Balancing Reliability and Efficiency in Urban Public Transport Timetables: A Passenger- and Operator-Oriented Multi-Criteria Evaluation

Sara Ghavamilahij



TU Delft

HTM





MSc Thesis Document

Balancing Reliability and Efficiency in Urban Public Transport Timetables:

A Passenger- and Operator-Oriented Multi-Criteria Evaluation

CIEM0500 – MSc Thesis

Sara Ghavamilahij (6218520)
MSc Traffic and Transport Engineering (TTE)
Faculty of Civil Engineering and Geosciences
Delft University of Technology

July 2026

Supervision:

TU Delft

Dr. Ir. N. (Niels) van Oort (Chair and First Supervisor)

Dr. Ir. A. M. (Maria) Salomons (Second Supervisor)

HTM Personenvervoer N.V.

Ir. D. (Dirk) Versluis (Company Supervisor)

Coverpage image source: HTM Vervoerplan (2026)

Summary

This research aimed to develop quality indicators that can be used to evaluate and compare public transport timetables from both passenger and operator perspectives. The study was applied in the urban public transport context of the HTM network, the public transport operator in The Hague. For this study, three tram lines were analysed, namely tram Line 9 (Vrederust - Scheveningen), 19 (Leidschendam - Delft) and 4 (De Uithof - Lansingerland-Zoetermeer), to capture different operating and route characteristics. The datasets used included planned and realised events, as well as passenger data from AVL and AFC systems provided by HTM.

Traditional timetable evaluations rely mainly on operator-oriented metrics, such as punctuality, which can underestimate the effects of service variations on passenger experiences. For passengers, unreliability arises from service disruptions, irregular headways, and schedule deviations, which can lead to increased in-vehicle and waiting times, thereby affecting their perceptions of the service. These perceptions may differ substantially from operator-based assessments, despite their importance for ridership and service attractiveness. On the other hand, service improvements may incur additional costs for the operator. For instance, a timetable with more layover time can make operations less fragile and provide a more reliable experience for passengers, while being less cost-effective because it may require additional vehicles and/or increase total operational hours. However, operators are typically required to comply with performance targets (e.g., punctuality and cancellation rates), and failure to meet these requirements can result in financial penalties. Consequently, measures that improve reliability for passengers and/or increase timetable robustness may also bring some financial benefits for the operators by improving operational performance, despite the associated costs. Another important aspect for passengers is the scheduled travel time, which can be improved through shorter scheduled in-vehicle times and/or higher service frequencies, resulting in lower average waiting times. This study seeks to address a methodological gap in the literature, namely the limited integration of the above-mentioned perspectives, and translates them into timetable quality indicators.

Passenger experience is quantified through reliability and efficiency indicators. Reliability refers to additional travel time due to service variability, while efficiency takes travelled distances and scheduled travel time (passenger speed) components into account. The operator perspective is represented through robustness and expenses indicators, respectively by incorporating performance-based penalties (e.g. for delays and early departures) and operational costs related to vehicle hours and fleet size. Four main performance indicators are thus defined: Passenger Reliability (PR), Passenger Efficiency (PE), Operator Robustness (OR), and Operator Expenses (OE).

To apply the indicators, a set of alternative timetables is created by altering a group of variables, namely layover time, trip time, and frequency. These alternatives include both increases and decreases in the variables in order to analyse their effects through the indicators. For OE and PE, calculations can be done directly on the schedules, since scheduled times, distances, operational hours, and fleet requirements are relevant. For PR and OR, however, simulation of realisations is needed; this is done by deriving and applying the empirical distributions from realised data in the base schedule, which includes the empirical distributions for initial delays, driving times, and cancellation probabilities. The number of simulation runs is determined based on relative error and stability checks of the results.

Results showed that increasing trip times and layovers generally improved passenger reliability (PR) and operator robustness (OR), while increasing frequency had the strongest positive effect on passenger

efficiency (PE), mainly through reduced waiting times. However, improvements in these indicators were typically associated with higher operational costs, resulting in a worse performance for operator expenses (OE). While the overall behaviour of the indicators was similar across the analysed lines, the effectiveness of specific timetable modifications varied. For example, increasing layover times was particularly beneficial for reliability on Line 4, whereas increasing trip times generally produced better reliability outcomes on the other lines. The amount of added slack also proved important, as additional slack improved reliability and robustness but could substantially reduce efficiency and increase costs. For instance, in the trip time scenarios for Line 9, even adding a small slack to the run times (1 second per segment) proved to be effective, without deteriorating the cost aspect significantly. Addition of 5 seconds had more reliability gains but came at the cost of more early departures, unless a holding scheme is applied that can mitigate these negative side effects.

The timetable scenarios were evaluated using a multi-criteria decision analysis (MCDA) approach under a range of weighting schemes. The results showed considerable sensitivity to the selected weights, and the preferred timetable modification differed between lines. Based on suggestions from literature, PR was recommended to be weighted three times more than PE. Regarding the relative importance of the passenger indicator compared to the operator indicator, an exact weighting was not recommended, as it requires further empirical study. This also depends on the case, for instance since HTM operates under the Metropoolregio Rotterdam Den Haag (MRDH), whose strategic goals focus on societal gains from public transport, it was recommended to apply a greater weight to passenger objectives.

Several limitations should be acknowledged. First, only a limited number of timetable modifications were analysed, meaning that the identified patterns should not be generalised too broadly. In addition, the simulation model was developed based on several assumptions, for instance the exclusion of passenger behavioural adaptations to delays and cancellations, such as choosing another mode or route. Furthermore, transfers were kept out of scope due to the modelling complexity and the time framework of the study, despite their important effect on the reliability outcomes for the passengers. Further research could incorporate transfers and passenger behavioural responses to service variability into the simulation model. In addition, including a more detailed holding model could improve the accuracy of the simulated operations and, consequently, the estimated indicator values.

Overall, the study provides a transparent basis for timetable evaluation through quality indicators by explicitly incorporating passenger impacts while accounting for operational constraints.

Preface

This thesis marks the completion of my Master's programme in Traffic and Transport Engineering at TU Delft. Conducting this research has been both a challenging and rewarding experience, allowing me to deepen my understanding of public transport operations while developing valuable research and analytical skills.

I would like to express my sincere gratitude to my supervisors at TU Delft, Niels and Maria, for their guidance, insightful advice and feedback throughout this project, and my supervisor at HTM, Dirk, for his guidance and the time he dedicated to discussing ideas and providing feedback, which was instrumental in shaping this thesis.

I am also grateful to the team at the planning department at HTM for providing me with the opportunity to carry out this research in a practical setting and for making me feel welcome throughout my internship. I would especially like to thank Tim, Mark, and Chantal for the time they dedicated, their support, guidance, and valuable feedback throughout this project.

I would also like to thank my friends. I am especially grateful to Rona for her support throughout my Master's, her kindness, and for all that I learned from her.

Finally, I would like to express my gratitude to my family, especially my mother, for their unwavering encouragement, patience, and confidence in me throughout my studies.

The Hague, July 2026

Sara Ghavamilahij

Contents

Summary **1**

Preface **3**

1 Introduction **7**
1.1 Research Background . . . 7
1.2 Research Problem . . . 8
1.3 Research Objectives . . . 9
1.4 Research Scope . . . 9
1.5 Research Questions . . . 10
1.6 Research Structure . . . 11

2 Literature Review **12**
2.1 Themes and Methodology from Literature . . . 13
2.1.1 Passenger-oriented Approach in Timetabling . . . 13
2.1.2 Efficiency and Robustness Balance . . . 13
2.1.3 Operational Reliability and Operational Efficiency . . . 13
2.1.4 Scenario Evaluation and Multi-Criteria Decision Analysis (MCDA) . . . 14
2.1.5 Methodology Overview from Literature . . . 14
2.2 Literature Gap . . . 15
2.3 Conclusion . . . 17

3 Methodology **18**
3.1 Methodology Overview . . . 18
3.2 Indicator Definitions and Conceptual Framework . . . 19
3.2.1 Operator Expenses (OE) . . . 19
3.2.2 Operator Robustness (OR) . . . 20
3.2.3 Passenger Efficiency (PE) . . . 20
3.2.4 Passenger Reliability (PR) . . . 22
3.2.5 Conceptual Framework . . . 24
3.3 Simulation Model . . . 25
3.3.1 Layover Time, Initial Delay and Delay Propagation . . . 27
3.3.2 Cancellation Model . . . 28
3.3.3 Holding Model . . . 29
3.3.4 Simulation Runs . . . 29
3.4 Timetable Scenarios . . . 30
3.5 Comparing the Alternatives . . . 31
3.5.1 MCDA formulation . . . 31

4 Case Study **33**
4.1 Description of the Case Study . . . 33

5 Experimental Setup **36**
5.1 Sample Selection . . . 36

5.2 Case-Specific Indicator Formulation . . . 38
5.2.1 Operator Expenses (OE) . . . 38
5.2.2 Operator Robustness (OR) . . . 38
5.2.3 Passenger Efficiency (PE) and Passenger Reliability (PR) . . . 39
5.3 Alternative Scenarios . . . 39
5.3.1 Scenarios with Altered Layover Time . . . 40
5.3.2 Scenarios with Altered Trip Times . . . 43
5.3.3 Scenarios with Altered Frequency . . . 43
5.4 Application of the Simulation Model . . . 44
5.4.1 Simulation Runs . . . 44
5.4.2 The distributions for real vs. simulated data . . . 46
5.4.3 Validation of Assumptions Regarding Initial and Propagated Delay . . . 49
5.4.4 Case-Specific Cancellation Model . . . 51
5.4.5 Case-Specific Holding Model . . . 51
5.5 Weighting Recommendations for MCDA Regarding the Case . . . 53

6 Results **54**
6.1 Line 9 . . . 54
6.1.1 Individual Indicators - Line 9 . . . 54
6.1.2 Results with Holding Applied . . . 58
6.1.3 Merged Indicator - Line 9 . . . 60
6.1.4 Results of Sensitivity Analysis with Holding Applied . . . 62
6.2 Line 19 . . . 62
6.2.1 Individual Indicators - Line 19 . . . 62
6.2.2 Merged Indicator - Line 19 . . . 64
6.3 Line 4 . . . 66
6.3.1 Individual Indicators - Line 4 . . . 66
6.3.2 Merged Indicator - Line 4 . . . 67
6.4 Comparing the Results for Different Lines . . . 69

7 Discussion **70**
7.1 Interpretation of the Results . . . 70
7.2 Comparison with the Literature . . . 71
7.3 Applicability of the Study . . . 72
7.4 Methodological Limitations . . . 73
7.5 Practical Implications . . . 75
7.6 Conclusion . . . 75

8 Conclusion **77**
8.1 Research Overview . . . 77
8.2 Main Findings . . . 77
8.3 Research Contribution . . . 78
8.4 Limitations and Recommendations for Future Research . . . 79
8.5 Conclusion . . . 79

Appendices **82**

A Block diagrams **83**

B Distributions of PR across alternatives **94**

C Empirical Distributions **99**

D Examples for Effect of Variables on Cost Elements **102**

Chapter 1

Introduction

1.1 Research Background

Public transport systems are the backbone of modern urban mobility, providing access for millions of people to employment, education, and social activities. These systems are critical for reducing traffic congestion, reducing carbon emissions, as well as providing mobility for people with limited access to private transport.

However, public transport operations are increasingly affected by uncertainty. Operational uncertainties stem from various sources, such as severe weather, interactions with other traffic, and rolling stock failures. Recent reports (NLTimes (2025)) show that operators face growing operational challenges, including budget constraints that threaten service levels and increase fares, as well as disruptions caused by infrastructure strain and staffing shortages. These factors complicate the provision of predictable services and risk negatively affecting passenger experience. Thus, maintaining service quality while ensuring operational and financial stability remains challenging.

To deliver its services, public transport planning is commonly divided into three interconnected layers (van Oort (2011); Liu et al. (2021)). At the strategic level, the network is designed, using ridership, budget and geographical characteristics as input. At the tactical level, the timetable is created, resulting in coherent schedules of trips, vehicle tasks, and crew duties. However, uncertainties and disturbances mean that operations do not necessarily adhere to the plan, leading to delays and disruptions. At the operational level, real-time monitoring and control measures are applied to cope with deviations from plans, such as vehicle holding or stop-skipping. These deviations are related to service variability, defined by van Oort (2011) as the distribution of output values (e.g., travel times or headways) from the supply side. Although this variability manifests itself during operations, it is strongly influenced by earlier planning and design decisions (van Oort (2011)). Thus, timetabling directly influences operational stability. Consequently, timetable quality indicators can be defined to quantify how effectively a scheduled plan utilises its resources and responds to unexpected operational disturbances.

In this context, the concepts of reliability and robustness are frequently used to evaluate public transport performance under uncertainty. In the literature, service reliability can be defined as the degree to which service characteristics, such as travel time (including waiting time), arrival time, and seat availability, are predictable relative to the timetable (van Oort (2016)). Iida and Wakabayashi (1989) define reliability as the probability that a transport service will perform a required function under given environmental and operational conditions for a stated period of time. On the other hand, robustness refers to a system's capacity to withstand adverse events or disruptions, absorbing disturbances with minimal impact on performance (Ge et al. (2022); Cats (2016)). Consequently, this study positions reliability as a passenger-focused measure of predictability, whereas robustness represents the system's capacity to resist disruption.

The effects of disturbances on passengers and operators can be described through a set of indicators for reliability and robustness. These effects are not the same, and passenger impacts are sometimes under-

estimated when described through traditional indicators, which mainly refer to the supply side. Studies that focus on the passenger perspective consistently demonstrate the central role of reliability in service evaluation. Regional surveys in European contexts show that passengers place high value on reliable departure times, clear travel information, and comfort, while crowding and missed connections reduce satisfaction (Departement Mobiliteit en Openbare Werken Vlaanderen (2025)). Empirical evidence further suggests that passengers perceive delays disproportionately negatively compared with scheduled in-vehicle travel time. For example, a stated choice experiment by Klinkhardt et al. (2025) finds that mean delay is valued 3.7 times more negatively than scheduled in-vehicle travel time. Similarly, Soza-Parra et al. (2022) show that reliability (in terms of predictability, thus accepting more planned travel time rather than unpredictable excess travel time) significantly influences mode choice and that its valuation varies across users and contexts. Thus, passenger experience and satisfaction influence willingness to use the service.

At the same time, robustness is particularly relevant from the operator perspective, as deviations from the schedule can lead to financial penalties and operational inefficiencies. Furthermore, studies show that some reliability-improving measures, such as adding slack time, can increase passenger travel times and operational costs (Mazloumi et al. (2012)). Therefore, the 'efficiency' of the timetable is also an important consideration, taking into account both operational costs and passengers' planned travel time costs.

Taken together, these aspects highlight the importance of incorporating passenger impacts alongside operational and financial considerations. Moreover, scheduling tools like HASTUS (8.1) allow planners to come up with multiple timetable alternatives; Choosing the optimum timetable requires knowledge of quantified effects of each timetable scenario on passengers and the operator. This further emphasises the need for an evaluation tool that accounts for both passenger and operator perspectives in timetable evaluation, assisting in choosing the best timetable given the priorities.

1.2 Research Problem

This section identifies the research problem in relation to the existing literature and highlights the gap that this study seeks to address. Existing research related to timetable quality addresses several important themes, including passenger-oriented reliability assessment, delay propagation, and the trade-off between robustness and efficiency. Studies emphasise that operator decision-making frequently relies on supply-side indicators such as punctuality and schedule adherence. However, these vehicle-based measures do not necessarily reflect passenger-experienced delays, such as additional waiting time, missed transfers, or increased in-vehicle time (van Oort (2011); Parbo et al. (2016)).

At the same time, a substantial body of work examines slack allocation, delay propagation, and the trade-off between robustness and efficiency. Improving reliability may come at the cost of reduced efficiency, as measures such as adding slack time increase scheduled travel times under undisturbed conditions (Lee et al. (2017)).

Despite these contributions, timetable design remains a complex decision-making problem, as it involves selecting among alternative configurations that affect multiple, and often competing, performance dimensions. Changes in timetable design, such as adjustments to runtimes, frequency, or vehicle deployment, do not influence passenger and operator outcomes in the same way, and improvements in one dimension may come at the expense of another.

However, these aspects are rarely assessed simultaneously within a unified evaluation approach. Without a comprehensive framework that takes into account competing perspectives, the effects of timetable design choices across different performance dimensions remain difficult to assess and compare, potentially leading to decisions that favour one objective while overlooking impacts on others.

Therefore, the research gap addressed in this study is the absence of an integrated evaluation framework

that captures passenger-experienced reliability and efficiency alongside operator-oriented performance measures at the timetable level. This study addresses this gap by developing a data-driven multi-criteria decision-making framework for timetable evaluation that explicitly reveals the trade-offs between passenger and operator objectives and supports more balanced decision-making.

1.3 Research Objectives

This research aims to evaluate timetable quality, by incorporating the passenger perspective while considering operational feasibility. To achieve this, the study develops timetable quality indicators for both passenger (demand) and operator (supply) perspectives, where each perspective is assessed along two dimensions. These indicators are: passenger reliability, passenger efficiency, operator robustness, and operator expenses. More specifically, passenger reliability measures service performance under disrupted conditions, while passenger efficiency relates to travel under undisturbed conditions. These indicators are defined in this way to account for the fact that design choices can affect these two conditions differently, as mentioned in section 1.2. Because the effects for passengers are also different among them, what is efficient/reliable for one passenger may not be the case for another. Therefore, passenger indicators also consider the proportion of passengers affected at each stop, segment, and regime; the detailed methodology is explained later in chapter 3. From the supply side perspective, robustness ensures that service levels remain acceptable by taking the penalties for underperformance into account, thus relating to the disturbed (realised) condition, while efficiency ensures that costs of the improvements are controlled, associating with the planned condition.

These indicators are then used to test and compare alternative timetable configurations within a multi-criteria decision analysis framework, revealing trade-offs. The passenger and operator indicators are aggregated separately, allowing customised weightings based on stakeholder priorities. The study is applied to the context of the urban public transport network of HTM in The Hague region, The Netherlands.

1.4 Research Scope

This research focuses on developing performance indicators for assessing timetable quality, analysing a set of alternative scenarios using empirical disturbance distributions from operational data, and applying these indicators to compare outcomes. Broader modelling efforts, such as, microsimulation, infrastructure analysis, or crew scheduling, fall outside the scope of this project. Limitations are discussed further in detail in section 7.4. The data used in this project is the realised events from (processed) AVL (8.2) and AFC (8.3) data, provided by HTM.

Figure 1.1 shows an overview of the framework used in this study.

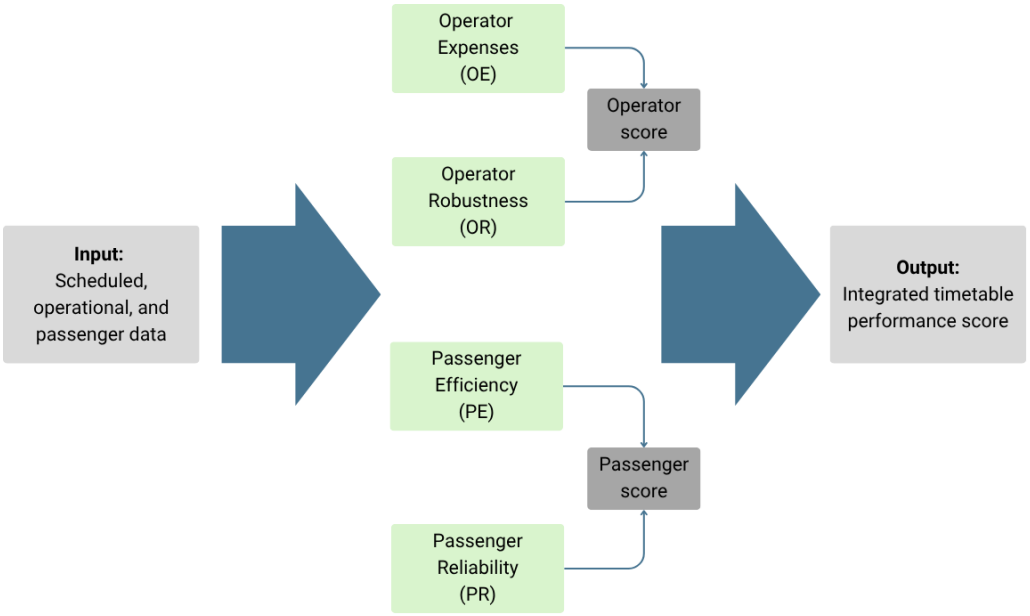


Figure 1.1: Overview of the study framework

As shown, the inputs consist of scheduled, operational, and passenger data, which are used to develop indicators from operator and passenger perspectives. The output is then the combined indicators, resulting in an integrated timetable performance score for comparing timetables.

The study starts with analysing Line 9 of HTM tram network (Vrederust - Scheveningen Noord), on which the methodology is tested and results are obtained, and then the framework is applied to Line 19 and 4 in the tram network. More details on the case study are provided in Chapter 4. Regarding the time interval of the data analysis, a homogeneous time period in which a certain timetable scenario was in use is chosen and it is ensured that there will be enough data to be used for simulation; more explanation on this will be given at subsection 5.1.

1.5 Research Questions

Main research question: To what extent do different timetable scenarios differ in quality, as evaluated using a composite data-driven indicator comprising passenger- and operator-oriented reliability, efficiency, robustness, and expense metrics under empirical disturbance conditions?

Sub-questions:

1. How are reliability/robustness and efficiency/expenses indicators quantified under the base scenario from the passenger and operator perspectives using planning and operational data?
2. How are operational data simulated under timetable scenarios for which no realised data exist?
3. How do timetable adjustments - changing the layover time, runtimes and frequency - affect timetable quality when evaluated through these indicators?
4. How do different indicator weightings influence the comparative evaluation of timetable scenarios?

1.6 Research Structure

This section explains how the research questions are structured and how they build on each other to address the research objective. The first sub-question quantifies the indicators for passenger reliability, passenger efficiency, operator robustness, and operator expenses. The second sub-question simulates the empirical distribution of disturbances on the alternative scenarios. The third sub-question applies these indicators to alternative timetable scenarios, enabling comparison across scenarios. The fourth sub-question examines how different weighting schemes affect the comparative evaluation of the scenarios. Together, these sub-questions answer the main research question by enabling a data-driven evaluation and comparison of timetable alternatives under different priority settings.

Chapter 2

Literature Review

This chapter provides an overview of the literature reviewed in this study, with the goal of establishing the theoretical foundation for the research and identifying the research gap that this study aims to address. First, an overview of the main themes is given. The main themes addressed include public transport reliability, particularly from a passenger perspective, the balance between robustness and efficiency, and operational costs in public transport systems. Then, the research gap and how it is addressed in this study are discussed and the chapter is concluded.

The review is based on keyword searches and snowballing using ScienceDirect, Google Scholar, and the TU Delft library. The search used terms such as ”public transport AND reliability AND passenger”, ”timetable AND robustness AND efficiency”, and ”passenger-oriented AND timetabling”. Relevant studies were selected by screening abstracts; for example, infrastructure-focused or behaviourally intensive studies were excluded. In addition, many studies are relevant to heavy-rail contexts rather than urban transport. When using these studies, the concepts of reliability, robustness, and efficiency were used, while it was noted that some part of the modelling efforts are different in those contexts, requiring extra constraints. Table 2.1 reflects a few example search results using the above-mentioned keywords.

Table 2.1: Literature search strategy and example results

Search Keywords	Example Papers
public transport AND reliability AND passenger	• van Oort (2011). *Service reliability and urban public transport design*.
public transport AND robustness	• Ge et al. (2022). *Robustness and disturbances in public transport*.
timetable AND robustness AND efficiency	• Khoshniyat and Peterson (2017). *Improving train service reliability by applying an effective timetable robustness strategy*. • Lee et al. (2017). *Balance of efficiency and robustness in passenger railway timetables*. • Zhou et al. (2025). *Robustness-based approach for slack time optimization in tram timetables*.
passenger-oriented AND timetabling	• Parbo et al. (2016). *Passenger perspectives in railway timetabling: a literature review*.
public transport AND reliability AND preferences	• Soza-Parra et al. (2022). *Public transport reliability across preferences, modes, and space*.

2.1 Themes and Methodology from Literature

This section provides an overview of the main themes found in the literature related to this study, and overviews how the methodologies that are used here are built on the existing studies.

2.1.1 Passenger-oriented Approach in Timetabling

Although many existing studies focus on optimising timetables, they predominantly adopt an operational perspective; that is, they assess timetable performance using indicators that reflect the outcomes from the supply side rather than the consequences experienced by passengers. Unreliable services result in longer and more uncertain travel times and lower seating availability for passengers (van Oort (2014)). These consequences, however, can strongly influence passenger satisfaction, and their perception of the travel mode, which impacts whether travellers continue to use public transport or switch to other modes. In addition to ridership impacts for operators, such shifts also have societal and environmental implications, as increased use of less sustainable modes can lead to environmental degradation, congestion, and related externalities.

Passenger impacts, however, are more complex to model because they involve individual choices and behavioural assumptions. According to van Oort (2011), traditional supply-side reliability measures (e.g., punctuality) often fail to capture the actual passenger impact of service variability that stems from uncertainties in operations and leads to unreliable service for passengers, resulting in extra travel time. This demonstrates that indicators such as additional travel time per passenger provide a more accurate representation of reliability from the user perspective. A broad literature review by Parbo et al. (2016) reinforces this gap between operator-oriented and passenger-oriented planning. They note that passenger delays are often significantly larger than the train delays that cause them, especially due to missed transfers, uneven passenger loads, and behavioural adaptations such as route or departure-time changes. These findings highlight the importance of considering the passenger perspective in timetable evaluation, where design decisions affect travel time and reliability as experienced by passengers.

2.1.2 Efficiency and Robustness Balance

A recurring theme in the literature is the fundamental trade-off between timetable robustness and operational efficiency. Robustness is the system's ability to withstand disturbances and disruptions with minimal impact on performance (Ge et al. (2022)). However, measures that improve robustness often lead to increased operational costs and planned passenger travel times (Mazloumi et al. (2012)). For instance, Lee et al. (2017) examined timetables with up to approximately 13% total supplement time in a real-world railway case study. They showed that adding time supplements increases robustness at the expense of longer scheduled travel times, although the optimal amount depends on the desired balance between efficiency and robustness. Friedrich et al. (2018) extended this line of work by presenting a simulation-based framework that evaluates public transport plans using operating cost, passenger travel time, and passenger-oriented robustness indicators derived from delay-propagation tests. Taken together, these examples illustrate the need for frameworks that explicitly assess the balance between robustness and efficiency. Focusing on one aspect in isolation risks overlooking how timetable adjustments may increase costs and travel times while yielding limited or no reliability gains. This further emphasizes the need for a multi-criteria framework, which is discussed in Section 2.1.4.

2.1.3 Operational Reliability and Operational Efficiency

Operational costs are an important constraint in timetable design, and reliability gains can come at the cost of increased financial pressure. On the other hand, unreliability of the operations can induce extra costs, as deviations from planned timetable can result in fines for the operator (Yap and van Oort (2018)).

2.1.4 Scenario Evaluation and Multi-Criteria Decision Analysis (MCDA)

Multi-Criteria Decision Analysis (MCDA) offers a structured approach for comparing alternatives which might have contradicting objectives. Nalmpantis et al. (2019) illustrate how such methods can integrate diverse evaluation criteria into a single comparison framework, enabling clearer decision-making in multi-objective contexts. Sensitivity analysis is a central component because results often depend strongly on criteria weights and model assumptions. Wiéckowski and Sałabun (2023) show in a review of more than 250 studies that small changes in input parameters such as criteria weights can lead to changes in the ranking of alternatives.

MCDA is considered to be a suitable approach for this study, because timetable evaluation inherently involves multiple, often conflicting criteria. In transport planning, decision-making commonly requires balancing objectives that cannot be optimised simultaneously (Broniewicz and Ogrodnik (2025)). Its ability to combine different perspectives in one framework, together with sensitivity analysis, makes it useful for this study.

2.1.5 Methodology Overview from Literature

Indicators Definition

The development of the indicator definitions is mainly based on the literature. Passenger reliability is quantified as the extra travel time experienced by passengers due to service unreliability, following van Oort (2011). In this study, reliability is limited to excess waiting time and in-vehicle time, weighted by the proportions of passengers boarding at stops and travelling along segments. Although missed transfers and the resulting additional travel time can also affect reliability, they are kept outside the scope of this study due to the complexity of the modelling. Furthermore, waiting time is generally perceived as more burdensome than in-vehicle time. Wardman et al. (2026) find, across various studies, an average weight of 1.7 for waiting time relative to in-vehicle time. Existing works in the literature (Section 2.1.2) have addressed the trade-off between travel time efficiency and robustness, where additional timetable slack can mitigate the effects of delays. Therefore, the passenger efficiency indicator is developed based on the rationale that increasing planned trip times may make the timetable less fragile, but also increases scheduled travel times.

The operator-side indicators are built based on the cost and robustness dimensions of operations, with operator efficiency representing direct operational costs and operator robustness quantified through the financial penalties associated with punctuality and trip cancellations, as a more robust timetable is expected to reduce the occurrence of these events. A more detailed description of the indicator quantification is provided in Chapter 3, while the case-study-specific assumptions are presented in Chapter 4.

Empirical Simulation

If a timetable was not in operation before, realised data do not exist. Thus, an empirical simulation is applied (Friedrich et al. (2018)), although without the infrastructure constraints that would be required for railway contexts. The empirical distributions of driving times and terminal departure times are used as inputs to the simulation (van Oort (2011)).

When scheduled parameters are modified, the actual driving times are still sampled from the empirical distributions derived from the realised data of the base timetable. This means that, for instance, increasing scheduled trip times increases the likelihood of vehicles operating ahead of schedule, potentially resulting in more early departures. Holding is a widely used operational control strategy for regulating public transport services and remains popular because of its simplicity of implementation (Laskaris et al. (2021)). In holding strategies, vehicles that arrive early at designated control points are required to wait until their scheduled departure time, preventing early departures and improving schedule adherence (van

Oort et al. (2012)). However, holding introduces a trade-off between reliability and efficiency. While preventing early departures reduces passenger waiting time caused by unreliability, it may increase in-vehicle travel time for passengers already on board. (Furth and Muller (2009)) describes holding as a trade-off between reliability and travel speed, showing that increased holding can improve schedule reliability while simultaneously increasing passenger riding time and operational costs.

Individual Indicators Response to Alternative Timetables

In order to observe the effects of timetable design decisions on the selected indicators, alternative scenarios are generated by varying layover time, trip time, and frequency. These variables were selected because they are among the principal timetable design parameters available to planners.

Layover time is defined as the buffer or recovery time provided at the terminal, which is included in the timetable to cope with deviations and improve reliability (van Oort (2011); Sindzingre (2019)). In practice, it can also serve operational purposes such as delay recovery, crew changes, or break time. Layover time and trip time influence the amount of timetable slack available to absorb disturbances and improve reliability ((van Oort (2011); van Oort et al. (2012); Lee et al. (2017))), while frequency affects passenger waiting times, service attractiveness, fleet requirements, and operational costs. Individual Indicators then can be computed for these alternatives and can be compared.

Merging the Indicators with MCDA

Finally, MCDA is a widely used approach for evaluating scenarios (Section 2.1.4) and is therefore used in this study for timetable alternative evaluation. To determine suitable indicator weightings, values reported in the literature are considered. Asensio and Matas (2008) state that travel time variability is valued 2.4 times more negatively than travel time savings on average. Similarly, Klinkhardt et al. (2025) find that mean delay is valued 3.7 times more negatively than scheduled in-vehicle travel time. Furthermore, Preston et al. (2009) report that, in the British rail context, one minute of lateness is typically valued by passengers as equivalent to approximately three minutes of scheduled journey time.

Although such valuations vary with transport mode, trip purpose, passenger characteristics, and the reliability measure used, these studies consistently indicate that reliability is valued more highly than travel time savings. Therefore, a weight of 3 for passenger reliability relative to passenger efficiency (after normalisation) is adopted as a representative value. To account for uncertainty in these valuations, a range of alternative weights is also tested through sensitivity analysis.

2.2 Literature Gap

This section presents the research contribution of the study. Many current studies primarily focus on the operator perspective, with a stronger emphasis on railways rather than urban transit systems. As mentioned earlier, timetabling decisions inherently involve multiple and often conflicting objectives, where the goal is not the optimisation of a single objective but a balanced consideration of various dimensions.

This research addresses a methodological gap by proposing a decision-support framework for timetable evaluation that enables transparent comparison of timetabling scenarios and explicit assessment of how costs and benefits are distributed across passenger- and operator-related objectives. It characterises timetable quality through a set of indicators that capture the relationship between passenger and operator priorities within an MCDA framework, which has seldom been applied to public transport timetabling, since existing MCDA applications in public transport tend to focus on high-level planning or strategic assessments and do not incorporate detailed timetable-based indicators. By employing sensitivity analysis in a multi-criteria setting, alternative design choices can be examined under different priority structures.

Table 2.2 presents the key literature grouped into three themes. Theme 1 covers passenger-oriented approaches to timetabling. Theme 2 includes studies applying multi-criteria optimisation, mainly addressing trade-offs between robustness (from a traditional perspective) and travel time efficiency and/or operational costs. Theme 3 addresses the effects of service quality on passenger satisfaction (and ridership). The conclusions highlight the most relevant insights from the reviewed studies. The table also indicates whether each study examines an urban transit system, as methods and findings may differ from those in heavy-rail contexts. The aim of this study is to bring these perspectives together and provide an integrated view of these themes.

This chapter reviewed the key literature on passenger-oriented reliability, the trade-off between robustness and efficiency, operational costs, and the application of MCDA in public transport planning, as well as the gap this study aims to address, providing a foundation for the analysis that follows in chapter 3.

Table 2.2: Literature Review Summary

Title; Authors (year)	**Urban Transit**	**Theme 1: Passenger-oriented Timetabling**	**Theme 2: Efficiency and Robustness Balance**	**Theme 3: Service Improvement and Ridership**	**Main Conclusions**
Service Reliability and Urban Public Transport Design; van Oort (2011)	✓	✓		✓	Traditional supply-side reliability measures can misrepresent passenger experience. The indicator "average additional travel time per passenger" provides a more accurate, passenger-oriented reliability measure. Tactical instruments can reduce additional travel time and are generally cost-effective.
Incorporating service reliability in public transport design and performance requirements: International survey results and recommendations; van Oort (2014)	✓	✓			Traditional practice lacks consistency in defining and measuring service reliability, with indicators largely focused on vehicle performance rather than passenger impacts. Different measurement methods yield non-comparable results, leading to misleading conclusions; a consistent, passenger-oriented indicator such as additional travel time is therefore recommended.
Data-driven Public Transport Ridership Prediction Approach; van Oort et al. (2015)	✓			✓	An elasticity-based approach enables a practical short-term demand prediction that captures traveller responses to service changes; generalised travel time cost affects the ridership.
Passenger Perspectives in Railway Timetabling: A Literature Review; Parbo et al. (2016)		✓			Passenger delays often exceed train delays due to missed transfers and crowding. Passengers place high value on reliability, and severe delays impose disproportionately high disutility.
Balance of Efficiency and Robustness in Passenger Railway Timetables; Lee et al. (2017)			✓		Timetables can be improved with small slack reallocations, reducing delay propagation. Time supplements are more effective than buffer times in dense networks. There is a clear efficiency–robustness trade-off, and the optimal slack level depends on disturbance levels.
Improving Train Service Reliability; Khoshniyat and Peterson (2017)			✓		Small timetable refinements can improve operational robustness by limiting delay propagation.
Robustness as a Third Dimension for Evaluating Public Transport Plans; Friedrich et al. (2018)	✓	✓	✓		Using empirical delay for tests provides realistic expected delays; vehicle circulation has a substantial effect on delay propagation; and, as expected, improvements in robustness through slack or waiting strategies increase planned travel time or operating cost.

Title; Authors (year)	Urban Transit	Theme 1: Passenger-oriented Timetabling	Theme 2: Efficiency and Robustness Balance	Theme 3: Service Improvement and Ridership	Main Conclusions
Driver Schedule Efficiency vs Public Transport Robustness; Yap and van Oort (2018)	✓		✓		More complex driver schedules reduce costs in undisrupted operations but greatly increase disruption costs, demonstrating a clear efficiency–robustness trade-off.
Public transport reliability across preferences, modes, and space; Soza-Parra et al. (2022)	✓	✓		✓	Passengers assign substantial disutility to unreliable services, with heterogeneity in preferences highlighting the importance of passenger-oriented reliability measures in planning and evaluation.
Reliability of Public Transportation Systems with Autonomous First-/Last-Mile Services in a Stated Mode Choice Experiment; Klinkhardt et al. (2025)	✓	✓		✓	Reliability improvements through autonomous first-/last-mile services significantly influence mode choice. Passengers value reliability in integrated multimodal chains, and improved reliability increases public transport attractiveness and ridership potential.
Robust Optimisation of Tram Timetables; Zhou et al. (2025)	✓		✓		Slack-time optimisation substantially improves robustness and punctuality; Operating cost constraints limit the feasible slack but has limited influence on robustness.
This study	✓	✓	✓	✓	Integrates passenger reliability, operator costs, robustness and MCDA within one timetable evaluation framework.

2.3 Conclusion

In conclusion, the literature shows that timetable design has a significant influence on reliability and passenger experience. Existing studies consistently demonstrate that improving robustness often requires additional slack time or resources, which creates a trade-off with operational efficiency in terms of costs, as well as increased scheduled travel times for passengers. It can also be concluded that reliability from the passenger perspective differs from reliability from the operator perspective. Despite these insights, the integration of passenger and operator perspectives within a single timetable evaluation framework remains limited. The literature therefore indicates a need for a decision-support approach that evaluates timetable alternatives using multiple criteria, aiming to explicitly balance passenger reliability and travel time efficiency, while ensuring financial feasibility and accounting for the benefits of robustness for operators through the avoidance of penalties and disruption-related costs. This gap forms the basis for the methodology developed in this study, described in Chapter 3.

Chapter 3

Methodology

This chapter describes the methodology used to develop timetable quality indicators and apply them in the evaluation of alternative timetable scenarios. After providing an overview of the methodology, the set of indicators is defined, and the required inputs and variables are related through a conceptual framework. Then, a set of alternative timetables is defined based on variables including layover time, trip time, and service frequency. For the reliability and robustness indicators, realised data is required, which is simulated for the alternative timetables. Subsequently, the methodology and assumptions used in the simulation process are described. Finally, the last section provides an overview of Multi-Criteria Decision Analysis (MCDA), which is adopted to compare the alternatives.

