California Transportation Plan (Jackson 2000; CALTRANS 1998) and recommended by Lomax et al (2003) and *NCHRP report 618* (Cambridge Systematics et al 2008).

3.4.4.3 Percentile travel time

The 95th or 90th percentile travel time (refer figure 3.11) is a simple and straightforward measure of travel time reliability. It can be used to predict how long a delay will be on a specific route during the heaviest traffic days. These worse-case travel days are generally caused by non-recurring congestion, eg traffic crashes, inclement weather, construction work or a special event. The percentile measure is not sensitive to the presence of outliers or rare events as it ranks observations from smallest to largest. Hence it is robust to the presence of any abnormality in the data.

Figure 3.11 Percentile travel time (Rashidi 2016)



Some commonly used percentile travel time measures are as follows (Wakabayashi 2012):

- TT_{95} : The 95th percentile travel time is a measure representing the first worst travel time (equals to the mean plus twice the standard deviation of travel time following a normal distribution).
- TT_{90} : The 90th percentile travel time is a measure representing the second worst travel time.
- TT_{50} : The 50th percentile travel time (equal to the median).
- $TT_{90} - TT_{10}$: The differences between 90th and 10th percentile travel time.
- $TT_{85} - TT_{15}$: The differences between 85th and 15th percentile travel time.
- $TT_{70} - TT_{30}$: The differences between 70th and 30th percentile travel time.
- $TT_{84} - TT_{50}$: The difference between 84th and 50th percentile (roughly equal to the standard deviation, assuming a one-tailed normal distribution (List et al 2014)).

Figure 3.12 shows how Washington State Department of Transportation uses a percentile travel time measure to inform customers of the worst-case travel times (ie 95th percentile). By allowing for the calculated travel time, commuters can expect to arrive at the end of the route, on time, 19 out of 20 working days a month (95% of trips). In this instance, the measure informs car travellers and does not appear to be in use for multi-modal or PT trips.

Figure 3.12 Washington State Department of Transportation provides reliability measures for traveller information (Source: www.wsdot.com/traffic/seattle/default.aspx?cam=9350#cam)

The screenshot shows the Washington State Department of Transportation website. At the top, there is a navigation bar with links for News, Search, Contact WSDOT, and WSDOT Home. Below this is a menu with categories: TRAFFIC & ROADS, PROJECTS, BUSINESS, ENVIRONMENTAL, and MAPS & DATA. The main heading is 'TRAVEL INFORMATION'.

On the left side, there are two sections: 'TRAFFIC LINKS' with links for Freeway Cameras, Traffic Conditions, Incidents, Travel Times, Construction Info, Map Archive, Pass Reports, and Questions; and 'MOST REQUESTED' with links for Highway Cameras, Puget Sound Traffic Flow Map, Washington State Ferries, Amtrak Cascades, Frequently Asked Questions, and Driver & Vehicle Licensing.

The main content area is titled 'Calculate Your Commute'. It shows the results for a route from Everett to Seattle, leaving at 7:00 AM. The results indicate a travel time of 59 minutes and a reliability of 19 out of 20 weekdays per month. A link to 'Modify your route' is provided. To the right, there is a map showing the route from Everett to Seattle, with a red line indicating the path. The map includes labels for Everett, Mukilteo, Lynnwood, Bothell, Woodinville, Redmond, Bellevue, and Seattle. A link to 'Check traffic cameras for your commute' is also present.

3.4.4.4 Skew statistic

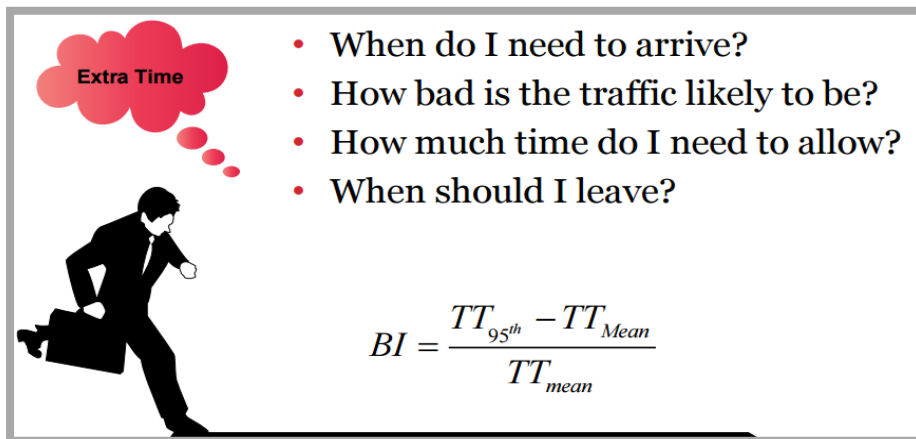
The skew statistic is defined as the ratio of the difference between the 90th percentile travel time and the median and the difference between the median and the 10th percentile (Margiotta et al 2008). It shows how much greater the range of the distribution is above the median compared with the range below the median (Dowling et al 2009).

3.4.4.5 Buffer index

Buffer time is the extra time required to arrive on time in addition to the travel time under average conditions. Traditionally, arithmetic average travel time is used to represent the travel time under average conditions, and as indicated in figure 3.13 the buffer index or buffer time index is defined by the difference between the 95th percentile travel time and the average travel time, normalised by the average travel time (FHWA 2009; Lomax et al 2003; Margiotta et al 2008).

$$BI = \frac{TT_{95^{th}} - TT_{mean}}{TT_{mean}} \quad (\text{Equation 3.2})$$

Figure 3.13 Buffer index (Rashidi 2016)



(Equation 3.3)

For example, a buffer index of 40% means that for a trip that usually takes 20 minutes a traveller should allow an *additional* eight minutes to ensure on-time arrival most of the time.

- average travel time = 20 minutes
- buffer index = 40%
- buffer time = 20 minutes × 0.40 = 8 minutes.

The eight additional minutes are called the buffer time. Therefore, the traveller should allow 28 minutes for the trip in order to ensure on-time arrival 95% of the time.

Meanwhile, Mazloumi et al (2008) applied different percentile-based measures including a buffer index to investigate travel time variability/reliability using AVL data. They preferred the buffer index and the coefficient of variation as they provide more information to users and planners. They argue that travel time interval selection can cause variability between different measures and shorter interval recommended. They also found that PT travel time variation trends differ compared with car travel time.

The buffer index is not a preferred measure when the travel time distribution is heavily right-skewed and median should be used instead of the mean (Pu 2011).

3.4.4.6 Modified buffer index

In line with the concern by Pu (2011) about distribution, a median-based variation of the buffer index is used by members of the Auckland Motorways Alliance. This measure, which they refer to as a modified buffer index, also uses 85th percentile travel time. As noted in the review of measures used for New Zealand PT, AT now also uses the modified buffer index, but they refer to the measure as 'reliability'.

$$MBI = \frac{TT_{85th} - TT_{Median}}{TT_{Median}} \quad (\text{Equation 3.4})$$

3.4.4.7 Frequency of congestion

Another reliability measure recommended by the US DOT guide is the frequency of congestion exceeding some expected threshold, typically expressed as the percent of time that travel times exceed a threshold (FHWA 2006). It assumes conditions are considered congested when speed is less than or equal to 50% of the free-flow speed, or, equivalently, travel time is equal to or larger than two times the free-flow travel time (INRIX 2011).

3.4.4.8 Travel time index

The travel time index is the ratio of actual average travel time to free-flow travel time. Strictly speaking, the travel time index is a congestion intensity measure rather than a reliability measure. To compare the relative scale of congestion intensity and travel time reliability, this report also includes this index:

$$\text{Travel Time Index} = \frac{TT_a}{TT_f} \quad (\text{Equation 3.5})$$

where TT_a is actual travel time.

3.4.4.9 Planning time index

Planning time is the *total* travel time (including buffer time) and is usually calculated as the 95th percentile travel time (FHWA 2006). The planning time index (or 'planning index') is the ratio of the 95th percentile travel time over free flow travel time. It expresses the extra time a traveller should budget in addition to free-flow travel time to arrive on time 95% of the time. If the 15th percentile travel time is considered as free-flow travel time, the planning time index can be calculated as follows (FHWA 2010):

$$\text{Planning Time Index} = \frac{TT_{95th}}{TT_f} \quad (\text{Equation 3.6})$$

Where TT_f is free flow travel time.

For example, a planning time index of 1.60 means that for a trip that takes 15 minutes in light traffic a traveller should budget a total of 24 minutes to ensure on-time arrival 95% of the time. For travellers who are familiar with congestion on a particular trip (eg regular commuters), the buffer index would be a preferred travel time reliability measure because it is based on average travel time; for those who are not familiar with that, the planning time index may be preferred because it is based on free-flow travel time (Lomax et al 2003).

3.4.4.10 Failure rate

On-time arrival estimates the percentage of time that a traveller arrives on time on the basis of an acceptable lateness threshold (Cambridge Systematics et al 2008). Failure rate = 100% – percent of on-time arrival (Pu 2011). The threshold travel time to determine an on-time arrival ranges from 110% to 130% of average travel time (Lomax et al 2003; Cambridge Systematics et al 2008).

3.4.4.11 Misery index (tardy trip indicator)

The misery index looks at how the average of the worst set of trips (ie 20 worst) exceeds the average of all trips. Comparing that with the average travel rate for all trips would give a measure of ‘how bad are the worst days’?

$$\text{Misery Index (MT)} = \frac{ATT_{20}}{ATT} \quad (\text{Equation 3.7})$$

where ATT_{20} is the average travel time for the longest 20% of the trips and ATT is the average travel time.

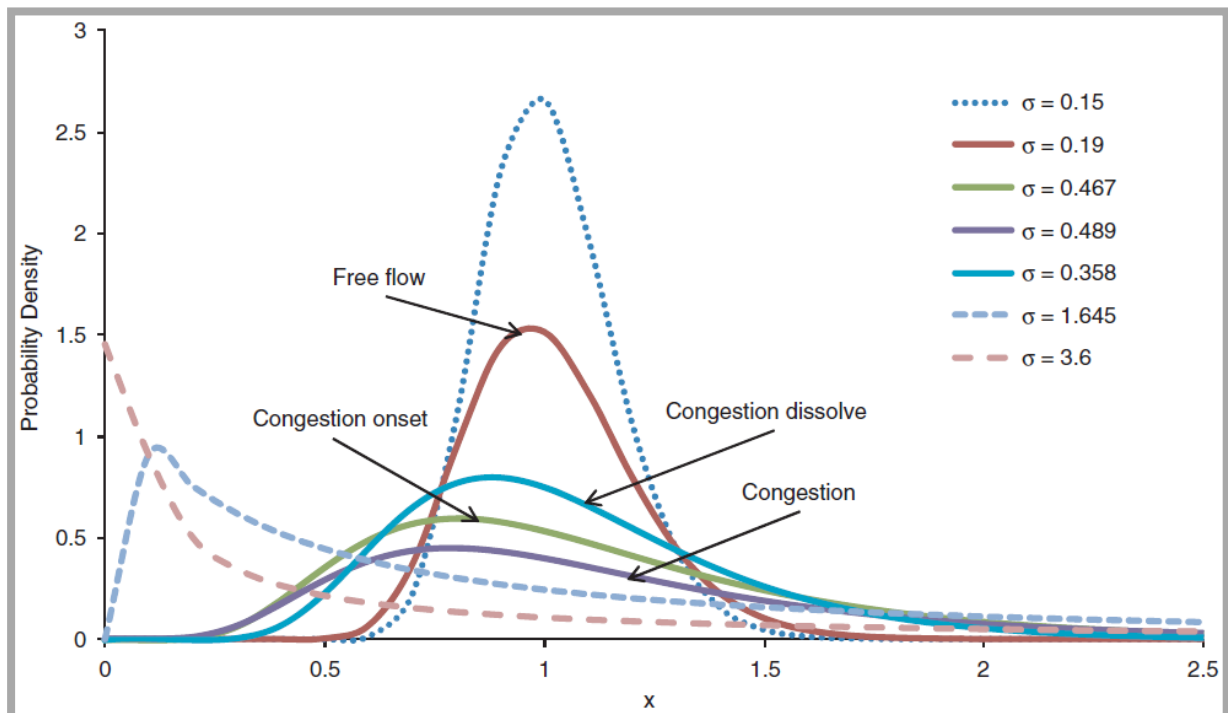
3.4.4.12 Distribution of travel time

Cambridge Systematics (2003) states that ‘the log-normal distribution is the closest traditional statistical distribution that describes the distribution of travel times’. Van Lint and van Zuylen (2005) and Van Lint et al (2008) depict travel time distributions with four different shapes based on traffic conditions:

- 1 Free-flow conditions (approximately symmetric, small spread, low median)
- 2 Congestion onset (right-skewed, higher median than the free-flow conditions)
- 3 Congestion (approximately symmetric or slightly right-skewed, wide spread, highest median)
- 4 Congestion dissolve (right-skewed, median similar to congestion onset).

These shapes of travel time distribution are similar to those of the lognormal distribution if different parameter values are used. Figure 3.14 shows probability density function of standard lognormal distribution for different values of standard deviation.

Figure 3.14 Probability density function of standard lognormal distribution (Source: Pu 2011)



Further investigation of the distribution of travel time has been undertaken by Park et al (2010) who argue that commonly used reliability measures, such as percentile travel time, travel time index and buffer index consider a single unimodal normal or lognormal distribution based on historical roadway travel times. The authors argue that, in fact, field travel time data indicate a ‘multimodal distribution’ and in response the

team developed a 'multistate mixture model' which supported the appropriateness of a two-stage model that could provide drivers with a probability of encountering congestion and information on the possible duration travel time if congestion is encountered. While this research is promising, the framework presented is self-described as 'novice' with 'further refinement of the model...required' (Park et al 2010, p82).

