Aalto-yliopisto
Insinööritieteiden
korkeakoulu

Petri Huhtanen

Process definition for selecting reliability indicators for transit service: Case of Helsinki Region Transport

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 29.04.2020
Supervisor: Assistant Professor Milos Mladenovic
Advisor: Assistant Professor Milos Mladenovic

Aalto University
School of Engineering

Aalto University
P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of master's thesis

Author Petri Huhtanen

Title of thesis Process definition for selecting reliability indicators for transit service: Case of Helsinki Region Transport

Master programme Spatial Planning and Transport Engineering **Code** ENG26

Thesis supervisor Assistant Professor Milos Mladenovic

Thesis advisor Assistant Professor Milos Mladenovic

Date 29.04.2020 **Number of pages** 62+16 **Language** English

Abstract

It is important to have information on the day-to-day operational level events and planning objectives in public transport. Performance indicators are needed to inform related decisions in these domains. As one aspect of public transport level of service, problems with reliability appear to passengers as delays in different parts of the journey, such as increasing overall travel times or adding extra waiting time at transfer points. Factors affecting reliability can be classified in many different ways, but the factors mainly consist of decisions and choices at the planning and operational level, as well as environmental variables. This work considers that reliability consists of four components: punctuality, regularity, waiting time, and cancellation. As Helsinki Region Transport (HSL) is undergoing a continuous effort of developing performance indicators, the aim of this thesis was to develop a process for defining and selecting reliability performance indicators for HSL case.

Development in this thesis. starts with defining alternative performance indicators for punctuality, regularity and waiting time. The selection of a single indicator to describe the reliability was performed using the analytical hierarchy process (AHP). The AHP method was used to create an assessment framework consisting of a subjective and an objective part. The purpose of the evaluation framework was to help assess the superiority and suitability of the different indicator alternatives. The objective part evaluated the technical and computational details of the indicators, while the subjective part evaluated the indicators from the perspective of the user's needs. Based on the evaluation, the selected indicators were subjected to spatio-temporal analysis to find the most effective aggregates. The analysis also outlined possible graphical visualizations for the indicators.

In addition, a workshop was organized for users to define the criteria for subjective evaluation and to evaluate the indicator options. The workshop was evaluated as positive for its involvement and information sharing, as well as development wishes for the way the assessment was collected and for exploring the alternatives in advance. In general, the AHP method and assessment framework were found to work well as part of the selection process. Performed spatio-temporal analysis yielded the desired result, but the analysis was also found to need improvement in certain points. For example, the partial combination of analysis with a workshop was found to be a necessary addition. Overall, the indicator selection process outlined in this work can be considered successful, as details of the process are almost at the level where they can be implemented in the ongoing HSL processes.

Keywords public transport, reliability, key performance indicator, analytical hierarchy process, AHP

Aalto-yliopisto
Insinööritieteiden korkeakoulu

Aalto-yliopisto
PL 11000, 00076 AALTO
www.aalto.fi

Diplomityön tiivistelmä

Tekijä Petri Huhtanen

Työn nimi Luotettavuuden indikaattorien valintaprosessin määrittäminen: Case Helsingin Seudun Liikenne

Maisteriohjelma Spatial Planning and Transport Engineering **Koodi** ENG26

Työn valvoja Apulaisprofessori Milos Mladenovic

Työn ohjaaja Apulaisprofessori Milos Mladenovic

Päivämäärä 29.04.2020 **Sivumäärä** 62+16 **Kieli** Englanti

Tiivistelmä

On tärkeää saada tietoa joukkoliikenteen päivittäisistä operatiivisen tason tapahtumista ja suunnittelutavoitteista. Suorituskykyindikaattoreita tarvitaan avuksi näihin asioihin liittyvissä kysymyksissä. Yhtenä julkisen liikenteen palvelutason näkökulmana luotettavuusongelmat näyttäytyvät matkustajille viivästyksinä matkan eri osissa, kuten kasvavina kokonaismatka-aikoina tai ylimääräisenä odotusaikana vaihdoissa. Luotettavuuteen vaikuttavat tekijät voidaan luokitella monin eri tavoin, mutta tekijät koostuvat pääasiassa suunnittelu- ja toimintatason päätöksistä ja valinnoista sekä ympäristömuuttujista. Työssä katsottiin, että luotettavuus koostuu neljästä osasta: täsmällisyydestä, säännöllisyydestä, odotusajasta ja ajamattomuudesta. Koska Helsingin seudun liikenne (HSL) pyrkii jatkuvasti kehittämään suoritusindikaattoreita, tämän työn tavoitteena oli kehittää prosessi luotettavuuden suorituskykyindikaattorien määrittelemiseksi ja valitsemiseksi.

Tämän työ alkaa vaihtoehtoisten suorituskykyindikaattorien määrittämisellä täsmällisyydelle, säännöllisyydelle ja odotusajalle. Yksittäisen indikaattorin valinta luotettavuuden kuvaamiseksi suoritettiin käyttämällä analyyttistä hierarkia prosessia (AHP). AHP-menetelmää käytettiin luomaan arviointijärjestelmä, joka koostui subjektiivisesta ja objektiivisesta osasta. Arvioinnin tarkoituksena oli auttaa arvioimaan eri indikaattorivaihtoehtojen paremmuutta ja soveltuvuutta. Objektiivisessa osassa arvioitiin indikaattorien teknisiä ja laskennallisia yksityiskohtia, kun taas subjektiivisessa osassa arvioitiin indikaattoreita käyttäjän tarpeiden näkökulmasta. Arvioinnin perusteella valituille indikaattoreille tehtiin spatio-temporaalinen analyysi tehokkaimpien koosteiden löytämiseksi. Analyysi hahmotteli myös indikaattorien mahdollisia graafisia visualisointeja.

Lisäksi tuleville käyttäjille järjestettiin työpaja subjektiivisen arvioinnin perusteiden määrittelemiseksi ja indikaattorivaihtoehtojen arvioimiseksi. Työpajaa arvioitiin positiivisesti sen osallistamisesta ja tiedon jakamisesta. Kehitystarpeita nähtiin arvioinnin keräystavassa ja tunnistettiin tarve saada tutkia vaihtoehtoja etukäteen. Yleisesti AHP-menetelmän ja arviointikehyksen todettiin toimivan hyvin osana valintaprosessia. Suoritettu spatio-temporaalinen analyysi tuotti halutun tuloksen, mutta analyysin todettiin myös tarvitsevan parannusta tietyissä kohdissa. Esimerkiksi analyysin osittainen yhdistäminen työpajaan todettiin olevan välttämätön lisä. Kaiken kaikkiaan tässä työssä hahmotettua indikaattorivalintaprosessia voidaan pitää onnistuneena, koska prosessin yksityiskohdat ovat melkein sillä tasolla, missä ne voidaan toteuttaa osana meneillään olevassa HSL:n suorituskykyindikaattorien kehittämisprosessia.

Avainsanat joukkoliikenne, luotettavuus, suorituskykymittari, analyyttinen hierarkia prosessi, AHP

Alkusanat

Tämän työn on lähes kahden vuoden aikana kehittynyt pitkän tien kautta. Alun perin tarkoituksena oli tehdä työ samassa aihepiirissä kuin kandidaatin työ, mutta useiden sattumien kautta työksi muotoutui valintaprosessin testaaminen. Työssä löytyy silti pieni osa alkuperäisestä ideasta: testitapaukset todellisilla linjoilla. Lopulliseen työn aiheeseen vaikutti voimakkaasti oman työnpaikan kautta tullut tarve prosessin kuvakselle ja prosessin testaamiselle. Työllä ei voi sanoa olleen puhtaasti ohjaaja, sillä aiheen rajaukseen Miloksen lisäksi tuli apua myös lähimmiltä työkavereilta ja omalta esimieheltä. Jos tätä työtä ei olisi rajattu muiden kuin kirjoittajan tarpeesta, olisi kirjoittamiseen mennyt vielä pidempää ja monta yksityiskohtaa olisi tullut lisättyä mukaan. Työn ohjaajan puute loi pohjan itseohjautuvuudelle ja opetti esimerkiksi, kuinka aiheen rajaamista pitää tehdä. Iso kiitos täytyy antaa Riikalle, joka jaksoi lukea koko tekstin läpi ja huomauttaa kaikista monikkovirheistä. Kiitokset täytyy myös antaa kandityön ohjaajalle, Virpille, joka jaksoi patistaa työtä eteenpäin juuri oikeassa kohdassa. And special thanks to Milos, who was one of the best teachers, that I have had.

Helsinki 29.04.2020

Petri Huhtanen

Table of Content

List of Figures

List of Tables

1 Introduction

Assessing the performance of public transport can be approached either at the network level or at the operational level (Vuchic 2017). At the network level, the quality produced by the planning is examined (Weckström et al. 2019) , while at the operational level, the approach is more in terms of day-to-day functionality. It is easier to approach the evaluation of functionality in network level when individual small phenomena do not need to be considered very carefully, because they disappear into the variation of the outcome. For this reason, network level reviews are often made based on planned schedules. On a daily basis, any small change or choice can have significant effects on performance, but with large amounts of data produced, the effect can fall within random variations. Due to the large amount of data, it can be difficult to produce information on the daily operation of traffic. Also, information on day-to-day performance should be linked to network level evaluations to bring the evaluation closer to reality. Indeed, the large amount of data and the connectivity to network level assessments place demands on day-to-day performance indicators. For this reason, it is necessary to consider in more detail which individual indicators to choose for performance metrics.

The aim of this work is to outline the process for finding, selecting and implementing appropriate performance metrics. The purpose of the process is to define alternatives for the required indicators as well as to evaluate individual indicators. Based on the evaluation, the most suitable performance indicator should be selected. The outlined process is tested in this work in the creation of the reliability key performance indicator for Helsinki Region Transport (HSL) as an example case. The example case can be used to evaluate the functionality of the outlined process in practice. In this work the simplified process is used to limit requirements of time and work. In reality, the outlined process has many different branches for other needs. Several steps in the process are also extensive and repetitive, so this work only goes through individual cases as an example. A more detailed delineation of the different parts of the process takes place separately for that part.

The literature is used to determine the options for the different performance indicators and also to identify possible limitations, strengths and weaknesses for the different options. This mapping provides information on how a single indicator option works, but also additional information on their selection and evaluation. The selection and evaluation of indicators emphasize the needs of their future users and therefore a workshop was held for users. In the workshop, the indicator options were both presented to users and users assessed their functionality in their own daily work. The suitability of the outlined process for other decision-making levels also had to be considered in the selection and evaluation. A good advantage of suitability is that the same process can be repeated without any changes at each level.

The selected indicators are subjected to a spatio-temporal analysis for their implementation. The main purpose of the analysis is to find out with which data aggregation levels would be practical to calculate and to present an individual indicator. The results of the analysis can also be used to outline simple graphs for different aggregate levels. As a by-product of the analysis, information is obtained on how the planning might affect the values of the indicator as a systematic error. This information can be used to take the error into account in interpreting the results of the indicator.

This work is organized in four sections after introduction. The section one contains the literature review that goes through how reliability can be defined as well as which factors affect

reliability. The same section also presents what indicator options are available as measures of reliability. In the second section, the methodology of this work is described. It includes how the evaluation and the selection of the process was done and a description of spatio-temporal analysis. The section three gathers information of the results of both evaluation and spatio-temporal analysis. That section also briefly presents as an exemplary analysis how the results of the spatio-temporal analysis look like when applied to real lines. In the last section includes discussion and conclusion of this work.

2 Literature

2.1 Reliability

2.1.1 Reliability in general

The reliability of public transport can be approached from many different perspectives, such as availability, convenience, travel time or economy. The most interesting aspects of reliability in viewpoint of traffic planning and production are adherence of a schedule, driving time and headway. Reliability problems which are caused by production variability, reflect directly into delays at departure, at travel time and at arrival time. The passenger is experiencing reliability through these delays and does not necessarily consider the headways as important as a reliability perspective. It is also good to take into account the perception and experience of potential passengers, as improving the reliability. Especially improved perceived reliability will attract more users to public transport. (van Oort et al. 2012; Cham 2006; Vincent et Hamilton 2008; Firew 2016; van Oort et al. 2015).

The reliability of public transport can be defined either from planning or from the passenger's point of view. It is difficult to define general reliability from a passenger's point of view due to different preferences of different travellers. Thus, the passenger's perspective is usually considered only when defining the indicators to measure reliability. The reliability of transport is defined in different ways from different sources. In principle, all definitions have the same basic idea behind them - the ability to produce a planned service. The variety of definitions of public transport reliability already shows that reliability should be a multidimensional service capability. With this interpretation and because the passenger experiences most of reliability through delays, reliability can be also interpreted to be the probability that the trip takes place between the origin and destination as planned. The probability is affected by the different attributes of the trip, such as the wait at the departure stop, the load of the departure and the passenger information. (Vincent et Hamilton 2008; Yaakub and Napiah 2011; Chen et al. 2009; Sorratini 2008; Benedetto et al. 2016).