3.1 Methodology Overview

This thesis adopts an integrated empirical (descriptive knowledge) and modelling methodology (Zwart and de Vries (2016)) to analyse the performance of an existing timetable and evaluate how timetable modifications influence both passenger experience and operator feasibility. The workflow consists of connected stages. This corresponds to a molecular methodology (Zwart and de Vries (2016)), in which different methodological cycles support different tasks but together address the main research question.

Stage 1: Empirical Descriptive Analysis answers the first research question: ”How are reliability/robustness and efficiency/expenses indicators quantified under the base scenario from the passenger and operator perspectives using planning and operational data?” This stage follows the empirical cycle, since the focus is on quantifying the indicators used to describe how a system performs.

Stage 2: Scenario Design, Simulation and Modelling answers the second and third research questions: ”How are operational data simulated for timetable scenarios for which no realised data exist?” and ”How do timetable adjustments influence timetable quality when evaluated through these indicators?”. A modelling and simulation approach is considered appropriate because the alternative timetable scenarios have not been operated, meaning no realised operational data are available to use to compute the indicators. This stage includes designing feasible timetable scenarios (using HASTUS), simulating realised operational data for these scenarios, and modelling their expected performance. The scenarios include altering layover time, trip times (driving times), and frequency. Section 5.3 details the modelling process for alternative scenarios. Using scheduled and simulated data, the indicators are computed for each timetable scenario, providing input for the final comparative evaluation model.

Stage 3: Optimisation Cycle – Trade-off Evaluation answers the fourth research question: ”How do different indicator weightings influence the comparative evaluation of timetable scenarios?” by following the optimisation cycle, since the research seeks to evaluate and compare alternative timetable designs based on multiple, potentially conflicting objectives. The indicator values obtained in Stage 2 for base and alternative schedules are normalised and combined using weighting sets that represent different stakeholder perspectives. The aim at this stage is to see how different priorities influence the perceived quality of each timetable modification. The indicator values are normalised and combined using Multi-criteria decision analysis (MCDA) principles to explore trade-offs between conflicting objectives,

acknowledging the sensitivity of outcomes to weighting choices (Wiéckowski and Sałabun (2023)). By doing this, the best design choices can be determined using the given weights. The formulations for this part of the study are depicted in Section 3.5.

3.2 Indicator Definitions and Conceptual Framework

This section illustrates how the quality indicators are defined and calculated. Using actual and scheduled departure times at each stop, a set of variables is computed to develop the main indicators. First, the formulations are described, followed by the conceptual framework that depicts the process and relationships between the variables, from the inputs to the final output of the study, being the timetable performance scores derived from the indicators.

To account for differences in disturbance patterns throughout the day, the analysis period is divided into time-of-day regimes. The precise stratification will be detailed in the case study chapter (4). The indicators are calculated separately for each regime to account for differences in demand and disturbance patterns. The results are then aggregated across regimes to define the final indicators.

However, by adding the corresponding indicators for each line, they can be applied to scheduling units to create metrics for comparing different scenarios including multiple lines.

3.2.1 Operator Expenses (OE)

Operator Expenses reflects the operational costs of a given timetable by capturing the resource requirements needed to operate the timetable.

When modifying a variable in a timetable scenario, the total operational hour (crew and vehicle) and number of required vehicles can be affected. In this study, crew scheduling is not performed. Instead, operational hour is approximated based on vehicle block durations (including the duration of required pull-in and pull-out trips between the depot and the line, opportunity trips, and regular trips). Thus the cost elements taken into account here include hourly costs of operation (energy, drivers), and vehicle costs. The block allocations determine the Peak Vehicle Requirement (PVR) and thus number of required vehicles, as well as altered number and duration of pull-in/out trips. An example of this case is given in Appendix D.

Let B_l denote the set of vehicle blocks assigned to line l. For each block $b \in B_l$, the block duration is defined as the time difference between the first pull-out event and the final pull-in event of that block. The total operating hours (in a day) are therefore computed as:

$$H_l^{oper} = \sum_{b \in B_l} h_b \tag{3.1}$$

where h_b represents the total duration of block b.

Assuming an average operating cost c^{oper} per hour (including crew and energy costs), the operating cost of the service per day is:

$$C_l^{oper} = c^{oper} \cdot H_l^{oper} \tag{3.2}$$

Finally, the required fleet size corresponds to the PVR, defined as the maximum number of vehicles simultaneously in operation. Let V_l denote this value. The vehicle purchase cost is converted to an equivalent daily vehicle cost based on the expected service life of the vehicle:

$$c^{veh,daily} = \frac{c^{veh}}{365 \cdot L} \tag{3.3}$$

where c^{veh} is the vehicle purchase cost and L is the expected service life in years.

The fleet cost component is then defined as

$$C_l^{fleet} = V_l \cdot c^{veh,daily} \tag{3.4}$$

The total operator cost is therefore approximated as

$$C_l^{total} = C_l^{oper} + C_l^{fleet} \tag{3.5}$$

The **Operator Expenses** indicator is defined as the negative of total operational cost:

$$OE_l = -C_l^{total} \tag{3.6}$$

The negative sign is used so that higher values indicate better performance (higher score) in this indicator.

3.2.2 Operator Robustness (OR)

A further component is estimating the effect of improved service performance on financial penalties. Besides the core cost components, which can change across the alternative scenarios, robustness of the timetable also affects operators. As shown by Yap and van Oort (2018), deviations from the planned timetable result in operational costs, including fines for early and late departures. Operators must comply with performance requirements set by the transit authority. These requirements may include measures related to punctuality (early and late departures) and trip cancellations. Penalties can be computed based on rates over a reporting period (e.g., annually or quarterly) and/or per vehicle-kilometre. Punctuality metrics can be an outcome of robustness. This indicator can include the fines that result from unpunctuality and/or service cancellations, which can be mitigated by a more robust timetable, and is addressed here as operator robustness.

Let P_l denote the penalties associated with late departures, P_e the penalties associated with early departures, and P_{canc} the penalties associated with cancellations. The **Operator Robustness** indicator is defined as:

$$OR = -\left(P_e + P_l + P_{canc}\right) \tag{3.7}$$

A negative sign is adopted so that lower penalty costs correspond to higher values of the indicator. The exact formulation is case-specific and is further detailed in Chapter 4.

3.2.3 Passenger Efficiency (PE)

Passenger Efficiency reflects the effective scheduled travel speed experienced by passengers under a given timetable, independent of realised operational variability.

For high-frequency services, passengers are typically assumed to arrive randomly at stops. Studies based on surveys and empirical observations suggest that passenger arrivals tend to be random when headways are around 10 minutes or less, with 10 minutes commonly identified as the threshold between random and non-random passenger arrivals (Fan and Machemehl (2002); van Oort (2011)). Then, if $H_{l,j,t}$ is the

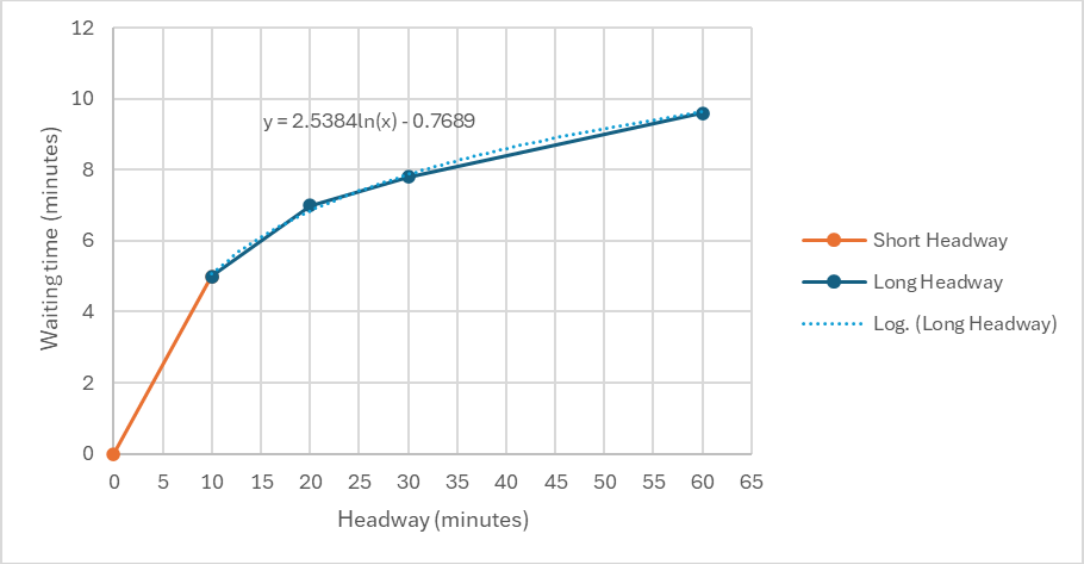


Figure 3.1: Relationship between scheduled headway and average waiting time for scheduled low-frequency services

scheduled headway at stop j for line l and time-of-day regime t, the associated scheduled waiting time $SWT_{l,j,t}$ is calculated as:

$$SWT_{l,j,t} = \frac{H^{\text{sched}}_{l,j,t}}{2} \qquad \text{for } H^{\text{sched}}_{l,j,t} \leq 10 \tag{3.8}$$

This relationship represents the expected waiting time under the scheduled timetable and does not account for service irregularities, as intended by this indicator.

However, for low-frequency services, passengers tend to plan their arrival at stops according to the timetable. Therefore, average waiting time is generally less than half of the scheduled headway. Ingvardson et al. (2018) show that when published schedules are available, average waiting time decreases to approximately 35%, 26%, and 16% of the headway for 20-, 30-, and 60-minute headways, respectively. Together with the assumption that a 10-minute headway corresponds to an average waiting time equal to 50% of the headway, these points were used here to estimate the relationship between scheduled headway and average waiting time using a logarithmic function:

$$SWT_{l,j,t} = 2.5384 \ln(H^{\text{sched}}_{l,j,t}) - 0.7689 \qquad \text{for } H^{\text{sched}}_{l,j,t} > 10 \tag{3.9}$$

where $H_{l,j,t}$ is the headway (minutes).

As stated earlier, days are stratified into different regimes. For line l, scheduled waiting time is computed per stop and per regime. To account for passenger exposure, SWT is weighted by the share of passengers boarding at each stop in each regime. Let $p^{stop}_{l,j,t}$ denote the proportion of passengers boarding at stop j in regime t. By definition,

$$\sum_j p^{stop}_{l,j,t} = 1 \tag{3.10}$$

Then, the passenger-weighted scheduled waiting time for line l is calculated as:

$$\overline{SWT}_l = \sum_t \sum_j p^{\text{stop}}_{l,j,t} \cdot SWT_{l,j,t} \tag{3.11}$$

Scheduled in-vehicle travel time for segment j in regime t for line l is calculated as:

$$SIVT_{l,j,t} = \frac{1}{n_{l,j,t}} \sum_{i=1}^{n_{l,j,t}} IVT^{\text{sched}}_{l,i,j,t} \tag{3.12}$$

Again to reflect passenger experience, SIVT is weighted by the proportion of passengers travelling on each segment (pair of stops) in each regime. Let $p^{seg}_{l,j,t}$ denote this proportion. By definition,

$$\sum_j p^{seg}_{l,j,t} = 1 \tag{3.13}$$

The passenger-weighted scheduled in-vehicle travel time is then:

$$\overline{SIVT}_l = \sum_t \sum_j p^{\text{seg}}_{l,j,t} \cdot SIVT_{l,j,t} \tag{3.14}$$

total travelled distance for line l is calculated as:

$$\overline{DIST}_l = n_{trips} \cdot \sum_j d_{l,j} \tag{3.15}$$

where $d_{l,j}$ denotes the length of segment j per direction and n_{trips} is the number of trips per direction.

In order to account for the effect of increased frequency on waiting times, average waiting times are used. If the total waiting time is used, the waiting time would remain unchanged, since although more trips operate within a given time period, the total accumulated waiting time remains similar while the average waiting time experienced by passengers decreases. On the other hand, passenger-kilometres travelled are aggregated using summation, such that higher frequencies result in larger total travelled distances. Finally, Passenger Efficiency for line l is defined as the ratio between passenger-weighted travelled distance and passenger-weighted generalised scheduled travel time. Scheduled travel time includes scheduled waiting time and in-vehicle time, with waiting time weighted 1.7 times more heavily than in-vehicle time, in line with the average value reported by Wardman et al. (2026):

$$PE_l = \frac{\overline{DIST}_l}{\overline{SIVT}_l + 1.7 \cdot \overline{SWT}_l} \tag{3.16}$$

Higher values of PE_l indicate more efficient passenger movement in terms of effective scheduled travel speed.

3.2.4 Passenger Reliability (PR)

This indicator is built according to van Oort (2011), where service reliability is measured by the additional travel time experienced by passengers. In general, additional travel time mainly consists of excess waiting time, excess transfer time stemming from missed transfers, and excess in-vehicle time, due to service variability. In this study, transfers are kept out of scope (more detail in 7.4), thus only waiting time and in-vehicle time are included. For each simulation run, the indicators are calculated and then averaged over all runs.

Additional Waiting Time

The additional waiting time (AWT) depends on passenger arrival behaviour at stops. As stated in section 3.2.3, a headway of 10 minutes is taken as the threshold below which services are considered high-frequency and passengers tend to arrive randomly. Under this assumption, expected passenger waiting time depends on both the mean headway and the variability of actual headways. The expected waiting time at stop j for line l and time-of-day regime t can be expressed as:

$$E\left(T_{l,j,t}^{waiting}\right) = \frac{E\left(\tilde{H}_{l,j,t}^{act}\right)}{2}\left(1 + CoV^2\left(\tilde{H}_{l,j,t}^{act}\right)\right). \tag{3.17}$$

Where $\tilde{H}_{l,j,t}^{act}$ is the actual (realised) headway for line l, stop j, and time-of-day t. The expected additional waiting time due to headway variability is therefore:

$$E\left(T_{l,j,t}^{add-waiting}\right) = \frac{E\left(\tilde{H}_{l,j,t}^{act}\right)}{2} \cdot CoV^2\left(\tilde{H}_{l,j,t}^{act}\right). \tag{3.18}$$

Which is the same as average additional waiting time:

$$AWT_{l,j,t} = E\left(T_{l,j,t}^{add-waiting}\right) \tag{3.19}$$

In this study, the coefficient of variation of headways is approximated by the Percentage Regularity Deviation from the Mean (PRDM) ($CoV \approx PRDM$), which is defined as:

$$PRDM_{l,j,t} = \frac{\sum_i \left| \frac{H_{l,i,t}^{sched} - \tilde{H}_{l,i,j,t}^{act}}{H_{l,i,t}^{sched}} \right|}{n_{l,j,t}} \tag{3.20}$$

Where $H_{l,j,t}^{sched}$ is the scheduled headway and $n_{l,j,t}$ is the number of services for line l, stop j, and time-of-day t. Substituting $CoV \approx PRDM$ yields:

$$AWT_{l,j,t} = \frac{E\left(\tilde{H}_{l,j,t}^{act}\right)}{2} \cdot PRDM_{l,j,t}^2 \tag{3.21}$$

If no systematic bias in headways is observed in the data, $E(\tilde{H}_{l,j,t}^{act})$ is replaced by $H_{l,j}^{sched}$. Thus:

$$AWT_{l,j,t} = \frac{H_{l,j,t}^{sched}}{2} \cdot PRDM_{l,j,t}^2 \tag{3.22}$$

For lower-frequency services, passengers are assumed to plan their arrival according to the timetable. In this case, additional waiting time depends directly on the realised departure deviation.

$$AWT_{l,i,j,t} = \begin{cases} H_{l,j,t}^{sched}, & \text{if } \tilde{d}_{l,i,j,t}^{departure} \leq -\tau_{early} \\ 0, & \text{if } -\tau_{early} < \tilde{d}_{l,i,j,t}^{departure} < \tau_{late} \\ \tilde{d}_{l,i,j,t}^{departure}, & \text{if } \tilde{d}_{l,i,j,t}^{departure} \geq \tau_{late} \end{cases} \tag{3.23}$$

where τ_{early} and τ_{late} denote the early and late departure thresholds.

As stated earlier in 3.2.3, here also to account for passenger exposure, AWT is weighted by the share of passengers boarding at each stop in each regime. Let $p^{stop}_{l,j,t}$ denote the proportion of passengers boarding at stop j in regime t. By definition,

$$\sum_{j} p^{stop}_{l,j,t} = 1 \tag{3.24}$$

The passenger-weighted additional waiting time for line l is

$$\overline{AWT}_l = \sum_{t} \sum_{j} p^{stop}_{l,j,t} \cdot AWT_{l,j,t} \tag{3.25}$$

Additional In-Vehicle Time

Additional in-vehicle travel time (AIVT) is computed at the segment level along a given line. For each vehicle i, the scheduled and actual travel times are measured between consecutive stops.

Let $IVT^{act}_{l,i,j,t}$ and $IVT^{sched}_{l,i,j,t}$ denote the realised and planned in-vehicle travel time of vehicle i on segment j of line l in regime t, respectively. The additional in-vehicle travel time is defined as:

$$AIVT_{l,j,t} = \frac{1}{n_{l,j,t}} \sum_{i=1}^{n_{l,j,t}} \left(IVT^{\text{act}}_{l,i,j,t} - IVT^{\text{sched}}_{l,i,j,t} \right) \tag{3.26}$$

AIVT is weighted by the proportion of passengers travelling on each segment in each regime. If $p^{seg}_{l,j,t}$ denote this proportion, by definition,

$$\sum_{j} p^{seg}_{l,j,t} = 1 \tag{3.27}$$

The demand-weighted additional in-vehicle time for line l is

$$\overline{AIVT}_l = \sum_{t} \sum_{j} p^{seg}_{l,j,t} \cdot AIVT_{l,j,t} \tag{3.28}$$

Finally, the Passenger Reliability indicator is defined as follows, including excess waiting time and in-vehicle time, with waiting time weighted 1.7 times more heavily than in-vehicle time (also in 3.2.3):

$$PR_l = -\left(1.7 \cdot \overline{AWT}_l + \overline{AIVT}_l\right) \tag{3.29}$$

A negative sign is adopted so that, after normalisation, higher values indicate better performance, thereby enabling consistent comparison across all indicators.

3.2.5 Conceptual Framework

In order to obtain the final output of this study, namely the merged indicator obtained by combining the four quality indicators whose calculation was explained earlier, the conceptual framework depicted in Figure 3.2 is used. This figure illustrates the relationships between the input data, the intermediate variables, the quality indicators and the final output. The input consists of AVL data, AFC data, as well as timetable alternatives (schedules), with the main variables being layover time, trip time, and frequency (blue boxes). Through a set of intermediate variables (yellow boxes), the main indicators

are defined and combined using MCDA (where w1, ..., w4 represent the weights) to build a score for each timetable, which forms the output of the framework. For the base scenario, both scheduled and realised data are available. The realised data (from AVL) are used alongside the scheduled data to build empirical distributions of driving times and departure time variability, which are then applied to the new schedules, enabling the calculation of the indicators of robustness (supply-side) and reliability (demand-side). The scheduled timetable is sufficient for computing the operator expenses, since it is related to costs that directly come from the operational hours and fleet requirements (An example for this is given in Appendix D). For passenger efficiency and passenger reliability, the AFC data is used as well, to account for the demands at the stops, segments within each pair of stops, and time-of-days. The indicator values can then be compared with the base scenario and the other alternatives. Finally, these indicators are combined using multi-criteria decision analysis (MCDA) to derive a score for each timetable and enable overall comparison. A range of weightings is applied to perform sensitivity analysis.

In relation to the research questions, the definition and calculation of the indicators using AVL, AFC, and scheduled data address sub-question 1. The simulation of realised operational data for alternative timetable scenarios addresses sub-question 2. The evaluation of timetable adjustments, including changes in layover time, trip time, and frequency, addresses sub-question 3. Finally, the application of MCDA with varying indicator weights and sensitivity analysis addresses sub-question 4, ultimately contributing to answering the main research question.

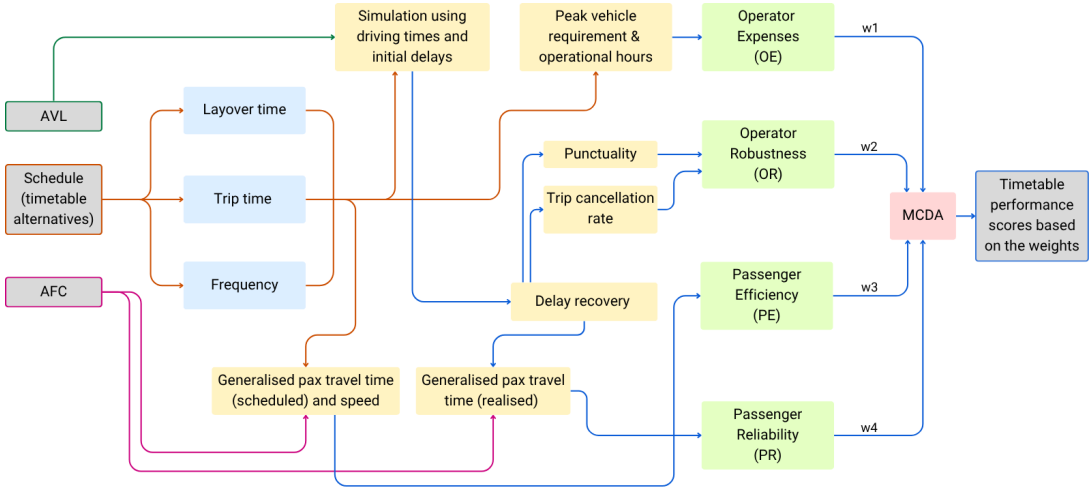


Figure 3.2: Theoretical framework linking timetable design and performance indicators.

3.3 Simulation Model

As described in the previous sections, for the passenger reliability and operator robustness indicators, realised data is needed. Since for the alternative scenarios actual data does not exist, realised events are simulated. For this, it is needed to know what source delays must be applied and how delays are propagated to the trips, as well as the cancellations that would occur in those trips. Given the distribution of source delays at terminals, as well as the driving times between each pair of stops, the disturbances can be simulated, and the actual events can be constructed (including the departure time at each stop and whether the event would be cancelled). Regarding driving times, they are sampled from an entire trip in the realised data rather than independently for each segment. This preserves the correlation structure between segment travel times within the same trip. Sampling each segment independently would remove these correlations and could lead to unrealistic trip trajectories, even if the marginal driving time distributions remain unchanged. Lastly, using the same driving time distributions as those observed for the realised data of the base timetable when simulating schedules with altered trip times may lead to a higher occurrence of early departures. Therefore, a holding model is applied to the

simulated departure times to adjust them accordingly. Further details are provided in Section 3.3.3.

The simulation uses realised AVL operational data. First, empirical distributions are constructed for:

- Initial delays at terminal points (more detail in Section 3.3.1);
- Driving times between stops;
- Service cancellations (more detail in Section 3.3.2);
- Holding (more detail in Section 3.3.3).

These empirical distributions are segmented by time-of-day regimes, as described in previous sections. Thus, for each time regime, a distribution is constructed for each of these elements. Using these distributions, realised operations are simulated for each timetable scenario. For each schedule, trips are first grouped based on vehicle blocks, since it is intended to apply the disturbances on trips and study how delays propagate to the next trip operated by the same vehicle. Then, within each block, the trips are sorted based on the order of the trips. By doing this, the disturbances are applied to the trips in order and then propagated to the next trip that a vehicle executes. For the alternatives without blocks being recreated, which are usually the ones with minor alterations, this ensures consistency, meaning that the differences between timetable alternatives result primarily from timetable structure and response to disturbances, rather than from different random realisations.

Additionally, after application of these distributions on the alternative scenarios, it is investigated whether the initial delays and driving times distributions follow the same distribution as the realised data from AVL or not. For this purpose, the Kolmogorov-Smirnov (KS) test is performed to assess whether the simulated and realised distributions differ significantly or not.

As stated earlier, trips are grouped into vehicle blocks. For each block, an initial delay is randomly sampled from the corresponding (time-of-day regime, line) empirical delay distribution, and then is applied to the first stop of the first trip. This determines the simulated actual departure time at that stop. For the remaining stops of the trip, driving times are sampled and used to generate realised departure events. The realised departure time at each stop is then obtained by adding the sampled driving time to the simulated departure time at the previous stop. Consequently, delays can be determined for all stops along the trip, including the final stop.

At the end of the trip, the delay propagation process is evaluated. Based on the available layover buffer time (described in Section 3.3.1), part or all of the arrival delay may propagate to the subsequent trip within the same vehicle block. The next trip may operate in the opposite direction or, in the case of interlining, on a different route. For this subsequent trip, a new initial delay is again randomly sampled from the corresponding empirical distribution. The propagated delay from the previous trip is then added to this sampled delay, resulting in the simulated departure time at the first stop of that trip. Again the driving time sampling procedure is repeated, which generates realised departure events and delays for all stops along the trip. This process continues sequentially until the final trip in the block has been simulated. Finally, the same procedure is then applied to all remaining vehicle blocks.

Once all realised departure times have been generated, the performance indicators can be calculated. These include delays, headway irregularity, punctuality, passenger waiting times, and the other indicators introduced in Chapter 3.

In addition to generating realised departure times, the model also simulates trip cancellations. After the delays have been determined, the cancellation module evaluates each simulated trip. Based on the realised delays and the cancellation rules which will be described later in Section 3.3.2, the model determines whether the remainder of the current trip or a subsequent trip within the same block should be cancelled. If a trip is cancelled, then the corresponding simulated departure times will be irrelevant (acts as a missing realised event). The resulting cancellation decisions are then incorporated into the calculation of the performance indicators.

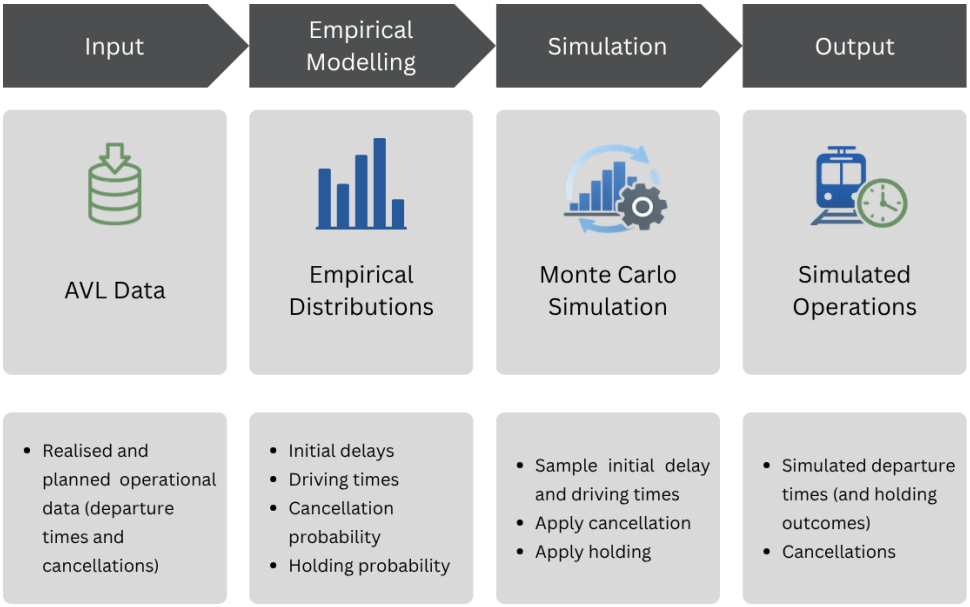


Figure 3.3: Simulation Model

Figure 3.3 provides an overview of the simulation framework adopted in this study. In summary, the process begins with the AVL data, from which empirical distributions of initial delays, driving times, and cancellation probabilities are derived. These distributions are subsequently used in a Monte Carlo simulation to generate operational events, including departure times, propagated delays, cancellations, and holding. The resulting simulated operations form the basis for the calculation of the performance indicators.

3.3.1 Layover Time, Initial Delay and Delay Propagation

Layover time is calculated as the difference between the last scheduled event of a trip and the first scheduled event of the subsequent trip operated by the same vehicle block, excluding any turnaround time (if present). This time can be used to absorb the delays at the end of a trip and prevent propagation to the next trip executed by the same vehicle. This time will give the realised available layover rather than the minimum planned (required) layover from the timetable, since in practice it will be important how much time is available between trips rather than how much minimum was required. Furthermore, labour agreements may require part of the layover time to be allocated to driver activities at terminal stops. This will be referred to as the drivers' activity time at the end point (in Dutch: 'keertijd' 8.5). It is assumed that this time does not contribute to delay absorption. Consequently, the layover buffer time is defined as the available layover between trips minus the 'keertijd'. If the available layover is smaller than the above-mentioned required time, the layover buffer is assumed to be zero. A more detailed modelling approach would require distinguishing between delays propagated from previous trips and newly generated source delays at the start of a trip.

As stated earlier, initial delay is defined as the delay at the start of a trip that is not caused by propagation from the previous trip:

$$\text{Initial Delay} = \text{Start Delay} - \text{Propagated Delay}$$

In addition to initial delays, delay propagation between consecutive trips within the same vehicle block needs to be considered. Propagated delay is defined as the remaining delay after layover time absorbs (part of the) delay from the previous trip, and it is applied to the subsequent trip consistently. Early departures are only allowed if sufficient layover time is available relative to the delay of the previous

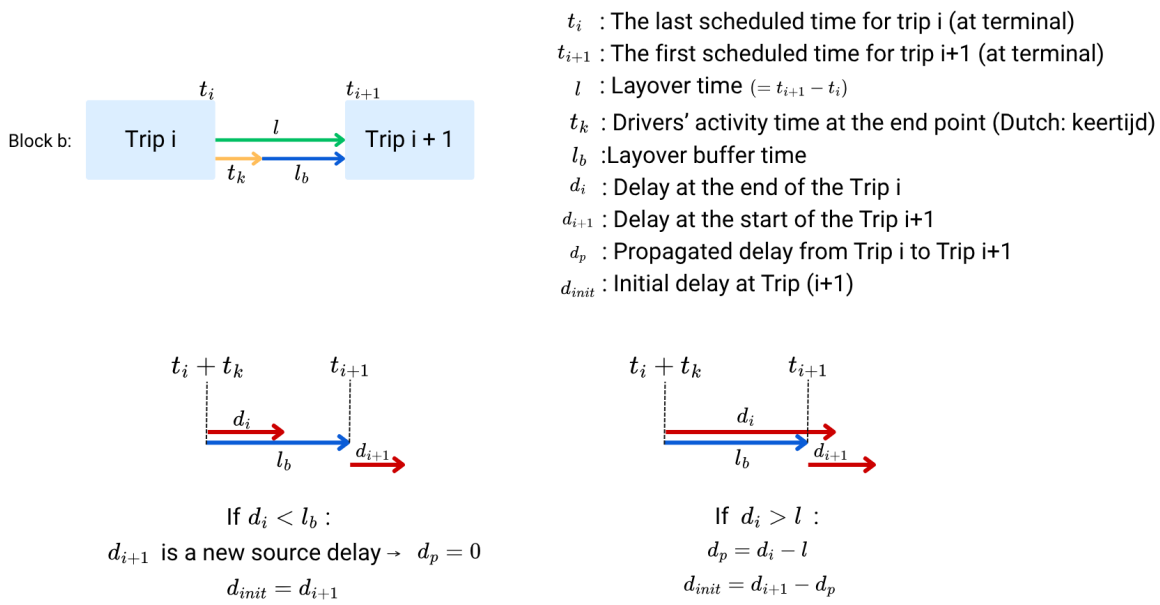


Figure 3.4: Illustration of the relationship between layover time, initial delay, and propagated delay

trip. More specifically, when there is delay at the end of the line, negative initial delays will be allowed based on how much buffer still exists in the layover time. Thus, when there is propagated delay (meaning that the layover buffer time was already used for partly absorbing the delay and the remaining is being propagated) no negative initial delay will be allowed.

Figure 3.4 summarises the above-mentioned definitions and illustrates the relationship between layover time, initial delay, and propagated delay, showing how delays are either absorbed or propagated depending on the available layover buffer time.

3.3.2 Cancellation Model

Besides delays and headway irregularity, service cancellations can represent an important aspect affecting both operators and passengers. Cancellations can affect the operator robustness indicator through financial penalties, while from a passenger perspective, missing services increase waiting times, which may correspond to one or more additional full headways. The assumption made in this study was that if service was cancelled, passengers did not choose for an alternative service or mode, and waited until the next available service.

There can be various reasons behind service cancellations, such as driver unavailability, severe weather conditions, or vehicle failures. Another reason can be excessive delays. In this study, since only delay-related cancellations can potentially be influenced through timetable design, the cancellation model is primarily based on the magnitude of delays. However, the location at which a cancellation occurs can also be relevant, as the decision to cancel a service may depend not only on the delay itself but also on where in the network that delay is observed.

To estimate these relationships, the delay immediately before a missing observation is analysed. Depending on the context, this delay is taken either from:

- the previous stop within the same trip; or
- the previous trip within the same vehicle block.

Delay values are discretised into quantile-based bins. For each bin, the empirical probability that the following observation is missing is calculated.

The number of bins is selected based on predictive performance using log-loss on a train-test split (70%-30%). Log-loss (cross-entropy loss) (Google Developers (nd)) measures how well predicted probabilities align with observed outcomes while penalising confident incorrect predictions more strongly:

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]$$

where $y_i \in \{0, 1\}$ is the observed outcome and p_i is the predicted probability. Lower values indicate better probabilistic predictions.

After choosing the best q, the resulting delay - cancellation relationship is then embedded into the simulation model. At each stop, the current delay as well as the relative location of the stop determines the probability of cancellation, after which a random draw determines whether the remainder of the trip is cancelled. The table below illustrates these probabilities, given the magnitude of the delay and the relative position of the stop.

Since delays must first exist before cancellations can occur, the cancellation process is applied after departure and delay generation within the simulation sequence.

For mid-trip cancellations, propagated delay handling requires additional assumptions. If a trip is cancelled midway, the delay immediately before cancellation still influences the following trip within the same vehicle block. However, this propagated delay is offset using the layover associated with the cancelled trip. This effectively assumes that the vehicle continues in the same direction as before cancellation by deadheading to the terminal. Then the delay that was observed in the last stop before cancellation, is subtracted by the layover buffer in that terminal, before starting the next trip (thus the delay that caused the cancellation does not disappear).

3.3.3 Holding Model

As mentioned earlier, using the original distribution of driving times while adjusting the trip times (particularly increasing the scheduled trip times between stop pairs) can lead to more early departures. This is because the schedule allows for longer driving times for a particular segment, but when drawing the driving times from the original corresponding distribution, the actual (simulated) driving times might be shorter.

The holding model is applied to a set of scenarios in order to mitigate the excessive occurrence of early departures. The analysis for all the scenarios is thus conducted first without holding, and for line 9 scenarios, a set is tested with holding, using an approach similar to the cancellation model. In this approach, delays are divided into arbitrary bins based on their magnitude, and the probability of holding is calculated for each stop and delay bin. If the average dwell time is denoted by t_{dwell}, when a vehicle arrives earlier than t_{dwell} than the scheduled departure time and thus has an actual dwell time greater than t_{dwell} (which corresponds to the average dwell time), meaning that the departure time occurs after the scheduled departure time, it is assumed that holding has taken place since the vehicle was too early.

3.3.4 Simulation Runs

In order to run the simulation in a manageable time, the required number of runs must be determined. This is estimated using (Wikipedia contributors (2026)):

$$rel = \frac{1.96 \cdot \text{std}}{\sqrt{n}} \Big/ |\text{mean}|$$

where rel denotes the relative error of the estimated Passenger Reliability (PR) and Operator Robustness (OR).

The required number of runs can be determined by choosing the number after which increasing it might be almost ineffective on reducing the relative error.

3.4 Timetable Scenarios

In order to evaluate the sensitivity and discriminatory power of the proposed indicators, a base timetable scenario and a set of alternative timetable scenarios are considered. The base scenario represents a timetable that was previously used in operation. The alternative scenarios are generated by modifying key timetable parameters that are expected to influence the indicator values, namely layover times, trip times, and service frequencies. These variables were selected because of their expected effects on timetable quality; For instance, additional layover time or trip time slack can improve reliability by providing more opportunities for delay absorption. On the other hand, increasing frequencies are expected to improve passenger efficiency, by providing more travelled distances and less average waiting time per passenger. These adjustments however are generally more costly for the operator, while they can improve the robustness of the timetable, and thus can help prevent facing financial penalties stemming from underperformance.

As far as possible, only one parameter is modified at a time in order to isolate its effect on the indicators and facilitate interpretation of the results. For each parameter, scenarios with both increased and decreased values are created. The feasibility and extent of the modifications however is dependent on the case. Therefore, the detailed construction of the alternative scenarios and the rationale behind the chosen modifications are described in the case study chapter (4).

Figure 3.5 depicts the expected effect of changing the chosen timetable variables on the indicators. For layover and trip time scenarios, intuitively adding slack help with delay absorption and is expected to improve robustness/reliability. For frequency scenarios however, it can be more complicated. First, OE is directly affected by the increase in operational hours when more trips are added. In addition, block assignments may change, increasing the total block hours or the number of blocks, and consequently the number of pull-in and pull-out trips. PE distances are proportional to the number of trips, while the average waiting time is expected to decrease because of the shorter headways. PR and OR, however, cannot be related only to frequency, since layovers are also changed. In particular, OR is strongly related to pull-out delays at the start of trips, while delay propagation is highly dependent on the available recovery time. PR, however, is expected to improve with increasing frequency, since in the case of a cancellation or early departure, passengers wait less on average because the next trip arrives sooner. The quantified results are presented in Chapter 6.

Variable	OE	OR	PE	PR
Layover Time ↑	↓	↑	—	↑
Trip Time ↑	↓	↑	↓	↑
Frequency ↑	↓	?	↑	?