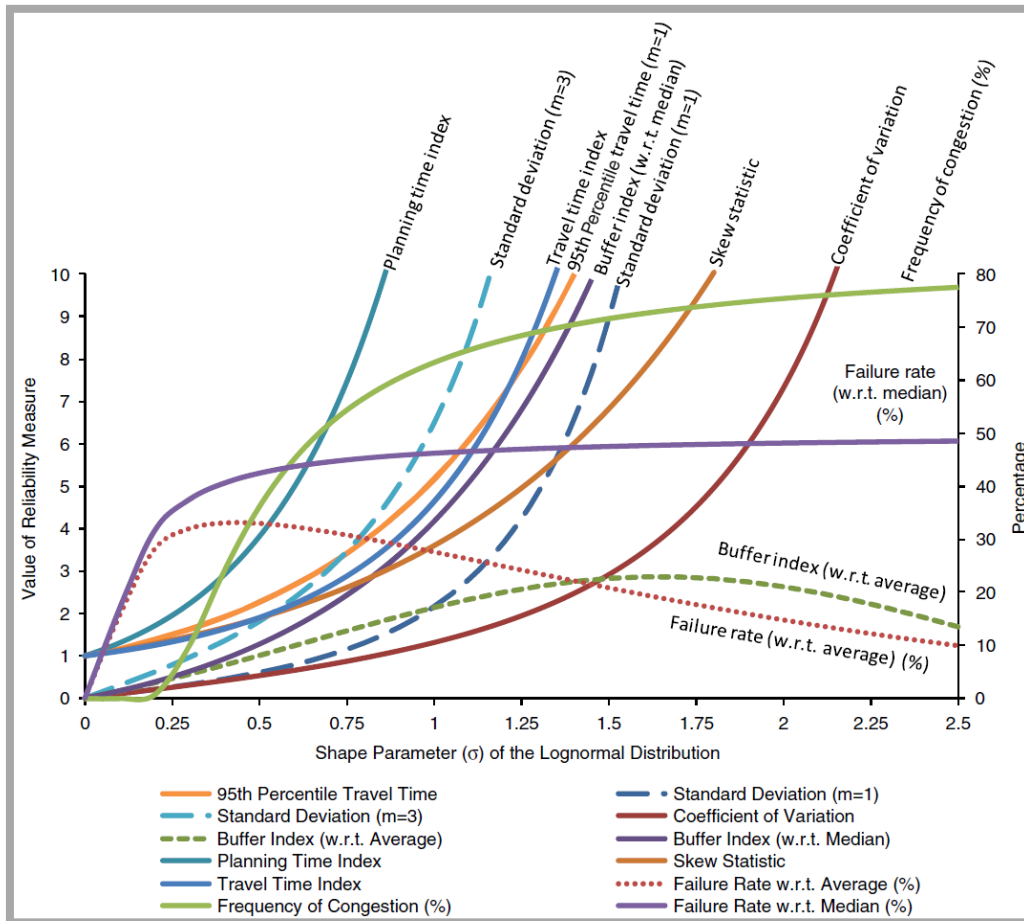
3.4.4.13 Exploring a single proxy measure of travel time reliability

Travel time reliability can be measured and calculated in many ways and from different perspectives. However, it would be more convenient and in some ways desirable, to find a single measure that would be representative or used as a proxy for other measures. Pu (2011) proposed that such a proxy should have at least the following three features:

- 1 There should be a well-defined traditional statistic so the proxy can be easily calculated with classical statistical methods (or software packages).
- 2 It should have the same varying direction (increasing or decreasing) as other measures when travel time distribution changes.
- 3 The magnitude relative to other measures should be rather stable.

Pu (2011) conducted analytical relationships between travel time reliability measures assuming lognormal distribution for travel time. Figure 3.15 presents travel time reliability measures as functions of shape parameter of standard lognormal distribution. Pu (2011) concluded that the coefficient of variation would serve as a good proxy for other examined measures. The standard deviation behaves similarly to the coefficient of variation regarding points a) and b) in the above list, but it does not meet c) because the scale parameter also plays a role in determining the value of the standard deviation. Pu (2011) also concluded that the average-based buffer index and failure rate (or percent on-time arrival) could potentially underestimate the unreliability when travel time distributions become heavily right-skewed (ie relatively large shape parameter σ value). These measures could even give unrealistic values if the shape parameter exceeds a certain value. To avoid that potential deficiency, the 'median-based buffer index and failure rate' are recommended by Pu (2011) instead. However, it should be noted that these recommendations were made in relation to road measures rather than PT.

Figure 3.15 Travel time reliability measures as functions of shape parameter of standard lognormal distribution (Source: Pu 2011)



3.5 Review of measures

As shown in table 3.8, the measures identified in the practice and literature review fell into four main categories: schedule adherence measures; statistical ranges; buffer time and tardy trip indicators. Table 3.8 provides an overview of the all measures reviewed and their associated advantages and disadvantages.

Originally it was envisaged that as part of this research project, a full separate framework would be used to evaluate the different measures to develop a shortlist of three preferred measures to further investigate. Instead, through agreement with the client, this preliminary shortlisting evaluation was undertaken through the review of literature which exposed advantages and disadvantages of different performance measures available. In particular, a key point for evaluating measures through the literature review was that an effective performance measure should, ideally, be applicable to users, easy to calculate, accurate and able to be clearly and consistently interpreted. Moreover, consideration of the measures' ability to meet the over-arching objectives of the research with regard to 'fitting' and being suited for intermodal and inter-regional aggregation was used for evaluation and shortlisting of measures.

Table 3.8 Summary of all measures reviewed

Indicator group	Indicators	Advantages/disadvantages
Schedule adherence	<ul style="list-style-type: none"> • Punctuality • Reliability • Lost customer hours 	<ul style="list-style-type: none"> • Punctuality measures are relatively simple for operators to understand but cannot be used to measure road network reliability (as private vehicles do not have scheduled departure and arrival times). • May incentivise slower speeds in timetables to increase schedule adherence performance. • Integration of customer component through compensation scheme (PTV 2016) and online webpage (MBTA 2016). • Lost customer hours take into account the passenger flow weighting for each line and time period, which places a relative importance on the critical, high-flow routes. Also, accounts for passengers' value of time, 'lost time' in stations (not just in-vehicle), and report on causes of lost customer hours.
Statistical ranges	<ul style="list-style-type: none"> • Standard deviation • Coefficient of variation • Percentiles • Skew statistic 	<ul style="list-style-type: none"> • Highlights extreme travel times. • Not sensitive to outliers. • Difficulties in aggregating from route level to regional level. • Outliers can be as a result of unforeseen incidents (crashes, weather) and hence not suited as an operational metric. However, this can be managed by using a lower level of percentile (ie 80th or 85th).
Buffer time	<ul style="list-style-type: none"> • Buffer index • Modified buffer index • Planning time index • Excess journey time • Excess waiting time 	<ul style="list-style-type: none"> • Customer oriented. • Relatively simple to understand. • Problem is defining the average trip that should be used as a benchmark. The average trip could be variable day to day. • There is flexibility in the formula in terms of which percentile to use. • High cost of data collection (excess waiting and journey time). • Flexibility in aggregation. • Excess waiting time is an operator-focused metric (does not take into account journey running time)
Tardy trip indicator	<ul style="list-style-type: none"> • On-time arrival • Misery Index 	<ul style="list-style-type: none"> • Useful as these are used in a variety of travel modes and similar to the punctuality factor in PT. • More research is required as to what is an acceptable threshold for 'lateness'. 10% has been used generically.

Although all PT organisations included in this review use some schedule adherence measures (deviation of actual time with schedule time by some pre-defined threshold) for PT performance, there is some variation in the measures, thresholds and terminology adopted. Some use more complex measures and some report the information in more understandable formats than others.

PTV uses service delivery and punctuality, while MBTA reports on reliability and Transport for London uses specific metrics depending on mode and frequency. As was found with the New Zealand measures, many of the international organisations reported use of some form of punctuality or reliability. These are primarily focused on the operator KPI perspective; however, both PTV and MBTA provide some means of customer engagement through the reporting of these measures. These include providing compensation for unreliable trips (PTV) and MBTA uses an online platform where customers can readily access this information. A key issue with schedule adherence measures, in the context of the present research, is that they cannot be 'fitted' to measuring road reliability, as private vehicle travel is not associated with any formal scheduling. One could argue that most drivers do have an intended time of arrival which one could compare as a 'schedule'; but given that transport management authorities lack quick and easy access to this personal information, it seems fair to say that schedule adherence measures are not appropriate for comparing predictability performance between PT and private vehicles.

Use of other, more sophisticated measures like Transport for London's 'lost customer hours' and AT's 'reliability' and 'delay' measures were less common (or less commonly made public perhaps). Transport for London used some buffer time type measures and has the greatest customer focus through the use of measures such as excess waiting time, excess journey time and lost customer hours, which even account for customers' value of time and perceptions of time for different aspects of their journeys. However, these measures are expensive to monitor because they comprise sophisticated data collection/monitoring including manual surveys. They are also intended for high-frequency services such as the underground although they could potentially be adapted and used for lower frequency services as well. Although these metrics are intended for customer purposes, they may not actually be that useful to travellers because they are not personalised and are typically aggregated at route level as opposed to being generated from origin to destination information.

In terms of academic literature, 'reliability' is the common terminology used to describe inconsistency, variability or predictability. The term 'predictability' was not in common use as a performance metric. Other key findings from the academic literature review included:

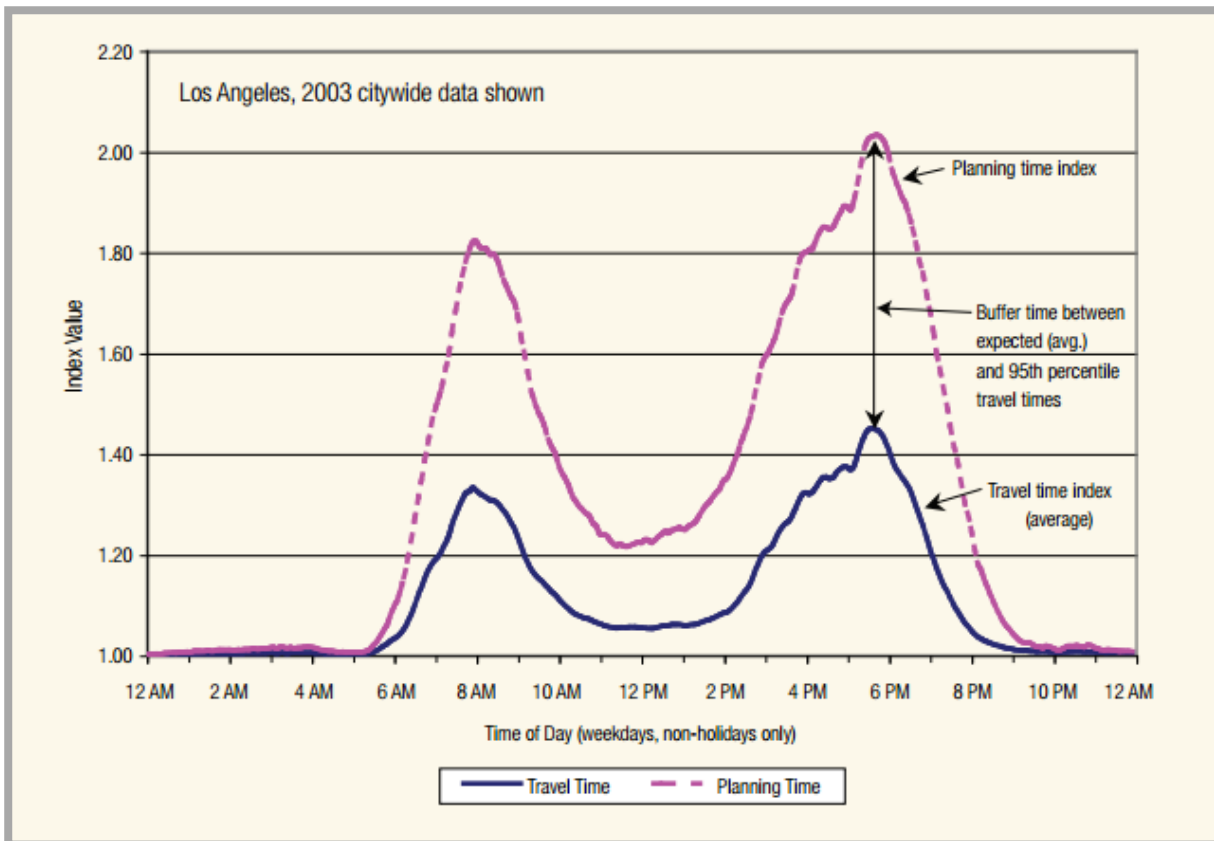
- The literature review did not reveal any reliability or predictability measures that were being applied to both PT services and private cars by one agency.
- PT reliability measures differ because they take into account passenger-related elements and schedules.
- Car travel time and PT running time can be compared using consistent measures; however, the literature shows PT services' users care more about out of vehicle time (eg waiting time).
- The literature review did not reveal any aggregation approaches that amalgamated the reliability of different PT modes.

The buffer time measures, in general, are popular in the United States among network planners, in operations, and are customer focused metrics that utilise the worst-case percentiles travel time to highlight the expected delay on top of 'normal' travel conditions. However, this 'normal' travel condition needs to be researched further with local context, application and outcomes in mind. A graphical depiction of the relationships between a buffer index, planning time index and average travel time is provided in figure 3.16.

In terms of evaluation of predictability measures, Mazloumi et al (2008) found no significant difference in outcome between different reliability and variability measures, but the variations in the performance of measures were due to the sensitivity of each measure to outliers. In particular, mean-based measures were more sensitive to outliers than median-based measures.

The literature review highlighted that particular attention should be paid toward defining thresholds for all the measures, eg lateness definitions vary from five minutes to three minutes, which can completely change the results. This matter was explored through sensitivity analysis as part of the data testing to assess modification potential of measures which is reported in chapter 4.

Figure 3.16 Reliability measures compared to average congestion measures (source: <http://mobility.tamu.edu/mmp/>)



3.5.1 Aggregation and 'fitting' to the Transport Agency road index

As noted above a key aspect of evaluating the measures available in order to create a shortlist of measures to trial using New Zealand PT data, was to evaluate the measures' ability to achieve two of the key overarching research objectives. These were:

To evaluate 'closeness of fit' to the existing road travel time measure used in New Zealand and assess the potential to modify preferred measures to 'fit' with the existing New Zealand measure, including the implications of modifications

and

To develop a method to aggregate comparable measures to create a national aggregate that will enable travel time predictability comparisons between modes and across regions.

In terms of 'fitting' measures to the existing road travel time measure used in New Zealand, chapter 4 examines a set of shortlisted measures alongside the Transport Agency index using real New Zealand PT data. In terms of shortlisting measures for this purpose, the practice component of the literature review found that the most commonly used PT predictability measures were based on schedule adherence, which as noted previously, relies on the presence of scheduled departure and arrival times. This type of measure

does not naturally closely 'fit' to the existing Transport Agency road index measure or other possible road measures as private vehicles on road networks do not have scheduled arrival and departure times. For this reason, one could not use schedule adherence as an aggregate measure to compare PT modes with private vehicles and the commonly used 'schedule adherence' measures are not recommended as a preferred measure. That said, schedule adherence measures could be used to aggregate and compare PT services across different PT modes.

Generally, the measures reviewed from the academic literature are better able to be used for aggregation and comparison of all modes as they largely did not rely on schedules for evaluation. That said, the literature review did not reveal any aggregation techniques between different modes of transport using a single measure. In fact, observations from the practice review indicated it is relatively common for a single PT authority to use different predictability measures for the different PT modes they manage. Aggregation through a *single* transport mode is common and regularly undertaken by PT agencies and roading authorities. Aggregation in this instance can be undertaken on a certain time period across all routes/lines or over a single route across all time periods. Some agencies such as Transport for London (2016b) use passenger flows as a weighting to ensure more critical high-demand routes have a higher weighting than others. This is also important when aggregating across different time periods where peak hours should have a high weighting than off peak periods.

After the literature review was completed, during a later stage of research (the validation testing workshops), it was revealed that AT now uses two performance measures that are similar to or consistent with the academic literature, and applies them to different transport modes including car and PT. Other terminology for these measures (or similar measures) are modified buffer index and planning time index, which AT refers to as delay and reliability. From a high level, it seems these measures can be used for comparison of predictability of performance between modes. Indeed, AT has reported using the measures for this purpose, although sometimes with adjustment in the actual measurements or thresholds used to fit the data. AT and other PT agencies' use of multiple measures demonstrates it may be more appropriate to use a combination of measures to provide a comprehensive understanding of route performance.

In aggregating data, selecting measures to apply to the data is one important aspect – another issue is preparing data for aggregate analysis. There may be issues with inconsistent data between modes. For example, it may be difficult to get spatial consistency between origin and destinations across all routes. Key routes covering car travel times through Bluetooth detector points may not necessarily align with key bus, train, or especially ferry, routes. Also, data standards between regions are likely to be of different formats. Therefore, any aggregation also needs to consider the limitations of the available data. This matter is explored further in the next section which documents shortlisted measures being applied to real data.