2.1.2 Effects of reliability

In a passenger's decision-making in route and mode selection, the reliability can be considered a decisive factor. Unreliability during the trip causes changes in the passenger's behaviour during the trip and in the experience of overall reliability. The passenger can also experience discomfort and unease due to varying travel time. Poor reliability reduces passenger counts and ticket revenue, but also increases operational costs. Poor reliability also makes scheduling more difficult, causing uneven loading between different units, but above all leading to longer and more uncertain travel times. (Vincent and Hamilton 2008; Firew 2016; Liu and Sinha 2007; Sorratini 2008).

Unreliability and scheduling deviations are an operational problem. The most typical way to improve reliability is to add timing points along the route. On the other hand, there is the possibility of increasing waiting time at the timing point and this should be taken into account in schedule planning. Other possible measures to improve reliability include, for example, better scheduling coordination. However, all improvements are more dependent on other variables, such as other lines and infrastructure. The best possible level of reliability improvements can be achieved when operational and planning level solutions are made together. (van Oort et al. 2012; Firew 2016).

Table 1. Different sorting methods for sources of unreliability.

Benedetto et al. (2016)	Improper service design Driver or supervisor failure Uncertain passenger load Uncontrollable external factor
Sorratini (2008)	Characteristics of traffic Characteristics of route Characteristics of passenger Operational characteristic
van Oort et al. (2015)	Driving and stopping time Dwell time

Different modes of public transport have different sources of unreliability. Reasons for unreliability include, for example, the general traffic situation, traffic lights, variation in demand and operating habits. In fully isolated transport mode, the effect of unreliability from other traffic is insignificant while for the bus transport it may even be the main reason for unreliability. (Cham 2006; Vincent and Hamilton 2008). When analysing sources of unreliability, the attributes can be grouped by in many ways. For example, Benedetto et al. (2016) divides the unreliability sources into categories of improper service design, driver failure, uncertain passenger load, and uncontrollable external source. These sorting categories are highly planning-oriented and somewhat neglect other sources of unreliability, such as the impact of other traffic. Sorratini (2008) suggest that the sorted unreliability sources are characteristics of traffic, route, passengers and operation. Of these, operational characteristics include all planning and operational level effects, such as scheduling design and driver behaviour. Traffic characteristics include, for example, congestion levels, traffic composition and route characteristics such as route length and street parking. Passenger characteristics are grouped by into two groups: passenger counts and passenger choices. Passenger counts have direct and immediate effects in planning and reliability. Passenger choices, such as route selection and arrival time to the stop, is affected by reliability but also are affecting in reliability. (Sorratini 2008; Saberi et al. 2013). Van Oort et al. (2015) divide the sources of unreliability according to the objects of influence. This division follows the separation that can be done to the causes of delays - driving and stopping time and dwell time. The driving and stopping time category include e.g. the structure of the service network (how many lines are on the street), the impact of the schedule structure and crossings on reliability. The dwell time category contains some of the same sources of impact as the driving and stopping time category, but also e.g. the impact of the stop design on reliability. (van Oort et al. 2015).

Table 1. summarizes the different categories of the sources of unreliability that has been shown here.

The direct benefit of understanding the sources of unreliability is improved reliability (also passenger perceived reliability) and passenger count growth. Understanding the sources of unreliability also allows the public transport agencies to use the sanction and bonuses to control the reliability in the operational level. On the other hand, creating a precise image of reliability can provide useful information that can influence transport policy decisions and decision-making. (Vincent and Hamilton 2008; Saberi et al. 2013; Sorratini 2008).

2.1.3 Components of reliability

In order to measure reliability, it is good to split reliability into the components that can be measured. Since the reliability experienced by the passenger is strongly linked to the trip and its various delays, reliability should be divided according to these delays and part of the trip. On the other hand, the components derived from the passenger's point of view can be difficult to measure and possibly heavily dependent upon passengers' preferences. This means that the components, which are defined this way, can be difficult to interpret or even misleading. However, to assess the reliability for the passenger, the components can be used as a guiding addition to explore the potential for improved reliability. Firew (2016) mentions the safety is one of the components of this type based on passenger's viewpoint of reliability. Safety, as part of reliability, can mean the technical safety of the vehicle itself or the sense of security experienced by the passenger himself. Of these, the technical security is easier to define and measure but the sense of security experienced by a passenger is influenced by e.g. the age and the life experience of the passenger.

Vincent & Hamilton (2008) divides reliability into four measurable parts: punctuality, cancelled departures, variability and variability of waiting time. The purpose of punctuality is to describe the adherence of schedules and can be measured e.g. as an average delay or percentage in the comfort zones. A comfort zone is defined as a predetermined desired delay time in relation to the schedule. With cancelled departures, it is desired to describe the probability of the arrival at the stop and e.g. the number of cancelled departures relative to the total number of departures or variation of headway that cancelled departures caused. The variability component is intended to describe the dispersion that formed between the scheduled and the actual event and often used to measure the standard deviation. (Vincent and Hamilton 2008).

The variability of waiting time is the most important aspect of reliability for the passenger. The value of waiting time has been shown to be about two times more valuable to the traveller than the travel time itself. On the other hand, there is also a reference to the fact that the perceived extra delay minute in departure time is about four times more valuable than the travel time. On this basis, punctuality would also be some significance to the passenger. The standard deviation is also used for the variability of the waiting time, the purpose of which is to describe the deviations of the expected waiting time from the scheduled waiting time. (Vincent and Hamilton 2008; van Oort et al. 2015).

Because the trip consists of travel and waiting times, Vincent & Hamilton (2008) also point out the unreliability of the various components that is involved in different parts of the trip. At the start of the journey, punctuality, cancelled departures, and variability in waiting time

are affected. Travel time is affected only by the variability, and the end of the trip is affected by punctuality and variability of the end of the trip itself. The end time is also affected by cancelled outputs, but through the time of departure and therefore the effect is not direct. It should also be noted that the departure variability and travel time variability together is the same as the arrival variability. This means that only the variability component will directly affect the end time. (Vincent and Hamilton 2008). It is good to note that departure time and its uncertainties strongly control the reliability of the whole trip and is therefore a good way to evaluate the reliability of trip. This is also the case of Chen et al. (2009). They show that it is sensitive to the reliability of the whole journey experienced by the passenger, especially when the events are handled at the stop level. (Chen et al. 2009).

2.2 Punctuality

2.2.1 Punctuality in general

Punctuality measures the difference between a planned schedule and an observation. It is one of the main ways to measure reliability. It is interesting from the point of view of planning to know how many departures take place as planned or within a certain time window. This kind of binary variable ignores whether the individual departure is early or late but categorizes it only as punctual or unpunctual. Often additional categories are created to differentiate between different cases. Such a breakdown is very operation-oriented, and it would be good to expand the categories with the passenger's point of view. There are two ways for the passenger to perceive punctuality - the punctuality of the waiting at the stop and the punctuality of the travel time. The first of these describes the timeliness of the schedules and the latter the duration of the travel time in relation to the planned. The punctuality of the stop is generally calculated as the difference between the observed departure time and the scheduled departure time. Whereas the punctuality of the travel time is calculated the difference between the observed departure and arrival time compared to the difference between the scheduled departure and arrival time. (Yaakub and Napiah 2011; Benedetto et al. 2015). In the selection of the indicator, it is good to note the possibility of using the indicator to calculate the punctuality of the stop as well as the punctuality of the travel time. As Yaakub & Napiah (2011) state, the calculated punctuality is dependent on the selected days and times. Also, the length of the line, the number of stops at the line and the number of passengers have a clear link to punctuality results. It is therefore advisable to calculate punctuality based on departure times, as dwell times can be quite long with large numbers of passengers. (Yaakub and Napiah 2011). Also, scheduling process and the location and numbers of timing points have great effect on punctuality (van Oort et al. 2012).

2.2.2 Options for indicator

The most commonly used indicator for punctuality is the number of departures within the set time window relative to the total number of departures. In general, the time window is one minute early to five minutes late, but it varies between different cases. (KFH Group 2013). The time window depends on set and desired target level. In addition, it describes the expected punctuality level of traffic. As previously mentioned, this kind of indicator does not see the difference between early and late events, so the indicator needs additional categories for those events.

Kho et al. (2005) describe one of the other possible indicators for punctuality. There are three different versions of the most commonly used punctuality indicator. In generalization, these three indicators calculate the ratio of punctuality dispersion to the scheduled or observed headways. This is one way to pay attention to the significance of the punctuality of different headways. (Yaakub and Napiah 2011; Kho et al. 2005). van Oort et al. (2012) presents the mean deviation from the timetable as an indicator of punctuality. This way is a clear way to describe the magnitude of punctuality. Saberi et al. (2013) starts with defining punctuality based on the distribution of delays. Of the presented punctuality indicator alternatives, the width index is directly related to the punctuality while the other indicators are more indicative. (Saberi et al. 2013). Chen et al. (2009) defines a punctuality indicator through probability, and this way can be considered as a mathematical formulation for the most commonly used indicator for punctuality.

2.3 Regularity

2.3.1 Regularity in general

Regularity is one of the keys to quality of service, especially on high frequency lines. On low frequency lines, regularity is not as important as punctuality is. There are also differences between different frequency lines in passenger arrival to the stop. On a low frequency line, the passenger tries to match the arrival time at the stop so that the waiting time is as small as possible. In this case, the adherence of the schedules is a determining factor in the experience of reliability. High frequency lines' schedules do not have such a great significance for a passenger who therefore arrives at the stop almost completely random. In this case, the headway and the regularity are a significant factor in the perceived reliability. (van Oort et van Nes 2004; Trompet et al. 2011; Strathman et al. 1999; Ruan 2009; Cats 2014). The difference between a high and low frequency, i.e., headway, is defined between 7 minutes and 12 minutes depending on the source. A smaller headway than this is interpreted as high frequency and a higher headway is interpreted low frequency. The definition is also influenced by the way in which schedules are planned and how the passenger is informed of the timetables for the line.

The common cause of irregularity can be said to be the uncertainties arising from the traffic environment. In the ideal situation, vehicles are evenly distributed across the line at even headways, but many practical factors prevent this from happening. For example, the number of passengers and its variation on the route will affect the dwell times and delays will vary at different times during the day. In addition, there is also no precise departures from the start terminal, which means that there cannot be absolute regular service even in the time of departure. Operational-level practices and instructions also affect the regularity, such as whether is it allowed for vehicles to overtake each other. It is also possible that the headway itself is the root cause of irregularity. For example, in the case of too frequent services, the possibility of bunching is great. That in turn leads to a more uneven load and longer trip times. (van Oort and van Nes 2004; Ruan 2009; Henderson et al. 1991; Golshani 1983; Cats 2014).

Regularity of traffic affects demand and supply. On the demand side, regular traffic increases the appreciation and attractiveness of the service. This is one way to get potential new passengers to use the service. On the other hand, on the supply side, good regularity enables efficient and economical use of line capacity. It is therefore important to manage and reduce

irregularities, especially at operational and planning levels. (van Oort and van Nes 2004; Strathman et al. 1999).

Irregularity in traffic can be affected in two ways: either preventing the occurrence of irregularities or correcting the irregularities that have arisen. For example, traffic signal priorities or own lanes may be required to prevent irregularities. The irregularities that have already arisen can be corrected by adjusting the headways, for example by extending dwell times. Whatever method is chosen and used to improve regularity depends on a number of factors. Such factors include, for example, the headway and the number of vehicles on the route, the possibility of communication between different vehicles and the characteristic of the route (e.g., length). The choice of the way of solving the regularity problems is also influenced by secondary factors such as the ratio of the headway and the dwell time and the capacity of the stop. (Golshani 1983). Cats (2014) mentions the use of holding points as the most common way to correct regularity. In this case, the units wait for a variable time in the holding stop so that the headway to the previous vehicle is close to the scheduled headway.

2.3.2 Quality of input data

Unlike in case of the punctuality, the quality of data is very relevant to the outcome in case of regularity. The challenge is increased due to the amount of data that needs to be cleaned of the errors. On the other hand, a higher amount of data provides a more reliable and precise value for the regularity. Trompet et al. (2011) shows how errors in data should be handled. If the observation is missing from the data, but the departure has been happened normally, the observation should be removed from the data. If the lack of data is due to the cancelled departures, it is replaced by the data of the next measured departures data. This corresponds well to the situation experienced by the passenger in the event of a cancelled departure**.** There is also a need to observe the overtaking vehicles in the observation data. From the passenger's point of view, the vehicles arriving at the bus stop are always in the correct order even if a single vehicle would have overtaken the other one. In case of overtaking vehicles, the observed arrival times should be arranged in order of observed arrival times not in scheduled arrival times. (Trompet et al. 2011). For situations where there is uncertainty about whether the departure is cancelled or if the departures measurement is missing due to technical problems, by default all events are cancelled departures. In this case, regularity is underestimated, but it also gives information about the lack of data quality.