Figure 3.5: Expected effect of variables on indicators

Moreover, given the line characteristics, the scenarios required to demonstrate the effect of the variables on the indicators can be determined. For instance, for a representative case with a regular service (e.g., not a peak service) and a range of service frequencies, a wide variety of scenarios can be applied. Adding

slack to layover or trip times can be tested using both a small and a larger increase to evaluate not only the effect of adding or removing slack, but also the impact of its magnitude. Similarly, frequency changes can be evaluated across multiple periods of the day. However, this representativeness should be regarded as an assumption, since no single line is likely to fully represent the entire network. Therefore, to capture other operating conditions, additional cases can be included. For example, for a low-frequency line, the effects of adding or removing slack and changing the service frequency can be evaluated. Another aspect to consider is infrastructure. Some lines operate on dedicated infrastructure, where the tracks are shared only with other public transport lines and have limited interaction with other traffic, except at crossings. For these lines, evaluating the effect of slack is particularly relevant because the influence of external disturbances is more limited. At the same time, increasing trip time slack can increase the probability of early departures, especially if holding control is not applied or at stops where holding is not implemented. Table 3.1 summarises the relationship between line characteristics and the scenarios to be evaluated.

Table 3.1: Relation of scenarios to line characteristics

Type of line	**Relation of the scenarios to the characteristics**
Representative Line	A wide range of scenarios can be evaluated to test the effect and magnitude of adding/removing slack in layover and trip times, alongside frequency adjustments.
Low-frequency Line	A targeted range of scenarios can be tested to show the impact of slacks and increasing/decreasing of service frequencies.
Dedicated Infra Line	Since disturbance from other traffic is limited, impact of adding or removing slack can be evaluated.

3.5 Comparing the Alternatives

3.5.1 MCDA formulation

This section describes the formulations needed to merge the individual indicators using MCDA. Each timetable scenario, is an alternative to the baseline (the base timetable scenario) that can be compared to this scenario and also together using this merged indicator.

Let $i \in \{1, \ldots, N\}$ denote timetable scenarios and $j \in \{1, \ldots, M\}$ denote performance indicators.

Indicator values for operator are expressed in monetary values (Euros), while PR (passenger reliability) is expressed in time value (seconds) and PE (passenger efficiency) has the unit of speed (km/s). Thus they need to be normalised to ensure comparability. For this purpose, a min-max normalisation approach (Han et al. (2012)) is used, which rescales all indicators to a common range between 0 and 1 while preserving the relative differences:

$$\hat{x}_{ij} = \frac{x_{ij} - \min x_{ij}}{\max x_{ij} - \min x_{ij}} \tag{3.30}$$

Operator's perspective: The operator-related indicators (both expressed in monetary terms) are aggregated into a single value for the operator (without weighting, and only directly accounting for the magnitude of the cost of each aspect):

$$O_i = \sum_{j \in \mathcal{O}} x_{ij} \tag{3.31}$$

where $\mathcal{O}$ denotes the set of operator indicators.

Passenger's perspective: Passengers may value reliability and efficiency differently. Passenger-related indicators are first normalised (since they have different units) and then combined into a single passenger indicator, using a set of weightings.

$$P_i = w_1 \cdot \hat{x}_i^{\text{eff}} + w_2 \cdot \hat{x}_i^{\text{rel}} \tag{3.32}$$

The aggregated operator measure and passenger measure are each normalised so that the indicators would be comparable. Then, using a set of weightings between passenger and operator objectives the timetable alternatives will be compared with each other.

As discussed in Chapter 2, the relative importance of reliability and excess travel time compared with scheduled travel time depends on the context, transport mode, trip purpose, and passenger characteristics. Based on the range of values reported in the literature for various contexts, a weight of 3 for the reliability to efficiency (PR/PE) appears reasonable and is therefore recommended in this study (thus, referring to figure 3.2, $w_1 = w_2$, and $w_4/w_3 = 3$)

For the passenger/operator (P/O) weighting ($(w_3 + w_4)/(w_1 + w_2)$), determining an exact value would require further empirical research on passengers' willingness to pay and stakeholder preferences. Moreover, the appropriate weighting depends on the decision-making context. For example, when the operator is not purely commercially driven, and if the policies under which the operator performs are driven by, for instance, making the service more desirable for passengers because of its positive externalities (such as modal shift or inclusion), considering a higher weight for passengers is more reasonable. However, rather than recommending exact weightings, this study performs a sensitivity analysis for the P/O weighting. This allows decision-makers to evaluate how timetable rankings change under different stakeholder priorities and select the timetable alternative that best aligns with their preferred weightings.

Chapter 4

Case Study

4.1 Description of the Case Study

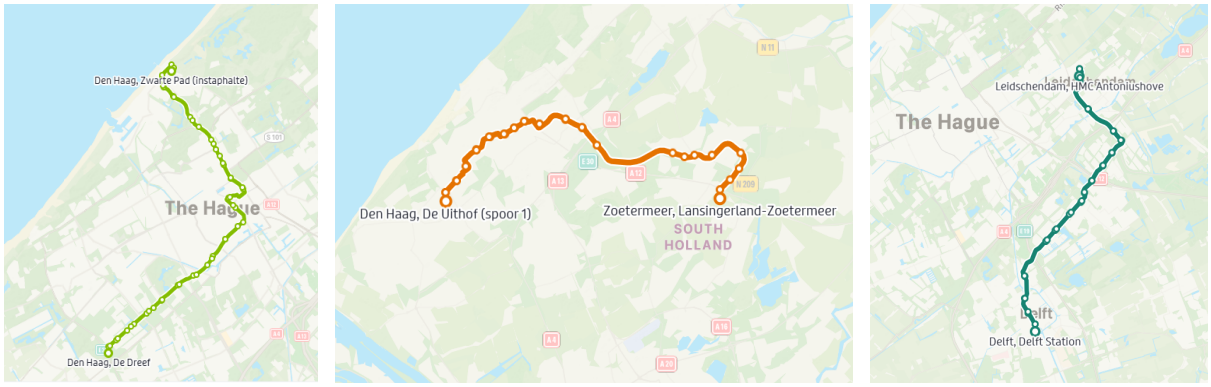
This study is conducted on the public transport network operated by HTM in the city of The Hague. HTM is one of the main urban public transport operators in the Netherlands and has been providing tram and bus services in The Hague and surrounding municipalities for more than 150 years. The network plays an important role in connecting residential areas, city centres, railway stations, and neighbouring municipalities within the metropolitan region. Currently, the network consists of 14 tram lines and 9 bus lines, including services which only operate in peak hours. The rail infrastructure is usually either shared with other traffic, dedicated with only shared crossings, or completely dedicated. Also, different tram lines vary in terms of the extent of shared infrastructure with other routes, with some operating on almost dedicated tracks for that line, while others have segments that are shared with multiple other lines.

As stated in section 3.4, the required scenarios for showing the effectiveness of the variables on the indicators varies based on the characteristic of the line. In this study, HTM tram lines 9, 19, and 4 are used as case studies. The methodology of this study is first developed and tested on tram line 9, operating between Scheveningen and Vrederust, consisting of 27 stops in each direction. This line was selected because of several representative operational characteristics. It serves both central and residential areas, has a moderate route length with a reasonable number of stops, and includes important transport hubs (Den Haag Centraal and Den Haag HS). In addition, line 9 operates at standard HTM service frequencies and carries a combination of commuter and leisure-related passenger demand. Most of the infrastructure Line 9 uses is dedicated, including only shared crossings, and a significant portion is used only by this line.

After validating the methodology on line 9, the analysis was extended to lines 19 (Leidschendam – Station Delft) and 4 (Den Haag De Uithof – Zoetermeer Lansingerland) to assess its applicability across the wider tram network. Line 19 was selected because its infrastructure characteristics are broadly comparable to those of line 9, particularly regarding the proportion of dedicated and mixed-traffic track sections. However, it also presents notable differences, including a lower service frequency with a 20-minute headway over all times of the day. Line 4 was selected because it operates largely on dedicated infrastructure, especially in Zoetermeer, where there are no road crossings. Nevertheless, some sections, such as the tram tunnel near Den Haag Centraal, are shared with several other tram lines. These differences in operational and infrastructure characteristics provide an opportunity to evaluate the results of the proposed methodology under varying conditions. Figure 4.1 illustrates the routes of the selected lines, and Table 4.1 summarises the main characteristics of these lines.

Throughout the year, different timetable scenarios are used in order to meet the varying demand more specifically. To set a base scenario, the timetable scenario in use in the weekdays between November 24 and December 12, 2025 was used, thus the operational data and passenger data from this time period

[0]Source: `https://transitapp.com`



(a) Line 9: Scheveningen - Vrederust

(b) Line 4: De Uithof - Lansingerland

(c) Line 19: Leidschendam - Delft

Figure 4.1: Maps of selected HTM tram lines in the Hague region

Table 4.1: Characteristics of the three tram lines analysed in this study

	Line 9	**Line 19**	**Line 4**
Line	Vrederust - Scheveningen	HMC Antoniushove, Leidschendam - Delft Station	Uithof - Lansingerland, Zoetermeer
Number of stops	27	19	33
Length (km)	13.4	12.7	34
Frequency (trips/hour)	4 - 8	3	4 - 6

were utilised for the analysis. The main data sources include the AVL data (including the scheduled and realised events) as well as the AFC data (including the check-in/out events of the passengers). The alternative schedules were made in the scheduling software (HASTUS) to be assessed against the base scenario through simulation.

Regarding the rolling stock, three main tram vehicle types are currently used by HTM: the GTL (operating line 1, 6, 10, 12), Avenio (operating line 2, 9, 11, 15, 16, 17), and Citadis (line 3, 4, 19, 34) vehicles. In the near future, the GTL vehicles are expected to be replaced by the TINA tram vehicles. The figures below depict an example of these vehicles. Since the operational and investment costs of these vehicle types differ, several approximations are made in the analysis. Regarding the bus network, the vehicles include electric buses that are replacing compressed natural gas (CNG) buses. An example is shown in Figure 4.6.

In a particular schedule, after creating trips, vehicle tasks can be created by assigning trips to vehicles. In most cases, a single vehicle operates a specific line. However, in some cases, interlining takes place, meaning that a vehicle operates across multiple lines within the same vehicle block, with Lines 15 and 16 as an example. A vehicle block is defined as the sequence of trips assigned to a vehicle from its departure from the depot until its return to the depot. Thus, it includes pull-in and pull-out trips (sometimes opportunity trips (8.4)) and the regular trips with passengers.



Figure 4.2: Avenio tram



Figure 4.3: Citadis tram



Figure 4.4: GTL tram



Figure 4.5: TINA tram



Figure 4.6: E-citaro Bus

Chapter 5

Experimental Setup

This chapter describes the experimental setup adopted in this study. First, the case-specific implementation of the methodology is presented, including the selection and preparation of the operational data, and choosing an operated timetable as the base scenario. Subsequently, the case-specific formulation of the performance indicators is given. Afterwards, the specificity of the methodology to the case is described. In order to take the different pattern of disturbances and disruptions throughout the day into account, as well as the variations in the demand patterns, the analysis is stratified into different time-of-day regimes. This stratification follows the trip time determination intervals defined in HASTUS, which are identical for all lines. Within these regimes, different headways may exist, while separate regimes can have the same headway. However, stratification based on headway would result in different time periods for different lines. Therefore, the stratification is done based on trip time determination intervals. The time intervals are defined as follows (in hours): 6–7, 7–8, 8–9, 9–12, 12–15, 15–17, 17–18, 18–18:30, 18:30–22, and 22–24.

5.1 Sample Selection

The primary data source is the operational dataset provided by HTM, which contains both scheduled and actual vehicle departure times (from AVL systems). A consistent period in which a particular timetable scenario was in use is selected to ensure homogeneity in scheduled departures, trip times, and headways (van Oort (2011)), and to define the base timetable for the analysis.

In this study, several time periods in which a consistent timetable was used were considered to be a homogeneous time period; From the last 2 years (2024, 2025), 4 different periods were selected. For each period, the number of available data points is compared with the number of required data points, which is calculated using (Montgomery and Runger (2018)):

$$n = \left(\frac{Z_{\alpha/2}\sigma}{E} \right)^2 \tag{5.1}$$

where:

- n is the required sample size,
- $Z_{\alpha/2}$ is the critical value of the standard normal distribution corresponding to the desired confidence level (1.96 for 95% confidence interval),
- σ is the population standard deviation,
- E is the allowable error (margin of error) of the estimate, defined as $E = p\mu$, where p is the desired relative precision and μ is the population mean.

The time interval that provides a sufficient number of datapoints is then selected.

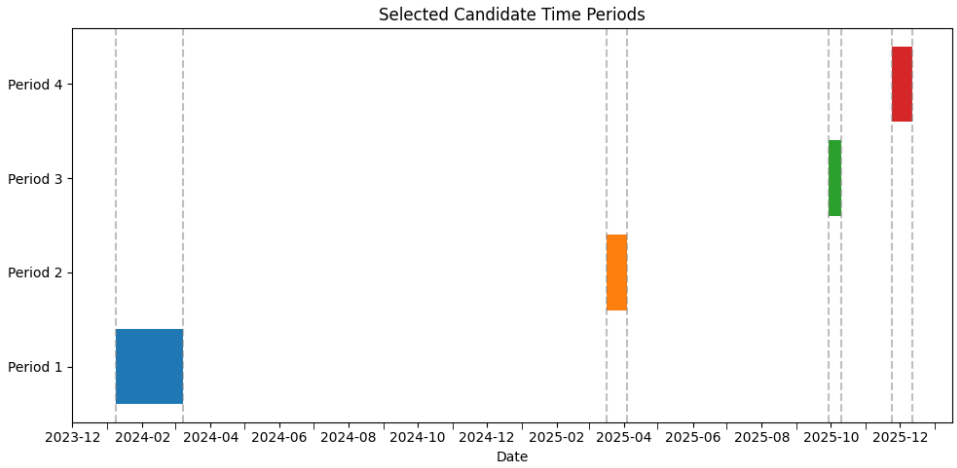


Figure 5.1: Selected Candidate Time Periods

The simulation includes applying initial delays at the first stops and then using driving time distributions to determine departure times at subsequent stops. Accordingly, the test is performed twice: once for the distribution of initial delays at the first stops (in both directions), and once for the driving time distributions between consecutive stops.

As stated earlier, four time intervals were evaluated for this purpose. The plot 5.1 shows these time intervals. Figures 5.3 and 5.2 present the proportion of regimes or regime–stops that, with 20% precision, have a sufficient number of data points. For this precision level, the November period provides enough data for initial delays; however, with a stricter precision requirement, this was no longer the case. The other time intervals do not provide sufficient data even with 20% precision. For driving times, none of the periods have enough data for all stop–regimes at 20% precision, although March and November perform relatively better. Overall, the November period (between November 24 and December 12) performs best and is therefore selected.

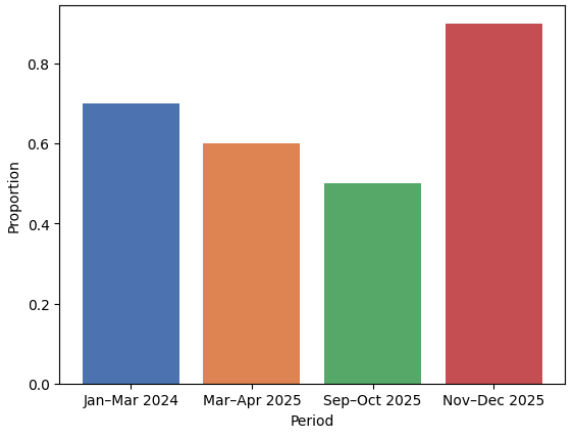


Figure 5.2: Fraction of regimes with sufficient data (initial delays distribution at terminals)

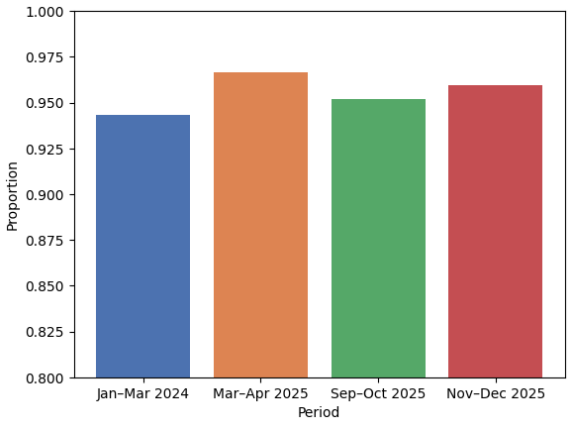


Figure 5.3: Fraction of stops-regimes with sufficient data (driving times distribution)

5.2 Case-Specific Indicator Formulation

5.2.1 Operator Expenses (OE)

As described in Chapter 3, this indicator consists of operating costs (driver and energy costs) and fleet requirements. The cost parameters were estimated using approximate values derived from data provided by HTM.[1] The hourly driver cost was estimated at €70, representing the total labour cost to the operator, including wages, overhead costs, and reserve staff. Trams were estimated to consume 70 kWh of electricity per operating hour, with an electricity price of €0.15 per kWh. The purchase cost of a tram was estimated at €3,000,000, assuming an expected service life of 30 years. Since the objective of this study is to compare the relative differences between timetable scenarios rather than estimate absolute operating costs, these approximate cost values are considered sufficient.

The resulting operating cost parameter is

$$c^{oper} = 70 + (70 \cdot 0.15) = 80.5 \text{ Euros/hour} \tag{5.2}$$

Similarly, the equivalent daily vehicle cost is

$$c^{veh,daily} = \frac{3\,000\,000}{30 \cdot 365} = 273.97 \text{ Euros/day} \tag{5.3}$$

Substituting these values into Equations 3.2 and 3.4 yields

$$C_l^{oper} = 80.5 \cdot H_l^{oper} \text{ (Euros/day)} \tag{5.4}$$

$$C_l^{fleet} = 273.97 \cdot V_l \text{ (Euros/day)} \tag{5.5}$$

5.2.2 Operator Robustness (OR)

HTM operates under a concession granted by MRDH[2], which specifies a set of contractual performance requirements. These requirements include criteria related to punctuality (late and early departures) as well as trip cancellations. Punctuality is evaluated on a quarterly basis. For tram services, punctuality is measured only at the first stop of each trip, whereas for bus services additional control points, such as major public transport hubs, are also considered. Cancellation penalties are calculated annually on a per-kilometre basis and are applied only to full trip cancellations, and only for trips for which the operator is deemed responsible. For these reasons, financial penalties resulting from service cancellations are assumed to be negligible and are therefore excluded from this indicator, and only the punctuality requirements are considered.

Punctuality metrics are defined over a quarterly evaluation period. In this study, OR is evaluated for each simulation run, after which the mean value is computed. Since the penalties are relevant per quarter, the resulting OR value is divided by the number of working days (assumed to be 60), such that the final OR corresponds to one day of operation.

Let l denote the percentage of late trips, defined as trips arriving more than 2 minutes late. The fine related to late departures is defined as:

[1] The detailed estimation procedure is omitted from this document due to confidentiality.

[2] Metropoolregio Rotterdam Den Haag

$$P_l = \begin{cases} \text{bonus}, & l < l_{bonus} \\ 0, & l_{\min} \leq l < l_{\max} \\ \text{penalty}(l), & l \geq l_{\max} \end{cases}$$

where l_{bonus}, $l_{\min}$ and $l_{\max}$ represent the bonus, and lower and upper thresholds for late trips, respectively equal to 10%, 12%, and 16%.

Regarding early departures, an early departure is defined as a trip for which the actual departure time is earlier than the scheduled departure time. Let e denote the percentage of trips that depart early at the monitored control stops. The fine related to early departures is defined as:

$$P_e = \begin{cases} 0, & e < e_{\min} \\ \text{penalty}_1(e), & e_{\min} \leq e < e_{\max} \\ \text{penalty}_2(e), & e \geq e_{\max} \end{cases}$$

where $e_{\min}$ and $e_{\max}$ denote the lower and upper thresholds for early departures, respectively equal to 0.5% and 6%.

As indicated by Equation 3.7, OR is defined as

$$OR = -\left(P_e + P_l\right) \tag{5.6}$$

5.2.3 Passenger Efficiency (PE) and Passenger Reliability (PR)

PE and PR are calculated according to the formulations given in 3.2.3 and 3.2.4. In order to assign weight to the excess travel time and scheduled travel time by passenger demands in different time-of-day regimes, the time-of-day stratification explained earlier in this chapter is used.

Moreover, when estimating the additional waiting time in low frequency services, τ_{early} and τ_{late}, which denoted the early and late departure thresholds need to be determined. The same values from van Oort (2011) are used, respectively, equal to 2 and 1 minute.

5.3 Alternative Scenarios

The alternative timetable scenarios examine the impact of modifying three timetable design elements: layover times, trip times (slack), and service frequencies. For each variable, scenarios with both increased and decreased values are defined. For every scenario, a new timetable is constructed and subsequently enriched with simulated operational data. This enables the effect of each timetable modification on the proposed quality indicators to be evaluated.

The extent of the scenario analysis differs between the three case-study lines. Line 9 serves as the primary case study, as introduced in Chapter 4. A relatively comprehensive set of scenarios is therefore investigated, including small and large changes to layover and trip times, as well as increases and decreases in service frequency. These scenarios provide insight into the sensitivity of the indicators to different magnitudes of timetable modifications.

For Line 19, a more limited set of scenarios is considered. For each design variable, one scenario with an increased value and one with a decreased value are evaluated to investigate whether the trends observed for Line 9 remain valid for a line with different characteristics (e.g., lower service frequencies).

Finally, Line 4 is analysed as a special case because of its dedicated infrastructure. The objective is to investigate whether the effects of adding or removing timetable slack differ on a line that is less affected

by mixed traffic conditions. Consequently, only layover- and trip-time modifications are considered for this line.

Line 4 Specific Considerations

Line 4 requires additional consideration due to its operational characteristics. A substantial part of the route operates on infrastructure shared with Lines 3 and 34 and with Metro Line E. Since Metro Line E has operational priority, its timetable effectively acts as a fixed constraint. Furthermore, the shared corridor does not provide overtaking opportunities. Consequently, isolated timetable modifications for Line 4, such as increasing trip times without corresponding adjustments to the other services, would not be operationally feasible. Therefore, the scenarios for this line should primarily be interpreted as a theoretical proof of concept rather than directly implementable timetable alternatives. Their purpose is to investigate how dedicated infrastructure influences the effectiveness of adding or removing timetable slack and to compare these effects with those observed on Lines 9 and 19.

5.3.1 Scenarios with Altered Layover Time

As stated in the previous chapter, from the total layover time available between a pair of trips operated by the same vehicle (in the same block), an amount of time must be potentially dedicated to the drivers' activities. For the case of HTM, according to the Collective Labour Agreement HTM Personenvervoer N.V. (2025), a minimum of 2 minutes is required as 'keertijd' to perform essential actions at the end point. Thus, this time may be used by drivers at the terminal points and the assumption here is that this 2 minutes will not contribute to delay absorption. The remaining layover time is considered available for absorbing propagated delay. If the available layover is already equal or below 2 minutes, the available buffer is assumed to be zero. A more detailed modelling would require knowing what share of the delays observed at the start of the trips are new source delays versus propagated.

Depending on the line and infrastructure layout, the first and last scheduled events of a trip may either correspond to passenger-service stops, where passengers can board and alight, or to terminal points used by vehicles when turning around and starting a trip in the opposite direction. Layover time is calculated as the difference between the last scheduled event of a trip and the first scheduled event of the subsequent trip operated by the same vehicle block. When terminal points are present (as is the case for Line 9), these terminal points are used as the first and last scheduled events. As a result, any driving time associated with the turnaround manoeuvre is excluded from the layover calculation by construction. For lines without separate terminal points (as is the case for Lines 19 and 4), no turnaround driving time is assumed to be required, and layover is therefore calculated directly as the difference between the last and first passenger-service stops of two consecutive trips operated by the same vehicle block.

The schedule is typically designed such that one terminal operates with the minimum feasible layover time, while the opposite terminal is provided with additional recovery buffer. For tram line 9, the minimum layover is applied at 'Zwarte Pad', whereas trips terminating at 'De Dreef' generally receive more layover time. For line 19, the minimum layover is applied at Delft, while additional buffer is typically provided at 'Leidschendam HMC Antoniushove'. However, for line 4, the minimum layover seemed to be usually applied at 'De Uithof', whereas trips terminating at 'Lansingerland-Zoetermeer' are generally scheduled with more layover buffer, but there were a few exceptions that the layover in Zoetermeer was more limited relatively.

Starting with increased layover time, several approaches can be adopted. Figures 5.4 show a simplified example of this case. In this example (Figure 5.4a), a set of trips are operated in 2 blocks, in both directions having a headway of h. For line 9, one possible approach is to increase the layover time at 'Zwarte Pad' by time t. This can be achieved by shifting all trips departing from this terminal (direction 1) by time t, while keeping the rest of the schedule unchanged. By doing this, the headways remain identical to those of the original timetable in both directions. However, the drawback of this approach

is that the layover time at 'De Dreef' decreases by time t. In other words, recovery time is effectively transferred from one terminal to the other. Figure 5.4b illustrates this, where only trips in one direction are shifted.

Another option is to keep the layover at 'De Dreef' unchanged, which requires shifting the trips in direction 2 by t as well. However, this adjustment results in irregular headways, as shown in the example of Figure 5.4c. To restore the regularity, whole vehicle blocks can be shifted in such a way that headways remain more regular, as shown in the simplified example in Figure 5.4d. In this example, an entire vehicle block (e.g., block j) is shifted by $t/2$ and the headways become equal to $h + t/2$, which differs from the original headway h.

Another alternative is to increase the layover times at both terminals simultaneously. However, this approach also introduces headway irregularity (Figure 5.4e), and resolving it again results in headways different from the original schedule (as the example in Figure 5.4f, where the headways become equal to $h + t$).

The aim here is to change only one variable at a time in order to observe its effect more clearly; therefore, changes in headways are not desired. For this reason, the first approach was selected, as it preserves the original regular headways while increasing the layover time at the terminal where the minimum layover was originally applied.

In the first group of scenarios considered in this study, the above-mentioned approach was followed by introducing a small amount of slack time, corresponding to a layover increase of approximately one second per stop (i.e., the number of route segments multiplied by one second). This group was only tested for tram line 9. The addition of one second per stop was chosen to assess whether a limited amount of additional recovery time could improve performance. Although the adjustment is relatively minor, the minimum layover requirement at the terminal with the larger buffer is still violated in some instances. As a result, slight shifts in trip departure times may be required, although this could lead to small changes in headways. An alternative approach would be to recreate the vehicle blocks while maintaining the original headways and/or introducing larger amounts of slack. This is explored in a separate group of scenarios.

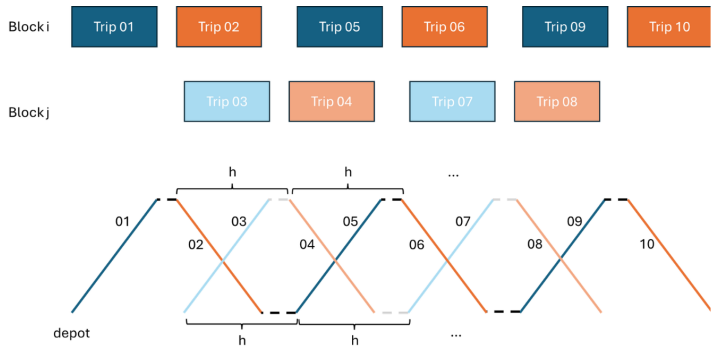
Furthermore, for the decreased layover scenario, the process works in the opposite direction. The trips in direction 1 are shifted backward, which increases the layover time at 'De Dreef'. Since the goal was decreasing the layover at 'Zwarte Pad' in which almost minimum layover was already in use, the layover will be violating the minimum layover at the 'Zwarte Pad', thus the minimum required layover is decreased as well.

Scenarios with larger layover modifications

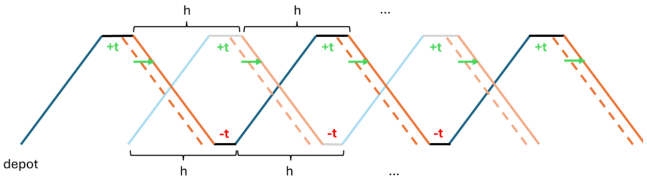
Another set of scenarios is introduced, involving more lines and more substantial alterations. The purpose is to examine whether larger changes lead to more significant effects. For instance, if layovers are increased by approximately 2 minutes rather than 30 seconds, how this affects the results. These choices were arbitrary; however, the underlying rationale included maintaining consistency across scenarios (e.g., a uniform increase of 5 seconds per stop), and ensuring that layover slack and total travel time slack remained equal. Also, in this group of scenarios, headways are kept as the original timetable, and when minimum required layover is violated, blocks are recreated.

The scenarios are defined as follows. For Line 9, which has 27 stops, a consistent increase of 5 seconds per stop is applied. As a result, the total layover slack equals 130 seconds (26×5 seconds). For line 19, this involved adding 90 seconds (18×5 seconds, since 19 stops), and for line 4, 160 seconds (32×5 seconds, since 33 stops).

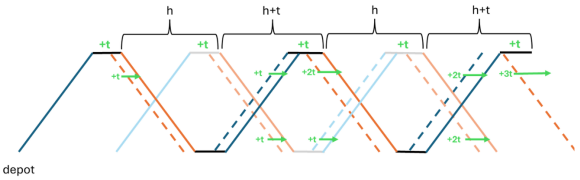
For the layover scenarios, trips in one direction are shifted by a time t, while the minimum required layover at both terminals is also increased by t. By doing this, an increase in layover of t is enforced



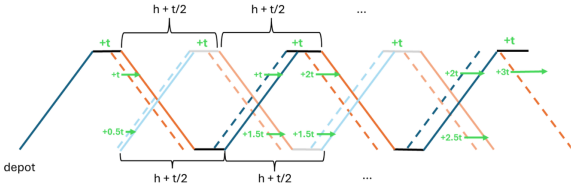
(a) Original Timetable



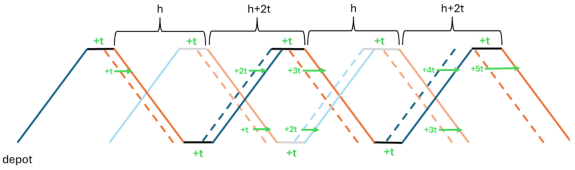
(b) Shifting trips in one direction only



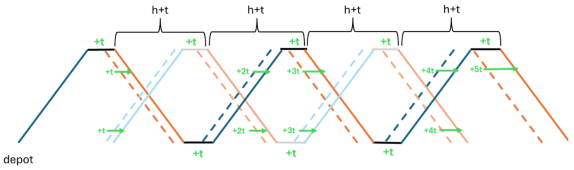
(c) Shifting trips in both directions, layover increased in one terminal and constant at the other



(d) Shifting trips in both directions, layover increased in one terminal and constant at the other, regular headways



(e) Shifting trips in both directions, increased layover in both terminals



(f) Shifting trips in both directions, increased layover in both terminals, regular headways

Figure 5.4: Simplified example illustrating different approaches regarding layover adjustments

at the terminal that originally had the lower layover. At the other terminal, since the minimum required layover is increased but by trips being shifted forward, the actual available layover is decreased. This requires the vehicle blocks to be recreated, which can affect how layovers are distributed between trips compared with the original schedule. However, overall, larger actual layovers will be available between the trip.

For the decreased layover scenarios, the minimum required layover is reduced by t, provided that the resulting minimum layover does not fall below 2 minutes, as this threshold is required by the labour agreement, as discussed previously. Trips in one direction are also shifted backwards by t. This does not violate the requirement, since the layover at the terminal with more buffer time is increased, while at the other terminal the minimum layover requirement is reduced by the same amount. To minimise operating hours, the vehicle blocks were recreated. In addition, manual adjustments (e.g., removing constraints that previously kept certain trips together) were made to obtain a more optimised block allocation with less operational hours.

5.3.2 Scenarios with Altered Trip Times

Considering the first group of alternatives with less significant slack times, a similar trade-off also arises when modifying trip times instead of layover times. Again, the case of line 9 is considered for this group. If the trip times in both directions are increased by a total of t' while keeping the layovers at both terminals unchanged, the headways must again be modified. With the objective of keeping the original headways as far as possible, with only slight changes to avoid violating the minimum required layovers, trip times are increased only in direction 2 (from 'De Dreef' to 'Zwarte Pad'), while trips in direction 1 are shifted forward by t'. In this way, the layover at 'Zwarte Pad' can be kept constant, but the layover at 'De Dreef' decreases accordingly by t'. When trip times are decreased, the trips in direction 1 are shifted backward, increasing the layover time at 'De Dreef' while keeping it constant at 'Zwarte Pad'.

By doing this, the alternatives based on trip time adjustments become conceptually comparable to those based on layover redistribution. In the layover scenarios, more layover time is applied at 'Zwarte Pad', while in the trip time scenarios the layover at 'Zwarte Pad' remains constant and the trip time itself increases. Also, to ensure comparability, it is set $t = t'$. Here, the trip time of each segment is increased by 1 second. Since there are 27 stops and therefore 26 segments for tram line 9, the total increase becomes $t' = 26\,\mathrm{s} = t$.

Scenarios with larger trip time modifications

This group of scenarios was considered for Line 9, 19 and 4. For the trip time scenarios, when trip times are increased by t' in one direction, trips in the opposite direction are shifted by t' in order to keep the layover unchanged at the minimum layover terminal. Although the minimum required layovers remain unchanged, this shift may violate the minimum layover requirement at the other terminal. Consequently, the blocks must be recreated, and the layovers between trips inevitably change relative to the original situation.

For decreased trip times, the travel time is reduced by t' in one direction, and trips in the opposite direction are shifted backwards accordingly. This can reduce the total blocking hours. In this case, the layover at the minimum layover terminal remains unchanged, while the layover at the other terminal increases by t'. However, recreating the blocks may result in a more efficient configuration, which can further affect the layover distribution.

Consequently, a limitation of the analysis is that the trip-time scenarios may also include changes in layover characteristics due to the modified block assignments, making it difficult to isolate the effects of trip-time adjustments entirely.

5.3.3 Scenarios with Altered Frequency

Another set of scenarios was defined to investigate the impact of service frequency changes. For this set, instead of applying a uniform change across the entire timetable, a specific hour of the day was selected (e.g., an hour in the morning or evening). The reason was the complexity of designing a timetable with major service frequency alterations. The selected hours were those immediately preceding a period with a different service frequency. As a result, adding or removing trips shifts the transition between frequency levels forward or backward in time. By doing this, it was attempted to avoid sudden changes in service frequency. For example, this prevents a situation where a period with 8 trips per hour temporarily increases to 9 trips per hour and then returns to 8 trips per hour shortly afterwards.

In the base timetable for line 9, there are five trips per hour in each direction between 07:00 and 08:00 and 8 trips per direction between 08:00 and 09:00. One scenario increases the frequency during this period by adding one extra trip per direction and adjusting departure times, where necessary, to achieve more regular headways. A second scenario reduces the frequency during the same period by removing one trip per direction. Consequently, the increased-frequency scenario contains two additional trips in

total, while the reduced-frequency scenario contains two fewer trips. To maintain headway regularity, trip departure times were shifted. These frequency scenarios were tested for both lines 9 and 19.

A further set of scenarios was developed to examine the effect of timetable changes in another time of day. For line 9, the service frequency decreases from eight trips per hour per direction between 18:00 and 19:00 to four trips per hour per direction between 19:00 and 20:00. To investigate whether a more gradual transition would improve performance, one additional trip per direction was inserted during the 19:00 – 20:00 period, increasing the frequency from four to five trips per hour per direction. Also, for the scenario with decreased scenario, a trip was removed per direction in the 18:00 to 19:00 interval.

Overall for frequency scenarios, since they involve adding or removing trips it was not feasible to keep the layover times unchanged. Thus, the results must be interpreted with care, as the improvement or deterioration of the indicator results would not only come from the frequency changes.

The block diagram for all the scenarios, as well as an example of the time/distance diagram for frequency scenario are depicted in Appendix A.

To simplify the presentation of the results, a naming convention is used for all scenarios. Scenario names consist of a line number, an intervention type, and an indicator of the direction and magnitude of the adjustment. For frequency scenarios, an additional identifier indicates whether the adjustment is applied during the morning or evening period.

The first two digits indicate the line number, followed by an abbreviation for the type of intervention (`LO`: layover adjustment, `TT`: trip-time adjustment, `FQ`: frequency adjustment). The sign (`+` or `−`) indicates whether the intervention increases or decreases the corresponding variable.

For frequency scenarios, the suffix `M` or `E` indicates whether the adjustment is applied during the morning or evening period, respectively. For example, `09FQ_M+` denotes the addition of one trip per direction during the morning period on line 9, while `09FQ_E−` denotes the removal of one trip per direction during the evening period.

For trip-time scenarios, the final number indicates the number of seconds added to or removed from each inter-stop segment. For example, `09TT+1` and `09TT+5` denote increases of 1 and 5 seconds per segment, respectively.

For layover scenarios, the final number indicates the number of seconds per inter-stop segment used to determine the total layover adjustment. Thus, `09LO+1` and `09LO+5` denote layover increases equal to 1 and 5 seconds, respectively, multiplied by the number of inter-stop segments on the line.

5.4 Application of the Simulation Model

This section presents the results related to the application and validation of the simulation methodology for the timetable scenarios. In order to ensure stable and reliable outputs, first the required number of simulation runs is determined. Next, the generated distributions of initial delays and driving times are compared with the realised data to assess whether the simulated operational conditions reasonably represent reality. In addition, the simulated cancellation rates are compared with the realised cancellation shares. Finally, the case-specific cancellation and holding model are described.

5.4.1 Simulation Runs

As described in Section 3.3.4, the required number of simulation runs was estimated by applying the simulation to the base schedule with different numbers of runs and investigating the relative error of PR and OR. Figure 5.5 presents the results, showing that increasing the number of runs decreases the relative error and improves the statistical reliability of the estimated outputs. Using 5% as an acceptable level of estimation precision, 100 runs were found to be sufficient, as the relative error for both indicators

dropped below this value. However, an additional check was performed to evaluate the stability of the results. Figure 5.6 shows the distributions of the PR indicator with different numbers of runs. By visual inspection, increasing the number of runs resulted in smoother histograms due to denser sampling of the stochastic distributions, while the standard deviation remained relatively similar. In the other alternatives (distributions in Appendix B), it seemed that the standard deviation will rise after 100 runs. Therefore, 100 runs were selected as a compromise between computational efficiency and statistical stability.