3.5.2 Shortlisted measures

From the full review of literature and practice review, including weighing up the advantages and disadvantages of all the measures reviewed it was decided that the preferred measures to further consider were:

- buffer index
- modified buffer index
- planning time index

These measures were then tested for modification, along with some of the existing, commonly deployed measures (Transport Agency road index and some punctuality measures) as described in the next section.

Another measure that would be interesting to try applying is Transport for London's measure of lost customer hours which could be adapted for use to compare modes as it largely relies on absolute travel times. However, this customer-centric measure is more time intensive to calculate and would require strategic thinking to adjust for different modes. It was not selected because it would likely be relatively time intensive and possibly too complicated for authorities to calculate.

4 Assess modification potential of measures using New Zealand PT data

The results of the literature review were presented to the study's steering group. It was agreed between the research group and the steering group that the shortlisted measures should be 'tested' using real PT data in New Zealand. Moreover, it was decided that to best understand the implications of the data testing, the shortlisted measures should be applied to the data along with the measures currently used for measuring road predictability (the Transport Agency road index) and PT reliability ('punctuality' and 'early' and 'late arrival'). The latter measures were solely included as a frame of reference to examine the shortlisted measures as they are currently commonly used to evaluate PT reliability in New Zealand and overseas. As outlined in chapter 3, this work involved four key stages: obtaining data, preparing data, applying the measures to the data and examining the results, then undertaking threshold sensitivity testing.

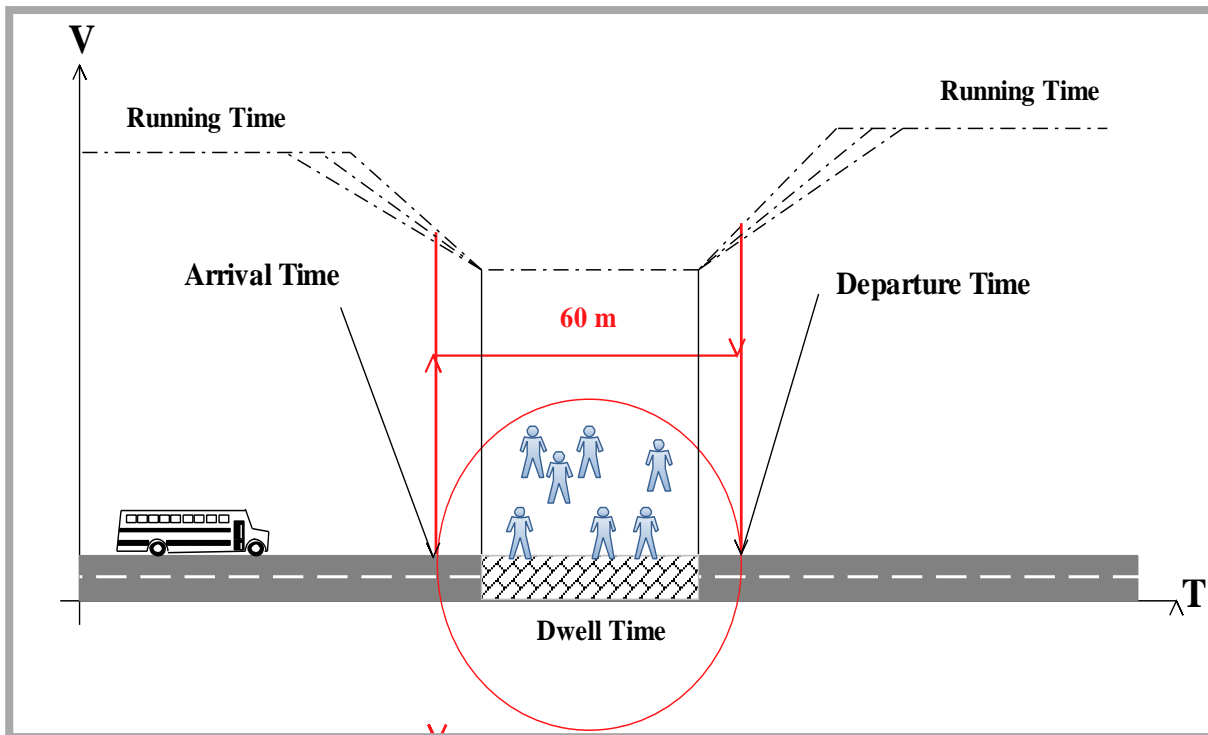
One point worth considering in relation to analysing PT data is that schedules tend to already include buffers that account for unexpected delays. This may have the effect of reducing variability in the data, improving reliability performance.

4.1 Preparing data: analytics, error detection and cleansing

In order to undertake the data testing, PT datasets were obtained from the three biggest cities in New Zealand: Auckland, Wellington and Christchurch. This involved contacting the responsible PT authority in each city and providing a detailed scope about the nature of the data required, including specific routes, time periods and ideal data formats. With negotiation, four key datasets for March 2016 were obtained: bus data for routes in Auckland, Wellington and Christchurch; and rail information for Wellington. No ferry data was provided. As is typically the case with data, the data was not in a format completely ready for analysis so some effort was required to prepare the data. This was deemed imperative in ensuring that the outputs were robust and meaningful. Specifically, this involved data analytics, error detection and cleansing. In order to 'clean' data, it may be helpful to first gain an understanding of how AVL data monitoring currently takes place.

Auckland has a signpost type of AVL system. Each signpost has its own unique code, which detects the buses within its range of influence. Through a transmitter, the system records the location of buses, time and vehicle identifier's number and sends the information to the dispatcher. Eventually, the collected information will enable operators to provide their customers with up-to-the-minute information of bus locations. Figure 4.1 illustrates how the system works. The AVL system records an arrival time as a bus enters the circle of influence of the AVL signpost and records the respective departure time as the bus leaves the circle of influence. The estimated values of 'bus dwell time' from AVL data represents the time spent by the buses while they are within a predefined circle of influence of the AVL signpost (usually 60 metres in diameter) at a bus stop. This system records the departure and arrival time of buses at different locations.

Figure 4.1 AVL system's working procedure to detect buses at a bus stop



As noted by Rashidi (2014), there are many instances where the AVL system cannot detect the buses or detects them by mistake. Some of these issues include:

- zero recordings of arrival and departure time
- duplicate recording of bus identification (ID)
- unreasonable arrival or departure records (eg departure times are less than arrival times)
- wrong stop ID records
- outliers (unreasonably small or large records).

There are two potential solutions to address these problems: the conventional method is removing the problematic observations (Furth et al 2006), and the second method is replacing those observations by imputation techniques. While some research suggests that imputation provides for more robust analysis, it is very time intensive so listwise deletion was instead used in this research.

4.1.1 Missing data treatment

Almost all the techniques that deal with missing values in a dataset (including listwise deletion) assume that the missing observations should not be dependent on other variables (SPSS 2011b). Therefore, it is important to investigate the pattern of missing data in advance. Different patterns of missing observations are: missing completely at random, missing at random and missing not at random (Saunders et al 2006; Streiner 2002). Missing completely at random indicates 'missingness' of a variable is not related to any other variables. The second and more common pattern - missing at random - means the omitted observation can be related to at least one variable but not to outcome variables. Missing not at random implies the presence of a systematic pattern, and means the missingness is related to one or more outcome variables. However, a different imputation technique can be safely tried as the AVL data abnormalities are due to technical issues and have not occurred systematically.

4.1.1.1 Listwise deletion method

In the listwise method, the entire row of the problematic observation is deleted. Despite its universality and simplicity, the listwise method will sacrifice a large amount of data and can cause a bias in correlation and significance of variables (Tsikriktsis 2005).

4.1.2 Outliers and extreme values identification:

Distribution analysis of bus travel time data shows the travel times are not normally distributed (Pu 2011). The modified z-score test is one of the recommended distribution free approaches to identify the presence of unusual data, as such data can disrupt the mean and variance of the original dataset (Rashidi et al 2014). Z-score value can be computed as follows:

$$z = \frac{(Y_t - \bar{Y})}{MAD} \quad (\text{Equation 4.1})$$

where, Y_t is the observed value at time t , \bar{Y} is the estimated value and MAD is the median value of the absolute deviation, which can be computed as follows:

$$MAD = \text{median}(Y_t - m) \quad (\text{Equation 4.2})$$

where m is the median value of the sample. For a normally distributed data sample, an observation can be labelled as an extreme value or outlier if the modified z-score is greater than 3.5. It is also important to pay attention to the minimum required sample size, and at least one year of data should be used to obtain a more stable value of reliability metrics (Patricio et al 2012).

4.2 Applying the measures

Once the data was 'cleaned', the shortlisted measures: buffer index, modified buffer index and planning index, along with the Transport Agency road index, punctuality, and early and late measures were applied to the data.

In this section, we provide a 'refresher' on what each of these measures are, how they were selected for testing and document the aggregation techniques used to calculate the results.

4.2.1 Shortlisted predictability measures

It is almost impossible to find a single measure which adequately describes the reliability or predictability of a journey. Different measures explain different aspects of journey consistency. A strong linear relationship between different measures indicates they can be used interchangeably. Therefore, in line with the literature review, our recommendation is to use a combination of measures to provide a comprehensive understanding of a route's performance.

To test this hypothesis and evaluate the performance of measures, data testing was undertaken for the shortlisted measures noted above. To assist in meaningfully interpreting the results of this data testing, a brief description of each of the measures applied is provided in table 4.1 below. A summary of the thresholds for each measure applied is provided in table 4.2. Note that for road-based measures, the thresholds applied in this stage of data analysis have not been adjusted.

Table 4.1 Shortlisted measures used for data testing and thresholds applied

Measure	Description
Transport Agency road index	This methodology calculates the proportion of journeys completed more quickly than a 'buffer time'. The buffer time is average travel time plus a fixed proportion of the average. The more journeys completed in less than the buffer time, the more predictable journeys are deemed to be.
Buffer index	<p>The buffer index represents the extra time (or time cushion) that travellers must add to their average travel time when planning trips to ensure on-time arrival. For example, a buffer index of 40% means for a trip that usually takes 20 minutes, a traveller should budget an additional 8 minutes ($20 \text{ min} \times 0.40 = 8 \text{ min}$) to ensure on-time arrival most of the time (95% of occasions).</p> <p>Green: reliable 0%–50% Yellow: moderately reliable 50%–75% Red: unreliable >75</p>
Modified buffer index	<p>Extreme values are included when calculating average travel time and can inflate it. Replacing the average travel time with median travel time removes this sensitivity toward extreme observations. Similar to the buffer index, the modified buffer index represents the extra time (or time cushion) that travellers must add to their normal travel time when planning trips to ensure on-time arrival. For example, a modified buffer index of 40% means that for a trip that normally takes 20 minutes, a traveller should budget an additional 8 minutes ($20 \text{ min} \times 0.40 = 8 \text{ min}$) to ensure on-time arrival most of the time (85 or 95% of occasions).</p> <p>Green: reliable 0%–50% Yellow: moderately reliable 50%–75% Red: unreliable >75</p>
Planning index	<p>The planning index represents the total travel time that should be planned when an adequate buffer time is included. The planning time index differs from the buffer index in that it includes typical delay as well as unexpected delay. Thus, the planning time index compares near-worst case travel time with a travel time in light or free-flow traffic. For example, a planning time index of 1.60 means, for a 15-minute trip in light traffic, the total time that should be planned for the trip is 24 minutes ($15 \text{ minutes} \times 1.60 = 24 \text{ minutes}$).</p> <p>Green: reliable 1–1.5 Yellow: moderately reliable 1.5–2 Red: unreliable >2</p>

Table 4.2 Threshold values used (for the three shortlisted measures, these come from private vehicles)

Colour code	NZ Transport Agency road index	Buffer index	Modified buffer index	Planning index
Reliable	>80%	<50%	<50%	<1.5
Moderately reliable	70–80%	50–75%	50–75%	1.5–2
Unreliable	<70%	>75%	>75%	>2

4.2.2 Aggregation technique

The data was analysed at two levels namely route and citywide. At route level aggregation, we analysed travel time data for the entire route from first to the last stop for each corridor. At city level aggregation,

we took mean and median of all routes/corridors selected based on patronage and frequency. The results of the latter analysis follow.

4.2.3 Aggregated results - bus network

An overview of the composite results for all bus routes in each of the three case study cities is provided in tables 4.3 and 4.4 which show the results using mean and median, respectively. Mean is a good measure of central tendency if the data is symmetrically distributed, and it can be affected by extreme observations. The median, on the other hand is the preferred method when the data is not symmetrically distributed and there are rare events or extreme observations in the data set. We have used both mean and median to test the impact of this issue. If the mean and median are significantly different, further investigation is recommended.

In tables 4.3 and 4.4 colour coding of the results is presented based on the performance against the thresholds presented previously. Green indicates 'reliable' predictability performance, amber 'moderately reliable' performance and red 'unreliable' performance. It is clear that under the Transport Agency road index, the buffer index, modified buffer index and planning index there was little differentiation in performance between cities. Conversely, using the schedule adherence measures, there was more variation with Auckland generally performing better than the other two cities.

Table 4.3 Composite results for all bus routes in each city based on means

Cities	Travel time	NZTA Index	Buffer Index	Modified BI	Planning Index	Punctuality	Early	Late
Auckland	33.5	81%	13%	10%	1.29	58%	45%	23%
Wellington	49.8	82%	11%	9%	1.25	31%	56%	18%
Christchurch	72.8	82%	10%	8%	1.20	37%	33%	48%

Table 4.4 Composite results for all bus routes in each city based on medians

Cities	Travel time	NZTA Index	Buffer Index	Modified BI	Planning Index	Punctuality	Early	Late
Auckland	29.0	81%	14%	10%	1.29	58%	46%	26%
Wellington	45.4	81%	11%	9%	1.26	32%	61%	9%
Christchurch	69.2	82%	11%	7%	1.20	38%	26%	54%

4.2.4 Aggregated results - train network

Tables 4.5 and 4.6 show the different performance results for Wellington's train network based on mean and median, respectively. In this instance use of median rather than mean showed big differences between different indices. For example, for probability of late train arrival there is a noticeable difference between mean and median. Further investigation revealed there is almost no late arrival record (0%), except three routes, which caused the mean of late arrival to be 5%.

Table 4.5 Composite results for Wellington's train network based on mean

Cities	Travel time	NZTA Index	Buffer Index	Modified BI	Planning Index	Punctuality	Early	Late
Wellington	28.1	85%	8%	8%	1.18	33%	47%	5%

Table 4.6 Composite results for Wellington's train network based on median

Cities	Travel time	NZTA Index	Buffer Index	Modified BI	Planning Index	Punctuality	Early	Late
Wellington	22.6	85%	8%	8%	1.16	29%	51%	0%

4.2.5 Observations

Some of the key observations from this high-level, comparative assessment of the networks include:

- Applying the Transport Agency road index yields almost the same results across all three cities.
- The punctuality-based measures reveal that Auckland's buses are performing the best, followed by Christchurch and then Wellington.
- Some variability can be observed in the percentile based measures, ie the buffer index, modified buffer index and planning time index, which shows that the Christchurch and Wellington bus networks are performing better than Auckland's.
- Wellington's train network is performing better than the bus network for almost all measures. (Data for Auckland's train network was not provided in a suitable format for analysis.)
- The Transport Agency road index '15-minute time interval' calculation can only be applied to high frequency routes. This could be addressed by changing the minimum time interval to 30 minutes. Lower frequency routes would continue to present challenges.
- Punctuality-based measures for complete trips including early and late arrival based on schedule arrival time at the last bus stop may lead to incorrect interpretation of trip reliability for travellers boarding after the start point or alighting before the end point.
- All the proposed nationally and internationally practised methods are sensitive to the thresholds used to evaluate performance. For this reason, some sensitivity analysis of threshold testing was also undertaken, as will be discussed.