2.3.3 Options for indicator

Henderson et al. (1991) presents two basic criteria for selecting a regularity indicator. First, the indicator should control the average headway, and second, normalize the result of the indicator to the zero-to-one scale. The average headway is problematic in that: It is hard to take into account the easily changing headway of the line, and it does not allow comparison of different headway lines. Scaling the indicator results to a standardized interface makes it easier to understand them. It also gives a clearer picture of how far away we are from an optimal situation. (Henderson et al. 1991).

As KFH Group (2013) defines a percentage-based indicator for punctuality, the equivalent can also be made for the regularity. The target level is generally a three-minute deviation from the planned headway, but also a percentage of the planned headway is used (Trompet et al. 2011). The same problems that there are in the punctuality percentage, are a problem

of regularity percentage. For example, the percentage indicator does not distinguish between high and low observed headways. The KFH Group (2013) itself suggests a variation coefficient of observed headways as a regularity indicator. For example, Cats (2014) mentions this indicator as a robust statistical indicator that can directly see changes. Cats (2014) also notes that the coefficient of variation is not generally intuitive and does not present the regularity experienced by the passenger.

Strathman et al. (1999) suggest the relationship between the observed and planned headway for the regularity indicator. The indicator is good for detecting bunching but ignores the effect of cancelled departures on regularity. (Strathman et al. 1999; Cats 2014). Golshani (1983) approaches regularity through an irregularity index, which is a modified version of the variation coefficient of headway. van Oort and van Nes (2004) suggest a relative mean difference between the planned and observed headway to be regularity indicator. This indicator approaches the regularity from the passenger's point of view and can be used to assess the perceived frequency. Regularity can also be approached through the unevenness of the headway with the Gini coefficient as Henderson et al. (1991) mentions.

In addition, regularity can also be expressed through probability as suggested by Ruan (2009). This probability-based indicator is consistent with the passenger-experienced headway, but the effect of very low headway is hidden. (Ruan 2009). Trompet et al. (2011) and Henderson et al. (1991) offers an excess waiting time as an alternative for the regularity indicator. This indicator accounts well for the passenger's experience and is independent of the average headway. (Trompet et al. 2011; Henderson et al. 1991).

2.4 Wait time

2.4.1 Wait time in general

The passenger waiting time at the stop is used as one of the criteria for assessing the quality of the services. This approach works especially in those situations where the user's point of view is emphasized. From the operator’s point of view, the waiting time can be an important performance indicator and can be used as a measure of production quality in tender contracts. In many transport projects assessment, waiting times are used as one of the reasons for investment. At the level of public transport planning, there is often interest in how a single passenger experiences waiting time and how long a passenger must wait on average before the departure happens. (McLeod 2007; Amin-Naseri and Baradaran 2015; Hess et al. 2004).

Traditionally, the way of measuring the waiting time is based on the relationship between the arrival time of the passenger at the stop, the scheduled departure time and the actual departure time. It is impossible to measure the exact arrival times of passengers because there has not been possibility to follow the exact movement of a passenger. The measures often also focus on average values, while the passenger-centred view emphasizes extreme values, which matter for positive or negative experience. Also, the traditional way of measuring waiting times does not always describe completely the impact of unreliability on the passenger. (KFH Group 2013; Furth and Muller 2006). It has been noticed and empirically demonstrated that the waiting time experienced by the passenger differs from the theoretically calculated waiting times and the observed waiting times. The passenger feels that waiting time has been longer than it actually has been. This is reflected in the fact that the passenger's waiting time at the stop is monetarily valued higher than the time spent in the vehicle

itself. For this reason, the passenger often underestimates the quality of the service. Thus, reducing the actual and perceived waiting time often leads to more satisfied passengers. (Mishalani et al. 2006; Furth and Muller 2006; Amin-Naseri and Baradaran 2015).

The basic calculation of waiting times is based on three assumptions: passengers arrive at the departure stops randomly i.e. evenly distributed, arrival times of vehicles is independent from each other and all the waiting passenger can board into first possible vehicle. In an ideal situation, the waiting time is half the headway, but the ideal situation requires e.g. completely reliable traffic and free capacity in the vehicle. There is also a difference between the observed waiting time and the calculated waiting time. The reasons of the difference would be partly that the assumptions are not necessarily fully consistent. (Amin-Naseri and Baradaran 2015).

Most of the uncertainty is related to the assumption of the distribution of passenger arrival times. Amin-Naseri and Baradaran (2015) mentions that there is the difference between the actual and the calculated waiting time, even in the case of the uniform distribution. Based on these facts, passengers have to be divided into three groups based on the time of arrival. First, a passenger who has no prior information on the line or route will arrive at the stop randomly. Second, if the passenger has information on the schedules of the line, the passenger will try to minimize waiting time by arriving at the stop shortly before departure. Third, if a passenger has a schedules and previous experience and/or knowledge of line reliability, the passenger can schedule arriving at the same time as the departure arrives at the stop. The line headway also affects the way the passenger arrives at the departure stop. Arriving at long headway cases is scheduled near the scheduled time of arrival, if the schedule is known. On the contrary, in short headway cases, arrivals are more random, even if the passenger knows the schedule. Differences between long and short headway lies somewhere between 7 to 12 minutes and exact value varying in the different sources. Various methods for calculating waiting times may be require for different headways. (Amin-Naseri and Baradaran 2015; Hess et al. 2004; Henderson et al. 1991). Henderson et al. (1991) states that the waiting time of a passenger is a function of the headway and the reliability.

The passenger's experience of waiting time seems to depend on whether a deliberate choice has been made to wait or whether the waiting is caused by external reasons. An external cause may be for example that trip departure is late. The waiting time experience is also affected by whether the passenger has activity while waiting. Moreover, the experience also is affected by real-time information about departure and its possible delays. Several studies show that the waiting time of the passenger who has access to the real-time information at the stop is smaller than for those without any information. This holds even in cases where the actual waiting time would not have changed itself. When the passenger uses the information of the arrival time, the passenger can use it more efficiently, such as choose another route for the trip leg, choose a completely different route or completely different transport mode. The passenger still tries to minimize waiting time by adjusting own behaviour. (Dziekan and Kottenhoff 2007; Hess et al. 2004; Mishalani et al. 2006). The actual waiting time at the stop can be reduced by increasing headway and/or improving reliability. The uncertainty of the waiting times is mainly due to problems of punctuality and regularity. The methods and plans to improve them is also way to reduce the uncertainty of the waiting period. Generally speaking, reducing uncertainties of waiting time can improve passenger’s satisfaction and ultimately increase the number of passengers. (Hess et al. 2004; Mishalani et al. 2006).

2.4.2 Options for indicator

As previously described in the section of regularity indicator alternatives, the waiting time indicator may also be selected by the same criteria for selection. For example, the indicator should regulate the mean headway and be scalable between zero and one. Of these two criteria, scaling is not a so decisive factor in the choice as it was in regularity cases. It may be also a good idea to display the wait time indicator in minutes or to relate it to the headway. The choice of indicator unit should therefore be determined on a case-by-case basis. For example, comparing the waiting times of two different lines, with two different headways, it is easier and more meaningful when waiting time is adjusted with headway. It also gives a chance to understand the significance of the time spent in waiting. On the other hand, communication with the passengers can be more clearly perceived by units of the minutes. The same pre-processing of the input data, which should be done for regularity indicators, must be done also in the waiting time indicators. For example, overtaking vehicles affects passengers same way in case of regularity and waiting time.

The average waiting time equation is well known. In different sources, its mathematical form is slightly different, but in practice, all equations are formulated in the same manner. For example, KFH Group (2013) presents it with a coefficient of variation. This equation assumes that the assumptions (arriving randomly, departure times is independent from each other and boarding to the first available unit) are correct. As noted in the previous paragraphs, assumptions may not always be correct. Amin-Naseri and Baradaran (2015) gives two different versions of the average waiting time equation, in which the initial assumptions are either partially or completely abandoned. They would be good to use, at least the version where some of the assumptions are abandoned, to achieve more accurate calculation results. KFH Group (2013) also introduces the budgeted excess waiting time as a waiting time indicator for larger headway lines.

The average wait time can be divided into two components: scheduled wait time and excess wait time. Of these, excess waiting time, KFH Group (2013) presents as the waiting time indicator. For example, Transport for London (TfL) uses this as an on-time performance indicator. Henderson et al. (1991) shows the version of excess wait time used by TfL, a standardized excess waiting time. This version contains several problems, as it has no upper limit and constant is arbitrary. The indicator is only in use as a TfL internal measure. Henderson et al. (1991) provides a waiting index as a waiting time indicator, which is a comparison of scheduled waiting time and the average waiting time.

.

3 Methodology

3.1 Evaluation framework

3.1.1 Definition

For comparing the reliability indicator options, an evaluation framework was outlined to determine the functionality and suitability of a single indicator in practice. The aim is to select the indicator that has the greatest impact on planning and decision-making. The need for comparison arises from the different needs of those involved in the various stages of public transport planning to measure reliability.

The aim was to make the evaluation process as objective and transparent as possible. The process could not be carried out in a fully objective manner, because the user perspective was also emphasized in the selection and comparison of the indicators. Part of the lack of objectivity lies in the transparency of the process but also in the division of the evaluation framework into two - objective and subjective. Of these, the objective evaluation includes all the technical requirements for the indicators and the evaluation was made in such a way that all parts of the evaluation are completely computational and independent of the views of the user or the evaluator. User views on the indicator were collected for the subjective part of the evaluation. The content and structure of the different parts of the evaluation are described in more detail in section 4.1.1. Created evaluate framework.

The analytical hierarchical process (AHP) was used as the method for constructing the evaluation framework, which was used to determine the weights of the individual evaluation criteria and components. The AHP is commonly used as decision supporting method in transport engineering (Mladenovic 2017). In reality, criteria and different indicator alternatives are dependent, so generalized form of AHP; Analytical Network Process (ANP) could also be a viable solution. This work was limited to the process of selecting the reliability performance metric, so AHP was chosen as a method for a slightly simpler and linear implementation. It is also possible to use the evaluation method to determine the relative importance of different reliability components such as punctuality. But it also can be used for evaluation of the suitability of different indicator alternatives for describing the component.

3.1.2 Evaluation method

AHP is a decision-making tool that breaks down decision-making problems into simpler forms and creates criteria for prioritizing solutions. The AHP supports multi-criteria decision-making and is able to address the different needs and views of different participants of evaluation process. In principle, AHP is designed to handle both rational and intuitive selection processes. The AHP consists of two steps, in which (1) the design phase generates a process objective, evaluation criteria and solution alternatives. These evaluation criteria are paired with each other (2) in the evaluation phase to give priority and weight to the different criteria. Solution options are also paired within each criterion, which gives priority to them.

The evaluation step can be done in three different scales - absolute, relative or comparative. The absolute scale compares the solutions to the set zero level, the relative scale compares the solutions to each other. The comparative scale sets one option as the baseline and compares the other options to it. Because AHP allows for inconsistency in evaluation, it follows

consistency through its own index. An index below 0.1 may be consistent, but small overshooting is not necessarily a sign of inconsistency. Saaty and Vargas (2012) describes the AHP method in more detail and presents several examples of its use.

Table 2. Different stages of building a hierarchy.

I	Define goal(s)
II	Define subgoals (if needed)
III	Define criteria
IV	Define subcriteria (if needed)
V	Specify participant
VI	Define participants' goals
VII	Define results and/or alternatives

The hierarchy does not need to be perfect. There is no need for a single criterion to work for each alternative solution, less important criteria can be dropped, and new important criteria can be added. Therefore, when choosing the criteria, attention should be paid to the environment affecting the problem to be solved, the factors influencing the problem and the solution to be chosen, and who wants to be involved in solving the problem. Defining the criteria for the selection process can therefore be said to be the most creative part of the definition process. Saaty and Vargas (2012) shows one way to construct a hierarchy. The presented approach goes through the different parts of the evaluation hierarchy from the goals to the alternatives. Table 2. shows part of the different steps of the Saaty and Vargas (2012) suggests.

3.2 Indicator selections

3.2.1 Selection in general

There should be some predefined requirements for indicators and their selection. It is good to choose the point of view of the indicators as it affects the selection. When describing the quality and functionality of a service, the indicators should look at the situation from the passenger's point of view. The most commonly monitored indicators - punctuality, regularity and driving time, can demonstrate the quality received by the passenger. It would also be a good idea to have other applications for indicators, such as the identification of bus bunching. (Sorratini 2008; Firew 2016; Saberi et al. 2013).