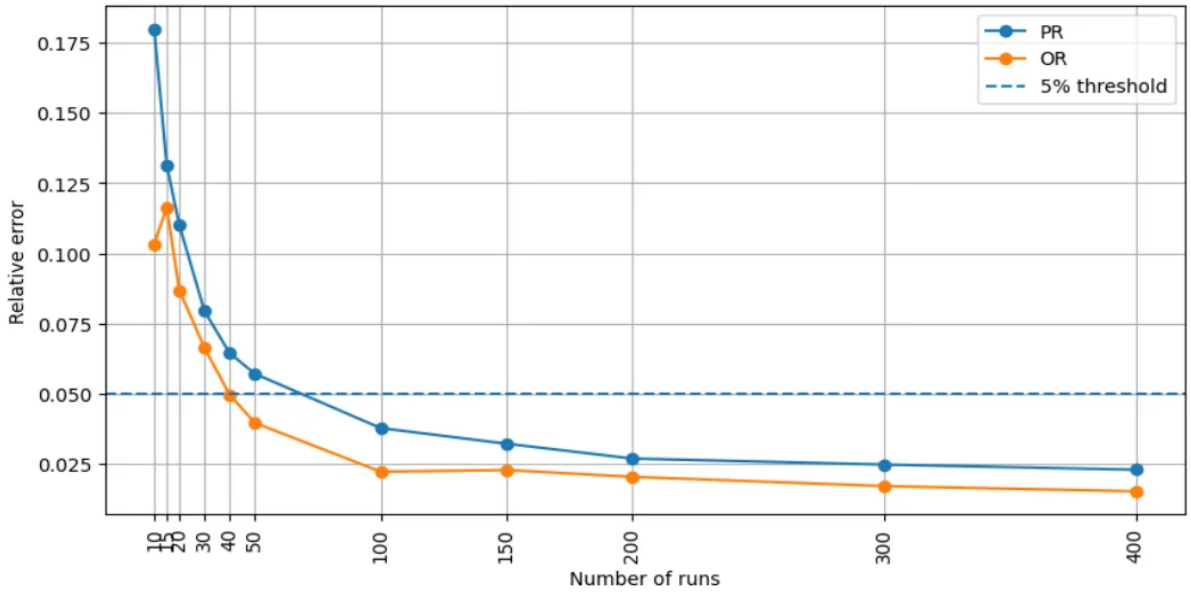


Figure 5.5: Relative error vs. number of runs

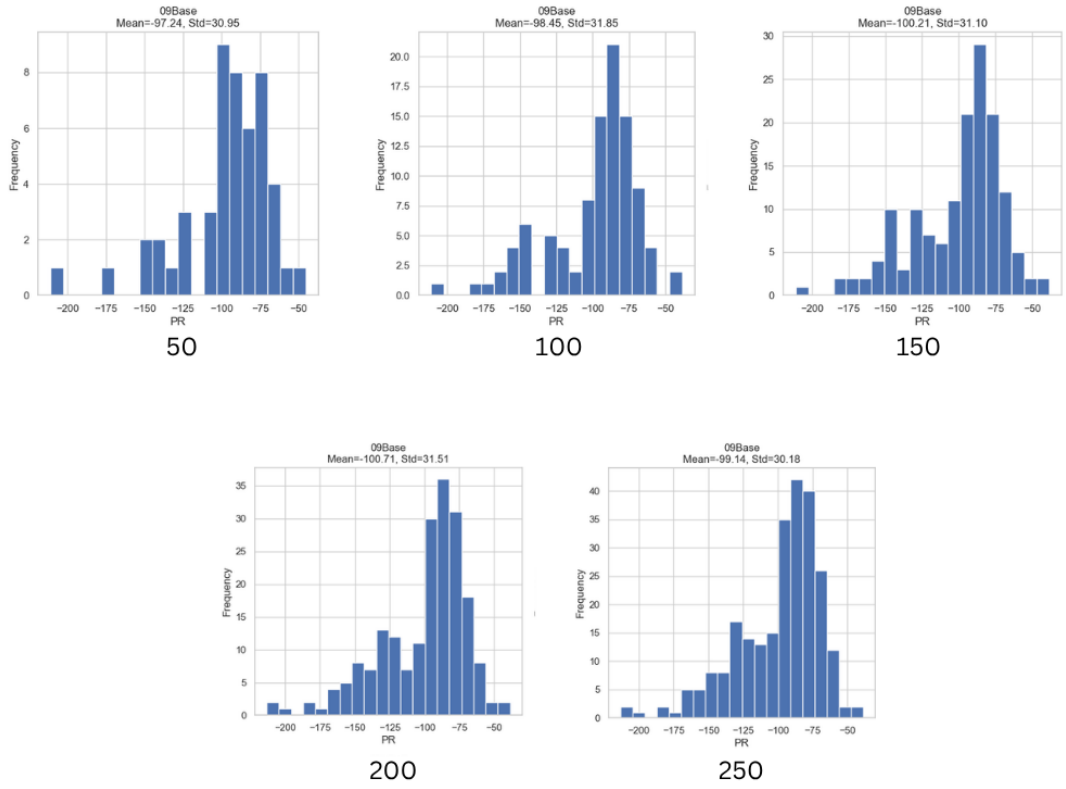


Figure 5.6: Effect of the number of runs on the distribution of PR - numbers (50 to 250) indicate the number of runs

5.4.2 The distributions for real vs. simulated data

In order to inspect whether the simulated disturbances (initial delays and driving time variations) follow the same distribution as the realised data, the Kolmogorov–Smirnov (KS) test was applied. The results are summarised in Table 5.1. Although the p-values for both distributions are very small, indicating a statistically significant difference between the distributions, this result should be interpreted carefully. Due to the large dataset size, the KS test has very high statistical power and is therefore sensitive to even negligible differences. The KS statistic (D) represents the maximum vertical distance between the two cumulative distribution functions (CDFs), in this case between the simulated and realised data. Values in the range of 0.05 – 0.09 indicate that the distributions differ by at most 5–9% at any point, which suggests a very small practical deviation. Therefore, while the null hypothesis that the distributions are identical is formally rejected, the observed discrepancy is minor, and the distributions can be considered practically equivalent. Therefore, the simulated distributions were considered sufficiently representative for the purpose of use in simulation analysis.

Schedule	**KS_init_stat**	**KS_init_p**	**KS_drive_stat**	**KS_drive_p**
09Base	0.005676	0.012919	0.088676	1.515570e-19
09LO+1	0.005676	0.012919	0.088489	1.824669e-19
09LO-1	0.005643	0.013710	0.089283	8.268318e-20
09LO+5	0.007393	0.000386	0.089564	6.242368e-20
09LO-5	0.005514	0.017185	0.096433	4.833784e-23
09TT+1	0.005676	0.012919	0.086573	1.195502e-18
09TT-1	0.005643	0.013710	0.090405	2.671988e-20
09TT+5	0.007146	0.000678	0.085265	4.214430e-18
09TT-5	0.006805	0.001427	0.099003	2.889088e-24
09FQ_M+	0.004568	0.076037	0.086309	1.454385e-18
09FQ_M-	0.006483	0.002805	0.090002	4.214555e-20
09FQ_E+	0.005434	0.019592	0.087976	2.848829e-19
09FQ_E-	0.005782	0.010755	0.089153	9.877662e-20

Table 5.1: Kolmogorov - Smirnov test results for Line 9 alternative schedules

Furthermore, Figure 5.7 and 5.8 compare the distributions of initial delays and driving times, respectively, in the realised operational data with those of the simulated base scenario. As expected, these distributions are very similar, since the simulation samples directly from the empirical distributions. This similarity is reflected both visually and in the corresponding means and standard deviations. The statistics for the remaining scenarios are provided in the Appendix C.

The Kolmogorov - Smirnov (KS) test was performed for the total delay values (including delay propagation). This comparison is only relevant for the base scenario against the realised data, as the alternative scenarios are intentionally designed to alter the delay distribution. The resulting KS statistic was 0.1106, indicating that the maximum difference between the cumulative realised and simulated delay distributions is approximately 11 percentage points. The corresponding p-value was effectively zero. However, given the large sample sizes (approximately 90,000 observations in the realised data and 600,000 in the simulated data), the KS test is highly sensitive and can detect even relatively small differences. Therefore, the KS statistic is considered more informative than the p-value. Based on the KS statistic and the visual similarity of the distributions shown in Figure 5.9, the simulated and realised delay distributions were considered sufficiently similar for the purposes of this study. After applying holding, the KS statistic decreased slightly to 0.1002, while the p-value remained effectively zero, indicating a small

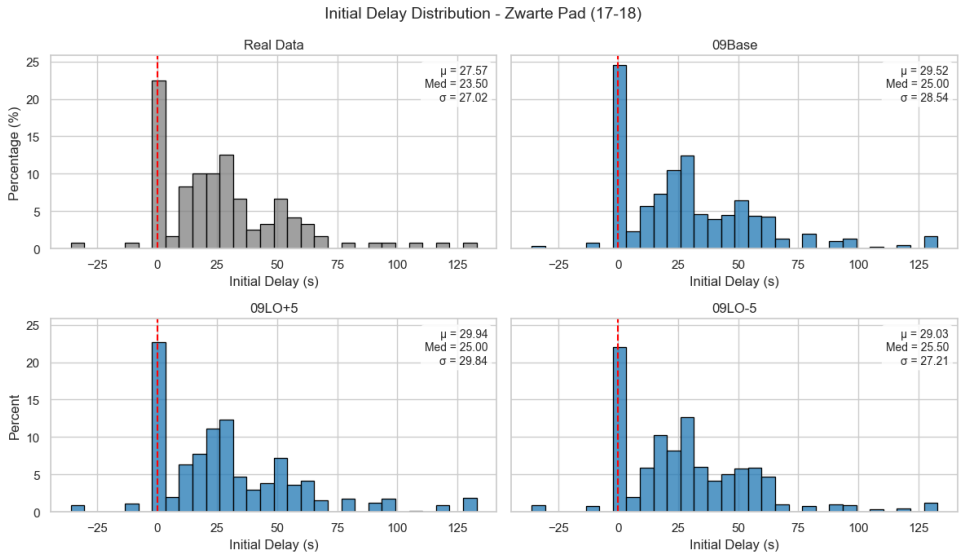


Figure 5.7: Distribution of initial delays for a particular regime (17-18) and terminal (start point), line 9

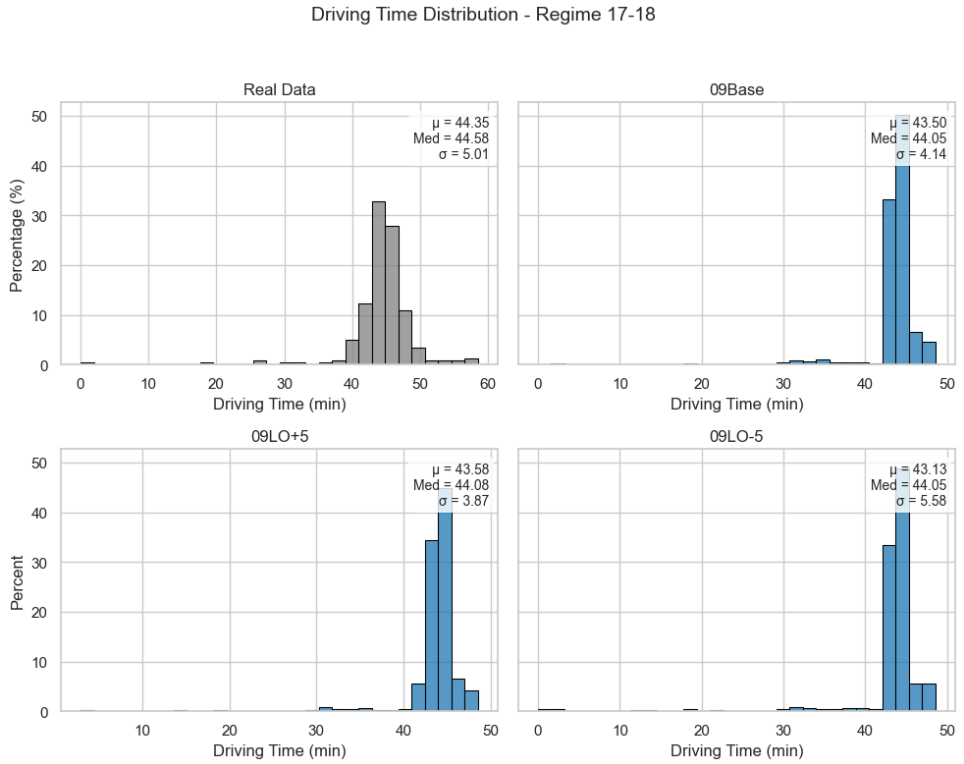


Figure 5.8: Distribution of driving times (full trip) for a particular regime (17-18), line 9

improvement in the agreement between the simulated and realised delay distributions. Figure 5.9 compares the total delay distributions of the realised operational data and the simulated base scenario. As expected, the similarity of the distribution of the base scenario and the realised data can be confirmed by visual inspection. Figure 5.10 presents the mean and standard deviation of the delay distributions for all scenarios, showing that scenarios with less slack time exhibit higher mean simulated delays, whereas scenarios with more slack time exhibit lower mean simulated delays due to delay propagation. Finally, Figure 5.11 compares the total delay distributions of the realised data and the base scenario with holding applied. Compared to Figure 5.9, the mean simulated delay for the base scenario is slightly closer to the mean realised delay, suggesting that the holding model potentially improved the realism of the simulation.

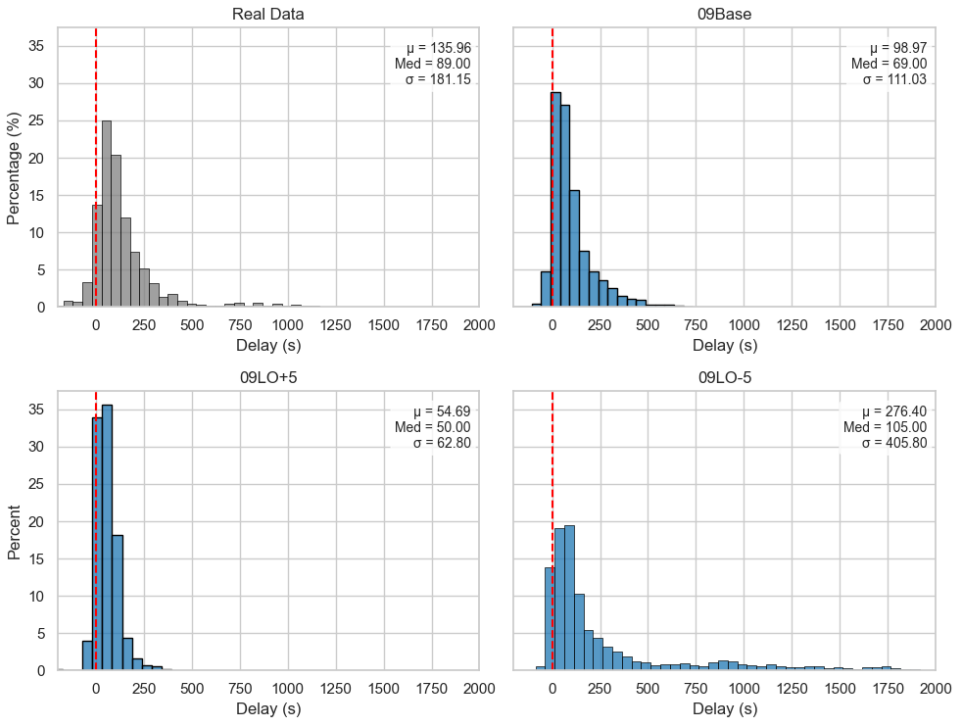


Figure 5.9: Distribution of delays for a particular regime (17-18), line 9

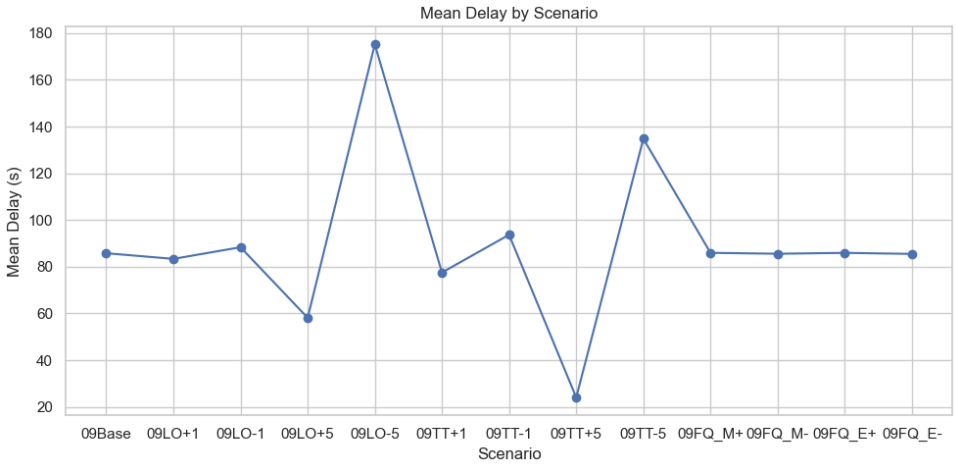


Figure 5.10: Mean simulated delay by scenario (line 9)

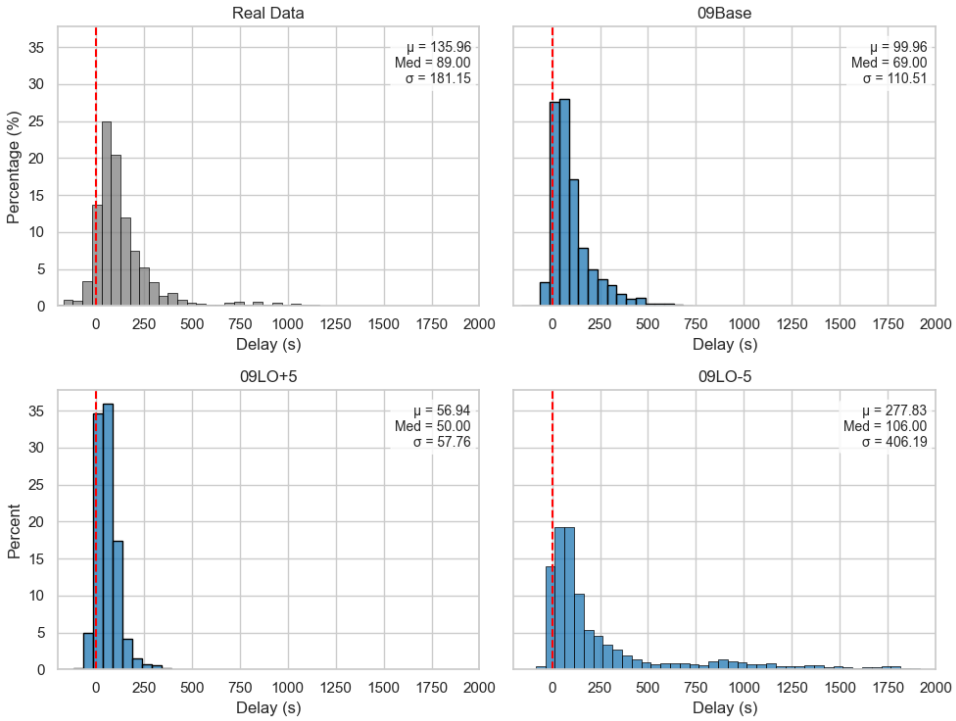


Figure 5.11: Distribution of delays for a particular regime (17-18) with holding strategy applied, line 9

Lastly, the cancellation rates in the realised and simulated data are investigated. The KS test was already applied to the simulated scenarios, which included the simulated cancellations. A KS test was not utilised to evaluate the cancellation model separately because cancellation data are discrete rather than continuous; instead, a visual analysis was preferred, which was used to capture specific time-of-day variations. Figure 5.12 depicts the rate of missed realised events in the realised data besides the cancelled events in the simulated scenarios. As can be seen, the rates are somewhat similar to the real data; however, especially for the first and last time-of-day regimes, somewhat larger differences can be observed. This is potentially due to the fact that in the simulation, based on the delay and position of the stop, the remainder of the trip is cancelled. On the other hand, in the real data, when a record is missing, the missing events can continue into the next trip(s), or only middle stops may be missing, potentially due to detouring. These assumptions made in the simulation can lead to differences between the cancellation rates in the simulated and realised data.

5.4.3 Validation of Assumptions Regarding Initial and Propagated Delay

As stated in the previous chapter, Section 3.3.1, one assumption made in the simulation was that a propagated delay exceeding the layover buffer would inherently prevent a vehicle's next trip from starting early. Looking at the AVL data, it was observed that if the full layover time acts as the buffer, early starts almost never occurred alongside propagated delays. The few exceptions that were observed included missing data, pointing to potential cancellations at the start of the trip (with service resuming from an intermediate stop) or detours. However, when the buffer was adjusted to the layover minus 2 minutes, more instances of minor early departures emerged. This indicates that, in reality, the effective buffer was sometimes larger, potentially because drivers simply bypassed their 2-minute break. However, because drivers retain the right to this break, the simulation assumed that these 2 minutes do not contribute to delay absorption. Consequently, when modelling the initial delay distribution, this 2-minute window was excluded from the buffer calculation to avoid separating the total layover into initial delay, propagated delay, and buffer time, since this would require knowledge of the relative contributions of initial and propagated delay at trip start points, which is not available.

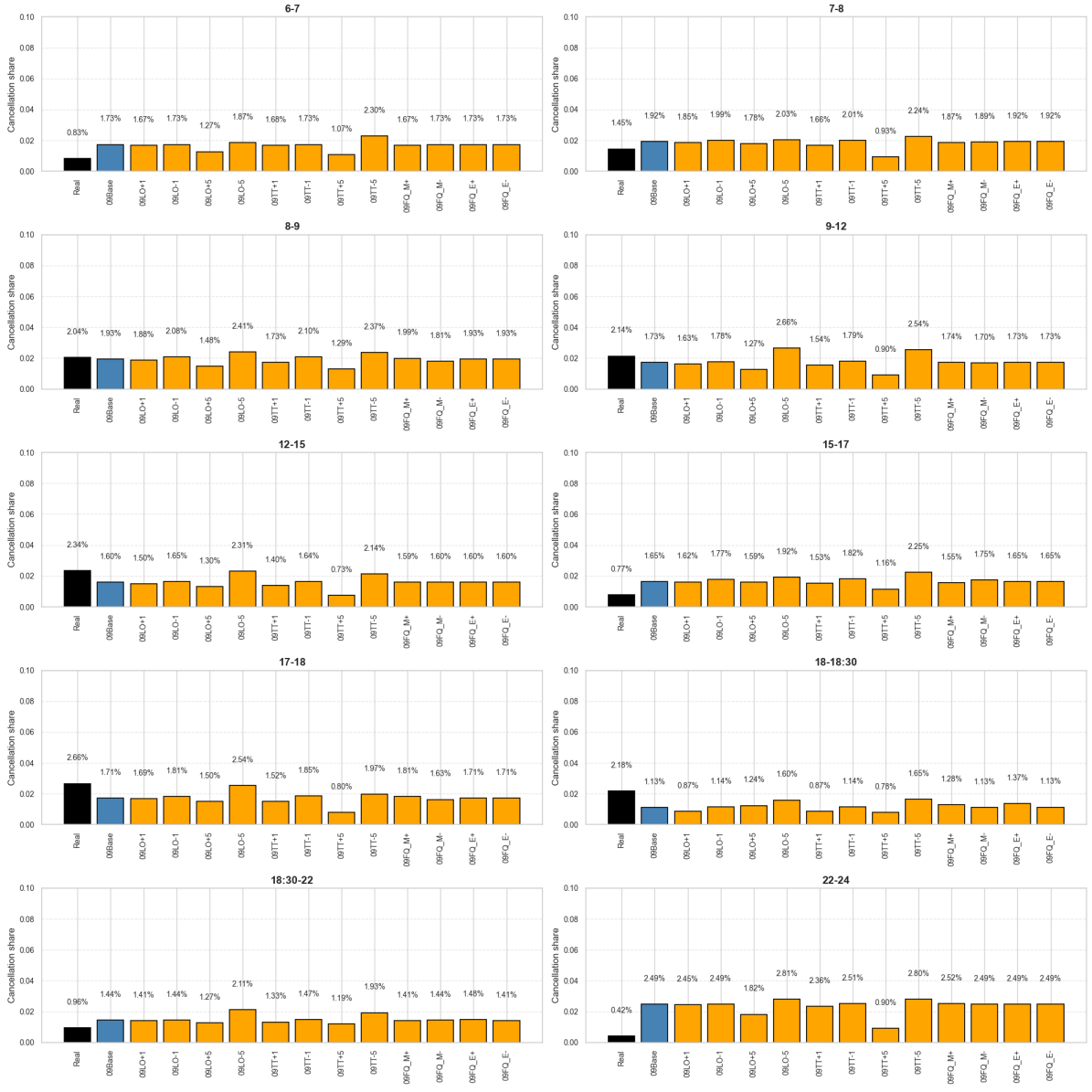


Figure 5.12: Cancellation rates

5.4.4 Case-Specific Cancellation Model

In the AVL data, some records are missing for certain events, for instance, the first or last stops of a trip, or a full trip, and occasionally, middle stops of a trip. These missing observations are assumed to correspond to service cancellations, including partial-trip cancellations, full-trip cancellations, or potentially detours. The modelling assumption is therefore that when a record is missing, the corresponding service was unavailable to passengers. As discussed previously (Section 3.3.2), service unavailability affects passengers by increasing waiting times, as passengers are assumed to wait for the next available service. This may correspond to one or more additional headways. Furthermore, only cancellations related to excessive delays are considered, since these are the cancellations that can potentially be influenced through timetable design.

The probability of cancellation is assumed to depend on both the magnitude of the delay and the location at which the delay occurs. Figure 5.13 shows the observed probability of trip cancellation for three groups of stops (beginning, middle, and end of the route), indicating noticeable differences between these locations. Consequently, the cancellation model is based on both the realised delay and the relative position of the stop.

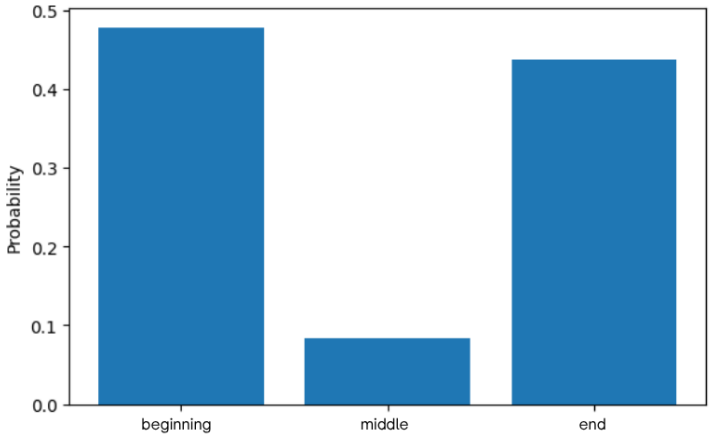


Figure 5.13: Probability of trip cancellation based on stop category

The delay used in the model is obtained either from the previous stop within the same trip or, in the case of a trip start, from the delay at the end of the previous trip after accounting for the available layover time. Delay values are grouped into bins, and cancellation probabilities are estimated for each combination of delay magnitude and stop position. The resulting probabilities are presented in Table 5.2.

5.4.5 Case-Specific Holding Model

The average dwell time (t_{dwell}) was found to be approximately 25 seconds. Figure 5.14 presents the estimated holding probabilities for Line 9. As can be seen, the probabilities are rather small, except for one start point (De Dreef) with probabilities close to one, suggesting that potentially holding generally takes place in this stop. This is consistent with the fact that this is the first stop in one direction, where control is likely to be more important from the supply-side perspective and for meeting the contractual requirements.

The holding model is therefore based on stop-specific probabilities conditional on the realised delay magnitude. During simulation, holding is applied according to the corresponding probability associated with the stop and delay bin.

Table 5.2: Cancellation probability lookup table by stop group

stop group	minimum delay	maximum delay	observations	cancellations	probability
beginning	1.000	23.000	5517	1	0.000
beginning	24.000	41.000	5278	0	0.000
beginning	42.000	62.000	5330	0	0.000
beginning	63.000	98.000	5311	1	0.000
beginning	99.000	1315.000	5288	8	0.002
end	1.000	56.000	5337	30	0.006
end	57.000	99.000	5238	24	0.005
end	100.000	144.000	5203	23	0.004
end	145.000	211.000	5181	31	0.006
end	212.000	3937.000	5240	48	0.009
middle	1.000	39.000	5127	3	0.001
middle	40.000	68.000	5006	3	0.001
middle	69.000	101.000	4953	7	0.001
middle	102.000	152.000	4911	7	0.001
middle	153.000	2057.000	4986	15	0.003

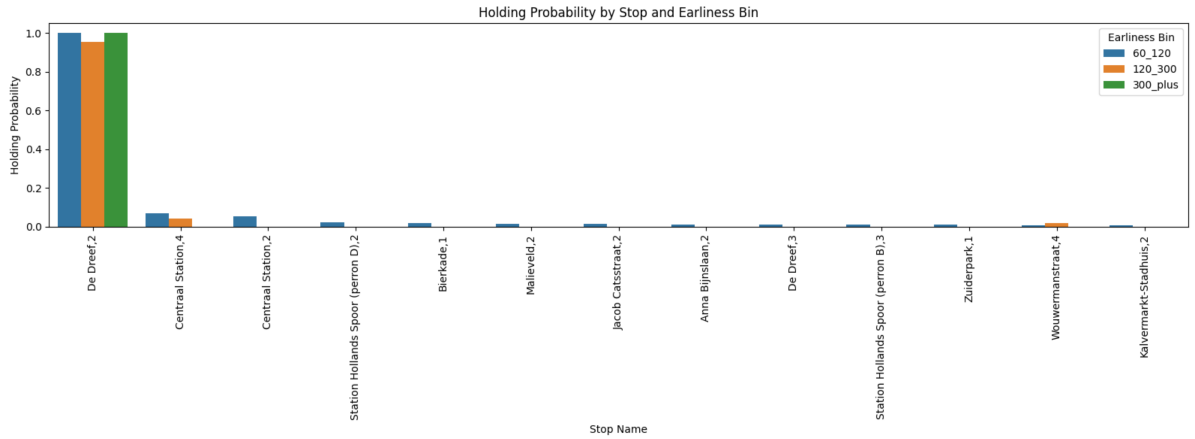


Figure 5.14: Holding shares per stop and delay magnitude bin

5.5 Weighting Recommendations for MCDA Regarding the Case

Since HTM operates as a public transport provider under the Metropoolregio Rotterdam Den Haag (MRDH), its service delivery is steered by strategic regional goals aimed at enhancing network accessibility, stimulating a modal shift away from private cars, and fostering social inclusivity (Metropoolregio Rotterdam Den Haag (2026)). Given this public service mandate, this study recommends a higher relative weight for the passenger objectives. In order to prevent the model from completely disregarding operational feasibility, a relative weighting of $P : O$ equal to $0.6 : 0.4$ to $0.75 : 0.25$ is recommended. Ultimately, pinning down on an exact weighting scheme would require further empirical research, such as passengers willingness to pay studies, and modelling the dependency on the line, to mathematically calibrate these weightings. Thus in conclusion, the weight is recommended to be heavier on the passenger's side but remains to be chosen by the decision-maker.

Chapter 6

Results

This chapter presents the results obtained from applying the methodology outlined in the previous chapter to the selected HTM tram lines. A set of timetable scenarios is analysed, including a base scenario (an existing and operated timetable) and a set of alternative scenarios in which layover time, trip time, and service frequency are modified. After quantifying the indicators, thereby addressing the first research sub-question, the quality of the different timetable scenarios can be evaluated using these indicators. Since realised operational data does not exist for the alternative scenarios, simulations are conducted in which initial delays, driving times, and service cancellations are introduced, thereby addressing the second research sub-question. In addition, holding is applied to mitigate the effect of early departures. The resulting indicator values are compared to evaluate how the timetable modifications affect the indicators, corresponding to the third research sub-question. Finally, a multi-criteria decision analysis (MCDA) is performed to assess the overall performance of the alternative timetables under different indicator weighting schemes, addressing the fourth research sub-question.

The indicators were calculated using the formulations presented in 3.2. After analysing Line 9, a more limited number of scenarios were tested on lines 19 and 4. The results are depicted in the following sections. Also, a holding strategy was not applied in the default analysis; it was applied to the available scenarios only for Line 9, as will be shown.

Furthermore, as shown in Section 3.5, the operator indicators are aggregated and subsequently normalised into a single operator score (O), while passenger indicators are normalised separately because they represent different dimensions and units. Passenger reliability (PR) and passenger efficiency (PE) are first combined into a passenger score (P), after which a second weighting step is applied between passenger and operator perspectives (P/O).

Since the selection of weights can influence the final ranking of alternatives, a sensitivity analysis was conducted to assess the robustness of the results. First, passenger reliability was assigned three times the weight of passenger efficiency ($PR : PE = 3 : 1$), following the recommendations from the literature discussed in Section 3.5. The analysis then investigates how the ranking of alternatives changes when different weightings are applied to passenger reliability and efficiency. For the weighting of passenger and operator objectives, several combinations are also considered, leaving the final choice open to the preferences of the decision-maker.

6.1 Line 9

6.1.1 Individual Indicators - Line 9

The first set of scenarios involves altering the variables in Line 9. Figure 6.1 shows the results of these computations before normalisation.

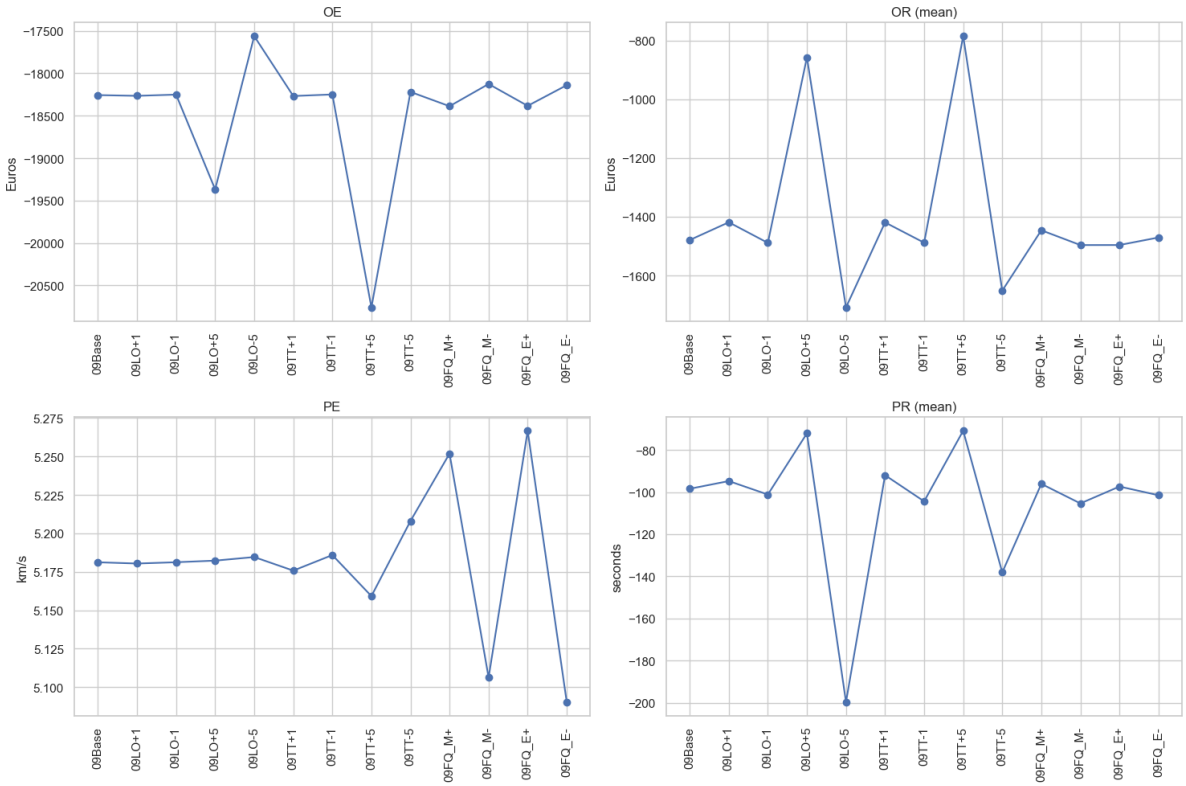


Figure 6.1: Indicator results

Operator Expenses (OE) primarily reflect the operational cost implications of the scenarios and therefore show expected trends. Increasing trip and layover times raises operational costs due to longer vehicle and crew utilisation, and vice versa. The effects become more pronounced for the higher-impact (± 5) scenarios. For instance, when trip times are increased by 5 seconds per segment (`09TT+5`), the fleet requirement increases by one vehicle, leading to a substantial deterioration in OE (the worst result among the scenarios). Conversely, increasing frequencies in both analysed scenarios did not require an additional vehicle; instead, the OE indicator deteriorated due to the increase in operating hours. In all the scenarios with decreased parameter, the same number of vehicles as the base scenario were still required for operations. However, decreasing layovers by 130 seconds (`09LO-5`) proved to be the most efficient approach from an operational perspective, as this configuration resulted in the fewest operating hours. The difference between increasing the frequency between 7:00–8:00 and 19:00–20:00 is relatively small, and a similar observation can be made for the decreased-frequency scenarios. Furthermore, increasing slack in the higher-impact scenarios was more costly than increasing frequency. This is likely because the slack increase was applied uniformly throughout the entire timetable, whereas the frequency increase affected only a specific hour. Also, the frequency increase did not affect the fleet requirements. Consequently, even adding two full trips did not increase costs as much as `09LO+5` and `09TT+5` scenarios.

Operator Robustness (OR) shows improvement for the increasing and deterioration for the decreasing layover- and trip-time-related scenarios, with larger effects for the higher-impact scenarios, which is expected. The best result was obtained for the increased trip times by 5 seconds scenario, despite the fact that more slack can also come at the cost of more early departures. Frequency-related scenarios affect layover distributions, and consequently, the propagation of delays. Therefore, the observed improvements and deteriorations in these scenarios cannot be attributed solely to frequency changes. However, the variation in this indicator for these scenarios was quite small, compared to higher-impact slack adding scenarios (± 5).