4.3 Threshold testing

The results were then critically examined and it became apparent that directly applying road-based measures to the PT data worked. However, similar to road-based measures these indices are sensitive to changes in thresholds. To better understand this issue, some sensitivity analysis, or exploration of adjusting thresholds, was then undertaken as part of this research stage. Table 4.7 shows the proposed arbitrary thresholds used for sensitivity analysis.

Table 4.7 Arbitrary thresholds used for sensitivity testing

Colour code	NZTA Index	Buffer Index	Modified BI	Planning Index	Punctuality	Early	Late
Reliable	>90%	<7%	<6%	<1.2	>95%	<23%	<26%
Moderately reliable	70%-90%	7%-21%	6%-16%	1.2-1.4	85%-95%	23%-42%	26%-52%
Unreliable	<70%	>21%	>16%	>1.4	<85%	>42%	>52%

Tables 4.8 and 4.9 demonstrate how changing threshold values affect the reliability scores for bus routes compared with tables 4.3 and 4.4. Bus reliability for all cities changed to 'moderately reliable' from the previous 'reliable' status.

Table 4.8 Thresholds changing effect on mean performance of buses in each city

Cities	Travel time	NZTA Index	Buffer Index	Modified BI	Planning Index	Punctuality	Early	Late
Auckland	33.5	81%	13%	10%	1.29	58%	45%	23%
Wellington	49.8	82%	11%	9%	1.25	31%	56%	18%
Christchurch	72.8	82%	10%	8%	1.20	37%	33%	48%

Table 4.9 Thresholds changing effect on median performance of buses in each city

Cities	Travel time	NZTA Index	Buffer Index	Modified BI	Planning Index	Punctuality	Early	Late
Auckland	29.0	81%	14%	10%	1.29	58%	46%	26%
Wellington	45.4	81%	11%	9%	1.26	32%	61%	9%
Christchurch	69.2	82%	11%	7%	1.20	38%	26%	54%

The effect of changing threshold values was also investigated for the train network in Wellington. Tables 4.10 and 4.11 show how these changes affect train network reliability. It can be observed that, with the exception of the modified buffer index, the reliability results for train networks change to moderately reliable, which confirms all measures rely on a threshold value.

Table 4.10 Thresholds changing effect on mean performance of Wellington train network

Cities	Travel time	NZTA Index	Buffer Index	Modified BI	Planning Index	Punctuality	Early	Late
Wellington	28.1	85%	8%	8%	1.18	33%	47%	5%

Table 4.11 Thresholds changing effect on median performance of Wellington train network

Cities	Travel time	NZTA Index	Buffer Index	Modified BI	Planning Index	Punctuality	Early	Late
Wellington	22.6	85%	8%	8%	1.16	29%	51%	0%

The implication of the threshold testing is simply that performance results are sensitive to the thresholds used. When the road-based thresholds are directly applied for PT, there is little variation in performance, indicating some work may be necessary to identify optimal thresholds which provide an accurate portrayal of different PT predictability performance.

4.4 Overall lessons from data testing

4.4.1 Data issues – obtaining data

One challenge to this project was obtaining the data and processing it into a suitable format to analyse. This is important to consider in developing strategies for ongoing performance assessments. There was some reluctance by transport authorities to release the data which also delayed the process and in some instances prevented analysis of data as part of this project entirely (eg ferry data and Auckland's train data).

The availability, quality and format of data varied considerably across modes of transport and the three cities. Considerable resource was needed to both source and 'clean' data before it could be used in the analysis. The following data limitations compromised the comprehensiveness of the baseline analysis:

- Wellington bus patronage data by route was unavailable due to concerns about commercial sensitivity; however, we were able to obtain an aggregated dataset.
- AT was unable to supply raw data for trains. Summary data was provided, but was not suitable for analysis.
- Christchurch bus data was provided but it was based on checkpoints. First stop, last stop and middle stop data was not available.
- Route length was provided for Auckland and has been included in the detailed tables of this report; it was used to calculate average speeds. This information was not available for Wellington and Christchurch but the focus of this report is predictability and speed is an efficiency measure.
- Ferry data was not analysed for any locations; Auckland and Wellington's ferry data is not available in the same format as bus and train data because it is collected manually. Since ferries do not experience the same predictability issues as buses (in particular) because timetables are mainly affected by the weather, driver shortages and occasional breakdowns, excluding ferries was not considered to have a material effect on the conclusions of this study.

4.4.2 Preparing data

As noted, data quantity (sample size) and completeness is imperative to reliable and robust results. Although a particular data format and template were provided, data quality differed between cities. AT's data was the 'best' data, in that it was the most ready for analysis. This was followed by Wellington, and Christchurch's data presented the most challenges to 'clean'. Each of these datasets was different and presented unique challenges for monitoring and analysis. This made the calculation of travel times and other measures inconsistent across different regions.

The datasets received had large omissions, eg missing observations for routes. Where data was missing or insufficient, routes were excluded from the analysis, which is a problem for anyone drilling down to understand the reliability of a particular route. Each of the datasets had some challenges which required some work in data preparation:

- The Christchurch dataset did not contain actual arrival or departure times for each bus stop. The only information provided was for particular timepoints, which had been arranged in a very difficult format to code and actually differed from the scheduled timepoints, making analysis challenging.
- The challenge with the Wellington data was that it was unclear whether the 'actual time' provided was arrival or departure from the bus stop or timepoint.
- Auckland data set had missing observations and there were some mistakes in the data for schedule adherence calculations (this index is the one which is used to define the punctuality and performance of the routes).

Despite these issues, when looking at the datasets in aggregate, sufficient information was available to compare PT data and private car data. From the experience of this data testing, for future predictability analysis it is recommended that the same cleaning algorithm be used for an entire dataset before feeding the data into the analysis software.

The challenge of accessing and using PT data in this project has provided insights into the possibilities and barriers for systematically monitoring PT performance. Future calculations would benefit from standardising the data format across entire regions and perhaps automating this data feed to the Transport Agency.

4.4.3 Conclusions

In this research, bus and train in-vehicle travel time data was used to investigate different reliability measures based on data collected in Auckland, Wellington and Christchurch. Different measures including the Transport Agency road index, buffer index, modified buffer index and planning index were used in the AM peak period from 7am to 9am for March 2016. Findings of this stage of the research can be outlined as follows:

- In this research phase the Transport Agency road index, buffer index, modified buffer index and planning index were used for in-vehicle bus travel time reliability comparison. Punctuality-based measures are not discussed in detail as they are not comparable with car travel time, but were included in the tables to provide a frame of reference for the performance of measures.
- It is almost impossible to find a single measure that adequately describes the reliability or predictability of a journey. Different measures explain different aspects of journey consistency. Strong linear relationships between different measures indicate they can be used interchangeably. Therefore, our recommendation is to use a combination of measures to provide a comprehensive understanding of route performance.

- Differences between measures are due to their mathematical formulation and their sensitivity toward extreme observations and rare events – notably whether they are mean or median-based measures.
- The buffer index should not be used as a standalone measure, but accompanied by a modified buffer index as the latter is not sensitive to abnormality in the data due to its reliance on median.
- These indices can be used by both passengers and decision makers. Users want to know ‘How much more time do they require over the free flow condition’, hence the thresholds typically represent the ranges when the user is happy, leading the road user to determine when the trip is unreliable.
- Another issue to consider is ‘How are the thresholds determined?’ Each measure should have clear logic to define thresholds as the usefulness of the measure relies on setting the right threshold. There are some defined thresholds for car travel time reliability in the literature; however, to the best of the authors’ knowledge there is no defined threshold for in-vehicle bus travel time reliability. Further investigation is required to propose thresholds for PT with statistical and engineering acceptance.

5 Validation testing – workshops

Following the previous stages of the research, three workshops were organised to discuss the preferred measures with key stakeholders. Originally in the proposal, this construct validity ‘testing’ was envisaged to be undertaken with PT customers; however, as the project progressed it became clear that in most jurisdictions, PT operators and management agencies are the actual users of the measures. Thus, these groups were targeted as the key stakeholders. This chapter documents three workshops undertaken with these key stakeholders; however, it is also worth noting that validation was also undertaken through a review of the literature review, a variation report documenting the data testing to assess modification potential, a peer review and a steering group presentation.

The review of the variation report and steering group also provided validation information.

One workshop was held in Wellington and another in Auckland. A third workshop was held with Environment Canterbury via teleconference (with a PowerPoint sent ahead of time) rather than an in-person workshop.

The workshops aimed to present the research and address the following key questions with stakeholders:

- whether the measures made sense to stakeholders (construct validity)
- if stakeholders thought the measure(s) were useful
- what they foresaw to be the implications of adopting the measure(s).

5.1 Participants

A range of stakeholders to whom a PT predictability measure could be relevant was invited to the workshops. This primarily included regional council staff who calculate and use performance measure data and PT operators from across modes (bus, rail and ferry). A summary of the attendees for each workshop is provided below in table 5.1.

Table 5.1 Summary of workshop attendees

Workshop	Attendees
Auckland	<ul style="list-style-type: none"> • 2 representatives from AT • 3 representatives from AT Metro • 2 representatives from the rail operator • 1 ferry operator • 3 bus operators
Wellington	<ul style="list-style-type: none"> • 3 representatives from GWRC • 1 representative from the Bus & Coach Association • 1 data analyst from the Ministry of Transport
Christchurch	<ul style="list-style-type: none"> • 3 representatives from Environment Canterbury

In Auckland, representatives from AT were from the network performance and operational planning team and had strong experience in undertaking detailed data analysis. Three managers from AT Metro also attended who were responsible for operations improvements, support and service performance. In addition, operators from all three modes of PT were in attendance.

The Wellington workshop had three representatives from GWRC including one responsible for customer experience, one transport analyst and modeller, and one PT planner who worked in timetabling and route

planning. In addition, the policy manager from the Bus & Coach Association which represents bus operators from around the country attended as well as a statistical analyst from the Ministry of Transport.

The teleconference with Environment Canterbury (ECan) included a business systems analyst, a senior manager of PT and the manager of PT strategy, planning and marketing. ECan offered an interesting perspective with their responsibility for PT services in Christchurch and also regional centre Timaru.

Each workshop was facilitated by two members of the Opus research team: a social scientist who was the lead facilitator and a data analyst who was an expert on the most technical aspects of the research.

5.2 Structure of the workshops

The Wellington and Auckland workshops were 2.5 hours in length and were designed to be interactive. A slightly abbreviated, but still interactive, teleconference version was held for Christchurch.

The structure of the workshops was as follows:

- personal introductions
- introduction to research
- short discussion – predictability/reliability in the relevant region/role
- tea break (Auckland and Wellington only)
- research outcomes (including review of international measures, data testing, evaluation of measures)
- feedback on research findings
- wrap up.

The first workshop held was in Wellington, the next was in Auckland and finally the teleconference with ECan was held. There was some slight iteration between workshops to try and optimise the later workshops based on the previous workshop experiences.

5.3 Outcomes from the workshops

There was a range of outcomes from the workshops. From a high-level perspective, the smaller number of participants in the Wellington and Christchurch sessions allowed for more focused discussions and specific outcomes. The Auckland workshop was relatively more challenging. It was difficult to focus participants' attention on the understanding and evaluation of the measures.

5.3.1 How PT predictability is relevant to stakeholders

Early in the workshops attendees were asked how predictability is important to what they do. The responses are summarised in table 5.2 below.

Table 5.2 What does public transport predictability mean in your role?

Role	What PT predictability means in my role
Public transport operators	<ul style="list-style-type: none"> • KPIs included in contracts are relevant and closely monitored (usually just punctuality measures). There are often financial implications. • 'Punctuality is our commitment to customers'. • The way ferries are judged is different from other modes. According to the ferry representative, ferries are now being judged in relation to the timing of the next service. • Train operators discussed how high-level measures (which can be broken down by different

Role	What PT predictability means in my role
	<p>time periods) have data on segments of travel at stations. Performance is used to build next round of timetables.</p> <ul style="list-style-type: none"> Predictability is relevant under the new PTOM for its effect on KPIs in contracts.
<p>Regional councils (GWRC and ECan) and AT/ AT Metro</p>	<p>Predictability measures are important for various timeframes of reporting: monthly, real time, annual etc.</p> <p>Questions were raised about the timeframe this research was targeting.</p> <p>For those responsible for customer experience:</p> <ul style="list-style-type: none"> Large share of customer complaints are associated with service reliability (late running or failed to run) Predictability measures are important from a qualitative perspective. RTI predictability is a key touch point, with customers in Wellington losing faith in the RTI system. Any improvements to RTI need to be future-proofed. Current issues with geo-fencing, bus bunching and drivers occasionally not turning on tracking – but customers do not understand these issues and instead have poor trust in the RTI system’s accuracy. Customer expectations are increasing with technological developments. To increase patronage – optimal to have real-time comparative travel time information across modes. Uber setting a customer experience threshold with transparency of travel time and money. Evolving predictability measures from a customer experience perspective need to consider the move toward mobility as a service (MaaS). There has been a shift in how customers plan journeys from looking at scheduled departure times (which are more relevant for a schedule adherence measure) to just showing up and looking at RTI. Customer experience is paramount and customers’ experiences are based on operational performance. Discussion about how algorithms that underlie RTI are based on both current conditions and also historical performance (AT). <p>From a network planning perspective:</p> <ul style="list-style-type: none"> Spitting out numbers, importance to strategic project (like ‘Let’s Get Welly Moving’) particularly through travel time averages (absolute times) and variability. For timetabling and service planning – use reliability information for comparison and prioritise reviews. However, issues with predictability-related measures do not always reflect a timetabling issue – so then look at percentiles. Timetabling strongly seeks to avoid early arriving (big issue for customers). Compare RTI and timetables. Look at some reliability measures operators will not even know about. AT punctuality is examined in terms of first stop, last stop, timing points. Running times – tracking used for timetabling and to understand variability, length of trips. Have an in-depth understanding of system’s operations. Look at timings dynamically.
<p>Ministry of Transport (MoT)</p>	<p>From a national, strategic perspective the MoT representative noted:</p> <ul style="list-style-type: none"> There could be potential issues with any big changes to measures used. The importance of taking into account increasing vehicle kilometres travelled, congestion and the impacts of these increases on predictability. Costs to customers of late or unpredictable services – personal (even emotional) and economic costs.

In the Auckland workshop, operators (and others) also discussed how the new PTOM is very regimented to punctuality so it is not necessarily providing the best outcomes for customers. In fact there have been many complaints about how services are being held to meet punctuality targets (that is, some services

may run slower to be more in line with timetables). GWRC also discussed how it builds buffers into timetables and believes it is better, from a customer perspective, for buses to leave a stop late than early. ECan noted that in some locations they do not want buses to wait (eg the hospital, where space is restricted) so the buses always have to be 'late'.