The “Traffic Capacity and Quality of Service” -manual (KFH Group 2013), which serves as a basis for performance metrics in a number of public transport planning organizations, suggests punctuality within tolerance limits, adherence to the headway and extra waiting time as indicators of service reliability. The manual also presents ideas on which stops the measurements are best to be made. The best way to keep track of normal situation is monitoring the stops where passengers’ boardings and alignments usually happen. But it is also good to take account the timing point stops. The use of data from other stops, than those that are

mentioned above, is most useful when carrying out a more detailed analysis of unreliability and its sources. (KFH Group 2013). For scheduling, it is also good to add stops that are used to define driving times. This allows an assessment of the success of the planned schedules and to identify when and what changes are needed to be done to improve reliability.

Saberi et al. (2013) note some shortcomings in the metrics proposed by the manual. The metrics 1) do not take into account the time amount of the delay but only indicate the number of delays, 2) do not adequately point out early departures and the impact of early departures on passengers, 3) evaluate the performance in the fixed comfort zones around the schedule. Thus, there is a need for complementary indicators that consider the characteristics of the interaction between the variables that affect reliability and the unreliability of the service provided. (Saberi et al. 2013). Sorratini (2008) suggests travel time variation, excess waiting time, regularity of service, and recovery time as a measure of reliability. Liu & Sinha (2007) outlines the reliability of travel time, headway and waiting times as reliability indicators. Cham (2006) suggest the distribution of travel times, the adherence of schedules, the distribution of the headway and the number of free seats. Diana & Daraio (2010) compiles reliability indicators from multiple sources but also other metrics, such as the economics of public transport. The listed reliability indicators follow similar categories as in other sources. (Diana and Daraio 2010).

3.2.2 Selecting reliability indicators

A workshop was held for all those working in public transport planning in Helsinki region transport to familiarize themselves with the indicators, to define subjective criteria with their weightings, and to evaluate indicator alternatives. The subjective criteria created in the workshop were intended to update the default criteria created by the smaller test group. Participants in the workshop were not presented in advance with the criteria produced by the test team so that they would not affect the outcome of it. Prior to the workshop, pre-material had been distributed to participants to raise ideas about what criteria the workshop should define. At the beginning of the workshop, a brief overview of the AHP method used to process the results was given, as well as the current state of definition of key performance metrics. The workshop itself consisted of three main parts: the determination of subjective criteria, the determination of the weighting of the criteria, and the evaluation of indicator alternatives by criteria. The last part of the workshop could have been done separately from the first two. Combining everything in one workshop makes it easier for the evaluator to remember what a single criterion meant.

In this work, punctuality, regularity and waiting time were selected as reliability indicators. The cancelled departures were ruled out in this work because it describes more the technical reliability of a single transport unit, such as a bus. Cancelled departures also influence, for example, the regularity experienced by a passenger, but regularity itself should consider the matter directly. In addition, data availability had its role in selection. Cancelled departures cannot be calculated in the same dataset as the other indicators. If at some point in future the data contains knowledge of cancelled departures, it should be included in indicators of reliability. For this work, one indicator was aimed to be selected for each reliability component. In actual selection, it is good to use the results of the evaluation framework of indicators. Due to minor difficulties in the evaluation of the indicators in workshop, it was decided to select the indicators in random for this work. In this chapter the selected reliability indicators were described with more details and options for individual indicators was presented.

Punctuality

In this work, the first of Kho et al. (2005) suggested punctuality indicators is chosen indicator for punctuality. This chosen indicator measures ratio of the difference between the planned and the observed event and scheduled headway. Relation to scheduled headway gives an estimation of the significance of the deviation and allows comparison for the different headway lines. Mathematical formulation for the punctuality index is as follow:

$$P_{Kho} = \frac{(\frac{1}{I}\sum_{i=1}^{I}(t_i - \tau_i)^2)^2}{h^2} \quad (1)$$

h – Scheduled headway
I – Number of operations
t_i – Actual departure or arrival time of i:th operation
τ_i – Scheduled departure or arrival time of i:th operation

The selected indicator for punctuality does not require pre-set acceptability time windows, but target levels can be still set for the results. The result of the indicator can also be used to estimate the average waiting time. Possible problem with the selected indicators is that it requires planned timetables to exist. In other words, if schedule is planned based on only headways, the indicator and punctuality in general is useless. There are problems with lines that have varying headways at different times of the day. This problem can be avoided by calculating punctuality, for example, only at peak times, when the headway is the same. Another way to solve the problem is to use the scheduled headway of the event that is under calculations.

It is also good to consider if punctuality indicator should be calculated separately for rush hour and daytime. Whether arrival or departure times are used to calculate punctuality depends on the scheduling method and process. It is a good idea to calculate the events by the order of arrival events, whereby the wrong order of arrival of the individual units does not affect the value of the indicator. To make the result of the indicator more understandable, the value of punctuality is good to change to a percentage form, whereby 100% means punctual and 0% unpunctual traffic.

Regularity

In this work, the regularity indicator is selected to be suggestion by Henderson et al. (1999). The proposed indicator is based on the Gini coefficient, which is used by economists and sociologists for evaluation the distribution of income in the society. In the case of regularity, the subject of the evaluation is the uniformity of the headway. Changing frequencies affects the interpretation of the indicators' result, so the ratio of observed and planned headway must be used in the calculation. The Gini coefficient is one way to measure the dispersion of probability distribution. The Gini coefficient is defined as integral:

$$g = 1 - 2 \int_0^1 L(x)\, dx \tag{2}$$

where L (x) is the function of cumulative distribution for the variable x to be measured. The Gini coefficient is scaled between zero and one, where the value of zero corresponds to a perfectly even situation and the latter completely uneven situation. When considering understandability, it is good to define the scale so that one corresponds to the optimum situation and zero to the worst possible case. Henderson et al. (1991) presents a simplified version of the Gini coefficient for regularity to make computation possible without integration:

$$R_{Gini} = 1 - \frac{2\sum_{i=1}^{n}(h_r - H)i}{n^2 H} \tag{3}$$

h_r – series of headways
r – Rank of the headways from smallest to largest
H – Average headway
n – Number of headway observation

Henderson et al. (1991) mentions few general advantages of Gini coefficient. For example, coefficient is independent from scales and subjective values, and it is normalized to be between zero and one, that helps interpretation. Indicator selection has also advantaged to investigate the effectiveness of irregularity interventions and to compare different lines with different headways in different situations. (Henderson et al. 1991). The Gini coefficient has some limitations that also affect the Gini-based regularity indicator. For example, with small amount of the data the Gini coefficient tends to gain too small values compared to reality.

Wait time

The budgeted excess waiting time presented by KFH Group (2013) was selected as the waiting time indicator. The indicator is designed more for larger headway lines, but in this work, it was also used for lower headway lines. In this way, the function of the indicator was studied at different headways. On the other hand, it is also worth remembering that at smaller headway lines, the waiting time is closer to half the headway and its impact on the passenger is not necessarily so significant. The indicator is calculated from the difference between the 2nd and 95th percentile of the actual arriving times. As an equation, the budgeted excess waiting time can be represented by:

$$\mathrm{BWT} = Arr\_time_{95.} - Arr\_time_{2.} \tag{4}$$

In calculation of budgeted waiting times, 2nd percentile set a limit for an amount of the data to get a reliable result. KFH Group (2013) presents a minimum requirement of 250 measurements. This limit constraints on the spatio-temporal dimensions of the results and its graphical presentations. The budgeted excess waiting time can be thought of to illustrate the

hardness of the trip to the passenger based on the passenger's experience and previous knowledge of reliability of that line. The excess wait time is the average time what is expected to wait after the scheduled departure time and is not very precise for early departures.

3.3 Spatio-temporal analysis

3.3.1 Definition

Once the key performance indicators have been selected, there is need to find the most appropriate ways to show the results of the indicators. The need also applies changes of the indicators that occurs over time. The way how selected graphs can present results, possibly the output data as well as the indicators themselves influence what and how the indicator can be used to show and visualize related phenomena. In addition, from the point of view of practical indicator applications raises the question of how the results of the indicators should be aggregated. The purpose of aggregating the results is to provide an overall picture of the observed situation and to locate the problem that is highlighted by the indicators. Spatio-temporal analysis (later STA) was created and was used as a tool in this work to solve the question. The analysis method was used to find possible aggregations levels that work for the indicators and to determine the limitations of the observed data.

The very often used input data and the results of the indicators calculations are multidimensional. For example, in the case of public transport data, it contains at least two observation-based dimensions (observed arrival and departure times) and seven fact-based dimensions (day, route, direction, stop, scheduled departure time from the first stop, scheduled departure and arrival times at stop). In addition, any additional variable that had been calculated from this data, such as the headway, increases the number of dimensions. It is very difficult or almost impossible to present such multidimensional data in any graph that keeps the graph understandable and clear. The number of dimensions of data must therefore be reduced.

To reduce multidimensionality, there are three different approaches that can be defined: aggregate data over some dimension, filtering the input data, or finding correlated dimensions and merging them into one. These ways of reducing dimensions can be used as an individual tool or combined toolbox. There are several methods for determining if the dimensions are correlated, such as principal component analysis (PCA). The filtering of the input data and the correlation of the dimensions mainly focus on the pre-processing of the input data. But for example, PCA can also be used to determine the dependencies of indicator values. In aggregations, the data is projected into some other existing dimension of data using, for example, an average. This work focused mainly on aggregation for the STA method for the reduction of the data dimensions.

The data aggregation for the dimension reduction can have a very limiting effect on how the different phenomena in the input data are reflected in the values of the indicators. On the other hand, for example, filtering the output data can also completely prevent certain phenomena from appearing in the results. As the dimensions decrease, the sensitivity of the indicator often decreases too, and the scatter of the results is distorted. The purpose of the STA is to determine what is the most accurate aggregation level that can be calculated to keep the sensitivity of the indicators as high as possible. As the STA also provides information on how aggregated data can be presented, the constraints and presentation requirements of the uses of indicators and end-users can also be considered.

Table 3. Phases of the spatio-temporal analysis.

1. Set goals for the analysis
2. Identify possible data and indicator constraints
3. Define the most significant dimensions of data
4. Generate the input data that is needed for the analysis
5. Define how the randomness is modelled into data
6. Define what the phenomena should to be modelled
7. Define the calculation parameters of randomness and phenomena
8. Use phenomena to determine the most appropriate levels of aggregation
9. Recommendations and limitations of aggregations

The number of possible aggregations is directly proportional to the number of dimensions contained in the input data. However, not all possible aggregations necessarily reduce the dimensions of the data. In this case, the one reason may be the dependence of the dimensions. The minimum number of dimensions that can be achieved by aggregation is one. This is practically just one number over the entire data and the number can be considered as an average value over the whole set. The maximum amount of dimension left after aggregation is affected by the interactivity of the presentation. Interactivity refers to the ability to move a graph on, for example, a computer screen to achieve new viewing angles. If interactivity is used, the largest number of the dimensions is, at best, seven dimensions. Without interactivity, the maximum amount is severely limited by the clarity of graphs. For example, graphs displayed in a plane usually do not use more than three dimensions.

When using the STA method to find suitable aggregation levels, the method can be described in a nine-step process. In Table 3. the different main steps are presented. The different stages of the process depend on the determinations and choices made in the previous steps. Therefore, it is always a good idea to return to the previous levels to refine the definitions and choices to address potential challenges and problems. The method can be considered iterative, although it can be presented as a straightforward process. Since each selected indicator can be considered as a unique case, a separate process description should be made for each indicator. If the input data remains the same, there is no reason not to use the exact same description for all indicators.

In the first stage, the goal is set for what is the aim of the analysis. For example, who will be the main users of the indicator and what the needs of the users will be. In the same context, it is possible to determine in which platform or format the results should be presented, as these may place restrictions on the data aggregations. These limitations will be identified in

a second stage. Constraints include known errors in the data that may affect the calculation of the indicator. In this stage, it is also worth mentioning any known dependencies in the input data dimensions.

The purpose of the third step is to find the primary dimensions in the point of view of the calculation. For example, the calculation of regularity can be performed from the arrival times of the stop, in which case the departure time from the stop is not a relevant dimension. Another characteristic of the primary dimensions is their mutual (at least partial) uncorrelation. If two different dimensions correlate, only one of them can be used to aggregate the data. The primary dimensions of the data that is defined at this stage can later be utilized to determine which aggregation are possible and useful.