For OR, it is important not only how many early and late departures occur, but also by how much they exceed the corresponding thresholds. As a result, changes in the indicator appear to be more stepwise due to the calculation method used for the metric (as shown in 3.2.2). Looking at the shares of early and late departures at the first stops in both directions shown in Figure 6.2, scenarios with more recovery opportunities perform better with respect to late departures. Although, since for instance in the scenarios with increased trip times the distribution of realised driving times remained unchanged, while the timetable provided additional slack, these scenarios allow more early departures to happen. Also, in the scenarios with increased layover time, because of less propagated delays, the trips after more generous layovers will have a higher chance of departing early, which was part of the assumptions behind the simulation as well. Thus, scenarios with more slack perform worse in terms of early departures. On the other hand, scenarios with reduced layover and trip times perform better in terms of early departures but worse in terms of late departures.

Another comparison that can be made is that reducing layover has a larger impact than reducing trip times, making `09LO-5` rank the worst in this indicator. This can also be explained by Figure 6.3, where the average propagated delay in scenario `09LO-5` is noticeably higher than in `09TT-5`. With shorter layovers and greater delay propagation, negative source delays are less likely to occur, shifting the overall departure delay distribution towards larger values. In contrast, reducing scheduled trip times does not alter the layover distribution (see Appendix A) compared to the base scenario. Consequently, the initial terminal delay distribution remains the same as in the base scenario. Fewer early departures in `09LO-5` shift the realised trips slightly later in time, further increasing the number of late departures. As a result, although both scenarios increase delays, reducing layovers has a stronger impact on increasing late departures than reducing trip times.

The frequency scenarios remain relatively similar to the base scenario, with only minor differences. This is likely because only two trips were added or removed within an hour rather than applying a uniform change across the entire timetable, which resulted in limited changes to propagation patterns.

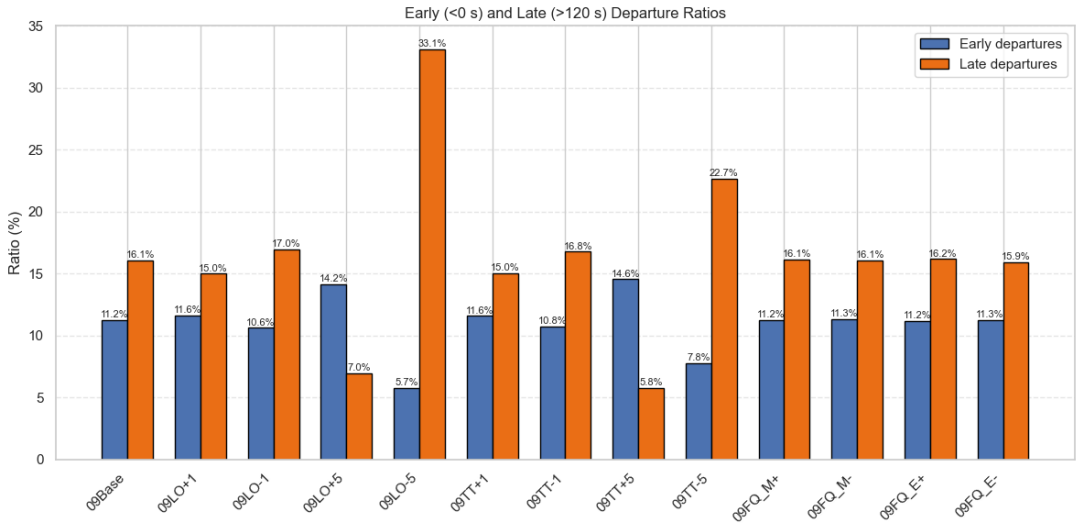


Figure 6.2: Early and late departure shares at the first stops per scenario (both directions)

For **Passenger Efficiency (PE)**, the results follow the expected trend. Increasing frequency and decreasing trip time improve performance (the best scenarios by this indicator), whereas decreasing frequency and increasing trip time worsen it (resulting in the worst scenario by this indicator). The improvement associated with higher frequency can be explained by shorter average passenger waiting times and higher total travelled kilometres, while reduced trip times contribute to lower in-vehicle travel times (according to the formulations in 3.2.3). The frequency changes between 19:00 and 20:00 seemed to have a

larger impact than those between 7:00 and 8:00 because passenger demand during the evening period was higher, thus this regime had a higher weight in the indicator calculation.

The relatively small differences between the layover scenarios and the base scenario, although less expected, can be explained by the shifting of trip start times. Although the headways remained almost unchanged and equally regular compared to the base scenario, the shifted departure times caused some trips to operate under different time-of-day demand regimes, thereby slightly altering passenger demand weights and potentially realised run times. Consequently, the PE indicator changes only marginally in response to layover modification. Increasing the frequency has a more significant effect due to more travelled distances.

Lastly, looking at **Passenger Reliability (PR)**, increasing both layover time and trip time generally improved reliability, whereas decreasing them resulted in a deterioration. The `09TT+5` scenario achieves the best PR value of all alternatives, while the `09LO-5` performs the worst; this can also be deducted from 6.3, where this scenario has the most average propagated delay because of reduced recovery opportunities. Consequently, passengers experience more late events and higher excess travel times due to delayed departures.

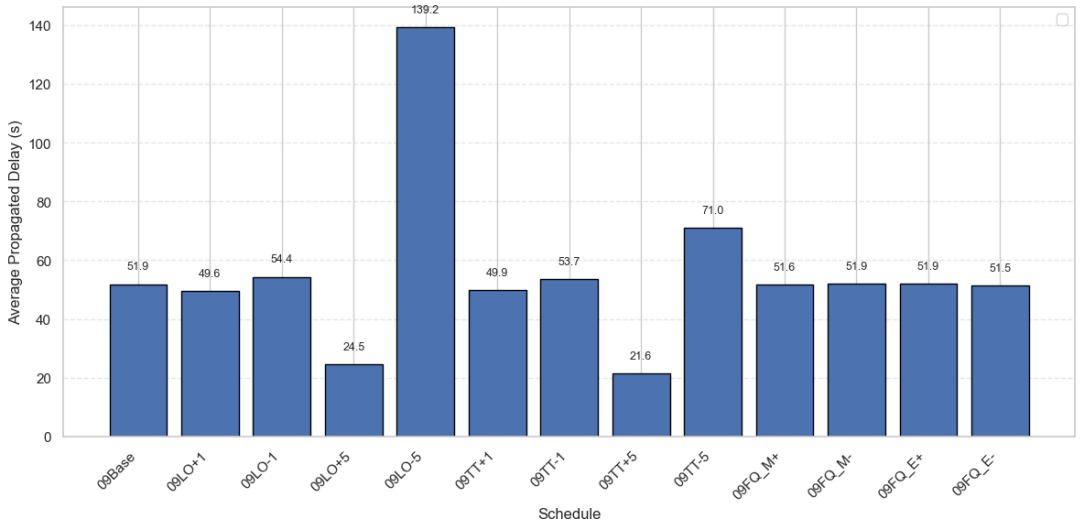


Figure 6.3: Average propagated delay per scenario - Line 9

For the lower-impact scenarios, the magnitude of improvement is larger for the trip-time scenarios than for the layover scenarios, despite both modifications introducing the same amount of additional slack (26 seconds per cycle). This difference is likely due to how the additional time is distributed within the timetable. In the layover scenarios, the extra time is allocated entirely at the terminal, meaning that recovery can only occur at the end of the trip. In contrast, the increased trip-time scenarios distribute the additional time along the route, which enables delays to be absorbed progressively throughout the trip and providing more continuous recovery opportunities.

For the higher-impact scenarios, the interpretation is more complex. In these cases, the modifications not only change the amount of slack but also alter the block allocation due to timetable reconstruction. Increasing or decreasing the minimum required layover shifts trip timings and changes the resulting vehicle schedules. Consequently, the layovers also are not unchanged when altering trip times, since trips are placed after each other differently in blocks.

A drawback of the scenarios with additional slack is the higher occurrence of early departures, as previously observed in Figure 6.2. While additional recovery opportunities improve reliability and reduce delay propagation, they also increase the likelihood that vehicles arrive ahead of schedule and depart

early when no holding strategy is applied. Yet the PR results seem to have overall better results for the scenarios with more slack.

Lastly, both increasing frequencies appear to improve while decreasing it seem to worsen the reliability results compared to the base scenario. However, these alternatives should be interpreted with caution, since as stated earlier, layover distributions are also altered, meaning that the observed effects cannot be attributed solely to frequency changes. In addition, reliability is represented through the excess travel time formulation, which does not explicitly account for improvements or deterioration in the level of service. Furthermore, changing the frequency affects the calculation of additional waiting time (refer to 3.2.4) because headways may shift between the long-headway and short-headway regimes used in the formulation. For example, a headway of approximately 16 minutes may become approximately 8 minutes after a frequency increase. Therefore, the observed PR changes are likely the combined result of delay propagation patterns and waiting times arising from passenger random and non-random behaviour, as well as the effects of altered headways.

Figure 6.4 summarises the indicator results in spider plots to allow for simultaneous comparison across indicators. To prevent overlapping, the visualisation is divided into three sets. This allows for an easy comparison of individual indicator effects across each variable. The scenarios with only minimal timetable changes (LO ± 1 and TT ± 1) largely overlap with the base timetable, which is expected since their modifications have a more limited effect compared with the higher-impact scenarios. The higher-impact increases in layover and trip time (± 5) perform best from the passenger reliability (PR) and operator robustness (OR) perspectives, whereas the corresponding decreases perform best from the operator efficiency (OE) and passenger efficiency (PE) perspectives. The small differences between the base timetable and the Layover scenarios are mainly caused by shifts in trip departures, resulting in some trips being assigned to different operating regimes and therefore receiving different demand weights, as discussed earlier. The frequency increase scenarios are positioned closer to the passenger efficiency perspective as expected, whereas the frequency decrease scenarios are positioned closer to the operator efficiency perspective. Overall, this figure provides insight into the relative changes of the scenarios across the individual dimensions.

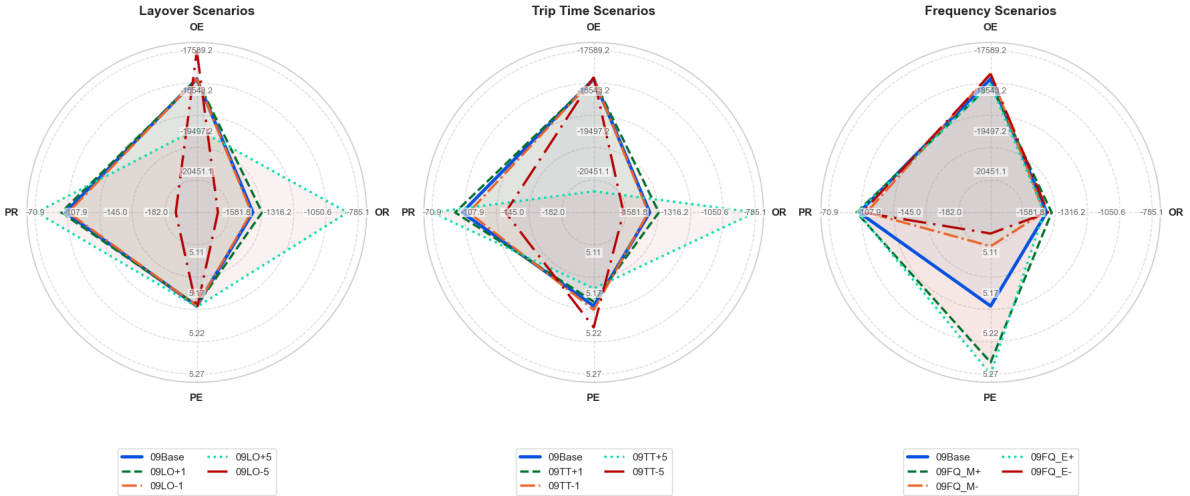


Figure 6.4: Indicator results in spider plots - Line 9

6.1.2 Results with Holding Applied

After the application of holding on Line 9, it can be investigated to what extent holding affects (potentially improves) the Passenger Reliability and Operator Robustness (as the only indicators to be affected). The results of these indicators are depicted in Figure 6.5. This figure shows the indicators calculation for the holding (right plot) case besides the previous case without holding (left plot), from which improve-

ments especially in PR can be observed for the case with holding applied. The results also suggest that operational measures such as holding may be as important as timetable design measures for improving robustness. For example, the `09LO-1` and `09TT-1` scenarios with holding achieved better PR results than the base scenario without holding. Furthermore, the base scenario with holding performs better in PR compared to the scenarios without holding but with small amount of slack (namely, `09LO+1` and `09TT+1`). This indicates that holding may compensate for a small reduction in slack as a timetable-level modification, or work better than small addition of slack. However, this effect was not observed for scenarios with larger slack reductions, suggesting that the magnitude of the slack reduction remains an important factor.

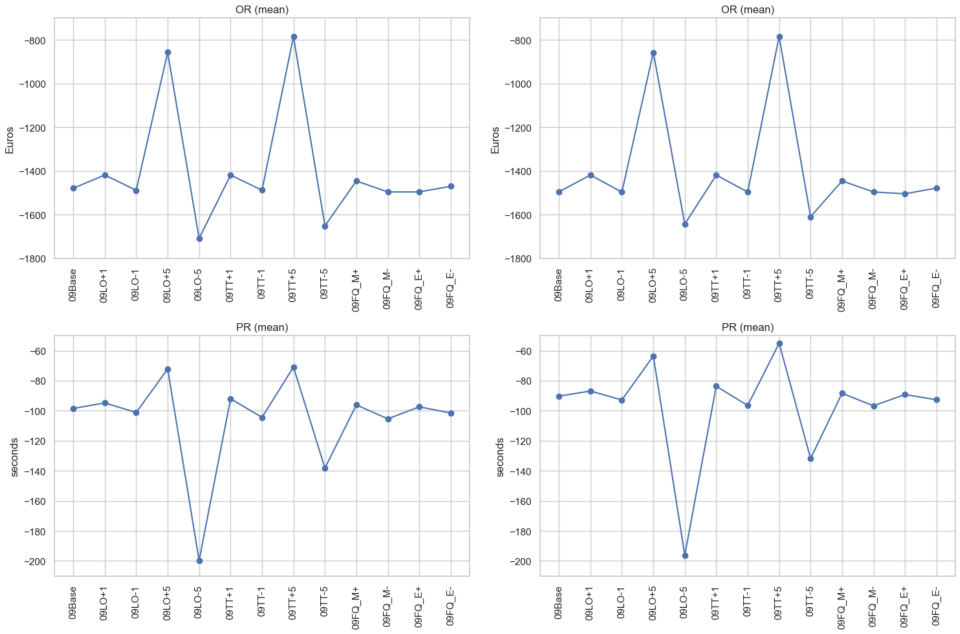


Figure 6.5: Indicators results for Line 9 without (left) and with (right) holding

Figure 6.6 depicts the percentage of improvements in PR and OR after the application of holding. As it can be observed, PR improved for all scenarios, since the holding strategy reduces the possibility of early departures. The effect was most pronounced for the trip-time scenario with higher slack (`09TT+5`), as these initially suffered the most from early departures and therefore benefited the most from the introduction of holding.

For OR, relatively small differences can be observed compared with the changes in PR. Despite the improvement in PR, OR generally deteriorated, except in the scenarios with a 5-second reduction in trip time and a 130-second reduction in layover time. A possible explanation is that these scenarios already have a relatively low probability of early departures. When holding is applied, it further reduces the occurrence of early departures, improving punctuality at the first stop. However, vehicles that would otherwise have departed early and recovered delay later in the route may instead depart on time and experience slightly higher delays, negatively affecting robustness outcomes. As stated earlier, OR is only evaluated at the first stop and changes in a stepwise manner, which may explain the relatively small differences observed between scenarios.

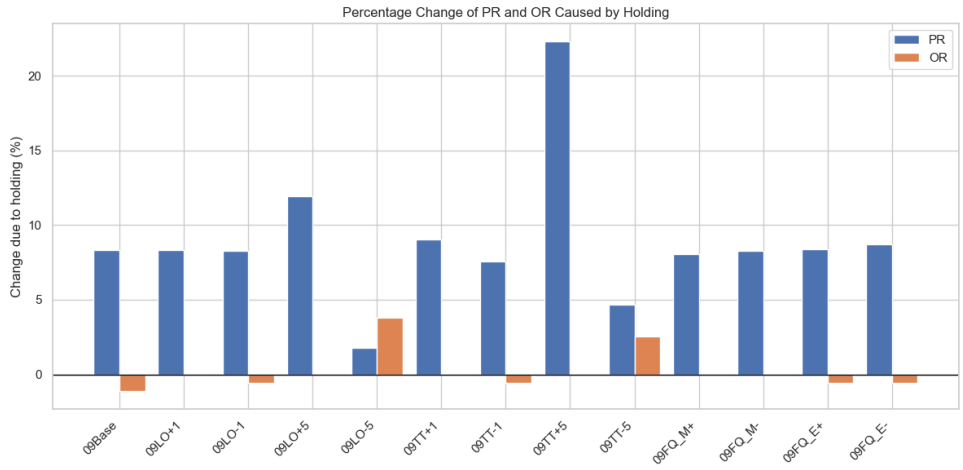


Figure 6.6: Percentage of improvement with holding for each scenario

Overall, the results for PR and OR indicate that the benefits of holding are primarily realised from the passenger perspective, through improved reliability. Although the impact on OR is generally neutral to slightly negative, these changes are relatively small. Since holding does not require additional timetable slack or extra operational resources, it can be considered a cost-effective operational measure. The results further suggest that, for moderate reductions in slack, holding can partially compensate for the loss in passenger reliability and may provide improvements comparable to, or greater than, small increases in timetable slack. However, for larger reductions in slack, timetable design remains the dominant factor, indicating that holding should be regarded as a complement to, rather than a replacement for, sufficient timetable slack.

6.1.3 Merged Indicator - Line 9

First, the set of scenarios for Line 9 is analysed. Figure 6.7 shows the rankings for different passenger - operator (P/O) weightings while keeping the $PR : PE = 3 : 1$ weighting fixed. Five different P/O weighting combinations were tested, and the scenarios were ranked for each of them.

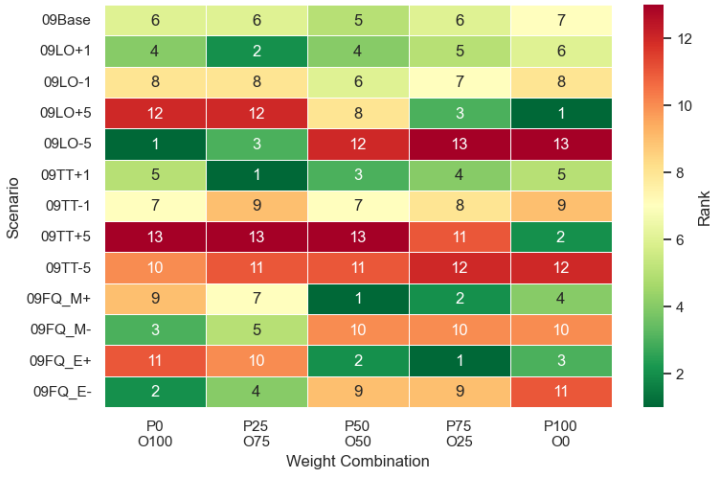


Figure 6.7: Comparison of the alternatives with PR/PE = 3 and P/O with different weighing schemes - Line 9; The colours rank the options from best to worst, starting with green as the top choice.

Several patterns can be observed from this figure. First, `09LO+5` ranks the best when passengers are fully prioritised ($P = 100$), while its rank deteriorates as more weight is assigned to operator objectives.

The opposite pattern can be observed for `09LO-5`. The base scenario is generally positioned around the middle of the rankings, indicating a balanced performance.

For the trip-time scenarios, it can be seen that scenarios with small adjustments (± 1) generally rank better than their counterparts with larger adjustments (± 5), suggesting that moderate changes in trip times provide a better balance between passenger and operator performance. This can also relate to the fact that increasing travel times besides increasing the costs for operator also has drawback for passenger by reducing the speed (efficiency).

Figure 6.8 shows the sensitivity of the rankings to the weighting of PR/PE across a range of passenger-operator (P/O) weightings. As expected, increasing the weight assigned to passenger efficiency over reliability favours the frequency increase scenarios, while the frequency decrease scenarios perform worse if the weight of reliability is higher. A similar trend can be observed for the trip time scenarios, where increased trip times become less attractive if the weight of efficiency is higher, due to their associated reduction in operating speed and passenger efficiency.

For a balanced P/O weighting of 50-50 for instance, `09TT+5` is typically among the worst-performing alternatives across most PR/PE weightings. In contrast, the smaller trip time and layover increases generally perform better, as they provide some reliability benefits while incurring more moderate costs and smaller reductions in operating speed. The frequency scenarios show particularly high sensitivity to the weighting assumptions. When efficiency is prioritised, reducing frequency tends to perform poorly because of the increased passenger waiting times. However, when reliability is given greater importance, frequency reductions become more competitive, as they involve relatively limited reliability losses while reducing operating costs. As a result, their ranking varies substantially depending on the selected weighting scheme.



Figure 6.8: Comparison of the alternatives with different weighing schemes - Line 9

Overall, the results show that the preferred scenario depends strongly on the selected weighting scheme. Scenarios that increase layover or trip times slightly tend to perform better when reliability is given more importance, whereas increasing frequency becomes more attractive when efficiency is prioritised. The results also suggest that small timetable adjustments are generally more favourable than larger interventions, particularly in the case of trip time increases.

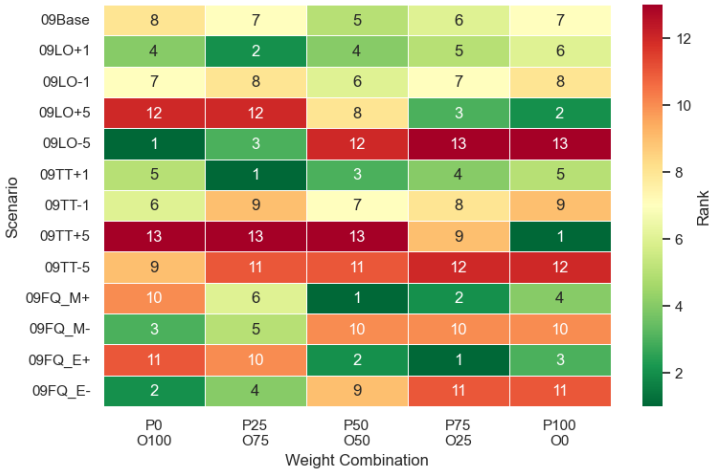


Figure 6.9: Comparison of the alternatives with PR/PE = 3 and P/O with different weighing schemes - Line 9, holding applied

6.1.4 Results of Sensitivity Analysis with Holding Applied

Figure 6.9 shows the sensitivity analysis for Line 9 when holding is applied, and weighting scheme $PR/PE = 3$. As it can be observed, when passengers' weight is high (75 or 100) the ranking of scenarios with more trip time slack (`09TT+5`) improves, resulting from the positive impacts of holding, which offset the negative impacts of the early departure rate, thus providing greater reliability gains.

6.2 Line 19

6.2.1 Individual Indicators - Line 19

Another set of scenarios concerns Line 19, and the resulting indicator values are shown in Figure 6.10.

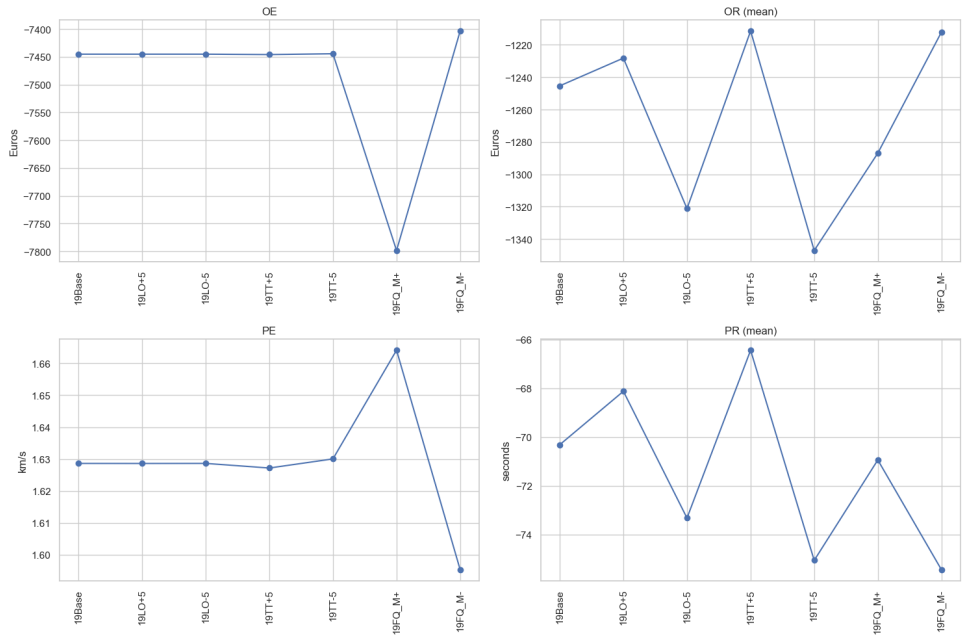


Figure 6.10: Indicator results for line 19

As it can be seen, **Operator Expenses (OE)** remained nearly unchanged for the layover scenarios.

Although the layover increase was implemented by shifting trips in one direction and increasing the minimum required layover time, the constraint was not violated in the terminal with higher layover. The effects of the timetable shifts largely cancelled each other out, resulting in no change in operating hours, and the fleet requirement did not change either. Increasing and decreasing trip times led to only minor changes in operational costs due to small variations in total operating hours. Since the trip-time adjustment was limited to 1.5 minutes, the overall effect remained small. In contrast, increasing the service frequency required one additional vehicle and increased total operating hours, resulting in higher costs. Decreasing the frequency reduced operating hours and therefore achieved the best OE performance.

For **Operator Robustness (OR)** and **Passenger Reliability (PR)**, trip-time increases had a more pronounced effect than layover increases. On the other hand, decreasing trip times had a more significant effect on OR than decreasing the layovers. The same pattern was observed for PR. For PR, running-time slack allows delays to be absorbed during the passenger journey, which is more effective than increasing layover time at terminal locations. Regarding frequency, reducing the frequency improved OR, potentially by increasing punctuality due to reduced operational pressure and greater schedule flexibility. However, PR deteriorated substantially because delays resulted in larger passenger travel-time penalties when services were less frequent, since for instance if a service is cancelled or delayed, passengers need to wait longer for the next service with increased headways. Increasing the frequency slightly reduced both indicators. This may be explained by increased delay propagation between trips (as shown in Figure 6.11, suggesting that less delay recovery took place, which negatively affected both the PR and OR indicators.

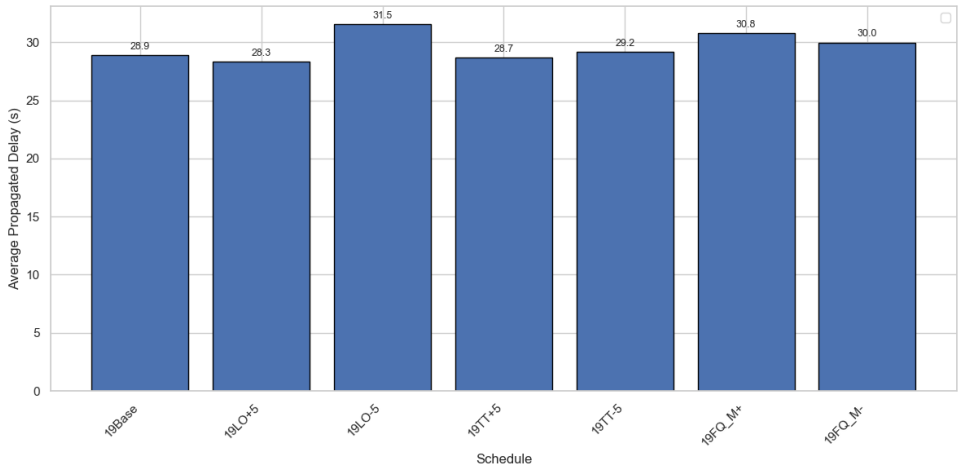


Figure 6.11: Average propagated delay per scenario - line 19

Passenger Efficiency (PE) was primarily influenced by service frequency. As expected, increasing frequency produced the best PE results because passengers experienced shorter waiting times and more travelled distances in total. Decreasing frequency had the opposite effect. Compared with the trip-time and layover scenarios, frequency adjustments had the largest impact on PE.

Figure 6.12 summarises the results above and provides a simultaneous comparison of the scenarios across the four indicators using a spider plot. As can be observed, `19TT+5` performs best in terms of Passenger Reliability (PR) and Operator Robustness (OR), with `LO+5` showing the next best performance, while both maintaining Operator Expenses (OE) as the base scenario. In contrast, `LO-5` and `19TT-5` both deteriorate PR and OR without providing any improvement in OE. Increasing the frequency (`19FQ_M+`) results in lower PR and OR than the base scenario and exhibits the highest OE, although it achieves the highest Passenger Efficiency (PE). Conversely, decreasing the frequency (`19FQ_M-`) yields the lowest OE and improves OR but does so at the expense of both PR and PE.

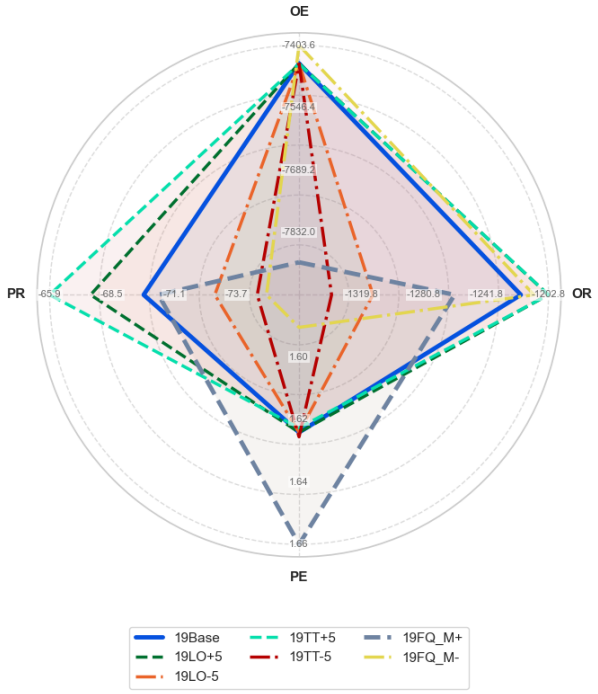


Figure 6.12: Indicator results in spider plots - line 19

6.2.2 Merged Indicator - Line 19

The MCDA results for Line 19 are presented in Figures 6.13 and 6.14.

Looking at Figure 6.13 considerable sensitivity for the frequency scenarios is observed. The decreased-frequency scenario `19FQ_M-` is the best-performing alternative when only operator indicators are considered ($P/O = 0/100$), due to cost savings and some robustness gains (Figure 6.10). However, its ranking deteriorates steadily as the importance of passenger indicators increases, eventually becoming the lowest-ranked alternative when passenger objectives receive full priority ($P/O = 100/0$), mainly due to lower passenger speeds. The opposite behaviour is observed for the increased-frequency scenario (`19FQ_M+`), which improves from one of the lowest ranks under a purely operator-oriented perspective to a relatively better rank when passenger indicators dominate.

The travel-time increase scenario (`19TT+5`) ranks best for most weighting combinations, except when $P/O = 0/100$. This is likely due to the reliability and robustness gains associated with this scenario. Conversely, the travel-time decrease scenario (`19TT-5`) ranks among the worst-performing alternatives, as it is associated with losses in both reliability and operational robustness. The layover scenarios show a similar pattern. The layover decrease scenario generally ranks below the base case, while the layover increase scenario is among the best-performing alternatives across most weighting combinations. One of the main reasons of the poor performance associated with decreasing layover and trip times is potentially the fact that they did not provide substantial cost-savings.

Looking at the second figure (6.14), the sensitivity of the rankings to the PR/PE weighting becomes apparent once passenger and operator objectives receive equal weight, and is most pronounced when passenger objectives receive full priority. In that case, the best-performing scenario shifts from the trip time increase scenario to the frequency increase scenario, as increasing frequency is particularly effective in improving passenger efficiency. At the same time, the trip time increase scenario drops from one of the best-performing alternatives to one of the worst as the weight of passenger efficiency increases, due to its negative impact on passenger speeds. Conversely, the trip time decrease scenario becomes the second-best alternative when $PE = 100\%$ and $P = 100\%$ (thus only considering passenger efficiency), as shorter scheduled travel times improve passenger efficiency. The layover increase scenario remains

among the best-performing alternatives, most often ranking second, while the layover decrease scenario generally performs worse than the base case.

Given equal weight to passenger and operator objectives, the best-performing scenarios across most PR/PE weightings are the increased trip-time and increased layover alternatives. This is likely because the improvements in reliability and robustness outweigh the additional costs and, in the case of increased trip times, the reduction in passenger speeds. In contrast, the increased-frequency scenario is often among the worst-performing alternatives, as the higher operating costs are not compensated by any reliability gains.

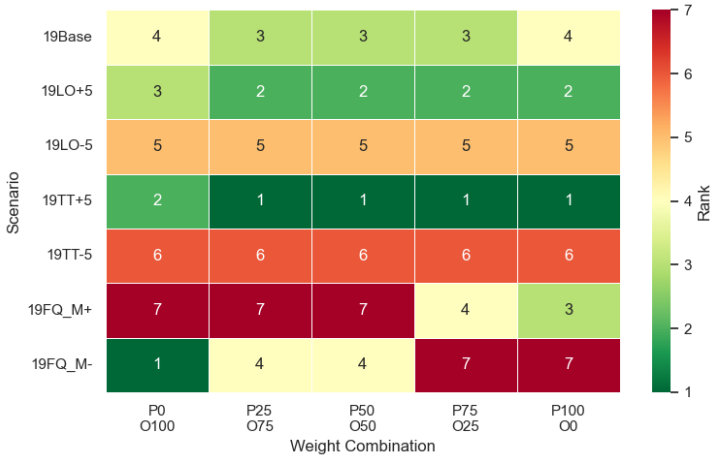


Figure 6.13: Comparison of the alternatives with PR/PE = 3 and P/O with different weighing schemes - line 19

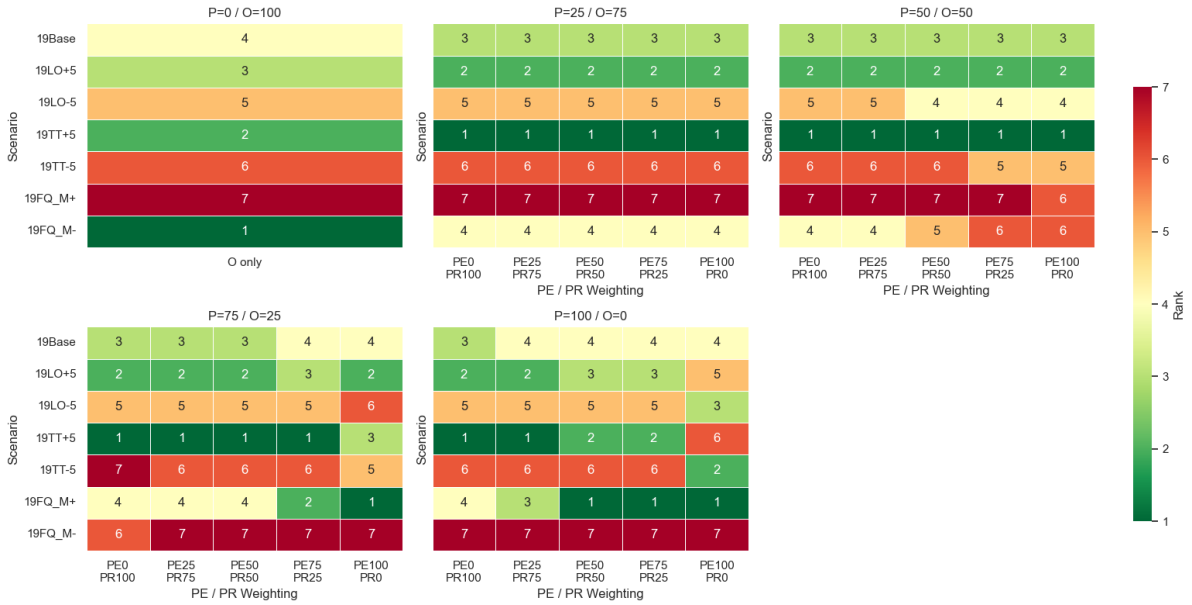


Figure 6.14: Comparison of the alternatives with different weighing schemes - line 19

Overall, the results show a clear trade-off between efficiency and reliability. While increasing layover and trip times perform well under most weighting schemes, increasing frequency becomes the preferred option when passenger efficiency is prioritised. In contrast, the reduced-frequency and reduced-slack scenarios generally rank among the lower-performing alternatives.

6.3 Line 4

6.3.1 Individual Indicators - Line 4

Lastly, layover and trip time alterations were tested on line 4. Here, the scenarios are limited to trip time and layover adjustments, as explained in Section 5.3.

The results of individual indicators are depicted in figure 6.15.

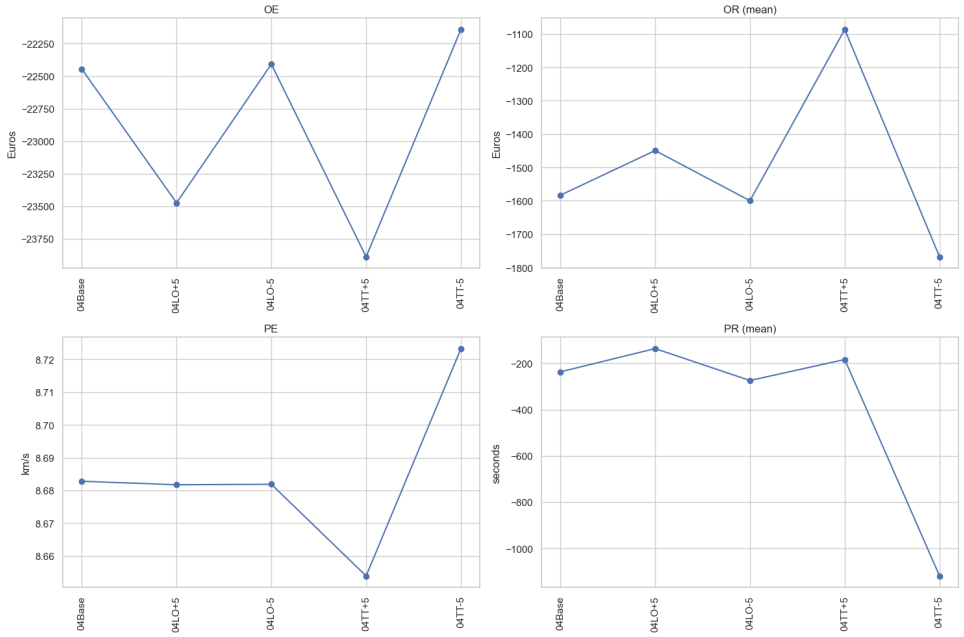


Figure 6.15: Indicator results for line 4

Operator Expenses (OE) behaved as expected, deteriorating when additional slack was introduced due to the increase in operational resources. **Passenger Efficiency (PE)** deteriorated when trip times were increased and improved when trip times were decreased. For **Operator Robustness (OR)**, increasing trip times again had a larger positive effect than increasing layovers, although the trip-time scenarios also altered layovers. For **Passenger Reliability (PR)**, however, increasing layovers resulted in better performance than increasing trip times. A possible explanation is the higher number of early departures observed in the increased trip-time scenario.