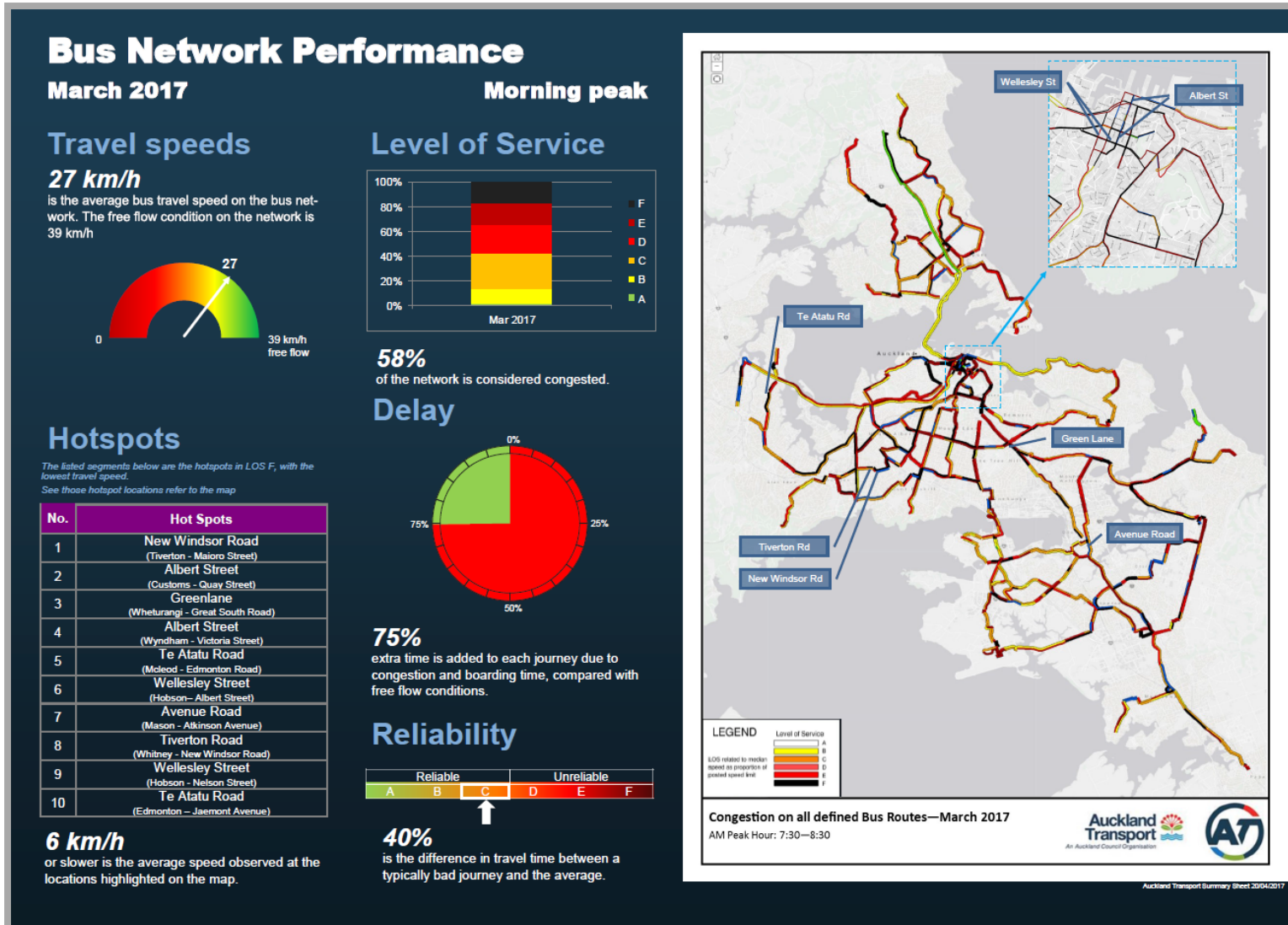
Interestingly ECan reported being dissatisfied with how predictability is currently measured in its area. When asked why, it said it 'tells a bad story' because it is currently measured by the occurrence of trips that are more than one minute late at several timing points. ECan suspects this is unnecessarily rigorous and beyond what customers reasonably expect. For this reason, ECan was particularly interested in our research and whether there might be better ways to report on predictability. In addition, ECan has a unique perspective on reliability in the earthquake recovery context. In organising the phone workshop, it was explained:

The ongoing earthquake repair work in Christchurch has meant that the reliability of our network continues to be compromised in places, and that we aren't always able to provide the level of service to our customers that we would like. It has also led us to question how we assess reliability, and whether our current measures tell the full story.

AT described its methods for undertaking predictability monitoring. It monitors predictability across modes (eg road network, bus network, pedestrian crossings, freight routes) and uses a variety of metrics. An example of one of its 'dashboards' for the bus network is provided in figure 5.1. Some of the measures used align with the shortlisted preferred measures presented at the workshop, albeit at times under a different name. AT suggested that the names it uses are more intuitive and may be preferable to use than the names used in academia. AT's communications with customers talk about reliability as 'worst to average travel speed', based on 85th percentiles. The level of service assessments is primarily based on average speeds and AT felt strongly that any predictability measure should include speed.

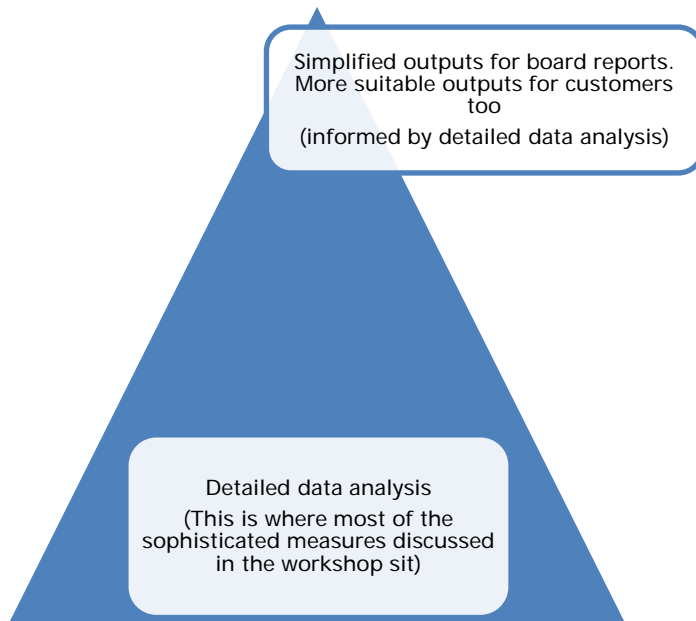
In addition to the bus network dashboard provided in figure 5.1, AT reported undertaking similar analysis for other modes using the same, or iterations of the, same measures. Examples of these for pedestrians, freight and road travel are provided in appendix A.

Figure 5.1 An example of AT's bus network performance dashboard



Workshop participants had varying levels of experience with the types of measures presented. Two of the representatives from AT's transport operations team were very familiar with the detailed data analysis undertaken for network planning and performance and explained how there is a 'pyramid' associated with reporting with detailed data analysis at the bottom of the pyramid and easy-to-understand simplified outputs toward the top which are typically presented in board reports and other forms for a more general audience. This is depicted in figure 5.2 below.

Figure 5.2 The performance measuring pyramid described by one stakeholder



5.3.2 Data sources for measuring predictability

The workshop presentation noted that main data sources used to measure predictability in New Zealand are AVL and RTI and asked stakeholders if there were other data sources. The ferry operator noted use of automatic identification system (AIS) which uses GPS-based broadcasting for marine traffic. Meanwhile the train operator noted they use balises for tracking; a balise is an electronic beacon or transponder located between the rails of a rail track. Ticketing was also mentioned as a data source. In the Wellington workshop, emerging data sources such as Bluetooth and GPS tracking on phones were mentioned.

Some stakeholders gave specific descriptions of data used to measure predictability. In the Auckland workshop, one of the bus operators reported having three to four AVLs on every bus from which they receive data points about every 10 seconds. This operator also mentioned they can compare RTI feeds with ticketing information to improve accuracy. They use these data sources to track vehicles to monitor their performance in meeting their contractual KPIs. Meanwhile the ferry operator also noted the main predictability measures they are concerned with are the KPIs from contracts. In terms of tracking their ferry services, the skippers usually record running times, though what time exactly the ferries depart from docks can be difficult to pinpoint because departing requires multiple steps. For this ferry sector, AT Hop data does not provide a departure time data source because of where the card readers are located – many passengers will swipe and then be waiting for some time before boarding a ferry and before the ferry departs.

5.3.2.1 Major disruptions

In the Auckland workshop, there was a brief discussion about some of the major disruptions that can affect reliability and how these might vary by mode. Some of the operators discussed how they predict disruptions based on historical disruptions. There was some sentiment that these sorts of disruptions needed to be considered in developing a preferred measure. Some of the points about disruptions include:

- In real time, customer communication about disruptions is a priority.
- Currently the quality of information available to customers varies by mode. It was acknowledged that bus users are particularly disadvantaged and there is a need to improve disruption communication for bus users.
- There are different kinds of disruption: unexpected vs planned disruptions (and there may be a need to classify further).
- There are different ways of looking at disruptions (real-time system vs annual reporting).
- Thresholds vary by sources of disruptions.

5.3.2.2 Data quality issues

In the workshops, it was noted by both facilitators and attendees that there can be limitations and issues with data such as missing values, inaccuracies associated with human error and potentially issues accessing data. Auckland workshop attendees noted the importance of using consistent data sources. The Wellington workshop felt that data quality issues might affect thresholds and/or measures adopted.

5.3.3 Overarching objectives of the research

Early in the workshops it was explained that the two overarching objectives of the research were:

- Can we produce a measure of predictable journeys for PT: bus, rail and ferry that can be compared to a similar measure that the Transport Agency uses for measuring the travel predictability by road?
- Can we aggregate measure across modes and across regions?

5.3.3.1 Key considerations

Participants identified several matters they felt should be considered in selecting a preferred PT predictability measure. These included:

- Geography – notably topography, populations and urban form
- PT customers' mindsets – it was noted that PT users have different mindsets to private car travellers. An example was provided that PT users might be more likely to expect delay. Another participant countered, however, that there is variability across different PT users by route and mode.
- Some stated the best measure would be one with an output relevant to customers.
- The question was raised of whether these measures are becoming increasingly irrelevant because of Google Maps and the move to MaaS, both of which provide customers with personalised options (and presumably inadvertently provide some intelligence to network planners).
- The AT analysts said a good measure needs to consider speed. They believed speed and trip duration need to be equally represented in a preferred measure.

Many of the attendees were largely focused on the value of the measures to customers, rather than to inform network operations. In all the workshops, the facilitator made an effort to re-focus the discussion around network planning objectives rather than thinking of the measures as being for customers. It was

acknowledged that measures could achieve both as long as outputs were translated appropriately as depicted in the 'pyramid' in figure 5.2.

5.3.3.2 Implications: network planning applications

There was a range of perspectives on how the measures might be used.

PT operators reported they were not overly concerned about what technical measures were being used by transport authorities and operation centres unless they formed part of their contracts, in which case they would become very familiar in understanding them.

Across a few workshops, stakeholders reported that the measures could be useful for high-level network planning, such as deciding where to prioritise putting in PT priority treatments like bus lanes and to build the case for funding. Another suggestion from an Auckland attendee was that the measure, especially if it incorporated speed, would be useful for monitoring projects once deployed, which would inform future project prioritisation and funding. Stakeholders noted that in applying the measures to compare routes or modes, it was important also to apply comparisons for similar time periods – such as peak with peak, weekend with weekend.

Environment Canterbury reported they would find it useful if a national measure were in place so they could compare their service predictability with that of other places. However, they noted that perceptions of what reliability means would vary by area. Even within Canterbury, Christchurch patrons have become accustomed to more service delays due to the prevalence of post-earthquake road works, whereas in Timaru, bus drivers and patrons complain over even the most minor delays as delays are so rare in a less dense town with little congestion. The implication of this may be that different areas would need to have local thresholds that are specific to the area even if the measures used are standardised between areas.

An ECan representative relayed that it would be more useful to compare PT services in different regions than to compare PT with car (saying something along the lines of, 'car is always going to win'). Furthermore, this respondent believed predictability should really be looked at 'within-mode' too, that it is unfair to compare buses with trains for example as you 'are not comparing apples with apples'. One participant did joke that if a measure made PT look good compared with private vehicles that would be beneficial.

With regards to the implications for customers, most participants thought the measures were not suitable for customers, but rather it was inferred that customers might indirectly benefit from any improvements to network planning. The perspective was that the general PT feed specification outputs already facilitated the provision of easy-to-digest information such as travel time estimates so outputs from a predictability measure might be redundant or just confusing for customers. The Wellington workshop noted that customers are accustomed to getting absolute journey times through modern apps (like Google) which are easily understood and reflect their intrinsic experiences on the ground.

Although most stakeholders did not think the measures, or outputs from the measures, were useful for customers, the contingent from Canterbury thought customers would probably be able to understand outputs of analysis if services were simply colour coded by level of predictability. So, for example, a MaaS app might show journey time estimates but delivered with colour coding to depict the predictability of the journey, eg there might be a 15-minute bus trip estimation, but with a red coding this travel time was likely to be affected by issues on the network, or green to indicate higher reliability of the travel time estimate. Similarly, some of the AT representatives thought outputs could be valuable to customers if more simple names for measures were offered and outputs were presented in real, easily digestible formats.

5.3.4 Assessing the measures

As noted in the previous sections there was some confusion about the measures. The only measures all attendees seemed to understand were schedule adherence-related measures like punctuality. A number of stakeholders had difficulty understanding the Transport Agency road index and some also had trouble understanding the shortlisted measures from the international literature.

AT reported it used some of the measures discussed but applied simpler, more easily understood names. After the workshop was held one of the attendees provided table 5.3 below which shows that AT uses a modified buffer index but refers to it as 'reliability' and uses an iteration of the planning time index which it calls 'delay'.

Table 5.3 Predictability performance measures used by AT and its equivalents

	Reliability	Delay
AT	85th percentile (peak) median TT (peak)	85th, median (peak), 15th free flow TT
Auckland Motorways Alliance/PT predictability research project	Modified buffer index	Planning time index
	85th percentile (peak) median TT (peak)	95th percentile (peak) free flow TT

The Christchurch contingent understood the measures relatively well. They did admit to getting a bit lost in the exact definitions of the measures but when we showed them the data testing imagery, they seemed to understand the colour patterns and implications. With patience and time, the Wellington group (or at least some of the attendees) were also able to grasp the shortlisted measures.

5.3.4.1 Transport Agency road index

All stakeholders found the Transport Agency road index the most difficult measure to understand. The measure was explained using the following three definitions:

Calculates the proportion of journeys completed quicker than a 'buffer time'. The buffer time is average travel time plus a fixed proportion of the average. The more journeys completed in less than the buffer time, the more predictable journeys are deemed to be.

The measure assesses deviations from mean travel times for 15-minute travel timeslots within a predefined 5% time buffer.

To calculate 'predictability', each peak hour per day during a monthly period (in 15 minute intervals) is assessed against this buffer and then allocated a value of zero if the observed travel time exceeds this buffer or one if it is below. Predictability is then expressed as a percentage and is equal to average of the sum of the ones and zeros in the peak hour period of that month. The percentage is directly proportional to the so called 'predictability' ie lower percentage indicates a lower predictability. The measure is then compared against the previous months for performance reporting purposes. Currently, the threshold is defined as the rolling 12-month average (average over last 12 months from current month) plus 5%.

However, none of the three definitions were easily understood by the workshop attendees. It was therefore difficult to gain feedback of the efficacy of either applying this measure directly to PT or of other measures to fit to it.