Because modelled data can be used as part of the STA method, the input data need to be produced as close as possible to observed data. And all the possible phenomena that can occur in observation has to be modelled in produced data. This creates the opportunity to validate the effects of environmental phenomena on indicators. In the process stages five, six and seven implementation of the randomness and the phenomena into produced data is defined. For example, the way which public transport timetables are made up can be consider a possible systematic error source. For modelled data to correspond to the actual observation, it needs to be produced as close as possible to normal production practices. In this case, non-environmental problems identified by the indicator can be identified as systematic errors. Adding randomness into produced data requires a distribution, for which the randomness is based, as well as the calculation parameters of the distribution. Before adding any phenomena into data, it should be determined which phenomena are needed to be modelled. Determining the intensity of phenomena is important because it defines how sensitive the indicators are in certain aggregation level. There is also a need to define how and where the randomness and phenomena are modelled into data.

In stage seven, the parameters of phenomena and randomness are determined. The purpose is to find so-called optimal values for the parameters. For randomness, the higher value of the parameter than optimum value causes randomness taking over indicator results. Therefore, after that point, the results of the indicator are associated with increasing unreliability. The best value of the randomness parameter is either the optimum point or a value slightly below it. In this case, the output data contains a maximum amount of randomness that does not clearly affect the results or compilation. The same logic works for the parameter values of the phenomena. But at parameter values lower than the optimal value, the effect of the phenomenon disappears and cannot be clearly distinguished in indicator results. Either optimal or higher values can be used for the phenomenon parameters values. The closer the selected value is to the optimum, the more sensitive it can be obtained from aggregation the indicator results.

The last two steps of the process of analysis use of the phenomena defined at the previous levels to test how the input data can be aggregated. To begin with, it is a good idea to start at the level from which the reduction of the dimensions is lowest. When such a level is found, it can be considered as a level capable of the maximum accuracy of the indicator. Reducing more dimensions reveals other compilations that work for the indicator. During these steps, potential systematic errors may often occur, which can be due to either an error in the production of the input data or an effect of the planning practices. There is also information on whether the indicator can detect certain phenomena at all. These steps of the process should

be performed for each indicator separately, as the functionality of the aggregate level is indicator selective. At the end of the method, information is compiled on the aggregation levels that have been proved to be the best for each indicator, on the functioning of the indicator under calculation, on the detection of various phenomena, and on the events limiting the use of the indicator. This gathered information can be later used, for example, in ad-hoc analyses.

3.3.2 Example of the analysis

The nine steps of the process presented in section 3.3.1. Definition has been applied in this work to perform a spatio-temporal analysis of the selected indicator. This section discusses in more detail how the different stages of the process were defined and how steps seven and eight were implemented in practice. Results of the steps seven and eight are presented in Section 4.2.1 Analysis parameters, and step nine in Section 5. Discussion & Conclusion.

1. *Set goals for the analysis*

In this work the goal of the STA method was to find few aggregation levels and graphs for the selected indicators by implementing nine step process. Also, the aim was to test how well created STA method works in practice and what improvements it needs for further development. Planners was set as end-user of the indicators and the practical use was set to be support for schedule planning and production. In selection of the graphs, the guideline was to keep the graph as simple and demonstrating as possible. The criteria that was obtained in the workshop was used, where applicable, for guiding the selection of graphs and aggregation levels.

For each indicator, the possible representations were limited to two graphs to limit the scope of the work. Another reason for limiting the number of the can be justified by the fact that repeating the almost identical process does not add value to the evaluate process functionality. The aim was to define the two selected levels of aggregation in such a way that each indicator's graph would become different. The choice of aggregation levels also emphasized the suitability of the aggregation for the indicator itself. One of the background reasons for choosing different cases was the need to get more knowledge how the analysis and implementation of STA method would work in different kind of situations.

2. *Identify possible data and indicator constraints*

Aggregate levels and indicator results are affected by the way the schedules are planned. This can be called a systematic error that already occurs at the planning level. The systematic error itself covers several different types of error sources, such as measurement errors due to way how the observation has been made. The method of scheduling in the HSL area is to divide the route into few segments within a few stops and assign defined driving times to the segments. The purpose of segmentation is to divide the route into similar sections where the delays and disturbances caused by the traffic environment are similar. The aim of segmentation is to reduce the deviation in driving times and the consequent uncertainty about the adherence to timetables. For stops within a segment, stop departure and arrival times are approximated by dividing the segment's driving time by a suitable factor, for example according to the stop distance.

In planning of timetables and driving times minute is used as precision level. This allows two stops to have exactly the same stop time, even though they actually have a difference in distance. This is likely to affect, for example, punctuality results at certain stops. On the other hand, the result of punctuality can also tell in certain situations how the successful of the approximation of driving times was. Another issue that is likely to affect the indicator's results through scheduling is planned vehicle tasks. A vehicle task contains all the departures what will be driven by one vehicle. Figure 3. shows an example of the structure of several vehicle tasks with multiple departures in two different direction. The structure of the vehicle task has an effect the delay from the previous departures exceeds the time between it and the next departures. In this case, the delay will move to the next departures. This delay transferring is repeated until there is enough time between the two departures to achieve a punctual departure. If there is a short time between two departures in vehicle task, the departures are generally late due to even a small random delay that occurs during the route.

Because the method of scheduling has a large effect on the observations, a theoretical line was created. This allows to detect possible systematic errors in schedule planning process. Using actual observations alone may not be possible. With real observation it not is easy to distinguish from the indicator results' deviations which is cause of deviation: the planning process or the traffic environment. By using modelled phenomena on an empty schedule, information on the success of the schedule structure as well as the functionality and planning of the vehicle task can be obtained as a by-product of the analysis. If there is considerable effect in indicator results caused by schedule planning, removing the effect should be considered. Without removing it, modelling different phenomena and detecting their effect on the indicator values becomes more ambiguous, and it may make it impossible to compare different phenomena with each other.

3. *Define the most significant dimensions of data*

From the observation data, the independent and primary dimensions were defined as "location", "date", "time" and "value". In Figure 1. the selected dimensions are presented as a three-dimensional model. For example, the dimensions route and direction were limited out by using a single route and dividing the graphs into two different graphs for each direction. Without filtering these dimensions out, location and direction is dependent, and calculating indicator's value through different routes are arbitrary.

The location-dimensions can be a random point or a stop along a route if there is available observation data supporting it. In this analysis, stop sequence as a measure of dimension. There is a clear difference between the two selected temporal dimensions. The date-dimensions identifies which schedule structure is in use and the time-dimensions which departure it is. The defined place, time, and date -dimensions can be considered as fact-based dimensions. Instead, the value-dimension is a variable that is measured at the point indicated by the other three dimensions. The difference between fact-based and variable data is that the fact-based data is already known before the time of event, but the variable data can only be obtained at the time of the event.

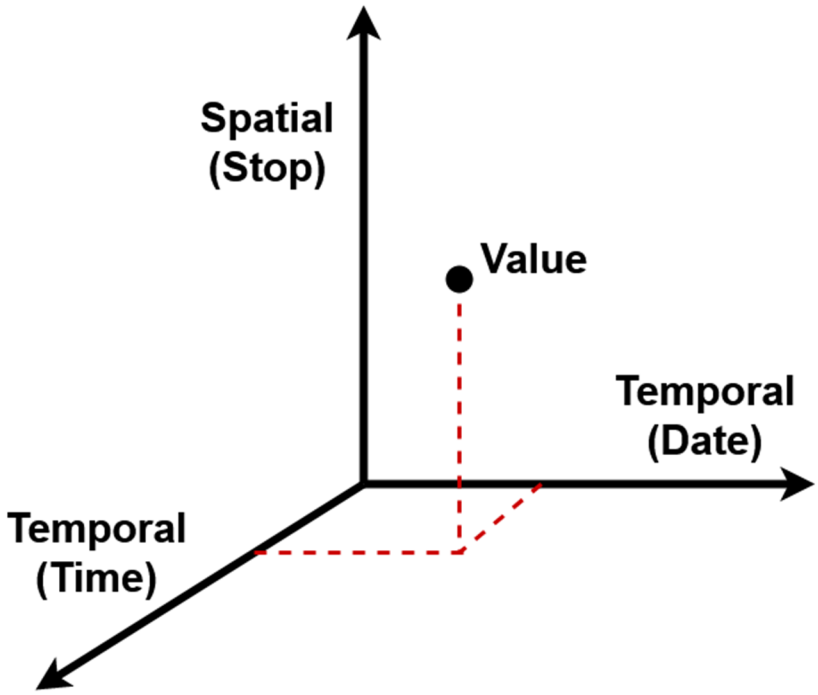


Figure 1. Primary dimension of used data.

Since the data defines the three dimensions as the primary, the most accurate level of aggregation can be considered the 4D-graph as in Figure 1. has been presented. Effective use of a 4D graph requires interactivity with the graph to make its use clearer. The Figure 1. is simplified version with only one datapoint so it can work well. In theory, each point of a cube bounded by three fact-dimensions has its own value defined by the indicator. In practice, this single value can necessary not be calculated without a partial aggregation of one dimension. The 4D-graph also requires the end-user to have some experience with using it.

The Figure 1., that shows the dimension structure of the data, made it possible to determine all possible aggregate levels of the indicator. Figure 1. is itself a possible representation, although there are some limitations in its calculation and presentation which depends on the indicator. Each face bordered by the two fact-dimensions on which the value dimension is projected produces three different type of graphs. In this case, the other fact-dimension perpendicular to the plane is aggregated into the value-dimension. A heatmap-styled graph can be considered as an example of this. A third possible way to aggregate is to project the other two fact-dimensions into a value-dimension, with value acting as the second axis. For example, traditional line and bar graphs are such projection. The fourth aggregation can be considered a situation where all the fact-dimensions are aggregated together with the value. In this case, the result is a single number that describes the value of the indicator over the entire data set.

Table 4. The fact- and value-dimensions which was used for each indicator.

Seq – Time – Date – Value	-
Seq – Date – Value	Wait time
Seq – Time – Value	Regularity
Time – Date – Value	Punctuality
Seq – Value	Punctuality
Time – Value	Regularity
Date – Value	Wait time
Value	-

For this work, two different aggregate levels were selected for each indicator. For each, one three-dimensional and one two-dimensional aggregation level was selected for which a model graph was created. Aggregation over the entire data set was omitted, as it does not clearly indicate where the potential problem is. Also, a graph in which all dimensions would be presented without compilation or partially compiled was omitted. The reason for this was the requirement for the interactivity of the graph that is required. In Table 4. it is shown which aggregate level and graph were tested for which indicator. With these choices, it was possible to get as versatile a picture of the functionality of the method. In the final implementation, all combination for each indicator should be reviewed.

4. Generate the input data that is needed for the analysis

Because the method of scheduling was identified as a limitation of data and indicator functionality, the input data had to be produced through the same design process as would normally be provided. By acting in this way, the possible effect of planning process on the indicator can be easily identified at a later stage. Scheduling begins with defining the route of the service line and the stops it uses. If necessary, the stop of the route can be set as a timing stop or a stop delimiting the segment. Twenty stops, including the start and end terminal, were randomly placed on the theoretical line made to produce the input data. In addition, one stop on the route was made a timing stop and another a stop delimiting the segment. Figure 2. shows a model of the theoretical line that was made. Stop distance were assumed to be the same, so the driving time is evenly distributed over them. In addition, extra dwell time of one minute was defined for the timing stop.

After defining the route, a driving time was created for the line, which is slightly longer in rush hours than in day traffic. In addition, added layover time, intended to prevent transferring delays to subsequent departures in vehicle task. After defining the driving time, the base schedule and vehicle tasks were defined for the line. A regular ten-minute headway was used, and the time of operation was defined as 05:00 to 01:00. In Figure 3. planned vehicle tasks are shown and from the picture it can be said that the line needs at least ten vehicles to operate. The vehicle tasks were later used as an aid in modelling the recurrence of phenomena. The terminals always used a turnaround time of at least 30 seconds, even if the departure

arriving at the terminal was late. The above data provided a base schedule in which all observed departures and arrivals were punctual. Randomness and phenomena were added to this base schedule in later steps of the analysis.