Overall, reducing slack had a larger negative impact than the improvements obtained from increasing slack, especially for trip time slack. One explanation is that because the line has dedicated tracks in some parts and traffic interaction is more limited, routine delays are relatively rare. But removing existing slack leaves the system unable to recover when rare disruptions do occur, which ultimately deteriorates the passenger reliability indicator.

Also, removing trip time slack had a more substantial effect on PR and OR rather than removing layover slack. In Figure 6.16 it can be seen that the average propagated delay is higher for `04TT-5` compared to `04LO-5`. This can be potentially due to the fact that more sufficient slack was provided in layover compared to trip times, thus removing trip time slack deteriorates these indicators more.

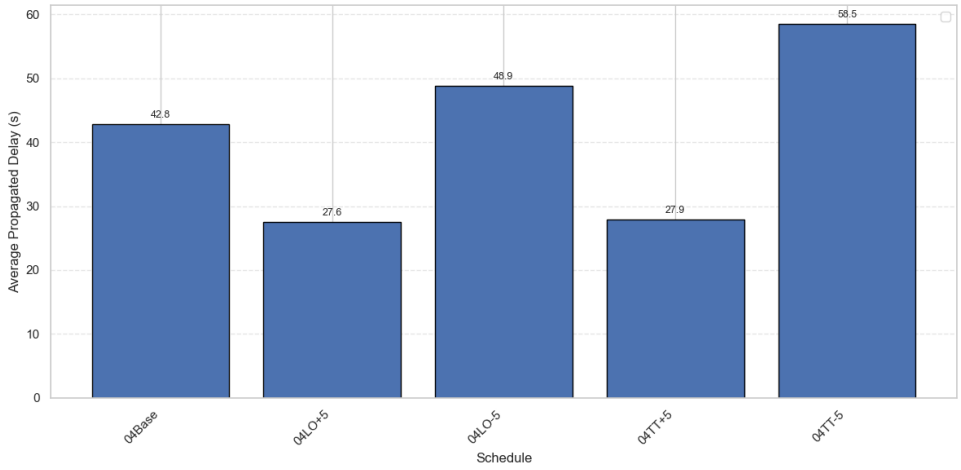


Figure 6.16: Average propagated delay per scenario - line 4

Figure 6.17 summarises the aforementioned details.

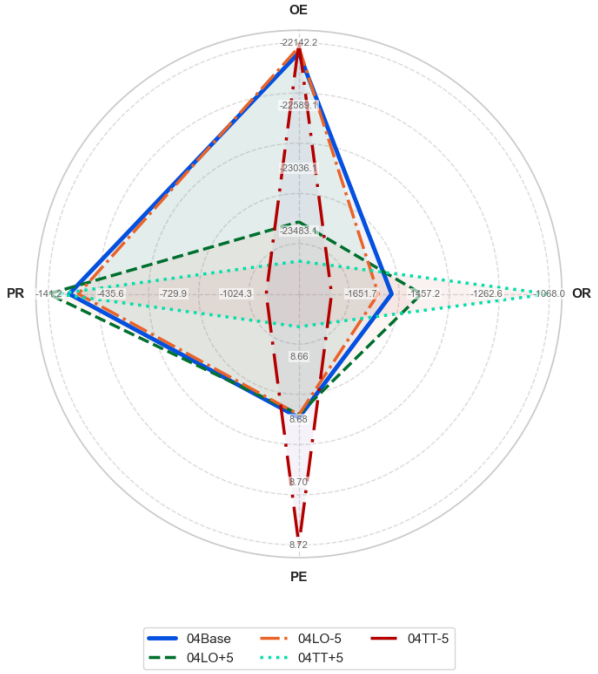


Figure 6.17: Indicator results in spider plots - line 4

6.3.2 Merged Indicator - Line 4

Lastly, for line 4 scenarios MCDA results are presented in the following figures (6.18, 6.19).

Looking at the first figure (6.18), in contrast to the other two lines, the base scenario ranks first or second in almost all weighting schemes. When passenger indicators receive a relatively low weight, the layover decrease and trip time decrease scenarios rank best, as expected due to their lower costs. When passenger indicators receive full priority ($P = 100$), the layover increase scenario appears to perform best, while the trip time increase scenario is not particularly favourable. For an equal weighting of passenger and operator objectives, the base scenario ranks first. When only operator objectives are considered ($O = 100$), the trip time decrease scenario performs best, but it becomes one of the worst-performing alternatives as the weight of passenger indicators increases. Its opposite counterpart, the trip

time increase scenario, also tends to rank among the worst-performing alternatives due to its higher costs and reduced passenger speeds, despite the reliability benefits it provides.

Figure 6.19 shows that the base scenario performs best when operator and passenger objectives receive roughly equal weight and when passenger reliability is weighted more heavily than passenger efficiency. When passenger efficiency receives a higher weight, the reduced trip time scenario tends to perform better due to the improvement in passenger speeds. However, it can also rank among the worst-performing alternatives when passenger reliability receives full weight and operator and passenger objectives are weighted approximately equally. The increased trip time scenario can rank second when passenger reliability is the only passenger objective considered, reflecting its reliability benefits. In most other cases, however, it ranks last or second last because it increases costs and reduces passenger speeds. Lastly, the layover decrease scenario generally performs better than the layover increase scenario. The main exception occurs when passenger objectives receive full priority, in which case the layover increase scenario ranks among the best-performing alternatives, particularly when passenger reliability is weighted more heavily than passenger efficiency.

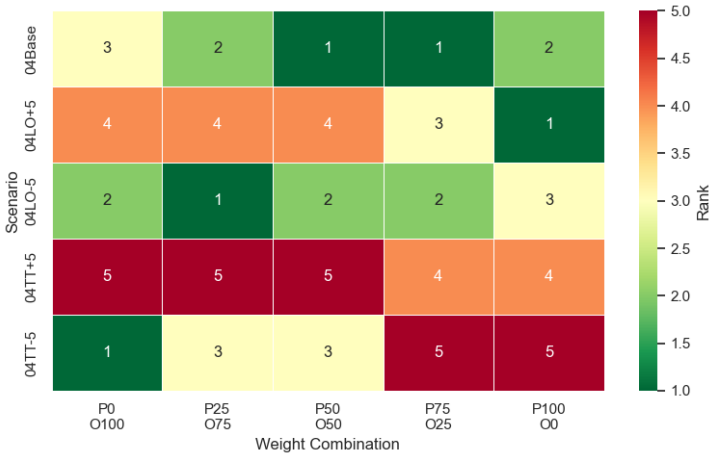


Figure 6.18: Comparison of the alternatives with PR/PE = 3 and P/O with different weighing schemes - line 4

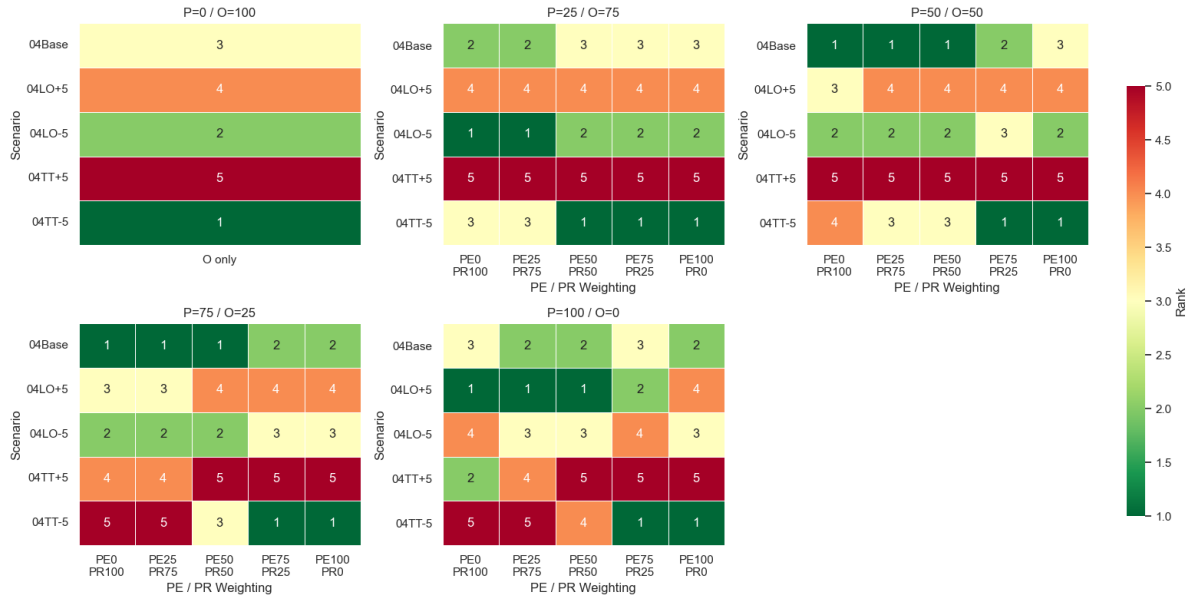


Figure 6.19: Comparison of the alternatives with different weighing schemes - line 4

6.4 Comparing the Results for Different Lines

Comparing the results across the three analysed lines shows both consistent patterns and also some line-specific differences. As expected, increasing timetable slack generally improves passenger reliability and operator robustness. However, the operational consequences of providing additional slack were not the same for all the lines. For example, on Line 19, increasing layover time did not increase the costs because the timetable structure allowed the additional slack to be introduced through timetable shifts without increasing operating hours or fleet requirements. This illustrates that the cost-effectiveness of adding slack depends on the existing timetable structure and vehicle scheduling.

Another difference concerns where additional slack is introduced. On Lines 9 and 19, increasing trip times generally produced larger improvements in PR than increasing layovers, suggesting that distributing recovery opportunities along the route is more effective for passengers than concentrating them at the terminals. In contrast, on Line 4, increasing layover time resulted in larger reliability improvements. A possible explanation is that Line 4 already contains relatively sufficient trip time slack, and adding more slack can result in more early departures and prevent further improvement of PR; thus providing slack at the terminal was more effective for PR for this line. Furthermore, Line 4 operates on dedicated infrastructure with less interaction with other traffic, making en-route recovery less critical than on the other two lines.

The effects of frequency modifications were considerably less consistent. As stated earlier, frequency adjustments also alter vehicle blocking, layover distributions, and passenger waiting times, making their influence on reliability more difficult to predict. For example, increasing frequency slightly (compared to other scenarios) improved passenger reliability and operator robustness on Line 9, whereas it deteriorated both indicators on Line 19. The latter can likely be explained by increased delay propagation resulting from the modified vehicle schedules. This further indicates that frequency changes should not be evaluated independently of their effects on the underlying timetable structure.

Finally, the MCDA results indicate that the potential for timetable improvement differs substantially between lines. While Lines 9 and 19 show clear benefits from selected timetable modifications, the base timetable of Line 4 consistently ranks among the best-performing alternatives across a wide range of weighting schemes. This suggests that the existing timetable of Line 4 is already relatively well balanced and offers less potential for improvement through additional timetable slack. The presence of partially dedicated infrastructure, together with the existing distribution of recovery time, likely contributes to this robustness.

Overall, the results suggest that the effectiveness of timetable modifications depends noticeably on the characteristics of the line and the existing timetable structure more than on the intervention itself. Consequently, although the general patterns are predictable, generic design rules, such as simply adding slack or increasing frequency, are unlikely to produce universally optimal outcomes. Also, the amount of improvement affects the trade-off, making quantified indicators necessary. With quantified effects for each aspect (indicator), and then merging via MCDA, the trade-offs are revealed, enabling choosing the best scenario among some alternatives given the priorities.

Chapter 7

Discussion

This chapter discusses the main findings of the study and places them within the broader context of public transport timetable quality. The results are interpreted from both passenger and operator perspectives, with particular attention given to the trade-offs between reliability, robustness, efficiency, and operational costs. The findings are then compared with existing literature, followed by an assessment of the applicability of the proposed framework, its methodological limitations, and finally the practical implications for timetable planning and evaluation.

7.1 Interpretation of the Results

As shown in the previous chapter (6), when the alternatives are assessed separately using individual indicators, the results are generally as expected. For example, adding slack through increased layover times or trip times tends to improve reliability and robustness, but at the expense of passenger efficiency, if scheduled travel times are increased, and operator efficiency, due to higher operational costs resulting from additional vehicle hours or fleet requirements.

The results further suggest that not all slack is equally effective. In many scenarios, increasing trip times produced larger improvements in reliability and robustness than equivalent increases in layover time. Although, this must be interpreted with caution since, increasing/decreasing the trip times also affects the block assignment and layover distribution in most cases. That said, a plausible explanation can be that when trip times are increased, running-time slack is distributed throughout the route, allowing disturbances to be absorbed progressively during the journey of the passenger. In contrast, layover slack is concentrated at terminal locations and can only contribute to recovery after the completion of a trip. This distinction is particularly relevant from the passenger perspective, since delays experienced earlier during the trip may already have affected waiting times, transfers, or arrival times before terminal recovery takes place.

Another observation is that reducing slack often had a larger negative impact than the positive effect obtained from increasing slack. In several cases, removing recovery time substantially worsened robustness and reliability, while adding equivalent amounts of slack resulted in comparatively modest improvements. This suggests that the relationship between slack and performance is not linear. Existing slack may already play a critical role in absorbing common operational disturbances. Consequently, removing this slack can significantly reduce the timetable's ability to recover from delays, whereas adding further slack may provide diminishing returns.

The frequency scenarios produced more complex results than the slack scenarios. From a passenger perspective, increasing frequency generally improved Passenger Efficiency (PE) through shorter waiting times and greater distances travelled. However, improvements in efficiency were not necessarily accompanied by improvements in reliability and robustness. In several cases, increased frequency slightly worsened both PR and OR. As noted earlier, in the analysed scenarios changes in the frequency inevitably affected the actual layover available between trips, after the blocks are recreated. One explanation is that the additional trips mostly reduced recovery opportunities between consecutive services,

thereby increasing delay propagation within vehicle blocks. Conversely, in a case reducing frequency improved OR while worsening PR. This was potentially the result that when a service is cancelled, for example, passengers need to wait even more for the next service when frequency is lower, but if the propagation of delays are smaller, then the punctuality improves, contributing to a better OR score. This finding illustrates that operator and passenger objectives regarding reliability and robustness concepts do not necessarily favour the same timetable structure.

The holding experiments provide further insight into the relationship between timetable design and operational control. The introduction of holding generally improved both PR and OR by reducing the occurrence of early departures. This result is particularly important because it demonstrates that adding slack can improve reliability by preventing late events, but early departures are also more likely to occur, which can affect passengers even more negatively if no control measure such as holding is applied. In some scenarios, modest reductions in slack combined with holding achieved better reliability outcomes than the base timetable without holding. This suggests that operational measures can partially compensate for reduced timetable recovery time. However, this compensation effect disappeared when larger amounts of slack were removed. This further highlights the positive impact of holding and suggests that better results can be achieved when holding is combined with additional timetable slack.

The final timetable score was found to be sensitive to the selected weighting scheme. A weighting of $PR/PE = 3$ was recommended based on the literature, while the relative importance assigned to passenger and operator perspectives was intentionally left to the decision-maker. The ideal weight distribution depends heavily on specific policy goals and institutional contexts. Because HTM operates under a concession granted by the MRDH, whose strategic goals emphasise accessibility and social inclusivity, a greater weight for passenger objectives (e.g. a passenger-to-operator weighting of 0.6:0.4 to 0.75:0.25) is recommended to align with these priorities Metropoolregio Rotterdam Den Haag (2026). While this weighting balances service quality with operational feasibility, future research, such as willingness-to-pay studies, is required for more precise calibration.

Different priorities can lead to different preferred timetable configurations. For example, under a 50-50 weighting between passenger and operator perspectives, the best-performing intervention for line 9 was an increase in frequency, whereas for line 19 it was an increase in trip times. For line 4, none of the analysed interventions consistently outperformed the base timetable. This further highlights the importance of operational context. A possible explanation is that line 4 already operates with a relatively well-balanced timetable and is less exposed to disturbances due to its greater share of dedicated infrastructure. As a result, additional timetable slack may provide fewer benefits than on lines that are more affected by interactions with surrounding traffic. However, further research involving a larger set of lines would be required to confirm this relationship.

Overall, the results demonstrate that timetable quality is inherently multi-dimensional and cannot be adequately represented through a single-objective indicator. This finding reinforces the motivation behind the framework developed in this thesis, namely that timetable evaluation should explicitly account for both passenger and operator perspectives rather than focusing solely on operational performance or benefits for passengers.

7.2 Comparison with the Literature

The results are broadly consistent with the literature reviewed in Chapter 2. A central theme in previous studies is the trade-off between robustness and efficiency. Lee et al. (2017) showed that additional timetable supplements improve robustness while increasing planned travel times. Similarly, Friedrich et al. (2018) found that measures designed to improve robustness generally increase either passenger travel time or operating costs. The findings of this study support these conclusions. Increasing layovers and trip times improved reliability and robustness but generally worsened efficiency-related indicators and operational expenses. The results of this study are also in line with van Oort (2011) and van Oort

(2014), who argue that traditional reliability measures often fail to represent passenger impacts adequately. Similarly, the findings support Parbo et al. (2016), who emphasise that passenger impacts may differ substantially from operational impacts. Operational unreliability can also increase costs through financial penalties (Yap and van Oort (2018)). Consistent with this, measures for improving reliability such as adding slack improved the operator robustness, which was representative of penalties that are associated with unpunctuality.

This study contributed through the development of a framework that evaluates timetable alternatives from both passenger and operator perspectives within a single decision-support approach. Rather than assessing individual indicators in isolation, the framework combines multiple, potentially competing, dimensions of timetable quality, including passenger reliability, passenger efficiency, operator robustness, and operator expenses.

A particularly relevant aspect of the framework is that the operator perspective is not limited to direct operational expenses. Measures aimed at improving reliability and robustness often require additional resources and therefore increase operating costs. However, improved reliability can also generate financial benefits for operators through reduced penalties associated with poor service performance. By representing operator robustness through penalty costs and combining these with direct operational expenses, the framework acknowledges that timetable decisions affect operators through both resource requirements and service quality. As a result, reliability improvements are evaluated not only as operational benefits but also through their financial implications.

Similarly, the passenger perspective incorporates multiple dimensions of passenger experience. Efficiency is represented through scheduled travel times and related travel distances, reflecting the speed of the service, while reliability is represented through excess travel time, capturing the effects of delays and service variability on passenger journeys. This distinction allows passenger experience to be assessed through both the planned performance of the timetable and its ability to deliver consistent service in operation.

Overall, this study sought to develop a set of timetable quality indicators based on concepts and measures discussed in previous studies in order to quantify the effects of timetable modifications. Modifications involving additional slack through layover and trip-time changes, as well as frequency alterations, were evaluated. By applying these indicators to a range of timetable alternatives, the study demonstrates how the impacts of timetable design choices can be assessed from both passenger and operator perspectives, and how these indicators can be combined under different stakeholder priorities.

7.3 Applicability of the Study

One of the objectives of this study was to assess whether the proposed framework could be applied beyond a single case-study line. The analysis of lines 9, 19, and 4 suggests that the framework is transferable to different operational contexts in tram networks, such as lines with varying service frequencies and infrastructure characteristics (e.g. mixed-traffic operation versus more dedicated infrastructure). The results showed that the general relationships between reliability, robustness, efficiency, and operational costs remained consistent. For example, increasing slack time generally improved reliability, but at the expense of higher operational costs. However, the magnitude of these effects varied between lines.

The results indicate that the effectiveness of specific timetable measures is highly context-dependent. For example, additional trip-time slack appeared more beneficial on line 19, whereas additional layover time often produced larger reliability improvements on line 4. Such differences are likely related to factors such as demand patterns, line length, operational characteristics, and disturbance profiles. However, the relationship between these contextual factors and the observed performance effects could not be quantified within the scope of this study and therefore requires further research.

Another important observation is that the impact of timetable modifications depends not only on the

characteristics of the line, but also on the structure of the existing timetable. For example, on line 9, increasing trip times caused layover constraints at terminal stations to be violated, requiring vehicle blocks to be reconstructed. As a result, changes in trip times also indirectly altered layover times. In contrast, for line 19, trip-time modifications could generally be implemented without requiring block reconstruction, allowing the effects of trip-time changes to be assessed more directly. This demonstrates that the practical implementation of timetable measures may differ substantially between lines, even when the same intervention is applied.

Infrastructure characteristics may also influence the effectiveness of timetable modifications. It was observed that line 4, which operates to a greater extent on dedicated infrastructure, appeared less sensitive to timetable adjustments, with the base scenario performing well under many weighting schemes. This may indicate that the timetable was already relatively well balanced or that additional recovery time provides fewer benefits when external disturbances are more limited. However, the number of analysed lines remains too small to draw definitive conclusions regarding the role of infrastructure.

The results therefore suggest that the framework itself is broadly applicable, while the effectiveness of specific timetable measures remains strongly dependent on the operational context in which they are implemented.

7.4 Methodological Limitations

The research is subject to several limitations related to practical feasibility, data availability, and time constraints. These limitations should be considered when interpreting the findings.

- Due to the complexity of modelling transfers within the available time frame, connections were kept outside the scope of this study. As a result, missed transfers and the associated additional waiting times were not included in the analysis. This may lead to an underestimation of passenger reliability impacts.
- Passenger behavioural responses were not modelled. More specifically, when using the realised data, missing events were assumed to indicate that a service was unavailable at a particular stop (e.g. due to detours or cancellations). Additional waiting time was then calculated based on the interval between the last available service and the first available service after the disruption. The resulting excess waiting time may therefore be equivalent to one or more full headways, depending on the number of missing events. This approach may overestimate excess waiting time, as passengers may choose alternative services when their intended service is unavailable. More generally, behavioural adaptations could influence both passenger loads and perceived timetable quality, but modelling such responses was beyond the scope of this research.
- The interpretation of the frequency scenarios requires caution. As acknowledged in the methodology, modifying frequencies also altered layover distributions and vehicle block assignments. Therefore, observed changes cannot be attributed solely to frequency modifications. Similar issues occurred in some trip-time scenarios, where changes in vehicle block allocation also affected realised layovers. In addition, frequency changes affect the level of service provided to passengers. Consequently, the impacts of frequency modifications reflect both service-level changes and operational reliability effects, making them fundamentally different from timetable modifications that only alter slack. However, in this study the effect of frequency alterations for passengers was simplified in reliability and efficiency gains.
- The simulation assumes that future disturbances follow patterns observed in historical AVL data from a specific time frame, namely working days between 24 November and 12 December 2025. Disturbance patterns during other periods (e.g., weekends or the summer season) may differ, meaning alternative empirical distributions would be required for those scenarios. Furthermore, while this specific window provides a realistic empirical baseline, long-term disturbance patterns

can shift due to infrastructure works, weather conditions, operational policies, or demand fluctuations.

- In the simulation, it is assumed that if a delay is propagated from the previous trip, the next trip cannot start early. The departure delay at the origin is therefore defined as the sum of the propagated delay and any initial delay. Early departures are only possible when sufficient layover time is available to fully recover delays from the previous trip. This assumption may lead to an overestimation of early departures for trips starting after generous layovers, as the modelling logic allows early departures to occur more frequently in such situations. Applying holding to some extent was able to mitigate these effects. Although the holding model used here was developed based on simplified assumptions, after applying holding, the KS statistic for the base scenario decreased slightly, indicating improved agreement between the simulated and realised delay distributions. Thus, the model was considered a simplified but sufficiently accurate representation of holding behaviour for predicting holding strategies and computing indicators closer to reality.
- It was assumed that driving times follow the same distribution regardless of whether scheduled run times are increased or decreased. This assumption was based on the premise that realised driving times are largely independent of the timetable, unless operational control measures such as holding are applied. To explore this effect, a holding experiment was conducted to assess its influence on the performance indicators. However, because of the given time frame of the project, this experiment was limited to line 9 and relied on assumptions regarding when holding occurred in the historical data. The probability of holding was estimated based on the magnitude of negative delays and the location of the stop.
- Initial delays are calculated using the offsets described in Section 3.3.1, where propagated delay is approximated as the difference between the layover time and the delay at the end of the previous trip. Consequently, the results depend strongly on the extent to which layovers absorb incoming delays, making it difficult to distinguish propagated delays from delays originating from other sources. It is assumed that layover time, minus two minutes required for driver activities, can be used to absorb delays. For the cases with the layover equal to or less than 2 minutes, this layover buffer was assumed to be zero.
- The demand distribution across stops and time-of-day regimes is based on check-in counts, assuming that these proportions are representative of total demand. This excludes certain passenger groups, such as travellers using alternative payment methods or travelling without checking in. Similarly, vehicle loads are estimated using check-in and check-out data. Estimates of passenger loads along route segments are sensitive to missing check-outs, which may introduce inaccuracies. Overall, these issues may affect the estimation of passenger distributions across stops and route segments.
- Crowding strongly influences passenger behaviour and satisfaction (van Oort et al. (2015)). Improvements in reliability may reduce excess waiting times, lower headway variability, and reduce bunching effects, thereby indirectly alleviating crowding. However, incorporating crowding would require detailed behavioural modelling and was therefore beyond the scope of this research. As a result, crowding effects are not modelled in this study.
- The number of analysed scenarios remains limited. The objective of this study was not to optimise timetables exhaustively but to demonstrate and evaluate the proposed approach. Consequently, the results should not be interpreted as identifying the optimal timetable configuration, as alternative timetable scenarios that outperform those analysed may exist.
- Time constraints form a broader limitation of the study. The twenty-two-week project duration restricted both the number of scenarios that could be analysed and the depth of the sensitivity analyses that could be performed.

The above limitations restrict the scope of the analysis and should therefore be considered when interpreting the results. At the same time, the adopted assumptions allowed the study to remain focused and feasible within the available project duration.

7.5 Practical Implications

The findings have several practical implications for public transport planning.

First, timetable evaluation should incorporate both passenger and operator perspectives. Reliance on a single indicator risks overlooking important impacts on either operational performance or passenger experience. By combining reliability, efficiency, robustness, and operating costs into a merged indicator, planners can assess the overall consequences of timetable modifications rather than optimising individual objectives in isolation.

Second, timetable planners should consider not only the amount of slack but also its distribution. The results indicate that trip time slack may be more effective than terminal layover slack in improving passenger reliability. However, this is not always the case; one drawback of increased run times is the higher probability of early departures if no control measure is applied. Thus, there is a trade-off between the improved disturbance absorption provided by running-time slack and its negative effects, such as longer scheduled travel times and a higher likelihood of early departures. In this regard, whether operational measures are in place can also be important. Schedule-based holding was analysed in this study and showed that modest reliability improvements can sometimes be achieved through operational control without requiring substantial timetable modifications.

Third, using the merged indicator can support planners during the timetable design process by enabling systematic comparison of alternative timetables before implementation. Planning tools like HASTUS allow planners to come up with multiple alternatives for timetables, and choosing the best timetable requires knowing the impact of each alternative's implementation, which can be described through the indicators and methodology used in this study. Rather than relying primarily on expert judgement or focusing on one-dimensional performance measures, planners can quantify the expected effects of changes in running times, layovers, or service frequencies on multiple stakeholder objectives simultaneously. This allows potential trade-offs to be identified at an early stage and supports more transparent decision-making.

Finally, as expected, the MCDA results demonstrate that there is no universally optimal timetable. The preferred alternative depends strongly on stakeholder priorities and organisational objectives. The framework aimed to make these trade-offs as explicit as possible. Using the recommended weighting of 3 between passenger reliability and passenger efficiency, the relative weighting assigned to passenger and operator perspectives can still substantially affect how timetable alternatives compare with one another. Using a weight of choice for this purpose, the framework can be used in order to assess the quality of alternative timetables.

7.6 Conclusion

In summary, the discussion highlights that evaluating public transport timetables requires a delicate balance between conflicting passenger and operator objectives. Also, interventions must be tailored to the operational context of each individual line or potentially scheduling unit, as the implications might differ. Despite limitations, this study aimed to provide a tool for assessing the timetable quality in a more balanced setting, that can be used to evaluate alternative urban public transport timetables.

Finally, the MCDA results demonstrate that there is no universally optimal timetable. The preferred alternative depends strongly on stakeholder priorities and organisational objectives. Even when adopting the recommended weighting of three between passenger reliability and passenger efficiency, varying the

relative importance assigned to passenger and operator perspectives can substantially alter the ranking of timetable alternatives. Consequently, the framework should not be viewed as prescribing a single best timetable but rather as a decision-support tool that enables planners to evaluate alternatives under different policy priorities and stakeholder preferences. This flexibility makes the framework suitable for supporting evidence-based timetable planning across a range of operational contexts.

Chapter 8

Conclusion

This chapter concludes the thesis by summarising the objectives, methodology, and findings of the study. The research focused on developing timetable quality indicators that integrate both passenger- and operator-oriented perspectives, and it was performed within the context of the HTM urban public transport network. By combining reliability, efficiency, robustness, and operational cost indicators within a Multi-Criteria Decision Analysis framework, the study aimed to provide a more transparent basis for timetable evaluation. This chapter first reflects on the answers to the research questions and the main findings, followed by the contributions of the study, limitations, and finally recommendations for future research.

8.1 Research Overview

The main research question of this study was:

To what extent do alternative timetable configurations differ in quality, as evaluated using a composite data-driven indicator comprising passenger- and operator-oriented reliability, efficiency, robustness, and expense metrics under empirical disturbance conditions?

To answer the research questions, the study first developed four indicators: Passenger Reliability (PR), Passenger Efficiency (PE), Operator Robustness (OR), and Operator Expenses (OE). The framework was applied to HTM tram line 9 (Vrederust–Scheveningen), 4 (De Uithof - Lansingerland-Zoetermeer) and 19 (Leidschendam - Delft), in order to represent more varied operating and route characteristics. Scheduled, operational, and passenger data were derived from AVL and AFC systems, provided by HTM. Subsequently, an empirical simulation was performed to generate realised operational data for a set of new timetable scenarios for which no observed data existed. Alternative timetable scenarios were created by modifying layover times, trip times, and service frequencies, in order to allow the effects of these changes on the indicators to be evaluated. Finally, the scenarios were compared using MCDA in order to combine the indicators and analyse how different weighting schemes influence the merged quality indicator of the timetables. Adopting a weighting of choice for passenger and operator objectives, the quality of the timetables can be assessed using this merged indicator.

8.2 Main Findings

The first research question addressed how reliability/robustness and efficiency/expenses indicators are quantified from the passenger and operator perspectives using planning and operational data from AVL and AFC systems. The development of these indicators was based on literature and concepts discussed in the previous studies, with some refinements. Operator Expenses (OE) represented resource requirements, including vehicle operating hours and fleet size. Operator Robustness (OR) represented operational performance through penalties associated with punctuality requirements. From the passenger perspective, Passenger Efficiency (PE) was developed as an indicator of effective passenger travel speed by incorporating scheduled waiting time (average), scheduled in-vehicle time (average), and travelled

distances (total). Passenger Reliability (PR) was quantified based on additional passenger travel time caused by operational variability, consisting of excess waiting time and excess in-vehicle time. Together, these indicators aimed to provide a more balanced overview and a more integrated representation of timetable quality than traditional measures or focusing on one perspective alone.

The second research question examined how operational data could be simulated for timetable scenarios for which no realised data existed, since the new alternatives were not in operation and need to be tested before being put in operation. To address this, an empirical simulation was performed using realised AVL data from the base scenario. The simulation process incorporated empirical distributions of initial delays, driving times, and cancellation probabilities in order to reproduce operational disturbances under alternative timetable configurations. This enabled the generation of simulated realised operations for the alternative schedules and therefore allowed the calculation of passenger reliability and operator robustness indicators for scenarios that had not been implemented in practice.

The third research question analysed how timetable adjustments, including changes in layover time, trip time, and frequency, affected timetable quality when evaluated through the developed indicators. The results demonstrated clear trade-offs between reliability, robustness, efficiency, and operational costs. Increasing layover times and trip times generally improved Passenger Reliability (PR) and Operator Robustness (OR), as additional recovery time reduced delay propagation. With less propagation, passengers suffer less from excess travel time. However, these improvements worsened Operator Expenses (OE). The results also showed that trip time adjustments might have a stronger influence on reliability improvements than layover adjustments, however they tend to reduce the Passenger Efficiency score due to reduced speeds. Frequency-related scenarios were more difficult to interpret since results stem partly from different layover distributions. Increasing frequency is associated with more costs but a better speed (since more travelled distances are provided for passengers, and less average waiting times), resulting from decreased average waiting time and more travelled distances. Furthermore, although the general patterns observed were similar across the analysed lines, the relative magnitude of the effects varied. The impact of a timetable intervention depends not only on the type of measure applied, but also on how the timetable is structured, the extent to which other timetable characteristics are affected, and potentially the operational context in which the line operates, such as the level of infrastructure segregation.

The fourth research question focused on how different indicator weightings influence the comparative evaluation of timetable scenarios. The MCDA results showed that the ranking of timetable alternatives strongly depends on the selected weighting assumptions. Timetable scenarios that performed favourably from a passenger perspective were often less favourable from an operator perspective, and vice versa. However, some alternatives performed relatively well under a wide range of weighting schemes because improvements in passenger reliability and operator robustness were aligned. To obtain a balanced overall indicator, MCDA was applied using multiple weighting combinations, while a constant weighting of $PR/PE = 3$ was adopted based on findings from the literature. Despite this, the results remained sensitive to the relative importance assigned to passenger and operator perspectives, which was intentionally left to the decision-maker. Even under a fixed 50-50 weighting between passenger and operator perspectives, different timetable interventions emerged as the preferred alternative for each analysed line. This demonstrates further the dependency on the context.

8.3 Research Contribution

This study contributes by developing an integrated timetable evaluation framework that combines passenger reliability, passenger efficiency, operator robustness, and operator expenses within a single methodology. Building on concepts and measures discussed in previous studies, the framework provides a structured approach for assessing multiple dimensions of timetable quality simultaneously and exploring the trade-offs between passenger and operator objectives.

8.4 Limitations and Recommendations for Future Research

Several limitations should be acknowledged for this study. First, passenger transfers and missed connections were not modelled and were thus excluded from the reliability calculations due to modelling complexity. Second, behavioural adaptations of passengers, such as departure time changes and route choice changes, were not considered. Furthermore, crowding effects and passenger comfort were not included in the passenger indicators in order to stay in the feasible scope of the study. Finally, the range of possible timetable configurations is endless, thus only a limited set of timetable alternatives was analysed.

Future research could extend the framework by incorporating transfer reliability and missed connections into the passenger reliability indicator, since this can significantly affect the excess travel time that passengers might experience and thus the reliability. Passenger behavioural responses and crowding effects could also be integrated in order to provide a more comprehensive representation of passenger experience in terms of reliability.

Further work could additionally incorporate more detailed operational constraints, including crew scheduling and depot operations, and line capacity, to improve the representation of operational feasibility and costs.

Another extension could be a more accurate modelling of holding and adjusting the actual distribution of the driving times distributions accordingly, since holding can have a rather significant effect on improving the performance of a timetable in operation by mitigating the negative effect of early departures for passengers and also preventing fines for the operator.

Finally, applying the framework to a larger number of lines and timetable scenarios would provide further insight into how operational context influences the effectiveness of different timetable interventions.

8.5 Conclusion

Overall, the study demonstrated that timetable quality is inherently multi-dimensional and depends on balancing competing passenger and operator objectives. The results showed that timetable modifications can have substantially different effects on reliability, robustness, efficiency, and costs, and that the preferred timetable configuration depends on both stakeholder priorities and operational context. The study developed a set of indicators and used a methodology that was capable of describing timetable quality, and provided a transparent and data-driven basis for supporting timetable evaluation and decision-making in urban public transport systems.

AI Statement

In this study, artificial intelligence (AI) tools were used as coding assistant, as a code refiner and optimisation tool. AI tools were also used for minor language refinement in order to improve clarity of the writing. The tools utilised in this study include ChatGPT (OpenAI) and Google Gemini. All outputs generated by AI tools were reviewed and adapted by the author. The author remains fully responsible for the analysis, methodology, interpretations and conclusions of this research.

Definitions

Definition 8.1 (HASTUS)**.** HASTUS (Horaires et Assignations pour Système de Transport Urbain et Semi-urbain) is a software suite used to generate transit timetables, vehicle blocks, and crew schedules.

Definition 8.2 (Automated Vehicle Location (AVL))**.** Automated Vehicle Location (AVL) is a tracking system that is used for automatically determining and transmitting the location of a vehicle.

Definition 8.3 (Automated Fare Collection (AFC))**.** Automated Fare Collection (AFC) is an integrated system that automates ticketing, fare validation, and revenue processing across a transit network.