Some attendees considered the measure was not intuitive or meaningful. Stakeholders thought it was definitely not suitable for any type of public understanding but rather was only potentially useful for modelling, but there seemed to be some scepticism for even that purpose. One of the managers from AT Metro seemed particularly opposed to the Transport Agency road index measure noting he had ‘serious concerns’ about it, though it was not entirely clear what these were. There were some concerns by workshop attendees about the potential consequences of the 15-minute interval measurements and how these might or might not have an impact on lower frequency PT routes. The only positive response to the Transport Agency road index was that it was already in use.

5.3.4.2 Punctuality

All respondents seemed to understand punctuality and thought it was a good measure for operators and the public alike. When it was explained that a major issue with punctuality was that it could not be used to compare PT with private vehicle travel which lacks scheduled departure and arrival times, stakeholders seemed to understand this.

In the Wellington workshop, the five-minute threshold for lateness was discussed with attendees querying where the five-minute threshold may have originated from – nobody knew for sure. It was noted that changing the threshold from five minutes could dramatically change the performance reporting.

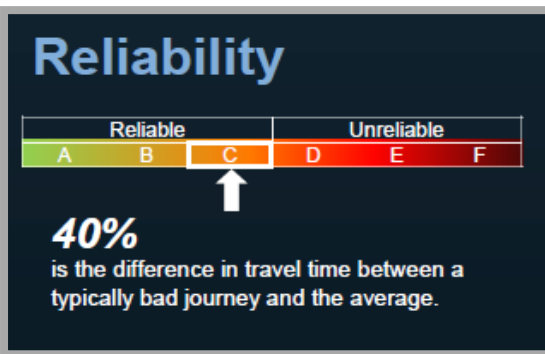
5.3.4.3 Buffer Index: 95th percentile, mean travel time

Stakeholders seemed to have a mixed understanding of a buffer index. Some reported understanding buffer time but then found the translation to percentages confusing. One stakeholder from ECan thought the buffer index was a useful measure because PT users definitely used buffers. He described how when he travelled to one destination he allowed more time based on his previous negative experiences with that service, whereas for another destination where he used a different service, he tended not to leave a big buffer. In this example the stakeholder described his buffer in terms of leaving for ‘one service’ earlier than he wanted to. This description is interesting because it shows that while our directive for this research was to focus on predictability in terms of in-vehicle travel time, service frequency certainly affects actual buffer times used by patrons. AT suggested it would be more intuitive to provide ‘worst compared to typical travel’ times.

5.3.4.4 Modified buffer index: 85th percentile, median travel time

AT already uses a modified buffer index as a predictability performance measure. As shown in figure 5.1 for the public, they call the measure ‘reliability’ and provide a percentage and define it as ‘the difference in travel time between a typically bad journey and the average’ and apply a spectrum of colours to help indicate meaningful performance of reliability. The fact that AT uses a modified buffer index (albeit under the name ‘reliability’) indicates they consider the measure suitable in terms of accuracy and statistical meaningfulness (for network planning), and to a degree, for customers. Despite the measure’s adoption by AT, in the workshops some attendees were concerned that it was still not an overly intuitive measure.

Figure 5.3 From the Auckland Trans bus scorecard. The modified buffer index is reported as ‘reliability’ and a colour spectrum of performance is provided

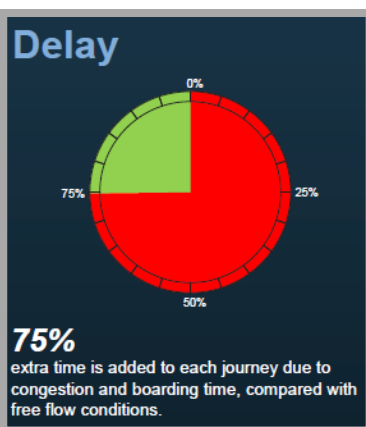


By AT’s own admission, much of its modified buffer index analysis was ‘low down in the pyramid’ so not necessarily performance information provided to the public. Some of the operators noted they had a sense this type of analysis was taking place and it would be good to have it more public to help evaluate performance of routes. Some AT stakeholders mentioned this analysis was currently used to support the case for infrastructure investment.

5.3.4.5 Planning time index: 95th percentile (peak), free flow travel time

AT also uses a measure similar to the planning time index, but using 85th percentile, median peak, 15th percentile, and like planning time index, free flow travel time. They refer to this measure as ‘delay’ (figure 5.4).

Figure 5.4 From the AT bus scorecard. Planning time index is reported as ‘delay’



Of the shortlisted measures presented, the Wellington workshop attendees seemed to have some preference for the planning time index measure. One of the issues they thought could be problematic for this measure, however, was the comparability between urban (more congested) conditions (where more stoplights might be present) and free-flow conditions as this measure was particularly focused on measuring the difference between free-flow and near worst-case travel time delays. One of the advantages of this measure in their opinion was the ability to fairly easily translate the outputs into actual journey times for people.

5.3.4.6 Other measures

In addition to the shortlisted measures described above, some workshop attendees suggested other measures that they thought might be more suitable as a predictability performance measure.

Absolute travel times

A timetabling and service planner from GWRC felt the optimal predictability performance measure would be to report on absolute travel times, including just reporting a range. For example, the performance measure for a bus from Island Bay to, say Willis Street in Wellington might take between 20 and 45 minutes depending on congestion and other variables. He argued that just using absolute travel times was easily comparable between modes and intuitive for customers, service operators, local authorities and wider groups. A greater range in the absolute travel times would indicate wider variability (and thus lower predictability). There was general agreement among the workshop attendees that this was a good measure particularly because it is relatable to customer experience of travel times. The idea of having absolute travel times shown as a range was seen to have real world applicability. Proponents of this measure argued that decision makers are lay people so an intuitive measure like this is more useful for them. Depending on the exact application and intentions of the Transport Agency, use of this measure could be a viable, useful suggestion. One limitation of this measure is that it does not provide a direct indication of speed (and thus distance) which is a factor T thought should be included in an optimal measure. With this measure, thresholds are also less relevant.

Speed

AT noted they looked at average and median speeds to assess corridors and believed it was imperative to examine speed.

5.3.4.7 A summary of scoring of measures

Each of the workshops was presented with a matrix of the shortlisted measures and five evaluation criteria with which to rate each measure: ease of understanding – operator, ease of understanding – customer, measure's accuracy, comparability with car, and cost/effort efficiency.

The attendees in the Auckland and Christchurch workshops did not seem to be comfortable enough with their understanding of the measures and/or the objectives of the research to complete this evaluation. In Auckland, this issue might have been exacerbated by the large number of attendees (11), which made it difficult to reach consensus in ratings. Moreover, two AT Metro managers voiced scepticism about the research objectives and therefore regarded that undertaking such ratings was inappropriate given their scepticism of the intentions of the measures. The Auckland group did not seem to consider any of the measures were 'better' than others (other than some strong scepticism about the Transport Agency road index), but rather the preferred measure to use really depended on exactly what one was trying to achieve. That said there was some scepticism that the measures might be limited in usefulness by not accounting for variability and speed. With regards to the evaluation criteria of 'accuracy' this group said this was much more dependent on the data available than the measure selected. In terms of 'ease of passenger understanding', the Auckland workshop said that really depended on how the measures were explained.

The PT operators at the Auckland workshop admitted they really did not understand the measures, that they were too complex and thus it was too difficult to evaluate them, but again noted that if they were integrated into their contracts as KPIs, they would thoroughly learn the measure(s).

Only the Wellington workshop was successful in rating each of the measures for each of the criteria. Their ratings are listed in table 5.4, which also notes any explicit reasons given for a rating. It is apparent from this table that the planning time index and modified buffer index performed best across the criteria. Interestingly the Wellington workshop suggested modified the planning time index's percentile from 95th to 85th, consistent with the modification used by AT. Some of the over-arching points made in undertaking this rating exercise in the Wellington workshop were:

- Problem definition is imperative to selection of an appropriate measure to use.

- Planning time index was perceived as more favourable due to a perception that it could be relatively easily transferred into absolute journey times (which was seen as the actual ideal output) and convertible across a variety of modes.

Table 5.4 Wellington workshop’s rating of different measures

Measure	Ease of understanding – operator	Ease of understanding – passenger	Measure’s accuracy	Comparability with car	Cost/effort efficiency
Transport Agency road index	Low	Low	Medium (Due to 15-minute interval issue)	Medium (PT has less data points; average of an average; is it comparable if the measure does not work for PT - due to 15-minute interval issue)?	Low Lots of math, and data
Buffer index	Medium	Low	Low (dependent on averages)	High	High (relatively straight-forward)
Modified buffer index	Medium	Low	High	High	High
Planning time index	High	Medium	Medium (Uses 95th percentile, 85th would be better)	High	High
Punctuality (Acknowledging this measure is not comparable with car-based measures)	High	High	Medium (binary)	Low	Medium (Wellington experience is there can be pains with this, measure not yet automated)

5.3.4.8 Use of multiple measures

In the Wellington workshop, we asked about the concept of using multiple measures under varying circumstances. The feedback we received was that this could be unnecessary, hard work, and result in too much information for the lay person and decision makers. In contrast, it was clear AT already used multiple predictability performance measures such as ‘delay’ and ‘reliability’ as discussed. The latter experience suggests that different measures can be useful for looking at different aspects of predictability.

5.3.5 Exploring thresholds

Workshop attendees were also shown the results of applying the shortlisted measures and different thresholds. This illustrated the sensitivity of threshold levels to the performance measures. A summary of feedback received about determining thresholds for the measures follows:

- In all three workshops, it was agreed that the private vehicle thresholds were not appropriate to apply directly for PT.
- Some participants (especially those accustomed to undertaking detailed analysis) were better able to understand the thresholds whereas some attendees reported not finding the outputs intuitively meaningful.
- Auckland workshop attendees thought care needed to be taken in determining the 'right' thresholds. AT was willing to share the thresholds used in the Auckland area.
- One participant thought thresholds should depend on journey length.
- Thresholds should consider 'acts of gods' (eg big storms/disruptions).
- The Christchurch contingent questioned whether it made more sense to have different thresholds in different areas or not. The advantage of consistent thresholds nationally would be the ability to compare services between regions but on the other hand, it might make more sense to have different local thresholds due to variation in geography, urban form and local tolerance to delay.

Sensitivity testing of thresholds revealed a lack of variability for three of the measures using the private vehicle thresholds. When these results were presented, in all three workshops, participants agreed it was not appropriate to apply directly private vehicle thresholds to PT. Moreover, in Christchurch, the representatives from ECan felt that different thresholds should be applied to PT compared with private vehicle travel because PT services must stop regularly for passenger boardings and alightings. One piece of valuable feedback from a modeller at GW was that to really understand the compatibility of the measures across PT modes *and* car modes, we should have applied the measures also to a private vehicle dataset to compare the PT travel data. This was an excellent suggestion but unfortunately was outside of the scope of the commission but it is recommended that this be undertaken as the next step in selecting a measure.

In the Wellington workshop, which was the first one to be held, we tried to discuss what sorts of thresholds might be appropriate to apply for each measure but it became somewhat evident in that workshop and more-so in the ensuing workshops that this was far too detailed of a question to ask. Though some exploration of thresholds was undertaken as part of this project, establishing appropriate thresholds is another piece of work that we recommend being undertaken.

Those from the Wellington workshop who had advocated for simply using ranges of absolute journey times to measure predictability suggested that adoption of this performance measure would also allow people to pick journey time thresholds, which would indicate they understood them.

5.4 Overall conclusions from the workshops

The workshops provided valuable research insights regarding the use of PT performance measures and their relevance to attendees. Some attendees were keen to learn about the overarching objectives of the research before giving feedback on the preferred measures. Others noted that the measure to use really depended on the aspect of reliability relevant to the service.

Some stakeholders who regularly undertake performance monitoring analysis on the PT network challenged the value of the research. The workshops enabled these participants to better understand the measures and the research objectives. That said, some still felt somewhat confused about the implications of the research and how the selected measure(s) would be used by the Transport Agency. There was not an overwhelmingly strong preference for any of the shortlisted measures but when pressed, some respondents from the Wellington workshop seemed to prefer an absolute travel time measure (with a

range to show variability). Everyone seemed to understand and like the punctuality measure, with which they were familiar, but given its non-applicability to the objective of comparability with private vehicle measures, their next preference of the shortlisted measures presented seemed to be the planning index and modified buffer index. The Wellington contingent thought the planning index seemed flexible in its applicability to convert into useful measures for both network planners and customers.

There was consensus across the workshops that the current Transport Agency road index was not seen as easily transferrable to PT and generally confusing to understand in the PT context. The implication of this was that it might not be sensible to try to 'fit' other measures to the Transport Agency road index but perhaps it would be more useful to use the shortlisted measure(s) to rate predictability for the road network.

Generally, the workshop attendees felt the measure(s) might be useful for network planning but the shortlisted measures examined were not appropriate for customers. That said, some participants did suggest that with further explanation, some of the measures might be informative for customers in a similar way to how AT reports the modified buffer index and a slightly different version of the planning index in their dashboard outputs. The fact that these measures were already being used across modes by AT indicated they were, at least, somewhat promising. Some of the Auckland attendees suggested the best measure would be one that was easy to understand and specific to the data sources available.

Table 5.5 provides a summary of the outcomes to the workshops' aim to present the research and address the following key questions:

- whether stakeholders understood the measure(s) being proposed
- if stakeholders thought the measure(s) were useful
- what they foresaw would be the implications of adopting the measure(s).

Table 5.5 Outcomes of key workshop questions

Key workshop question	Finding
Whether stakeholders understood the measure(s) being proposed	There was some variation in whether or not attendees understood the measures. A concerning number did not understand the measures, particularly the Transport Agency road index. Some seemed to understand the measures but not the discussion on thresholds. Some attendees had a rough but not detailed understanding.
If stakeholders thought the measure(s) were useful	<ul style="list-style-type: none"> • Generally speaking, the shortlisted measures were not seen to be useful by most of the workshop attendees, except, to a degree for network planning. • There seemed to be a sense among some of the 'bottom of the pyramid' attendees accustomed to doing more detailed data analysis that with some minor adjustments there might be potential to use some of the measures (eg under different names and with an adjustment to the percentile used in the planning index). • AT indicated that some form of some of the measures was already in use. • Workshop attendees did not seem to support 'fitting' shortlisted measures to the Transport Agency road index, a measure which they did not see as particularly useful due to its complicated nature
What they foresaw would be the implications of adopting the measure(s)	<ul style="list-style-type: none"> • The shortlisted measures were largely seen only to be appropriate to network planning, • For the measures to have any significance to customers either just an absolute travel time/range was seen to be appropriate, simple colour coding used, or certainly more simple names for the measures. • From the vantage of operators, measures only became significant if they would affect contractual KPIs. • The thresholds adopted could affect the implications of the measures used.

There seemed to be two important areas for future research emerging from the workshops:

- Development of appropriate thresholds for the preferred measure(s). This will involve some consideration of the factors identified above (geography) and whether it makes more sense to have national thresholds or locally specific thresholds.
- Application of the preferred measure(s) and potential thresholds developed to equivalent road datasets to compare the performance across modes and ensure that appropriately transferrable measures and thresholds have been selected.

6 Conclusions and recommendations

6.1 Summary of outcomes

The primary aims of the research were:

- Identify and develop the best measure for PT travel time predictability based on a review of literature, a practice review, ascertaining the benefits and limitations of measures, comparison with the current Transport Agency road-based measure and validation testing.
- Develop a nationally aggregated set of travel time predictability information across regions, travel modes and different times of day to 'trial' shortlisted measures.