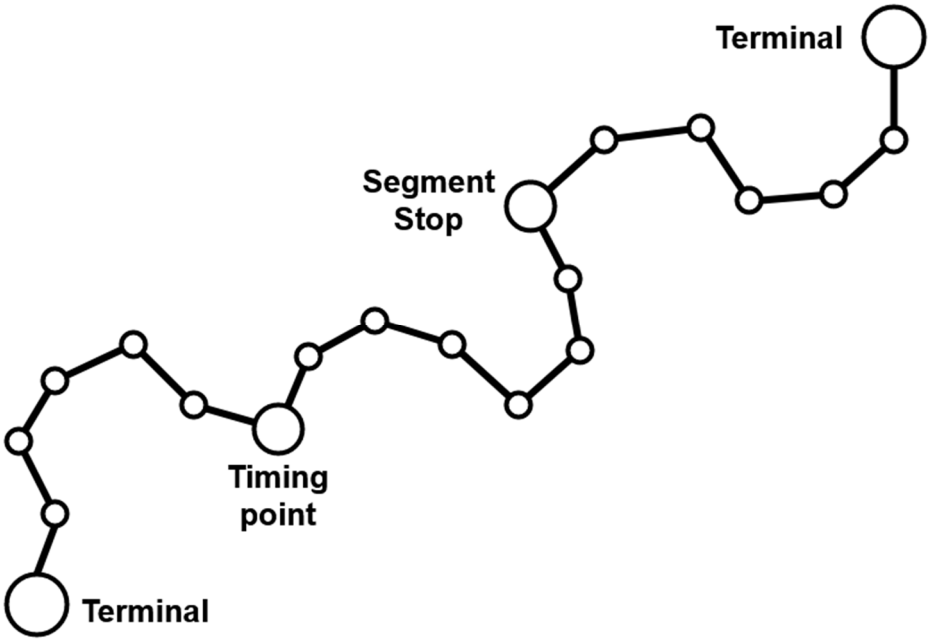


Figure 2. The theorical line and its stops

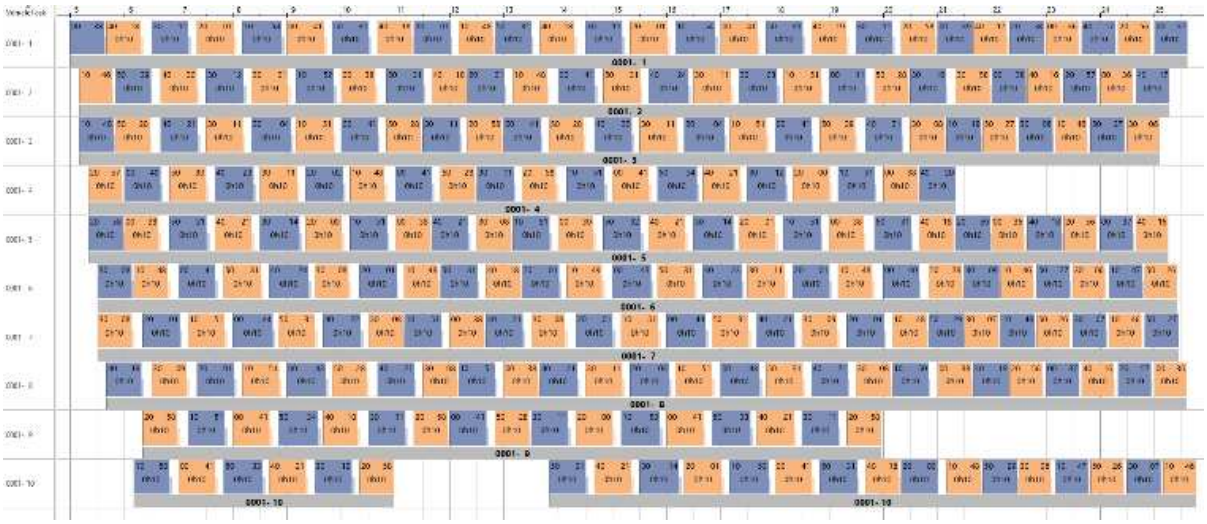


Figure 3. Vehicle tasks which was planned for theoretical line.

5. Define how the randomness is modelled into data

In the base schedule, which was produced in the previous phase, randomness was added to both the driving times between stops and the stop times. In both cases, it was randomly extracted values from the normal distribution. Since the normal distribution also produces negative values, some changes have to be made to the extracted values. In the case of driving time, negative randomly selected values are limited so that the driving time between stops cannot be less than half of the scheduled driving time between them. The stop time cannot have negative random values at all, so all the negative values was ruled out. These two assumptions provided a sufficiently good description of the randomness in reality for purpose of the analysis. Those assumptions also preventing some impossible situation to occur. The parameters σ and μ values are defined in section 4.2.1. Analysis parameters.

6. Define what the phenomena should to be modelled

The following phenomena were chosen to be modelled on the base schedule: cancellation of departures, increased driving time in rush hour, generally increased driving time, vehicle bunching, and a version containing all the above. These phenomena were modelled on thirty-day data in five different versions: random single day, once a week, every other day, weekdays, and every day. By modelling the phenomenon into monthly data, it was possible to test how often the phenomenon must occur in order to be able to detect it from the indicator and its graph. In addition, randomness has always been added each day, regardless of whether the phenomenon was added to that day. In this way, it can be illustrated whether the phenomenon that occurred differs from normal randomness.

The modelling of the increased driving times in rush hour was done by adding additional driving time to both morning (07:30 – 09:30) and evening (14:30 – 17:30) rush time. The increase in driving time was made the same amount for all section, and it was made for stop events that occur on schedule at specified rush times. Increasing driving time could also have been done, for example, as a time-varying amount of increment, or increasing driving time for some stopping sections randomly more than for others. Increasing driving time throughout the schedule day was done in the same way as increasing rush time driving time, but without limiting a time period. The vehicle bunching was modelled by adding a constant amount of delay to every odd departure. This modelling method does not involve or add any extra randomness in the departures, but certain departure events are only moved to occur later. For this reason, there is no direct breaking point for bunching. So, the value of the optimum parameter for vehicle bunching must be determined based on e.g. the planned headway.

The cancellation of departures was modelled based on the cancellation probability, which depends on the amount of the possible delay of the departure as well as the constant probability. The constant was intended to describe, for example, the accidental and unexpected breakdown of a vehicle. Because cancellation is modelled as random probability, the number of cancellations can vary between different data sets and dates. The probability of cancellations can be also determined from actual observations for the phenomenon parameter. In this work, the value of the parameter was determined according to the latter option. The last modelled phenomenon was the combined effect of all other phenomena. In this context, the parameters determined for other phenomena alone were utilized. It must be noted that the

interaction between different phenomena is probably very strong, so the parameter values need to be reduced somewhat from the initial values.

3.4 Example case

As an exemplar study, two lines, 500 and 510, are used to present the results of the spatio-temporal analysis. The lines operate part of the route in a dense urban environment, part in a sparse urban environment and part in a highway environment. The impact of the traffic environment is therefore expected to be reflected in the results of the indicators. The selected lines are transverse trunk lines and in Figures 4. and 5. the routes of the lines are shown. The trunk lines have their own brand with e.g. own specific colour theme. The results of the indicator on trunk lines are most affected by the lack of ticket sales on vehicles and longer stop distances. The lines do not have their own lanes and have only partial own traffic light benefits. Light benefits are limited because other forms of public transport intersect a lot with these lines. Full benefits would have a detrimental effect on the reliability of other traffic. Although the lines have different head numbers, driving times and schedules are planned as single line. In this case, a frequent headway can be provided with the common section of the lines. For this reason, the lines are not considered in individual lines in this example analysis. This approach is also how most passengers see the lines. Demand is concentrated, in terms of time, during peak hours and spatially on the common section of the lines, and the congestion direction of the lines is in the morning from the centre outwards and in the evening in the direction of the centre of the lines.

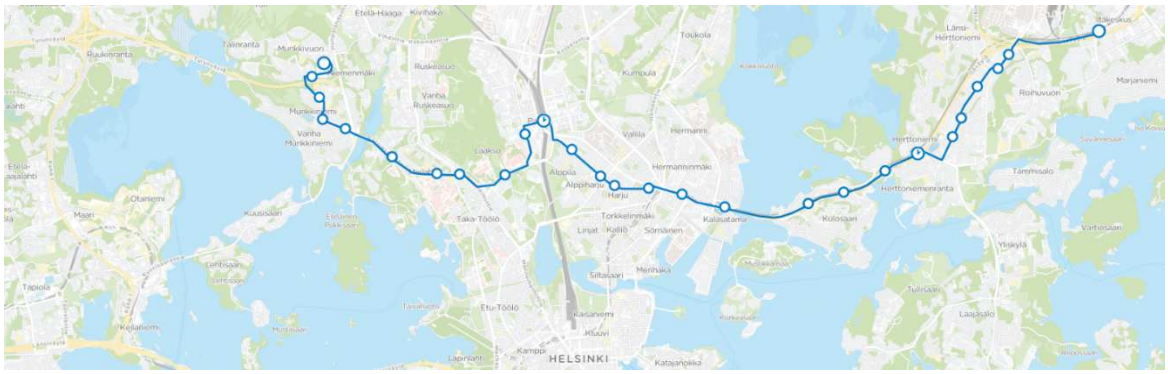


Figure 4. Line 500 route in direction one (HSL 2020).

The combined headway is 3 minutes in rush hour and in 6 minutes in day traffic. Individual lines have headway that is half of combined headway. Stop events from 2020-01-01 to 2020-04-10 were used in the analysis. The same events are also used in driving time planning, so the findings could be used directly to aid in driving time and schedule planning. The exemplar analysis focuses only on weekday events. The observed arrival and departure events are generated by the stop radius. A timestamp is registered for the event when the limit, specified by the stop radius, is exceeded. There may be different events within the radius that distort the actual moments of events. For example, traffic lights and congestion caused by other means of transport within a radius appear as phenomena that alternate stop time. Towards the end of the observation data, the method of identifying stop events was started to change

from location-based to door event-based identification. This is likely to affect the temporal comparability of the data. In some stops the affect can be significant.

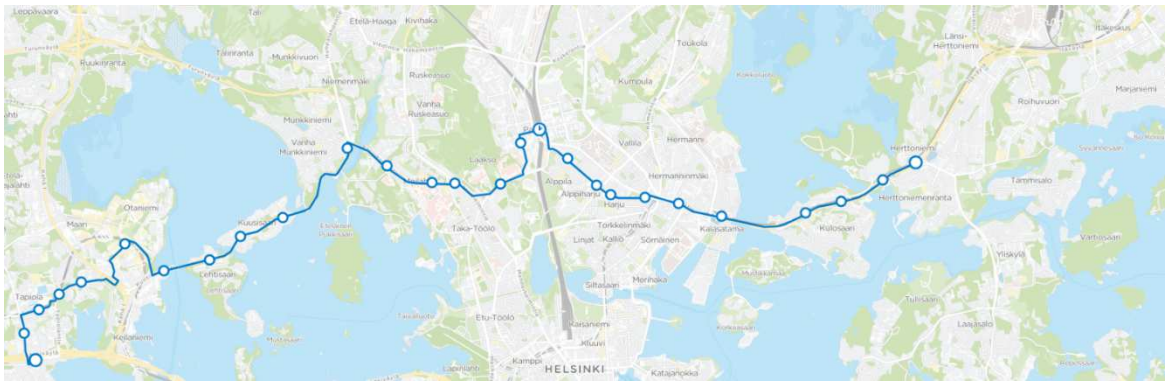


Figure 5. Line 510 route in direction one (HSL 2020).

4 Results

4.1 Evaluation framework

4.1.1 Created evaluation framework

In this work, public transport planners and the transport data from HSL were used to construct and define evaluation framework. Also, a smaller group of planners, who specialize in data processing and ad-hoc analysis, was used to test creation of framework in practice. The main objective of the evaluation was set to be the definition of a set of reliability key performance metrics as well as the indication of which indicator alternatives fit best in each of the defined reliability components. Additional objective was to define the general technical requirements for the indicator, to identify the best option for the user as an indicator, and to identify the indicator that has the greatest impact on planning and decision-making. The result of the evaluation should be one indicator for each defined component of reliability and a few possible supporting indicators.

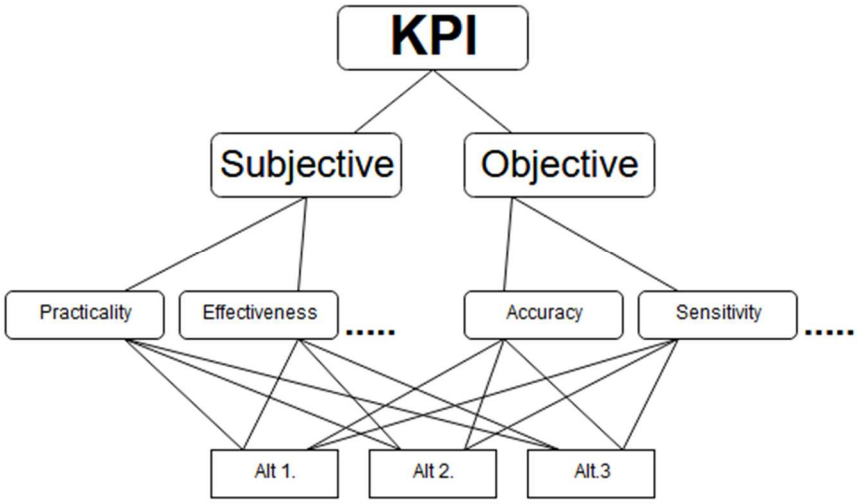


Figure 6. The hierarchy created by the AHP method is used as the basis for the evaluation frame.