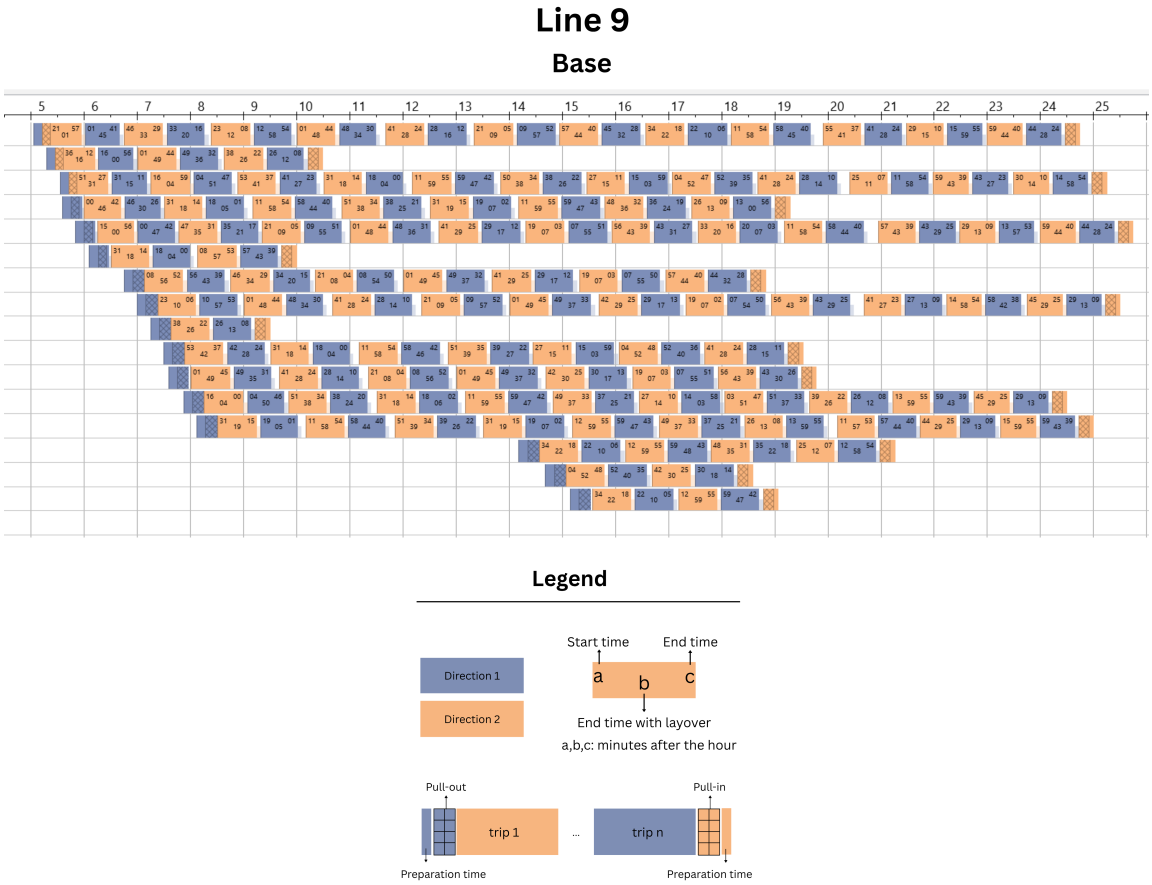
Definition 8.4 (Opportunity Trips)**.** Opportunity trips are operated primarily for vehicle positioning between depot and terminal, while also carrying passengers.

Definition 8.5 (Drivers' activity time at the end point)**.** Time reserved in the schedule for drivers to take their mandatory breaks, as required by labour agreements (Dutch: keertijd).

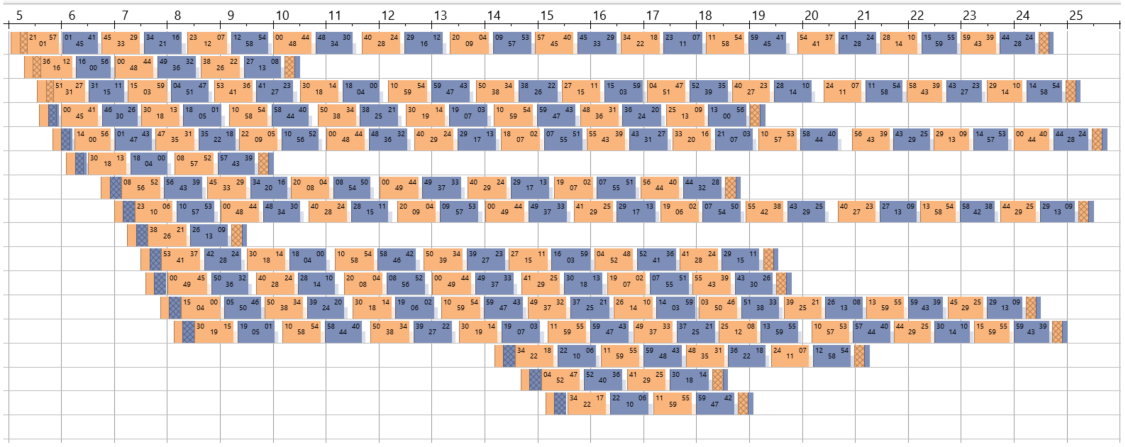
Appendices

Appendix A

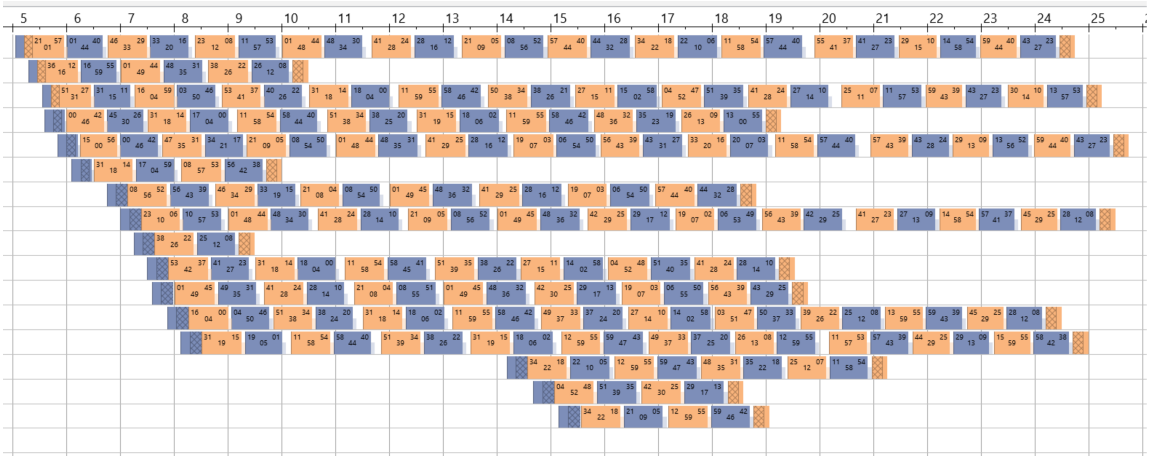
Block diagrams



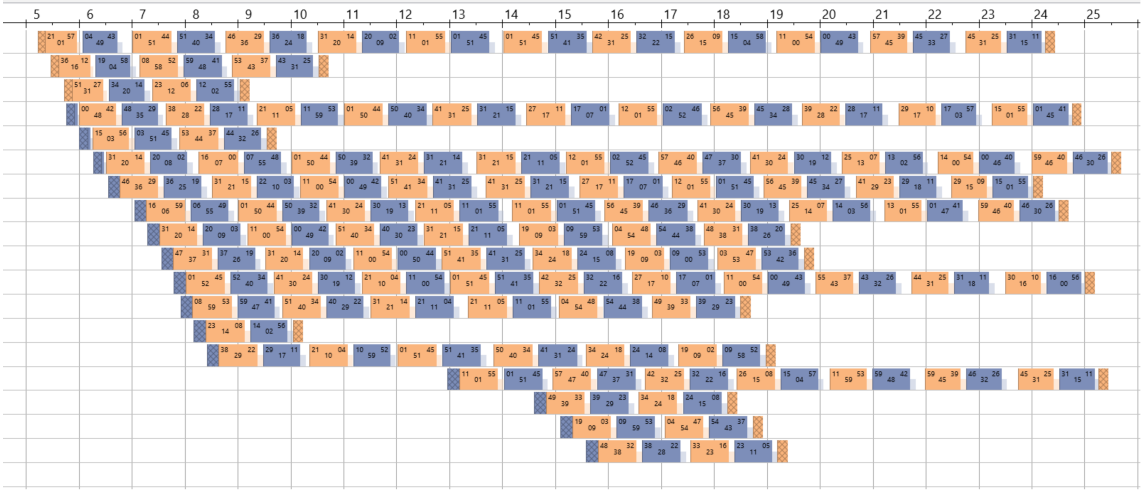
09LO+1



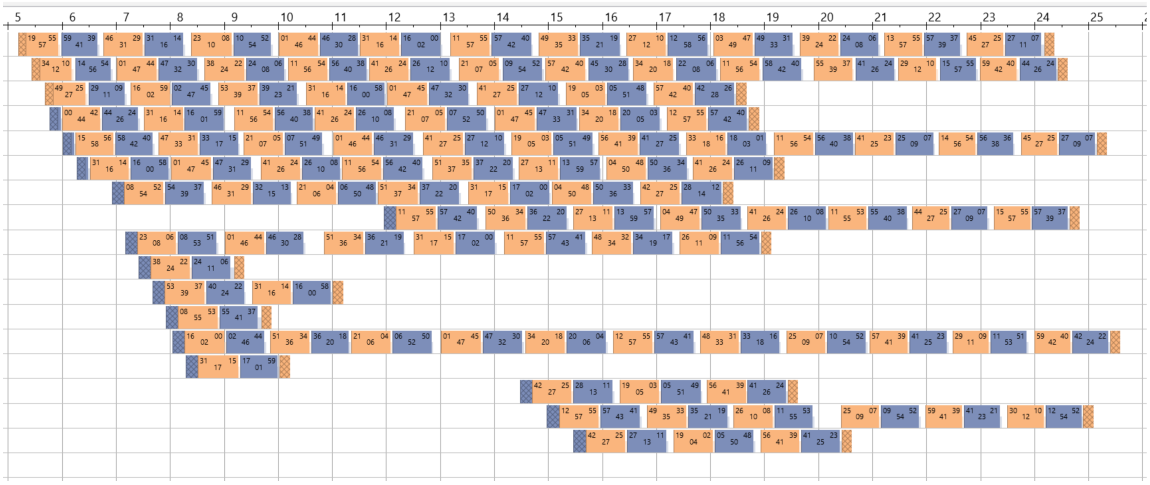
09LO-1



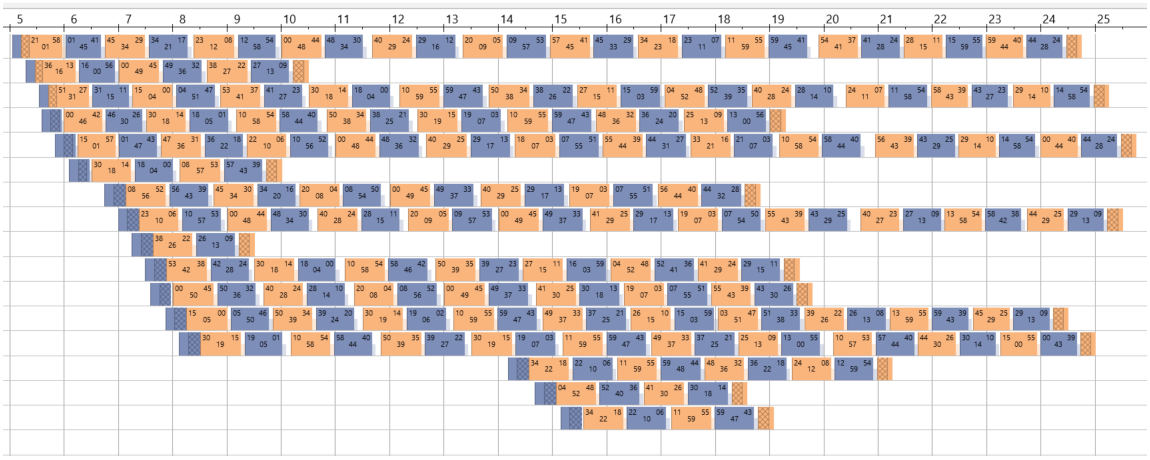
09LO+5



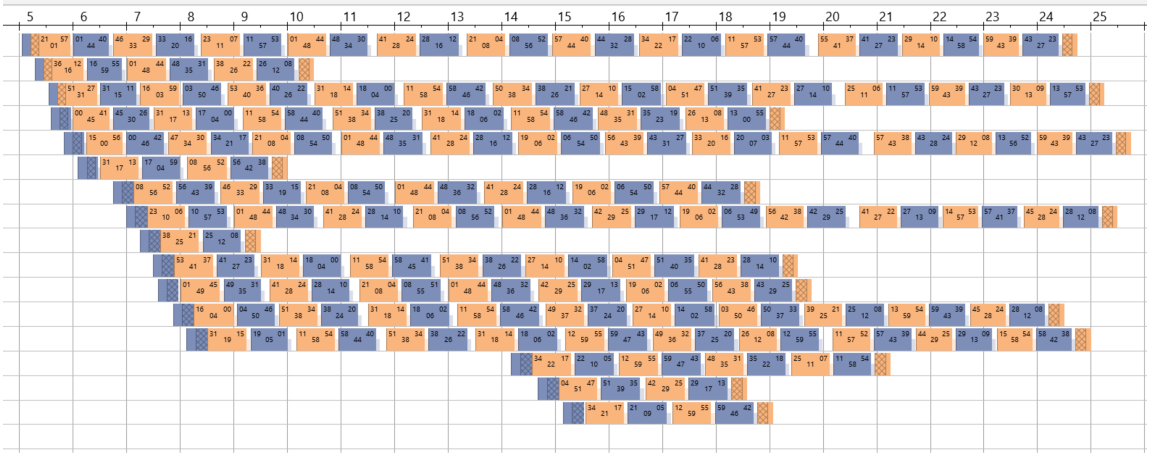
09LO-5



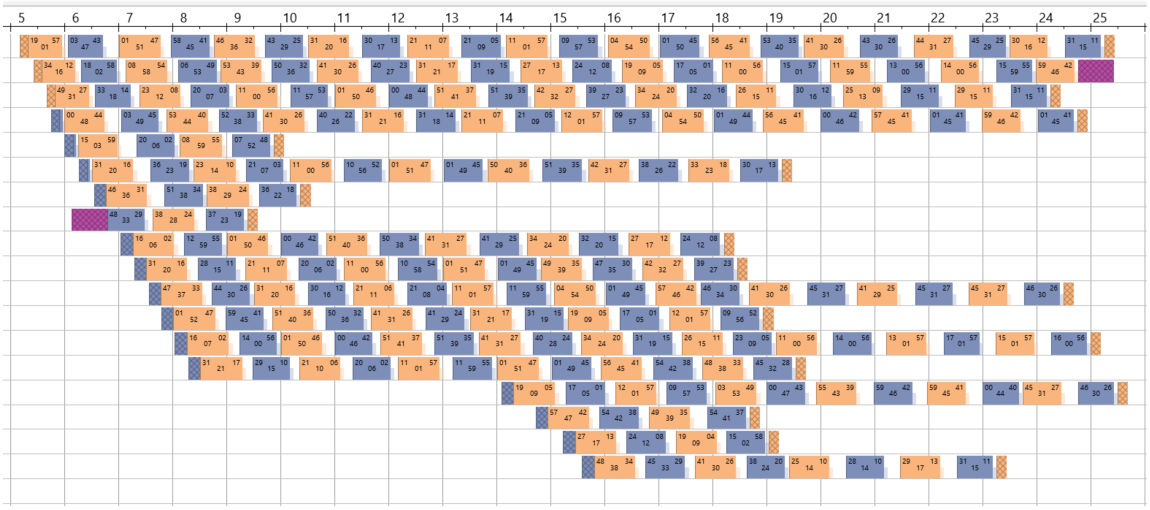
O9TT+1



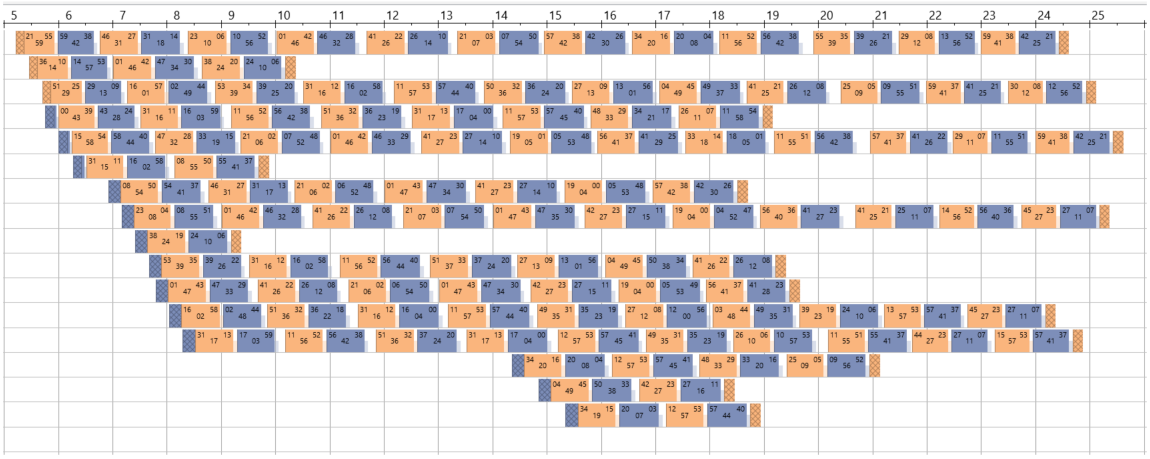
O9TT-1



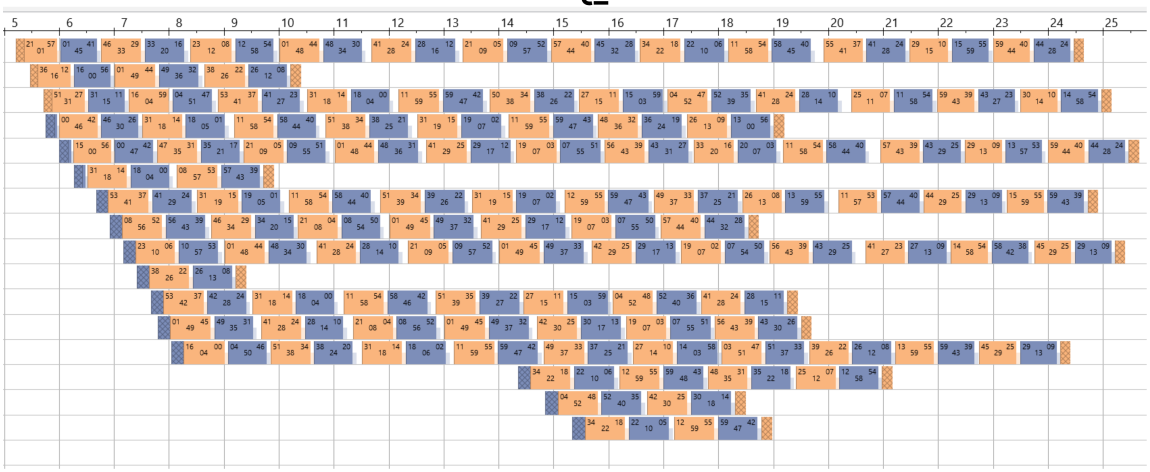
09TT+5



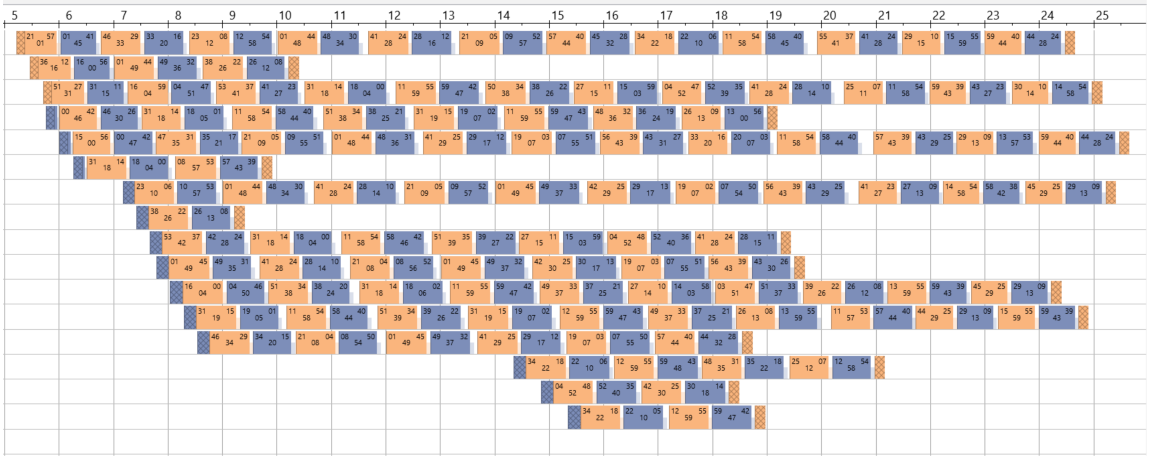
09TT-5



09FQ_M+



09FQ_M-



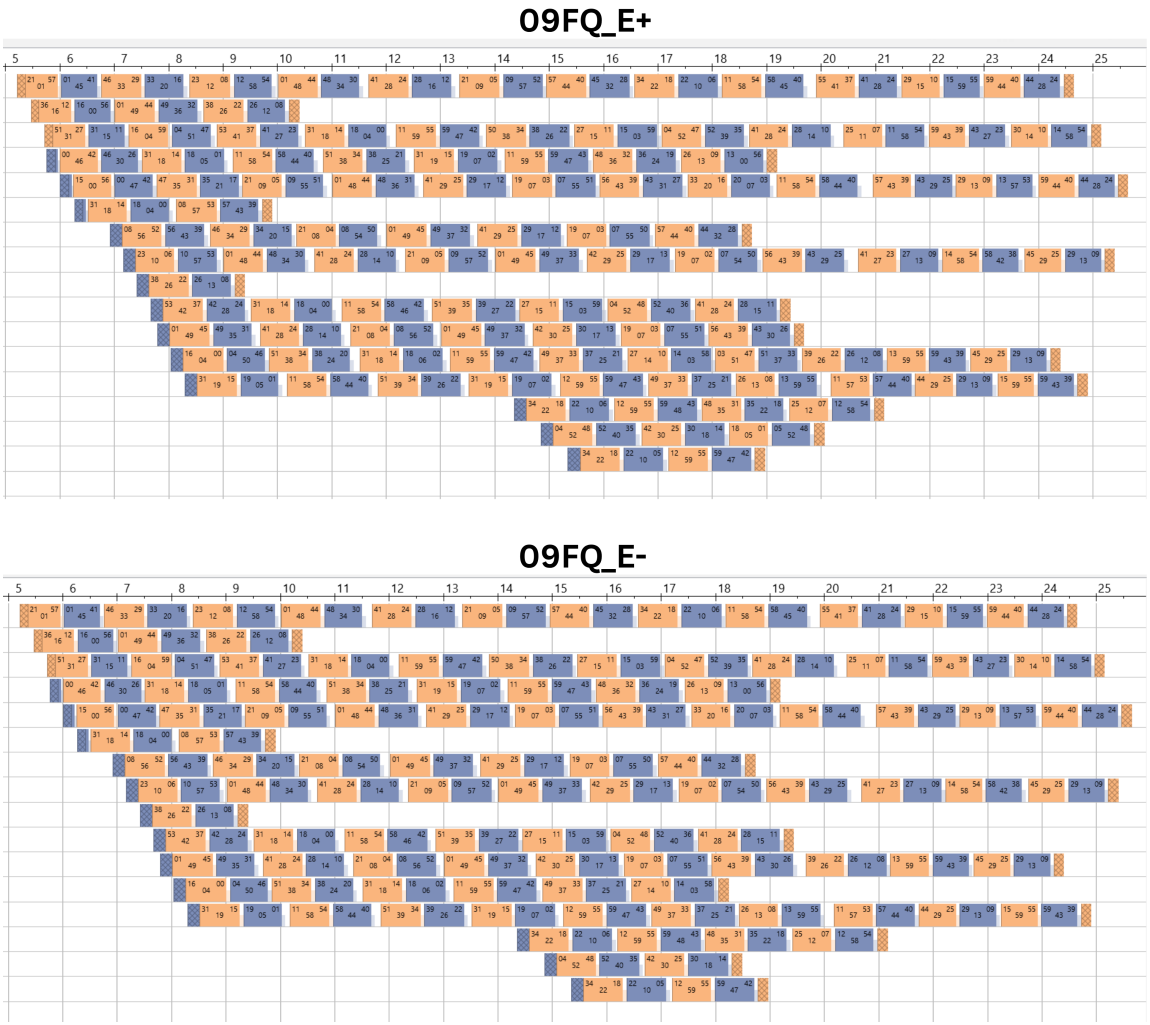
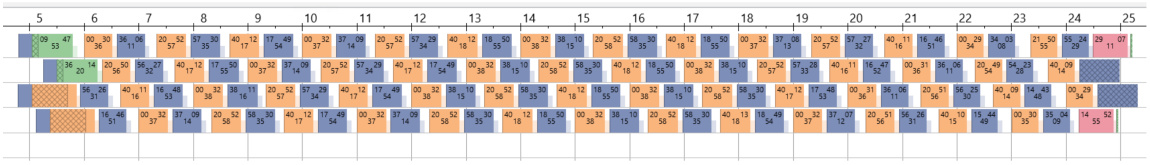


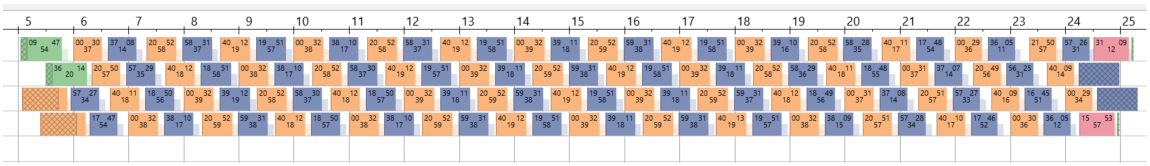
Figure A.1: Line 9 Blocks

Line 19

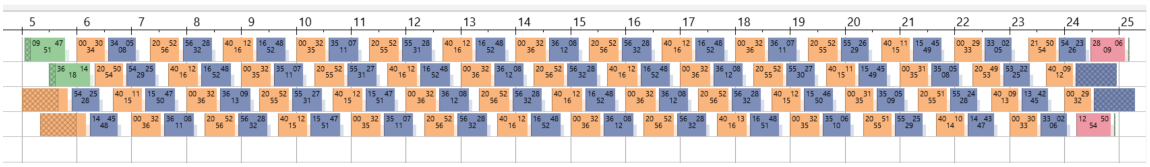
Base



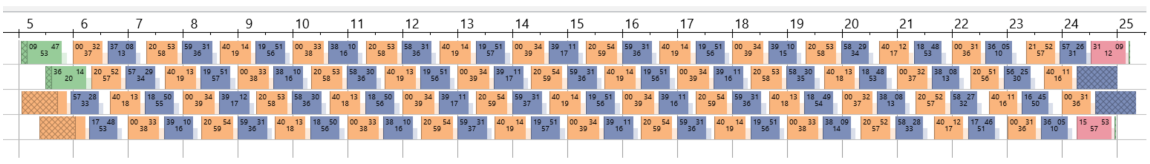
19LO+5



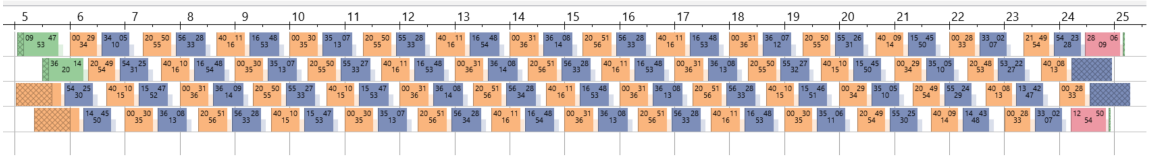
19LO-5



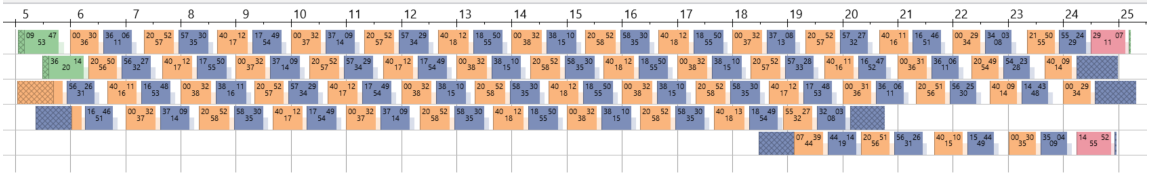
19TT+5



19TT-5



19FQ_M+



19FQ_M-

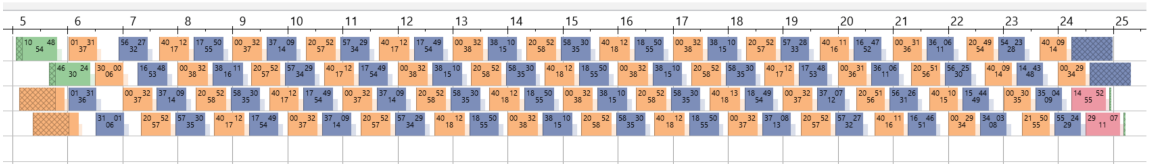
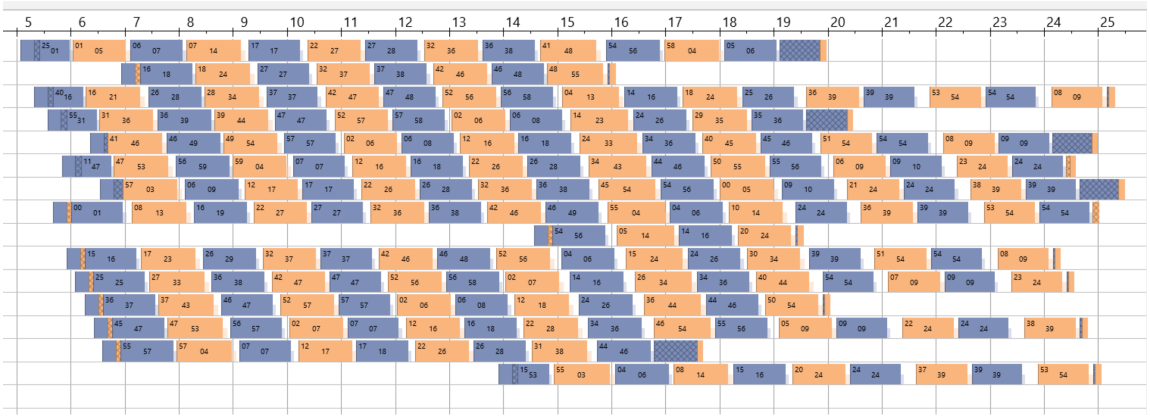


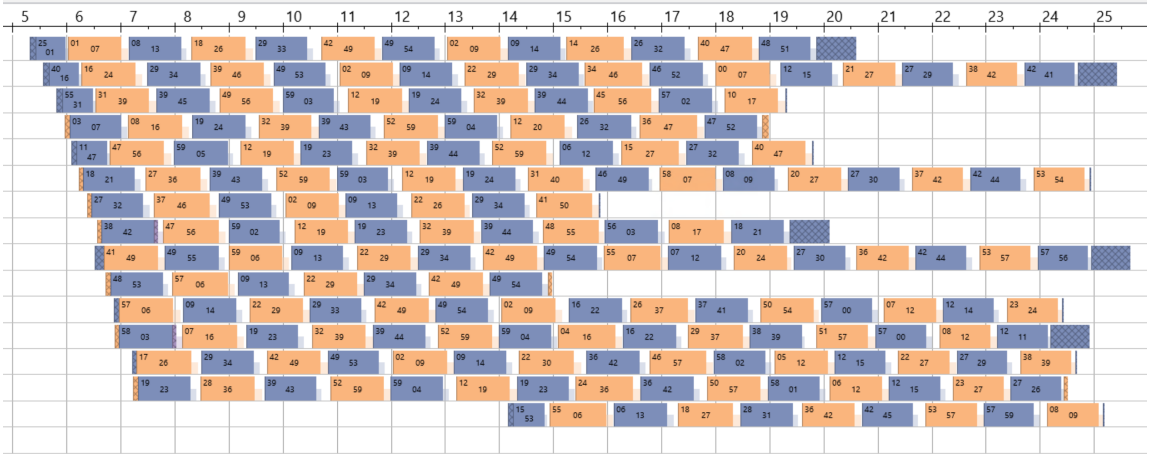
Figure A.2: Line 19 Blocks

Line 4

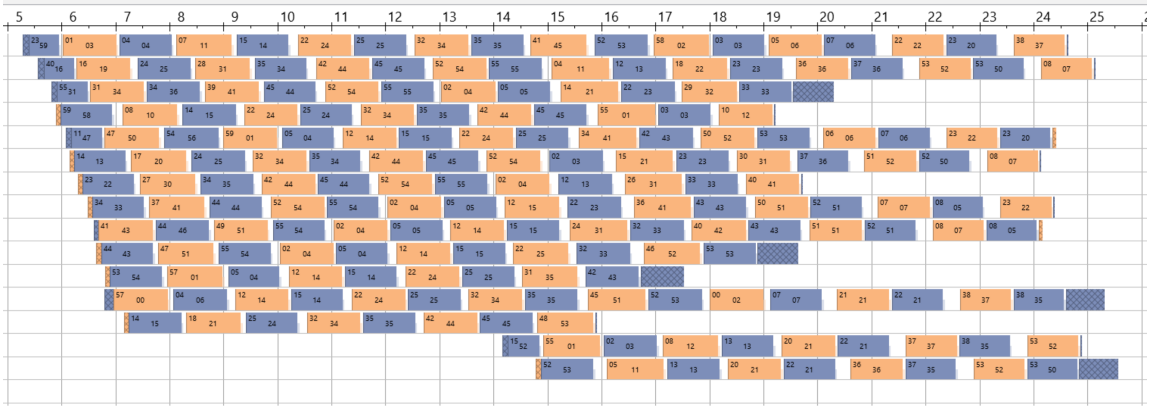
Base



04LO+5



04LO-5



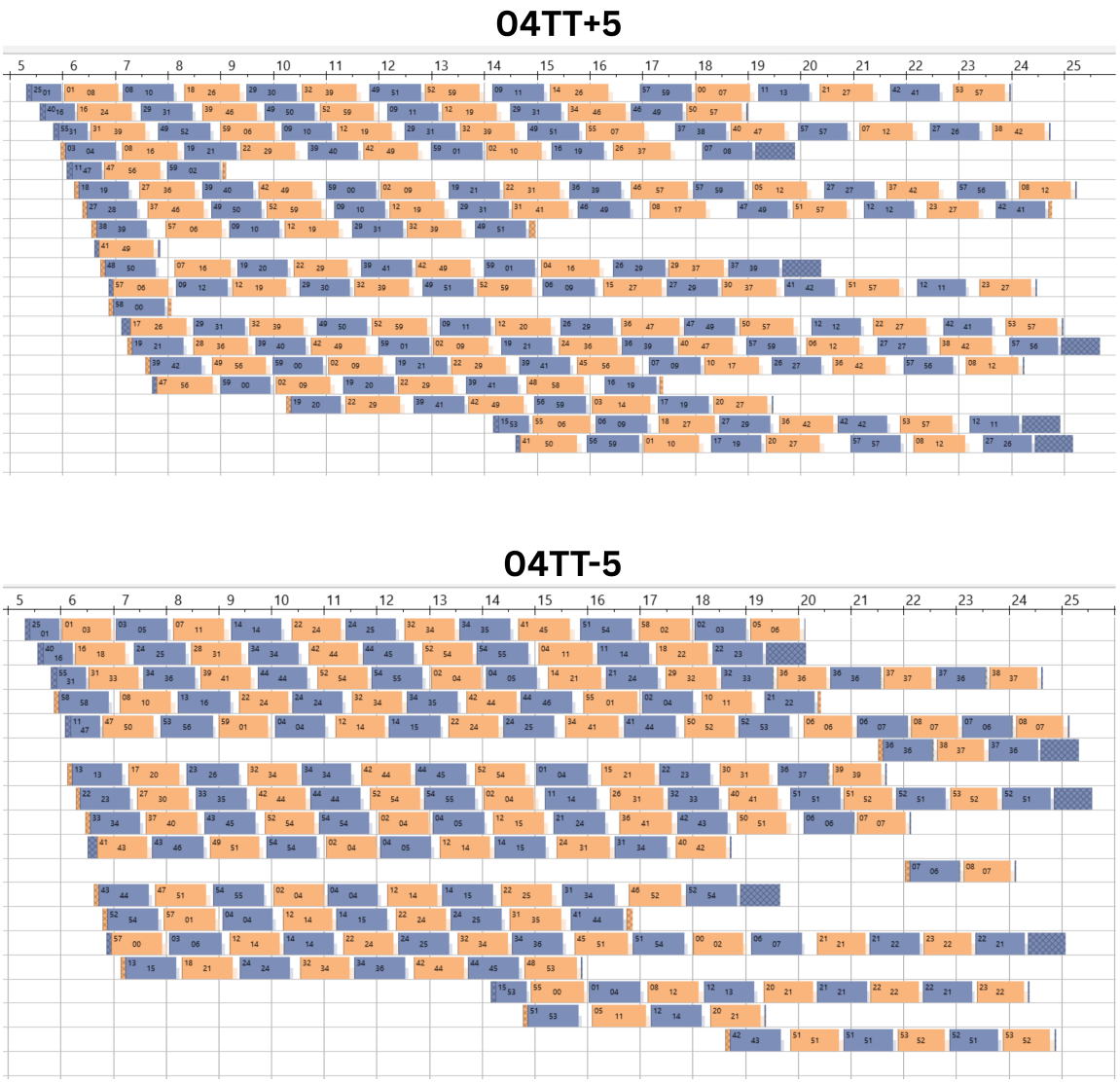
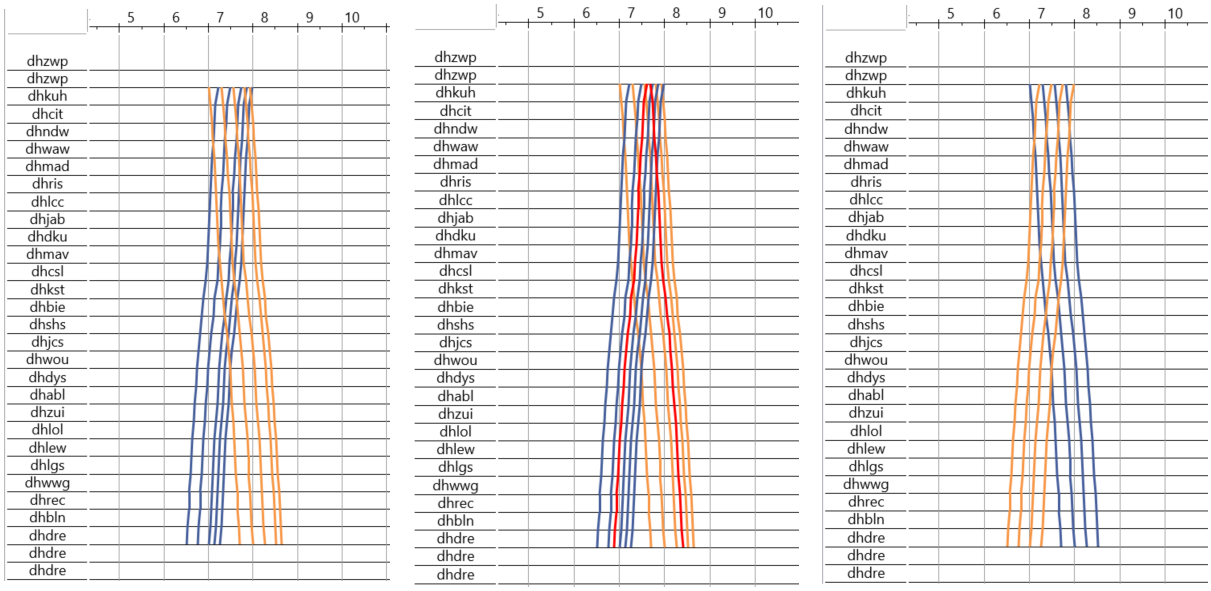


Figure A.3: Line 4 Blocks



(a) Base schedule (b) Increased frequency (c) Decreased frequency

Figure A.4: Time-distance diagrams for the frequency scenarios during 7-8. The selected trips correspond to the added trips in the increased-frequency scenario.