Overall, the research project was largely able to achieve these overarching aims. A review of literature exposed which reliability measures were in use in New Zealand and worldwide for PT as well as for private vehicle travel. The literature review also provided some evaluation of the measures by documenting their advantages and disadvantages. This evaluation resulted in a 'shortlist' of measures to apply to a nationally aggregated set of PT travel data across regions and PT modes in New Zealand. This provided a 'trial' of shortlisted measures to assess the 'fit' to the Transport Agency road index and modification potential. This data testing revealed that the measures were all linearly related and there was not a strong case for one measure or another to be used based on modification or fit. It is worth noting that while datasets were obtained for bus and rail, the project team was unable to obtain ferry data within the timeframes given so the measures were not directly tested on ferry data. However, it is expected there would generally not be any unique issues with applying the measures to examine ferry performance, even though there may be different obstacles to reliability (eg wind, chop, waves, as opposed to perhaps congestion).

Validation of the research was undertaken through three stakeholder workshops, a steering group review of the literature review, a variation report documenting the data testing to assess modification potential, a peer review and a steering group presentation. The validation workshops revealed that stakeholders felt selecting any measure really depended on what exact aspect of reliability one wanted to examine and that care and consideration needed to be taken in comparing modes and developing thresholds.

6.2 Evaluation of measures

The practice side of the literature review revealed that schedule adherence measures were the most common type of predictability measure used by PT authorities in New Zealand and abroad. These measures are particularly useful for assessing operators as KPIs. Unfortunately, this type of reliability measure is unsuitable for comparison with existing road-based reliability because the road based measures are not based on a prescriptive schedule. There was some evidence of more buffer time measures being used by AT and Transport for London for PT and, more commonly, for road reliability analysis. Three buffer time measures were shortlisted and applied to aggregated PT data alongside the Transport Agency road index to evaluate the 'closeness of fit'. The shortlisted measures included the buffer index, modified buffer index and planning time index. These measures were better suited for modification to perform aggregate analysis across travel modes and hence were applied to the New Zealand datasets.

In reviewing the results of the data application, it becomes apparent that all shortlisted measures can be used to assess reliability across modes. All shortlisted measures are linearly related, producing

comparable results across different measures. There is no 'right' or wrong measure but rather different measures can show variability from slightly different perspectives:

- With their reliance on average rather than median, the Transport Agency road index and buffer index are more sensitive to extreme observations than median-based measures like the modified buffer index.
- The buffer index can be useful for looking at how much fluctuation occurs on average along a route.
- The planning index can be relatively easily converted to total journey times, and offers an absolute minimum and maximum.
- Punctuality is a common and easily understood measure used to evaluate reliability for PT but does not meet the objective of being comparable to private vehicle travel.
- The Transport Agency road index is not seen to be appropriate for PT due to its reliance on 15-minute time intervals.

Again, in terms of selecting which of the buffer time-based measures is most useful, there was not a strong case for any particular one over another from the data testing alone. They were all linearly related which is consistent with the findings by Mazloumi (2008). This strong linear relationship between different measures indicates they are closely related and can be used interchangeably. In general, there is no significant difference in outcomes between the measures tested. The major factor in variation between the measures is whether they are mean-based or median based with the former being more sensitive to extreme observations in the data series than median-based measures. This finding is also consistent with previous research and is why Pu (2011) suggests median-based measures are preferable. Which measure to use depends on which aspects of reliability in particular one is interested in. A summary of some of the key measures examined in this research are provided in table 6.1 below.

Table 6.1 Summary of some of the key measures examined in this research

Measure	Better for measuring	Less optimal contexts	Explanation and additional comments
Transport Agency road index	Road reliability Detailed data analysis	PT generally but especially low frequency routes. General public reporting.	The Transport Agency road index's current use of 15-minute interval readings makes it inappropriate to use for lower frequency PT services. Feedback from the workshops was that the Transport Agency road index is very difficult to understand and there would be a strong preference not to use this measure.
Buffer index	Typically used for private vehicle travel	Reliance on mean makes this measure less optimal when there are extreme observations	This measure can be used across modes and provides similar results to the Transport Agency road index but thresholds need to be considered.
Modified buffer index	Good for measuring all modes	Thresholds may need adapting across modes/regions. Presentation to the public would benefit from a more intuitive name and an easy explanation of the measure.	This measure can be used across modes and provides similar results to the Transport Agency road index but thresholds need to be considered. It is already in use by AT.
Planning index	Good for measuring all modes	Thresholds may need adapting across modes/regions. Presentation to the public would	Research demonstrates this can be used across modes and provides similar results to the Transport Agency road index but thresholds

Measure	Better for measuring	Less optimal contexts	Explanation and additional comments
		benefit from a more intuitive name and a simple explanation of the measure.	need to be considered. A variant of this measure is already used by AT under the name 'delay'.
Lost customer hours (Transport for London)	Currently used for PT	Less optimal for quick and easy calculations, requires sophisticated data capture.	Not currently used for private vehicle travel but with some modifications could be used. Time intensive to calculate.
Punctuality	<ul style="list-style-type: none"> PT arrival and departure times To detect early and late running 	Not appropriate to use for private vehicle travel which lacks specific schedules.	Private vehicles lack specific scheduled arrival and departure times, so punctuality is not appropriate to use for multi-modal comparisons.

The outcomes of the data testing indicated that the buffer measures 'fit' was comparable to the Transport Agency road index but were sensitive to the thresholds adopted. The road-based thresholds initially applied did not provide for enough differentiation in interpreting performance to be useful. We have demonstrated that the performance of services is largely affected by which thresholds are applied. It is unclear whether it would be more sensible to have consistent national thresholds which would allow for eased comparisons of PT predictability between regions or whether it would make more sense to have different local thresholds due to variation in geography, urban form and local tolerance to delay. It is worth noting that it may be difficult to develop consistent national thresholds that suit the PT services of different cities; however, doing so would also be potentially beneficial should there be interest in making the predictability reporting public.

6.3 Implications

Based on the evaluation and application of measures in this research, any of the three shortlisted measures would be appropriate to use for aggregate comparisons of reliability across modes. Because the shortlisted measures originate from road-based measures they inherently 'fit' road measures but rather may require some modification of the thresholds to appropriately 'fit' and be useful for PT performance analysis. More research is recommended to determine appropriate thresholds. The Transport Agency road index as it is currently structured was not seen as appropriate for measuring PT reliability. Moreover, it became apparent through the validation workshops that the existing predictability measure for road travel was not easily understood or respected. Feedback seemed to strongly suggest that the Transport Agency road index not be used for PT. There was some scepticism about whether PT and road travel should even be compared. On the other hand, AT reported that it already applied two of the shortlisted measures (or similar renditions of them) to measure reliability and delay among different transport modes. This provides an excellent model for how this can be done in practice.

6.3.1 Data preparation

Data preparation is a fundamental aspect of aggregation: the quality of data will impact on the robustness of the outcomes. Our application of the shortlisted measures to PT data involved a great deal of work in data preparation, such as obtaining the data and removing missing observations. This removal of

observations is less robust than undertaking imputation. On the other hand, imputation is time -intensive. We recommend use of imputation in circumstances where robustness is preferable and time and budgets are not overly constrained.

6.3.2 Service frequency and predictability

One point worth noting is the finding that emerged through the research that predictability/reliability measures become less important as service frequency increases. As discussed in the literature review, Currie et al (2012) reviewed ten indicators and found one of the most preferred measures of reliability to be waiting time for a service, which is highly influenced by the frequency of services. In terms of the customer perspective of reliability, especially schedule adherence, it becomes less relevant as service frequency increases.

6.3.3 Advanced technology and predictability

It is worthwhile to consider the implications of this research in the context of rapidly advancing technology. There are three main aspects to consider:

- Customer expectations are changing with higher expectations of communication about travel disruptions and RTI.
- Big data is emerging and enabling real-time responsiveness and better data for monitoring.
- New technology-enabled transportation services are increasingly providing a 'first/last mile' connection to PT for customers. Some transport agencies overseas have entered formal partnerships to provide a connecting service. These new services include bikeshare, transportation network companies (eg Uber), carshare, dynamic carpooling and demand responsive or pop-up transit eg Via. The increasing diversity in shared-use transport services is driving more multi-modal trips; reliability is important to customers across all trip legs. What does predictability monitoring look like for these sorts of services and who is responsible for it?

It would be useful to consider these three trends in the decision of a preferred multi-modal reliability performance monitoring path into the future.

6.4 Recommendations and next steps

From these findings, the following recommendations are proposed:

- For network planning use either (or both) the modified buffer index and planning time index for analysis, both of which are statistically buffered from extreme externalities.
- Undertake further research applying each of the shortlisted measures to private vehicle travel; ideally this should be done for the same reporting period as for PT data.
- Set up regular PT monitoring workstreams: define appropriate thresholds, streamline data acquisition and cleaning process, develop software to undertake regular monitoring (eg quarterly).
- Provide simple travel times or a range of absolute travel times for customers. Alternatively, the modified buffer index and planning index outcomes can be presented using more easily understandable names and outputs, modelled on what AT currently produces.
- It might also be useful to test the measures on ferry data to confirm their applicability.

Finally, in considering the future of aggregating to compare modes, there is potential to weight modes based on the number of people being moved. This premise underlies the lost customer hours measure

used by Transport for London. However, such a weighting may prove difficult to assign and needs to be carefully considered in line with the intended use of the aggregated measure. Therefore, the correct context needs to be established to determine this, eg if the outcome of the measure is to evaluate total customer delays then higher weighting should be allocated to modes with the highest number of passengers (eg a high-capacity rail route). Such a customer-centric rendition might also consider *perceptions* of travel time for different aspects of a journey (eg waiting for services can be experienced as longer than in-vehicle time so penalties could be applied on the basis of service frequency). Or, if network productivity was a stronger strategic driver, then more weight should be added to road traffic routes that carry large volumes of freight. These matters would benefit from further consideration and potentially research. There is also potential for further theoretical research in the area, such as investigation of the ability to fit PT reliability measuring into the car-based multistate models of travel time reliability recommended by Park et al (2010). This has the potential to provide customers with two-stage forecasts of travel time and delay of PT services.

7 References

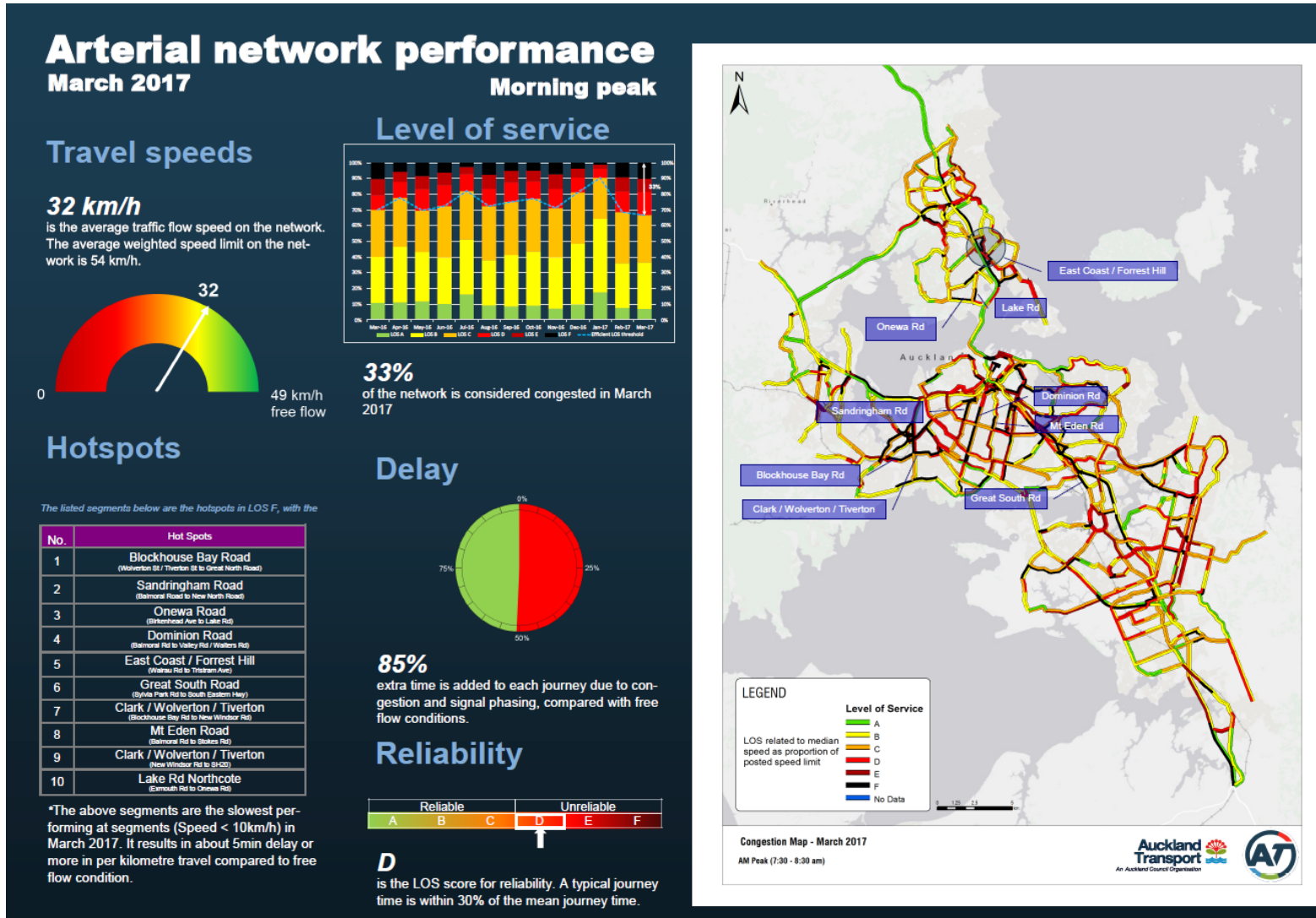
- Alemazkoor, N, W Burris and RD Santosh (2015). Using empirical data to find the best measure of travel time reliability. *Journal of the Transportation Research Board* 2530: 93–100.
- Auckland Transport (AT) (2016) *Monthly transport indicators report 2015/16. June 2016*. Accessed September 2016. <https://at.govt.nz/media/1911050/item-11-1-attachment-2-auckland-transport-monthly-indicators-report.pdf>
- Auckland Transport (AT) (2017) *Monthly transport indicators report. April 2017*. Accessed August 2017. <https://at.govt.nz/media/1973778/item-122-monthly-indicators-report-april-2017-covering-paper.pdf>
- California Department of Transportation (CALTRANS) (1998) California transportation plan: transportation system performance measures. *Transportation system information program final report*. 32pp.
- Cambridge Systematics (2003) Providing a highway system with reliable travel times. *National cooperative highway research program (NCHRP) report 20–58[3]*. 136pp.
- Cambridge Systematics (2009) Performance measurement framework for highway capacity decision making. *Strategic highway research program report S2-C02-RR*. 113pp.
- Cambridge Systematics, Dowling Associates Inc, System Metrics Group Inc and Texas Transportation Institute (2008) Cost-effective performance measures for travel time delay, variation and reliability. *National cooperative highway research program (NCHRP) report 618*. 70pp.
- Chen, X, L Yu, Y Zhang and J Guo (2009) Analyzing urban bus service reliability at the stop, route and network levels. *Transportation Research Part A: Policy and Practice* 43, no.8: 722–734.
- Chien, S and X Liu (2012) An investigation of measurement for travel time variability. Pp.21–40 in *Intelligent transportation systems*. A Abdel-Rahim (Ed). Intech. 206pp.
- Christchurch Traffic Operations Centre (CTOC) (2015) *Key route travel time monitoring measurements – memo to National Transport Operation Centre Team*.
- Christchurch Traffic Operations Centre (CTOC) (2016) *Key route travel time monitoring measurements*. 6pp.
- Cramer, A, J Cucarese, M Tran, A Lu and A Reddy (2008) Performance measurements on mass transit New York City Transit Authority case study. *Journal of the Transportation Research Board* 2111: 125–138.
- Currie, G, NJ Douglas and I Kearns (2012) An assessment of alternative bus reliability indicators. *Australasian Transport Research Forum*. Perth, September 2011.
- de Jong, GC and MCJ Bliemer (2015) On including travel time reliability of road traffic in appraisal, *Transportation Research Part A: Policy and Practice* 73: 80–95.
- Department of Transportation, State of Wisconsin (nd) *MAPSS performance scorecard*. Accessed 4 October 2016. http://wisconsin.gov/Pages/about_wisdot/performance/mapss/goalmobility.aspx#mobility
- Do, CB and S Batzoglou (2008). What is the expectation maximization algorithm? *Nature Biotechnology* 26: 897–899.
- Donders, ART, GJM Gvd Heijden, T Stijnen and KGM Moons (2006) Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology* 59: 1087–1091.