As mentioned earlier, the criteria for selecting the indicator were divided into two separate parts. The objective part is intended to evaluate indicators through technical criteria and the subjective part to evaluate indicators for the user. Figure 6. shows the hierarchy structure that was created and used in the evaluation. With the smaller test group, both parts were defined. The main purpose of the test group was to produce the objective evaluation criteria and weightings used in the evaluation. The group was also used to create default for subjective criteria and weightings, and to define the relationship between subjective and objective evaluation. The test group was able to evaluate the functionality of the method and possible areas for development before taking it to the workshop. The purpose of the workshop was to use a larger group to define the final subjective criteria with their weightings and to make the evaluation of the indicators.

In both parts of the evaluation framework, the consistency ratio is monitored so that the evaluation given by evaluators are consistent at some level of hierarchy. The ratio was also included in the evaluators' own evaluations so that they can keep track of whether the evaluation is good or whether it needs to be reviewed. The subjective part of the evaluation was made by the AHP paired comparison, but the objective part was done by applying the AHP method. This meant that, in the objective part, a single indicator was scored from one to five to describe its performance in the technical requirements. The advantage of this applied approach is that the evaluation can be made for only one indicator at a time and to find out whether the indicator itself requires some rethinking. For indicator alternatives, the evaluation grade can be changed to the same scale as the subjective part, if necessary.

In addition to the creation of objective criteria, for the whole objective evaluation part was created an automatic calculation that is not user dependent. This ensured that the user's or evaluator's own opinion did not influence the evaluation or its outcome. Attention was also paid to the reproducibility of the evaluation using only one input data which was created for the objective evaluation. The test group defined six different criteria as the criteria for objective evaluation: repeatability, accuracy, objectivity, stability, sensitivity and efficiency. In Table 5. presents the criteria for the objective part and their weighting factors that was worked out in test group.

Table 5. Evaluation framework weightings created by the AHP method with the test group.

	Weight		**Weight**
Objective	0.35	**Subjective**	0.65
Repeatability	0.11	Effectiveness	0.38
Accuracy	0.11	Understandability	0.22
Objectivity	0.22	Practicality	0.06
Stability	0.26	Descriptiveness	0.26
Sensitivity	0.21	Comparability	0.09
Efficiency	0.10		
λ_{max}	6.07	λ_{max}	5.10
CR	0.01	CR	0.02

The repeatability criterion compares the dependence of different indicators on the amount of input data. At best, a reliable value for an indicator can be achieved with only one measurement. At worst, an indicator requires a very large number of measurements. Accuracy in the criterion is an estimation of which of the indicators can be computed most accurately in the spatio-temporal dimension. Ideally calculation can be made at any point anywhere along the route. The objectivity criterion estimates how much arbitrary prior information the indicator needs. The stability criterion seeks information about the impact of a single piece of incorrect or incomplete information, whereas the sensitivity criterion seeks to know how small changes can be reflected in an indicator's results. The efficiency criterion estimates the amount of time used for calculation with different amounts of data. The efficiency criterion is partly dependent on the way the indicator is coded. On the other hand, there is usually

only one calculation method that is optimal for the indicator, so the result of this criterion can also be used to evaluate the coded algorithms and check if it needs re-structuring.

In defining subjective evaluation, the purpose of the test group was to practice the method and find areas for development. The two most significant areas for improvement were found in the response form and in the evaluation scale. The response form should be as easy to complete as possible and any unnecessary information should be cut out of it. It would also be a good idea to include a consistency ratio into it. For this reason, the response form was changed to a response on the website. On the page, it was possible to weed out all unnecessary calculations for the respondent, to give an estimate of the consistency and to simplify the answering format. Another significant development target, the evaluation scale, was found difficult to perceive and use in the test group. Therefore, the scale was changed to use the range -5 to +5 instead. The values were changed to a form more suitable for calculation in the background. In addition, the change in the scale affected the structure of the response form. The test team defined five criteria as subjective criteria: effectiveness, understandability, practicality, descriptiveness and comparability. In Table 5. presents the criteria for the subjective part and their weighting factors that was worked out in test group.

The subjective part of effectiveness criterion is used to evaluate which indicator is most influential in decision-making. The criteria of understandability and practicality assess the calculation results and calculation of the indicator based on how easily they can be understood and communicated internally and externally. The descriptiveness criterion is used to map how well the indicators is describing the phenomenon and its changes in the user's point of view. As a last criterion, the comparability in the subjective part assesses how easily the result of the indicator can be applied in comparison or can be used in other calculations.

4.1.2 The workshop

Three different steps were used to define the criteria, at the end of the steps the subjective criteria would have been defined, including their descriptions. At the beginning of the definition of the criteria, the participants were divided into groups of three. Within the groups, the criteria were discussed with the help of guiding questions what a good indicator is and what kind of indicator participants would use. The purpose of the discussion was to gather ideas, which had been aroused by the preliminary material and the presentations. The groups came up with individual words that could be the basis for the criteria. In the second step of the definition, the groups were mixed together so that at least one member of the group was from the original group. This way, at least one member of the group knew what the words in question meant and what was the idea behind them. The new groups were given the task of grouping individual words into higher themes. At the same time, the group considered initiating a description with a theme that could be used to define what the theme addresses. A new name was invented for the theme, if none of the words that were part of the theme, the theme described the theme well enough in general. At the end of the definition of the criteria, the themes that the groups had assembled were discussed throughout the workshop. From these proposals, four main themes were created, for which precise descriptions were defined at the end. Weightings were defined by participants in workshop for the subjective criteria thus formed and used to evaluate the indicator options presented in the workshop.

Table 6. Evaluation framework weightings of subjective criteria created by the AHP method in the workshop.

	Weight
Subjective	0.65
Relevance	0.15
Controllability	0.19
Legitimacy	0.43
Intelligible	0.23
λ_{max}	4.00
CR	0.00

As a result of the workshop, the subjective criteria became relevance, controllability, legitimacy and intelligible. In Table 6. the determined weights of the subjective criteria produced in the workshop have been compiled. Relevance aims to describe the necessity of the indicator for the work and the comparability of the results. The purpose of the controllability criterion is to assess how easily the results of the indicator can be used in one's own work and how one's own activities can affect the result of the indicator. The legitimacy assesses how well the use of the indicator is acceptable in decision-making and how it can be used to communicate to outsiders. The last criterion, intelligible, evaluates the ease of use of the indicator, how its results can be understood and whether the results are unambiguous.

After the workshop, participants were asked for feedback on development ideas for future workshops. A good aspect of the workshop was the selection and inclusion of indicators for information sharing and inclusion. The main criticisms were the evaluation and the familiarization with the indicators. In making the assessment, the comparison of one indicator between the two criteria was found to be very cumbersome and easily confusing. Also, the time to get acquainted with the indicators was not enough and the comparability of the results given by the indicators to reality could not be easily understood. Subjective evaluation obtained for these reasons may be distorted and their use is not recommended at this point.

Based on the feedback received, changes were made to the way the workshop was run and evaluated for the next workshop. The evaluation method is changed so that each indicator option is given one rating for how well it meets the criterion. A scale of 0 to 5 is used as the value scale and the result of the evaluation is converted into the form required for the calculation in the background. In the future, the pre-material will include detailed description of the indicators so that participants can learn more about them. Indicators will also be added to the evaluation page for preliminary research. At the same time, it would also be good to make modeled data that will allow the planners to study the impact of different traffic phenomena on the value of the indicator. The final phase of the workshop, the evaluation of the indicators, will be renewed later for better results with a revised evaluation form. No new workshop will be held during this work.

4.2 Spatio-temporal analysis

4.2.1 Analysis parameters

In the search for the optimal values of the parameters, the deviation from the schedules was used as an output parameter. The output parameter's value was determined as the average over all stops. The change of output parameter and the moving average over ten output parameter's values were also used to determine the optimal value. With the help of these three calculated tools, the value of the optimum parameter can be determined analytically and visually. The approach is in favour of punctual indicators. For example, if the average change in headways had been monitored, indicators of regularity would have been in favour instead. In case of wait time, both the deviation from the schedule and the change in headway should have been monitored. The output parameter that was used to determine the optimal values was identified as one of the method development's targets in the future. Since the effect of phenomena is very dependent on the structure of the schedule, each different schedule structure must have its own values of the phenomenon parameter. Only one schedule structure was used in this work, so only one optimal parameter could be defined.

Parameters of randomness

The mean parameter of the selected normal distribution was set in zero. The mean parameter can be used to describe a systematic error that was not intended to be modelled through randomness in this work. Determining the values of the randomness' parameters thus focused on determining the scattering parameter of the normal distribution. Each increase in randomness with the same scattering parameter produces a different result for the output parameter. Therefore, the same scattering parameter of randomness was tested several times to achieve the result.

The mere effect of randomness on the schedule and its sustainability clearly distinguishes the point at which the effect of randomness on the mean deviation changes. In Figure 7. the effect of one the test run in output parameters is shown. In this case, the effect of the increase output parameter has changed by about the value of the scattering parameter 40. The final optimum value for the scattering parameter was defined as 21.7, at which point randomness began to strongly affect the result. After that point, the variance of the value used to determine the output parameter begins to increase sharply. With a scatter parameter value of 21.7, the total effect on the mean deviation of the schedule is about one minute. The standard deviation corresponds to about one tenth of the headway in the used schedule structure.

The mean deviation of about one minute achieved can be considered as a generally good tolerance level. Even a larger deviation could be allowed, but the purpose of modelling randomness must be remembered. It was to add a small inherent variation to schedules. Larger tolerances can be used to identify potential problem areas in the schedule structure. To reach a benchmark for randomness optimum parameters, the same method was tested with a different schedule structure. From that experiment, the value of the optimum parameter was found to be 16, which is slightly lower than for the first test case. The reason for the difference is probably the difference in the vehicle task and the different headways in the schedules. Roughly, however, it can be estimated that decreasing the headway has a decreasing effect on the optimum value. The effect of the headway changes on the optimum value should be tested in more detail. If more detailed testing reveals that the headway has a dependence on the optimum value, the relationship can be used to determine the optimum value

more quickly. It should also be noted that the number of stops and the way the randomness is modelled, influence the determined optimum value.

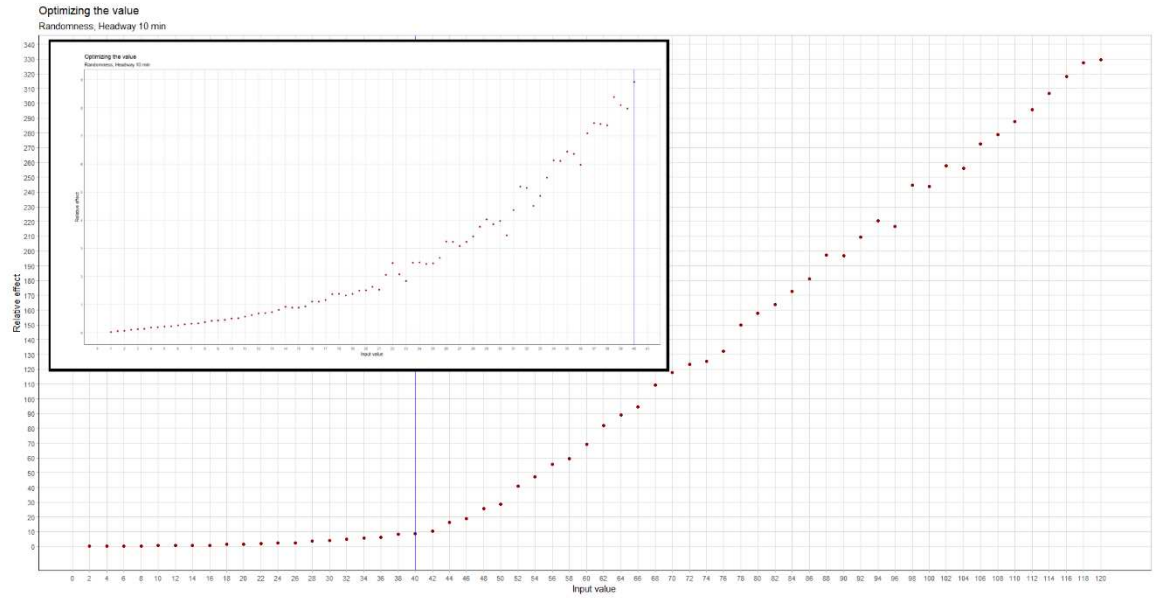


Figure 7. The result of one test run for the finding the optimum value of randomness.