Appendix B

Distributions of PR across alternatives

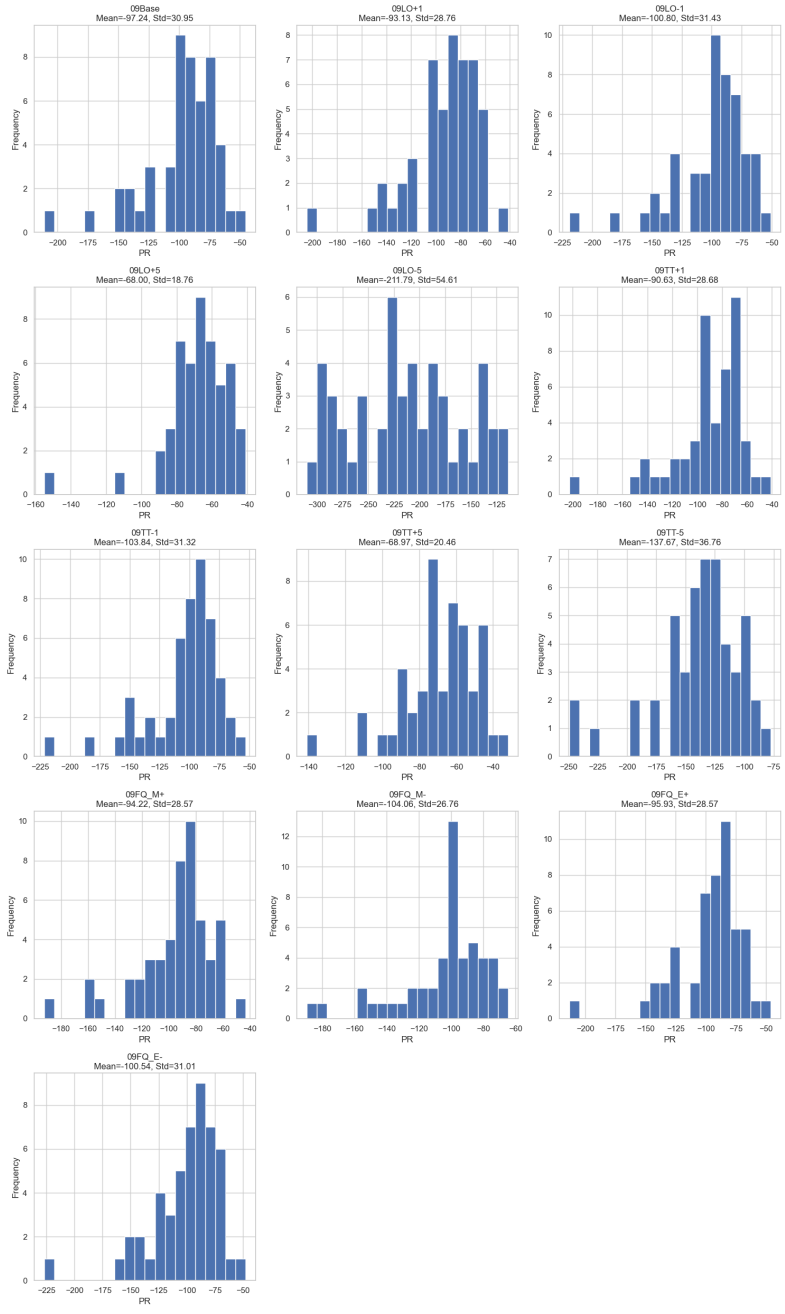


Figure B.1: Distribution of PR with 50 runs

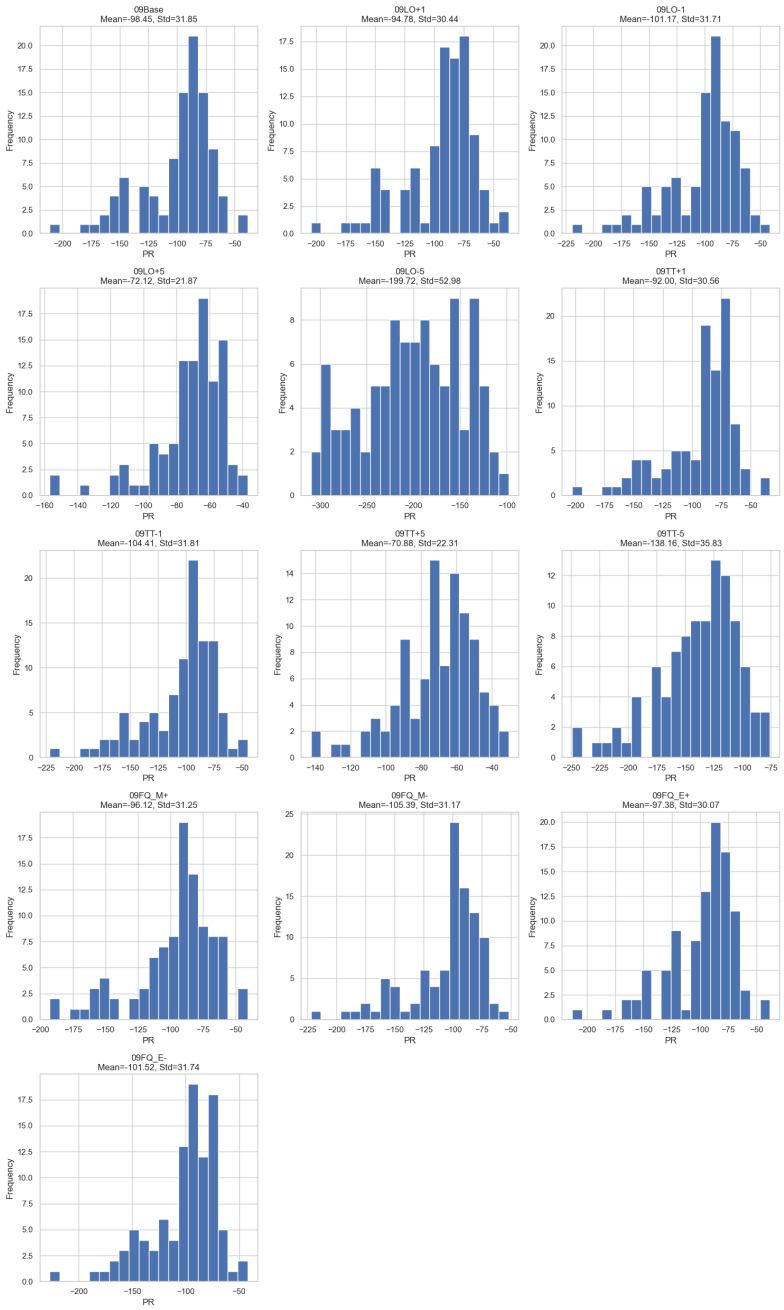


Figure B.2: Distribution of PR with 100 runs

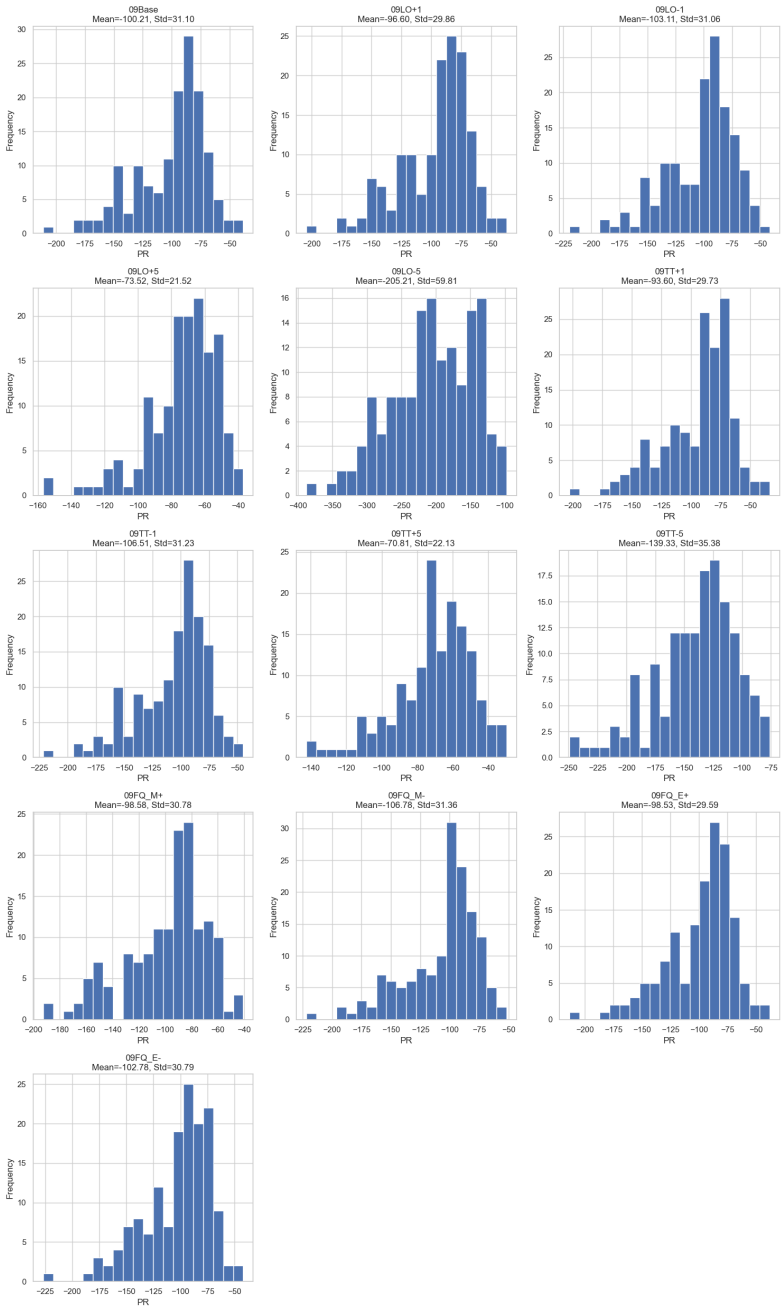


Figure B.3: Distribution of PR with 150 runs

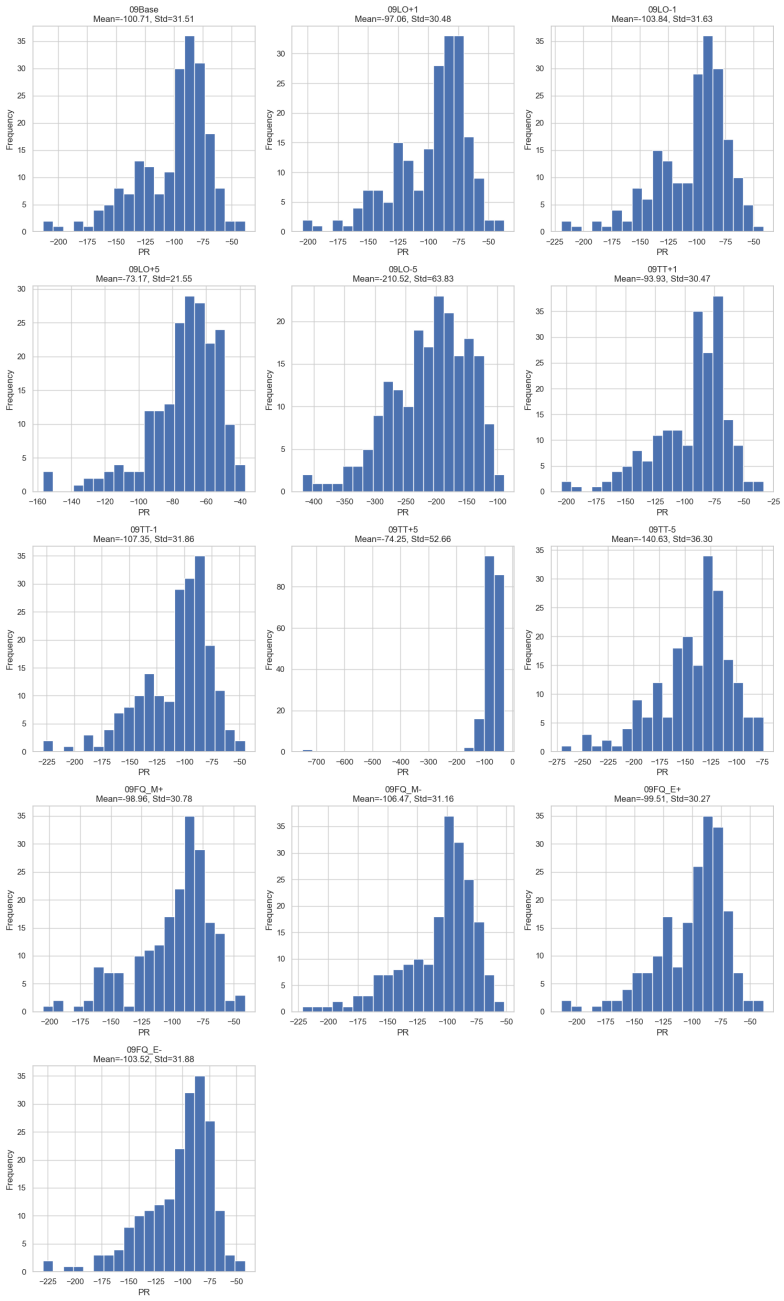


Figure B.4: Distribution of PR with 200 runs

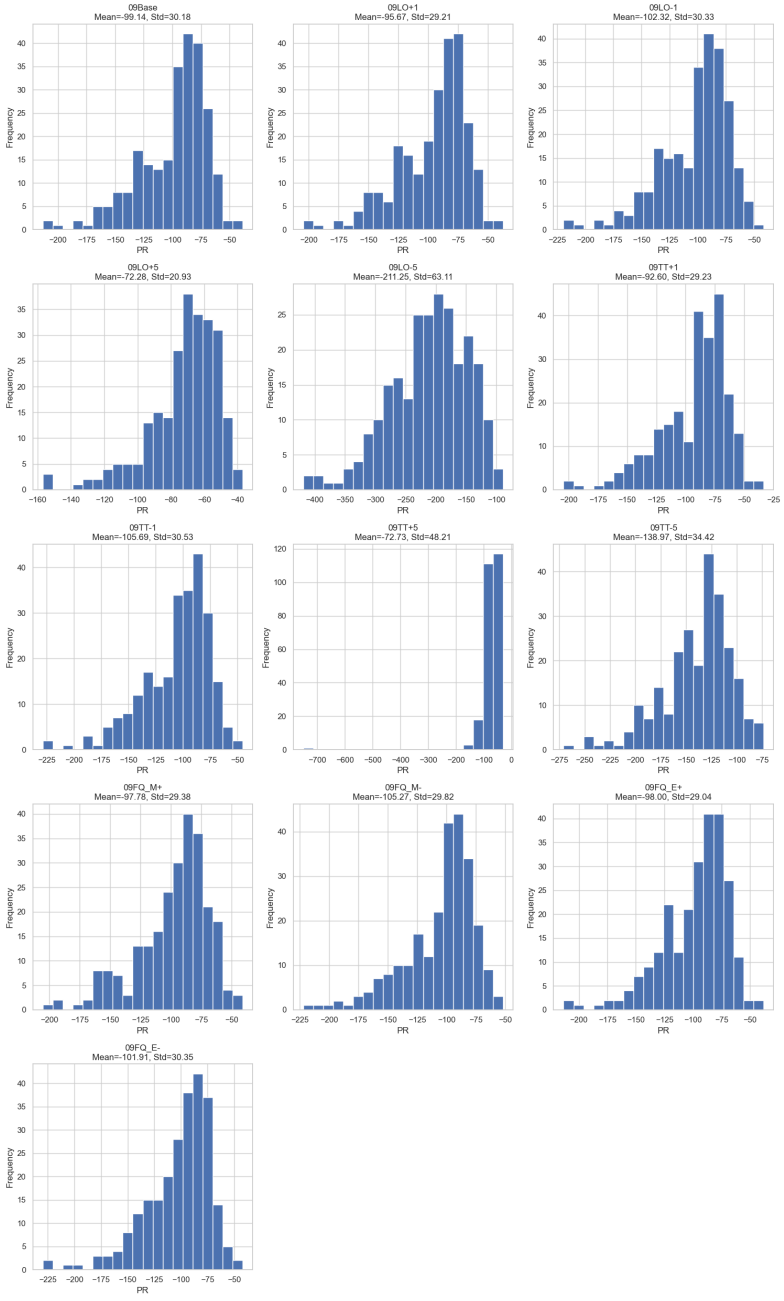


Figure B.5: Distribution of PR with 250 runs

Appendix C

Empirical Distributions

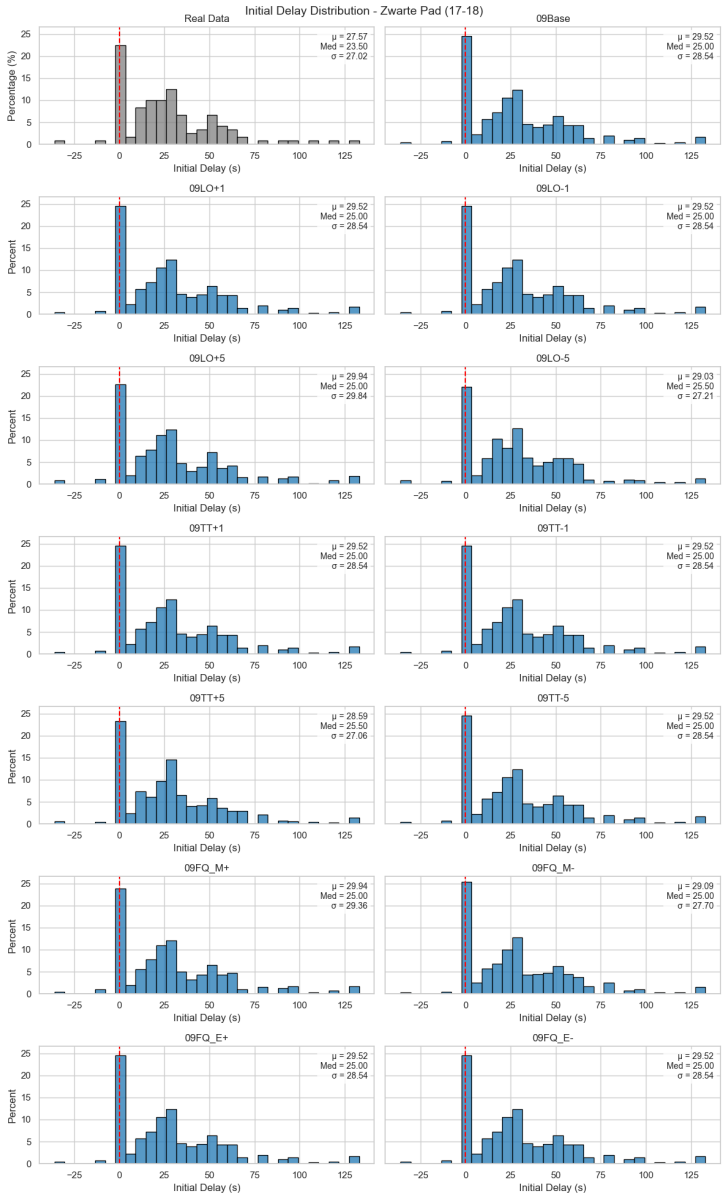


Figure C.1: Distribution of initial delays

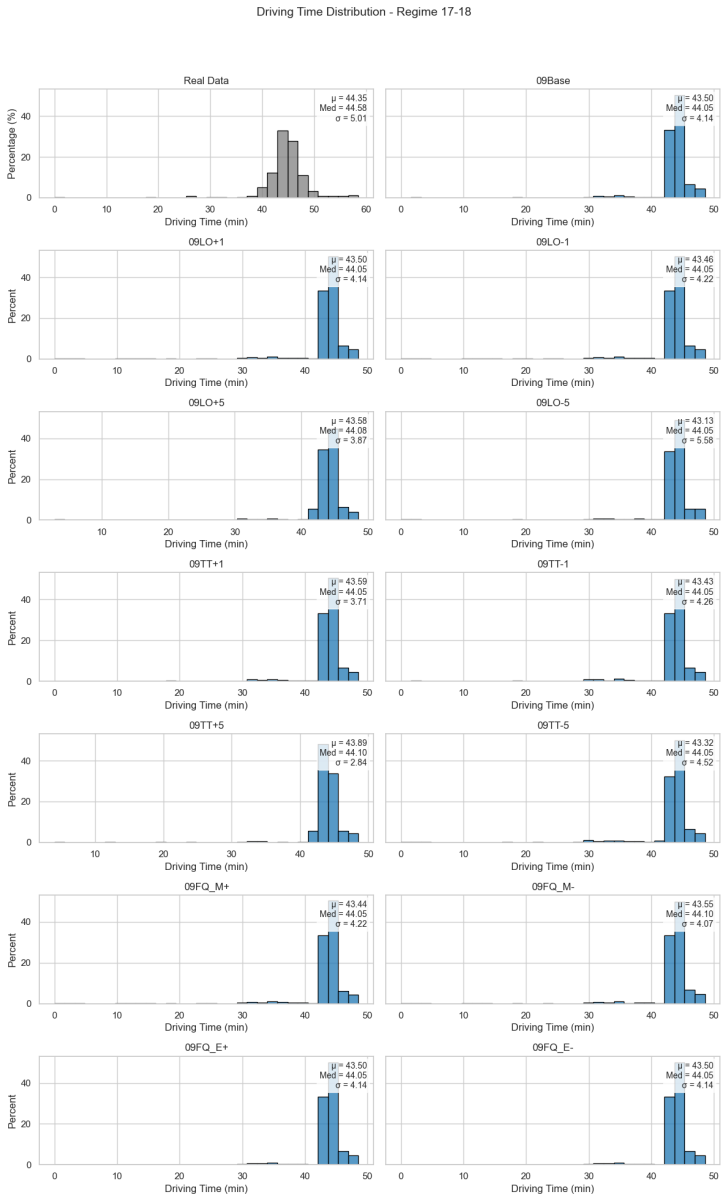


Figure C.2: Distribution of driving times

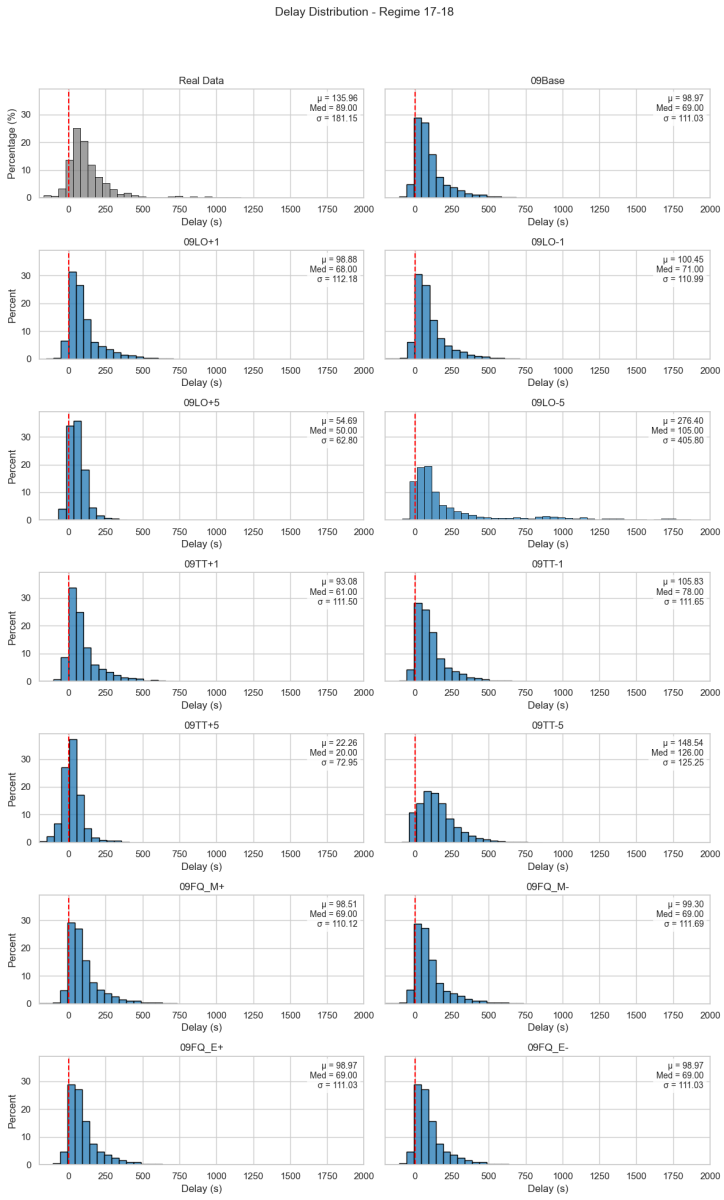


Figure C.3: Distribution of delays

Appendix D

Examples for Effect of Variables on Cost Elements

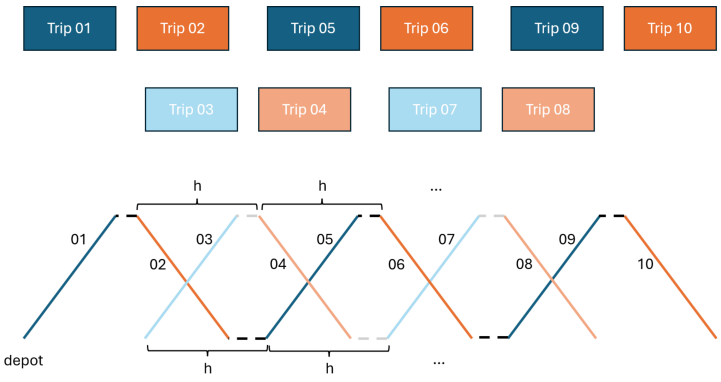


Figure D.1: A simplified example of a timetable with trips in two directions, trips executed in two blocks (The maximum number of simultaneous blocks is two, thus two vehicles are sufficient to execute the trips)

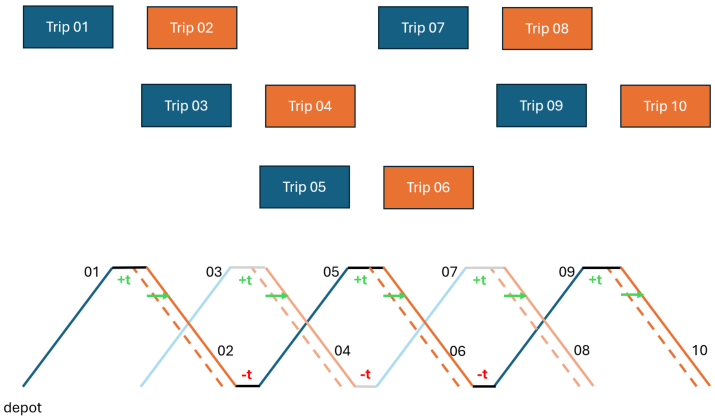


Figure D.2: Increasing the layovers in one terminal by shifting trips in one direction, minimum layover has been violated in one terminal; Reassignment of the blocks results in more operational hours and need for one more vehicle.

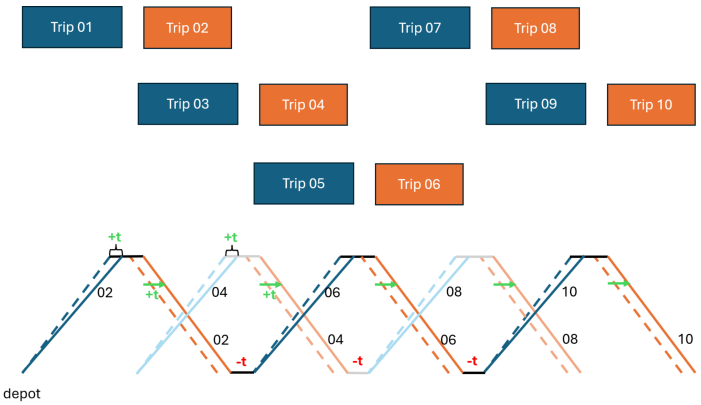


Figure D.3: Increasing the trip times and then keeping the layover unchanged in one terminal by shifting trips in one direction, minimum layover has been violated in one terminal; Reassignment of the blocks results in more operational hours and need for one more vehicle.

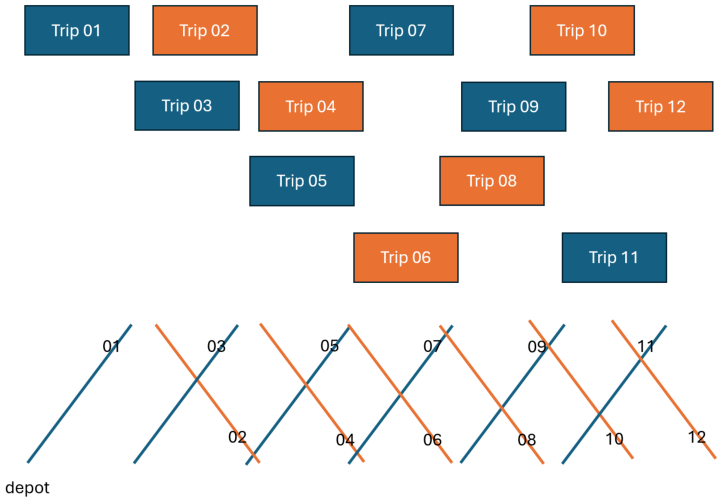


Figure D.4: Increasing frequency by adding one trip per direction and shifting trips for headway regularity; This resulted in more operational hour and need for 2 more vehicles.

References

Asensio, J. and Matas, A. (2008). Commuters' valuation of travel time variability. *Transportation Research Part E: Logistics and Transportation Review*, 44(6):1074–1085.

Broniewicz, E. and Ogrodnik, K. (2025). Application potential of mcdm/mcda methods in transport—literature review and case study. *Sustainability*, 17(17):7671.

Cats, O. (2016). The robustness value of public transport development plans. *Journal of Transport Geography*, 51:236–246.

Departement Mobiliteit en Openbare Werken Vlaanderen (2025). 77% van reizigers de lijn is tevreden over hun rit. https://www.vlaanderen.be/departement-mobiliteit-en-openbare-werken/pers/persberichten/19-februari-2025-77-van-reizigers-de-lijn-is-tevreden-over-hun-rit. Accessed: 2026-02-11.

Fan, W. and Machemehl, R. (2002). Characterizing bus transit passenger waiting times. *Proceedings, Annual Conference - Canadian Society for Civil Engineering*, 2002.

Friedrich, M., Müller-Hannemann, M., Rückert, R., Schiewe, A., and Schöbel, A. (2018). Robustness as a third dimension for evaluating public transport plans. In *ATMOS 2018: Algorithmic Approaches for Transportation Modelling, Optimization, and Systems*, pages 4:1–4:17.

Furth, P. G. and Muller, T. H. (2009). Optimality conditions for public transport schedules with timepoint holding. *Public Transport*, 1(2):87–102.

Ge, L., Voß, S., and Xie, L. (2022). Robustness and disturbances in public transport. *Public Transport (Heidelberg, Germany)*, 14(1):191.

Google Developers (n.d.). Machine learning crash course: Log loss. https://developers.google.com/machine-learning/crash-course/logistic-regression/loss-regularization. Accessed: 2026-05-04.

Han, J., Kamber, M., and Pei, J. (2012). 3 - data preprocessing. In Han, J., Kamber, M., and Pei, J., editors, *Data Mining: Concepts and Techniques (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems, pages 83–124. Morgan Kaufmann, Boston, third edition edition.

HTM Personenvervoer N.V. (2025). Collectieve arbeids overeenkomst htm 1 januari 2024 t/m 30 juni 2026. Official document from HTM Personenvervoer N.V. Versie 8 december 2025, gevestigd te 's-Gravenhage.

Iida, Y. and Wakabayashi, H. (1989). An approximation method of terminal reliability of road network using partial minimal path and cut sets. In *Transport policy, management & technology towards 2001: selected proceedings of the fifth world conference on transport research*, volume 4.

Ingvardson, J. B., Nielsen, O. A., Raveau, S., and Nielsen, B. F. (2018). Passenger arrival and waiting time distributions dependent on train service frequency and station characteristics: A smart card data analysis. *Transportation Research Part C: Emerging Technologies*, 90:292–306.

Khoshniyat, F. and Peterson, A. (2017). Improving train service reliability by applying an effective timetable robustness strategy. *Journal of Rail Transport Planning and Management*, 21(6):525–543.

Klinkhardt, C., Kandler, K., Kagerbauer, M., and Vortisch, P. (2025). Reliability of public transportation systems with autonomous first-/last-mile services in a stated mode choice experiment.

Laskaris, G., Cats, O., Jenelius, E., Rinaldi, M., and Viti, F. (2021). A holding control strategy for diverging bus lines. *Transportation Research Part C: Emerging Technologies*, 126:103087.

Lee, Y., Lu, L.-S., Wu, M.-L., and Lin, D.-Y. (2017). Balance of efficiency and robustness in passenger railway timetables. *Transportation Research Part B: Methodological*, 97:142–156.

Liu, T., Cats, O., and Gkiotsalitis, K. (2021). A review of public transport transfer coordination at the tactical planning phase. *Transportation Research Part C: Emerging Technologies*, 133:103450.

Mazloumi, E., Mesbah, M., Ceder, A., Moridpour, S., and Currie, G. (2012). Efficient transit schedule design of timing points: a comparison of ant colony and genetic algorithms. *Transportation Research Part B: Methodological*, 46(1):217–234.

Metropoolregio Rotterdam Den Haag (2026). Mobiliteitsvisie: Dromen, durven, doen. Policy document, Metropoolregio Rotterdam Den Haag (MRDH).

Montgomery, D. C. and Runger, G. C. (2018). *Applied Statistics and Probability for Engineers*. Wiley, 7 edition.

Nalmpantis, D., Roukouni, A., Genitsaris, E., Stamelou, A., and Naniopoulos, A. (2019). Evaluation of innovative ideas for public transport proposed by citizens using multi-criteria decision analysis (mcda). *European Transport Research Review*, 11(22).

NLTimes (2025). Drastic cuts looming for public transport: higher fares, cuts to key routes. `https://nltimes.nl/2025/03/27/drastic-cuts-looming-public-transport-higher-fares-cuts-key-routes`. Accessed: 2026-02-11.

Parbo, J., Nielsen, O. A., and Prato, C. G. (2016). Passenger perspectives in railway timetabling: A literature review. *Transport Reviews*, 36(4):500–526.

Preston, J., Wall, G., Batley, R., Ibáñez, J. N., and Shires, J. (2009). Impact of delays on passenger train services: Evidence from great britain. *Transportation research record*, 2117(1):14–23.

Sindzingre, M. (2019). *Detecting and quantifying bus operation impedance: the balance between reliability and speed*. PhD thesis, Massachusetts Institute of Technology.

Soza-Parra, J., Raveau, S., and Muñoz, J. (2022). Public transport reliability across preferences, modes, and space. *Transportation*, 49:621–640.

van Oort, N. (2011). *Service Reliability and Urban Public Transport Design*. PhD thesis, Delft University of Technology, Delft, The Netherlands.

van Oort, N. (2014). Incorporating service reliability in public transport design and performance requirements: International survey results and recommendations. *Research in Transportation Economics*, 48:92–100.

van Oort, N. (2016). Incorporating enhanced service reliability of public transport in cost-benefit analyses. *Public Transport*, 8(1):143–160.

van Oort, N., Boterman, J., and van Nes, R. (2012). The impact of scheduling on service reliability: trip-time determination and holding points in long-headway services. *Public Transport*, 4(1):39–56.

van Oort, N., Drost, M., Brands, T., and Yap, M. (2015). Data-driven public transport ridership prediction approach including comfort aspects. In *CASPT 2015 Conference*.

Wardman, M., Murphy, P., Ojeda-Cabral, M., Mitrani, A., Neki, K., and Humphreys, R. M. (2026). Value of time multipliers for walk time, wait time, service headway, displacement time, reliability and congested travel time: review and meta-analysis of worldwide evidence. *Transportation Research Part A: Policy and Practice*, 207:104925.

Wiéckowski, J. and Sałabun, W. (2023). Sensitivity analysis approaches in multi-criteria decision analysis: A systematic review. *Applied Soft Computing*, 148:110915.

Wikipedia contributors (2026). Margin of error - wikipedia, the free encyclopedia. Accessed: 2026-06-08.

Yap, M. D. and van Oort, N. (2018). Driver schedule efficiency vs. public transport robustness: A framework to quantify this trade-off based on passive data. In *CASPT 2018 Conference*.

Zhou, W., Teng, J., and Chen, E. (2025). Robustness-based approach for slack time optimization in tram timetables. *Urban Rail Transit*, 11:90–107.

Zwart, S. D. and de Vries, M. J. (2016). Methodological classification of innovative engineering projects. In *Philosophy of Technology After the Empirical Turn*, pages 219–248. Springer.