- Dong, J and HS Mahmassani (2009) Flow breakdown and travel time reliability. *Journal of the Transportation Research Board 2124*: 203–212.
- Douglas, N, G Currie and L Ferreira (2006) *Monitoring bus reliability: a review of ITSRR surveys*. Independent Transport Safety and Reliability Regulator, NSW Government. 76pp
- Dowling, RG, A Skabardonis, RA Margiotta and ME Hallenbeck (2009) Reliability breakpoints on freeways. *88th Annual Meeting of the Transportation Research Board*, Washington, DC, January 2009.
- Federal Highway Administration (FHWA) (2006) *Travel time reliability: making it there on time, all the time*. Accessed 15 September 2016. http://ops.fhwa.dot.gov/publications/tt_reliability/
- Federal Highway Administration (FHWA) (2010) *The urban congestion report (UCR): Documentation and definitions*. Accessed 15 September 2016. http://ops.fhwa.dot.gov/perf_measurement/ucr/documentation.htm.
- Fosgerau, M and A Karlström (2010) The value of reliability. *Transportation Research Part B 44*: 38–49.
- Furth, PG, B Hemily, THJ Muller and JG Strathman (2006) Using archived AVL-APC data to improve transit performance and management. *Transit Cooperative research and management (TCRP) report 113*. 83pp.
- Gaffney, J (2006). Understanding network performance information provided to users. *PIARC International Seminar on Intelligent Transport System (ITS) In Road Network Operations*, Kuala Lumpur, Malaysia, August 2006.
- Greater Wellington Regional Council (GWRC) (2015) *Annual monitoring report on the regional land transport plan*. 38pp.
- INRIX Inc (2011) *National traffic scorecard annual report 2010*. INRIX. 15pp.
- Jackson, D (2000) *Reliability as a measure of transportation system performance*. Master of Science thesis. Texas A&M University, College Station.
- Kimley-Horn and Associates (2011) Guide to integrating business processes to improve travel time reliability. *Strategic Highway Research Program 2 (SHRP) report S2-LO1-RR-2*. 41pp.
- Kristoffersson, I (2013) Impacts of time-varying cordon pricing: validation and application of mesoscopic model for Stockholm. *Transportation Policy 28*: 51–60.
- Lily, E and C Xiao (2007) Review of definitions of travel time reliability. *86th Annual Meeting of The Transportation Research Board*. Washington DC, January 2007.
- List, G, B Williams and N Roupail (2014) Handbook for communicating travel time reliability through graphics and tables. *SHRP 2 reliability project L02*. 58pp.
- Lomax, T, D Schrank, S Turner and R Margiotta (2003) *Selecting travel time reliability measures*. Texas Transportation Institute. 47pp.
- Margiotta, R, T Lomax, S Turner and M Hallenbeck (2008) Analytic procedures for determining the impacts of reliability mitigation strategies. *SHRP 2 Project L03 interim report*. 256pp.
- Massachusetts Bay Transportation Authority (MBTA) (2016) *Dashboard*. Accessed October 2016. www.mbtabackontrack.com/performance/index.html#/detail/reliability/commuter_rail/
- Mazloumi, E, G Currie and M Sarvi (2008) Assessing measures of transit travel time variability and reliability using AVL data. *Transportation Research Board 87th Annual Meeting*. Washington DC, January 2008.

- Mazloumi, E, G Currie and G Rose (2010) Using GPS data to gain insight into public transport travel time variability. *Journal of Transportation Engineering* 136: 51–60.
- Nakanishi, YJ (1997) Bus performance indicators: on-time performance and service regularity. *Journal of the Transportation Research Board* 571: 3–13.
- New Zealand Transport Agency (2016) *Investment performance measurement: list of measures*. Accessed September 2016. www.pikb.co.nz/assets/Uploads/Documents/KB-list-of-performance-measures-version-2016-3-30.pdf
- Office of Rail and Road (nd) *NRT data portal*. Accessed October 2016. <http://dataportal.orr.gov.uk/displayreport/report/html/458181f0-979d-42df-b25d-5cf685cea885>
- Park, S, H Rakha and F Guo (2010) Calibration issues for multistate model of travel time reliability. *Transportation Research Record: Journal of the Transportation Research Board* 2188: 74–84.
- Patricio, A and H Mohammed (2012) Time-variant travel time distributions and reliability metrics and their utility in reliability assessments. *Journal of the Transportation Research Board* 2315: 81–88.
- Pu, W (2011) Analytic relationships between travel time reliability measures, *Journal of the Transportation Research Board* 2254: 122–130.
- Public Transport Victoria (PTV) (2016) *Daily operational performance reports*. Accessed October 2016. www.ptv.vic.gov.au/about-ptv/ptv-data-and-reports/daily-operational-performance-reports/
- Public Transport Victoria (PTV) (2013) *Operational performance*. Accessed February 2018. www.ptv.vic.gov.au/about-ptv/data-and-reports/operational-performance/
- Ramachandran, KM and CP Tsokos (2009) *Mathematical statistics with applications*. Academic Press. 848pp.
- Rashidi, S (2014) *Bus dwell time and travel time modelling using data mining methods*. PhD thesis. University of Auckland.
- Rashidi, S and P Ranjitkar (2014) Estimation of bus dwell time using univariate time series models. *Journal of Advanced Transportation* 49: 139–152.
- Rashidi, S, R Pant, A Hooper and C Baker (2016) Travel time reliability measure using global positioning system data. *IPENZ Transportation Group Conference*. Auckland. March 2016.
- Reed, S (2013) Transport for London – using tools, analytics and data to inform passenger journeys. *Journeys*, September 2013: 96–104.
- Ruben, VL, R Piet and B Martijn (2011) Travel-time reliability impacts on railway passenger demand: a revealed preference analysis. *Journal of Transport Geography* 19, no.4: 917–925.
- Saunders, JA, N Morrow-Howell, E Spitznagel, P Doré, E Proctor and R Pescarino (2006) Imputing missing data: a comparison of methods for social work researchers. *Social Work Research* 30: 19–31.
- Schil, M (2012) *Measuring journey time reliability in London using automated data collection systems*. Master thesis. Massachusetts Institute of Technology.
- SPSS (2011a) *Estimation methods for replacing missing values*. IBM Corporation.
- SPSS (2011b) *IBM SPSS missing values 20* (user manual). IBM Corporation. 91pp.
- Streiner, DL (2002) The case of the missing data: methods of dealing with dropouts and other research vagaries. *Canadian Journal of Psychiatry* 47: 70–76.

- Transport for London (2011) *Developing a reliability metric for LU customers*. Customer research conducted by 2CV for TfL.
- Transport for London (2016a) *London buses: network performance second quarter 2017/18*. 24 June–15 September 2017. Accessed August 2016. <http://content.tfl.gov.uk/network-performance-latest-quarter.pdf>
- Transport for London (2016b) *London underground performance report 3 2016/17*. Accessed February 2018. <http://content.tfl.gov.uk/lu-performance-report-period-3-2016-17.pdf>
- Trompet, L, X Lui and D Graham (2011). Development of key performance indicator to compare regularity of service between urban bus operators. *Journal of the Transportation Research Board* 2216: 33–41.
- Tsikriktsis, N (2005) A review of techniques for treating missing data in OM survey research. *Journal of Operations Management* 24: 53–62.
- Van Lint, JWC and HJ van Zuylen (2005) Monitoring and predicting freeway travel time reliability: using width and skew of day-to-day travel time distribution. *Journal of the Transportation Research Board* 1917: 54–62.
- Van Lint, JWC, HJ van Zuylen and H Tu (2008) Travel time unreliability on freeways: why measures based on variance tell only half the story. *Transportation Research Part A* 42: 258–277.
- Wakabayashi, H and Y Matsumoto (2012) Comparative study on travel time reliability. *Journal of Advanced Transportation* 46: 318–339.
- Xume, G, Y Lei, Z Yushi and G Jifu (2009) Analyzing urban bus service reliability at the stop, route, and network levels. *Transportation Research Part A: Policy and Practice* 43: 722–734.

Appendix A: Multimodal dashboards from Auckland Transport



Freight route performance

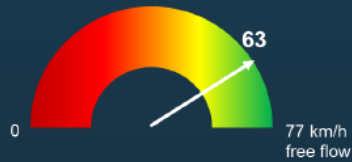
March 2017

Inter-peak

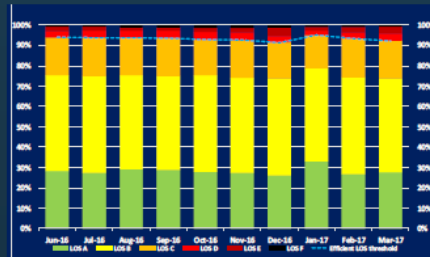
Travel speeds

63 km/h*

is the median speed on the freight network.



Level of service



8%

of the network is considered congested.

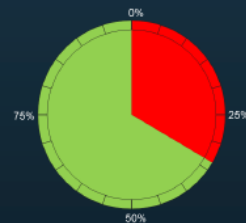
Hot Spots

The listed routes below are the hotspots with LOS E or F. Freight hot spots are highlighted on the adjacent map.

LOS	Hot Spots
F	Central Park Dr *
F	Inner Eastern (Sylvia Park Rd to SH1)
F	Rosebank Rd (BMW Ash St and Great North Rd)
F	Lincoln Rd (Siel Peacock Dr to Great North Rd)
F	Northwestern Motorway Westbound * (Quay St to Beach Rd)
F	Upper Harbour Motorway Eastbound (Caribbean Dr to SH1)
F	Upper Harbour Motorway Westbound (SH1 to Caribbean Dr)
F	Freight - Neilson (SH1 to Great South Rd)
E	Rosebank Rd (Great North Rd to Blockhouse Bay Rd)
E	Rosebank Rd (Great North Rd to Ash St)

* Segments longer than 0.5km

Delay



31%

Extra time is added to each journey due to congestion and signal phasing, compared to free flow conditions.

Reliability

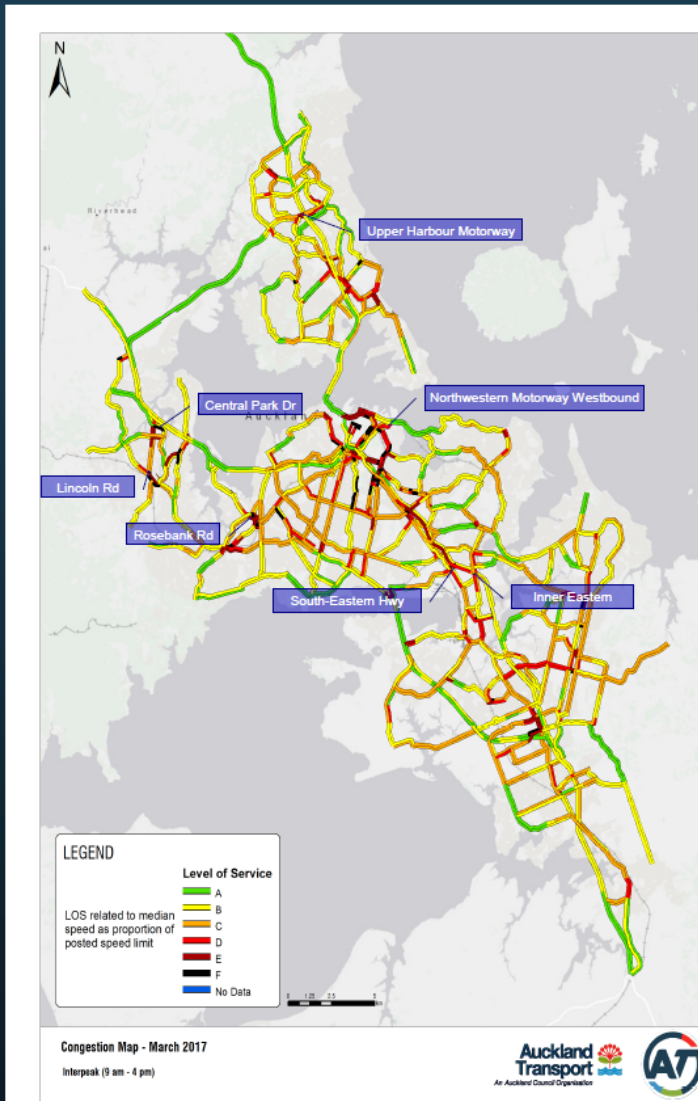


C

Is the LOS score for reliability.

17 km/h

is the average speed observed at the locations highlighted on the map.



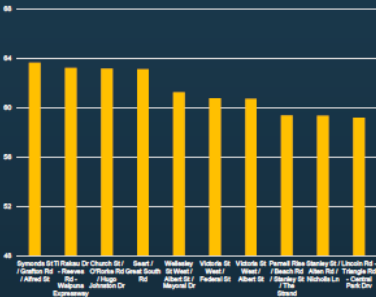
Pedestrian Crossing Performance

March 2017

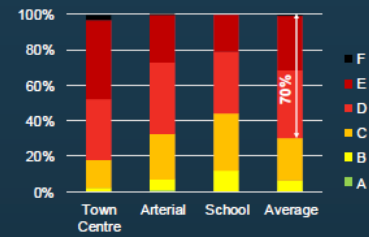
Inter-peak

Worst Intersection

Symonds/Grafton Rd/Alfred St is the worst intersection for Pedestrians on the network.



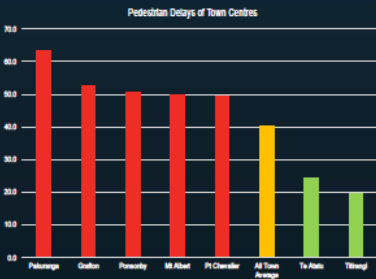
Level of service



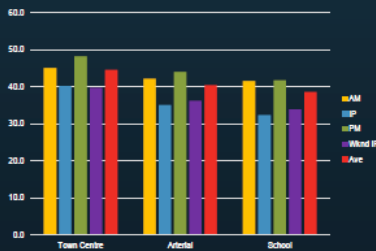
70% of the network is operating below desired LOS for pedestrians.

Town Centre Performance

63.2 seconds is the average pedestrian delay in Pakuranga town centre, the worst performing town centre in Auckland.



Pedestrian Delay



36.0 seconds of average delay is added to each pedestrian journey due to congestion and signal phasing at the intersection, compared with free flow conditions.