Parameters of phenomena

There is also similar behaviour in determining the optimum value for increased running times in rush hours (later rush) than was in case of randomness. Because the effect of rush is tested with randomness, several determinations test runs must also be performed to determine the optimal parameter value. The value of the output parameter used in the monitoring of the effect of a phenomenon seems to have exponential correlations. This is clearly different from the randomness, where there was a clear optimum point in the output parameter. The optimum value determination methods used gave a value of about 10.0. Problems in determining the value arose from finding a clear point of change and therefore the determined optimum value is a rough estimate. The estimate is based on the point where increasing one unit of the strength of the phenomenon changes the value of output parameter faster than the linear increase would require.

The time limitation of the rush hours and the output parameter calculation method cause the optimum value for rush phenomena to be considerably large. This is because the rush hours are only a fraction of the whole day data and the output parameter is calculated as the average of its values for the whole dataset. Temporal delimitation is presumably the cause of exponential dependence. It should therefore be noted that the defined duration of the rush time influences the value of the optimum parameter. It is therefore advisable to use reality-based rush times, during which the values of the indicators will be able to see the potential impact of rush-phenomena.

The increase in driving time throughout the whole day (later delay) eliminates the time-limited problem that made it difficult to determine the optimum value for rush-phenomena. Thus, in determining the optimum value of the delay, the same type of point of change can

be observed as in the output parameter of randomness. In the case of a delay, the change point occurs when the buffer times between vehicle tasks have elapsed and the departure delays are transferred to the next departure within the vehicle task. The specified optimum value for the delay is about 2.75 and is significantly less than the optimum value for rush-phenomena. The smallness of the optimal value of delay was influenced by the schedule's structural problem during the late evening traffic. This point is the most challenging in terms of compiling vehicle tasks. The reason for the challenge is the reduction of the defined driving time as well as the decreasing number of available vehicle task groups. This causes the time between departures to shrink to a minimum. This observation is also likely to be reflected in the definition of the aggregation levels of the indicators. If, for example, the extra buffer time of the evening traffic problem was added, it is likely that the optimum value of the delay would increase.

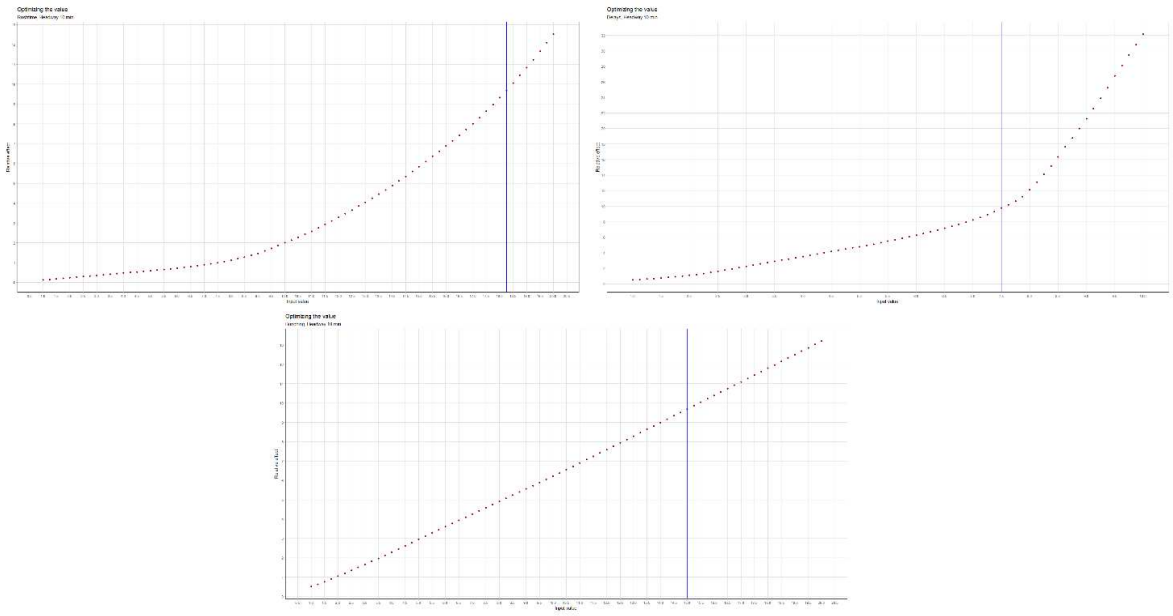


Figure 8. Example of defining a phenomenon parameter in one test run.

In the case of determining the optimal parameters for bunching, the method of modelling the phenomenon became a problem. If the determination of the optimum value were done in the same way as in the case of rush and delay, the optimum value would have been infinite. The reason for infinite value is the linear change in the output parameter. It is very logical, as the modelling of phenomenon shifts only half of the departures by the value of the parameter to happened later. In Figure 8. examples of one test run of optimum values for rush, delay, and bunching are presented. Different behaviour of phenomena can be observed in the Figure 8., such as a linear dependence of bunching. The bunching phenomenon was intended to study the aggregation of the regularity indicator, so the output parameter is not a good figure to monitor in this case. The optimum value for chaining was determined based on the headway, so the optimum value was set to 3.5.

The actual percentage of cancelled departures from reality was used to determine the value of the cancellation parameter. This value was found to be too small to be sure to cause even one cancelled departure for a single day. The value of the parameter determined based on

reality had to be increased to one. This corresponds to one cancellation of about every hundred departures. The determination of the cancellation parameter should therefore be tied to the number of departures of the data. This can affect how many departures are cancelled in one modelling session.

Table 7. Determined optimum parameters value for randomness and phenomena.

Optimum values	Randomness	Rushtime	Delays	Bunching	Cancellation
Individual	21.7	10.00	2.75	3.50	1.0
Combined	14.5	2.00	1.38	1.75	0.5

When analysing the combined effect of all phenomena, it was found that the optimum value defined for an individual phenomenon cannot be used. Phenomena easily accumulate and caused the entire schedule structure to disintegrate. Problems with cancellation were also caused by the additional effect of accuracy used in modelling. The values of the parameters of the individual phenomena had to be reduced somewhat. In the case of rush, the reduction was considerable because other phenomena easily accumulated with it. In Table 7. the defined optimal values of the randomness and phenomena parameter are presented. Those optimum values were used in the modelling of phenomena into base schedule. Examples of the effects of the phenomena on a single schedule are collected in appendix 1.

4.2.2 Punctuality

The aim was to find two aggregation levels for the punctual indicator: 3D and 2D levels. The first of these included the time, date and value-dimensions and the latter the location and value-dimensions. At the 3D aggregation, three different cases were created to present the indicator. Different cases were defined as the aggregation of the location-dimension, the partial aggregation of the location and time-dimensions, and the removal of the bias from the aggregation. The aim of removing the bias is to eliminate the effect of the schedule structure on the result of the indicator and to produce a clearer graph. In the calculation, it is done by using events involving mere randomness to set the baseline to which data containing phenomena were compared.

When drawing the graphs, attention was also paid to the colour scale used. When choosing a colour scale, it is good to think about which one is wanted to be emphasize in the different situations. Against a darker background, a lighter deviation is easier to read and clearer to detect at a quick glance. In the opposite situation, readability suffers quickly when a brighter light colour dominates the graph. In addition, one additional division was added to the graphs. The location dimension was grouped according to the segments used in the runtime planning. This division was intended to improve the usability of the results in the context of runtime planning.

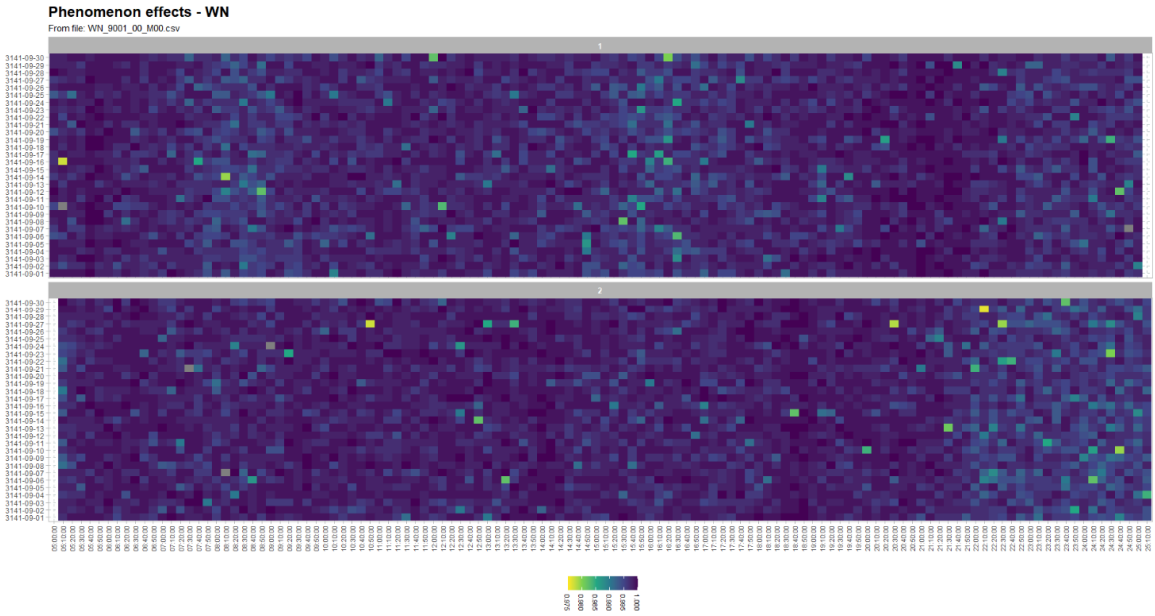


Figure 9. The effect of randomness in the date-time graph in both directions.

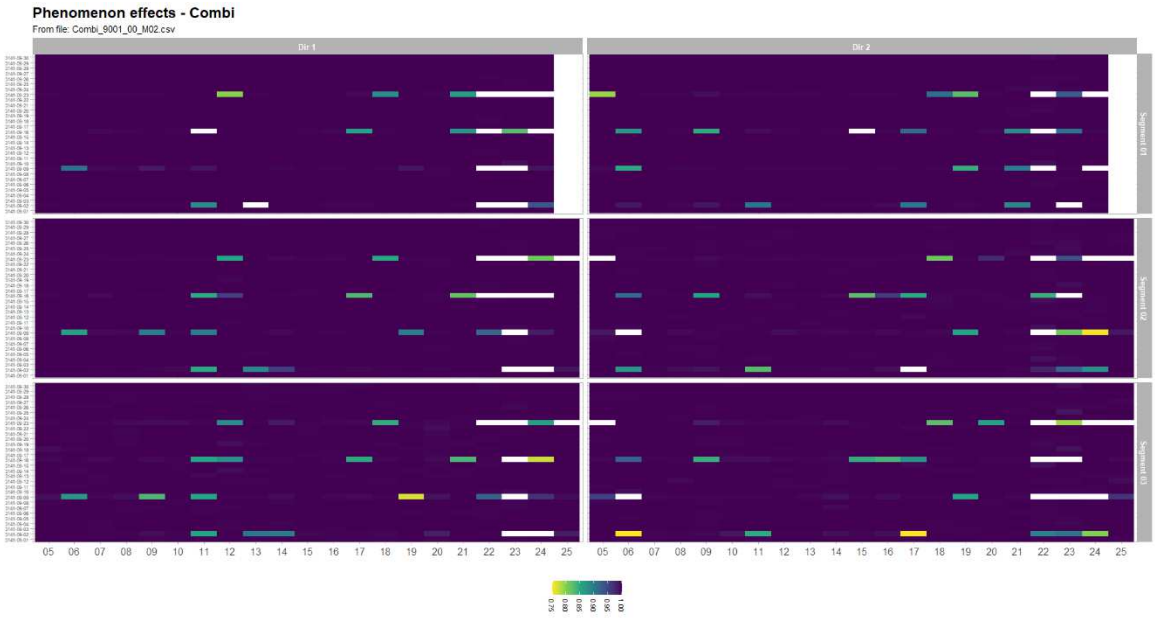


Figure 10. Punctuality in date-hour group graph were the effect of the schedule structure is removed.

Aggregation of the location-dimension proved to be practically impossible for the indicator without even partially aggregate another dimension. The reason for the challenge was the headway which was required to calculate the value of the indicator. Each stop has its own distance from each other, based on which the headway to be calculated is arbitrary. Either the stop distance must be standardized everywhere, or the observed stop driving times must be proportional to the planned time for successful aggregation. Therefore, partial aggregation of the time dimension was used as one option to study the 3D-aggregation.

From the mere graph of randomness (Figure 9.), one can clearly see the problem of punctuality in rush time and evening traffic. The problem is caused by vehicle tasks during those times and this is likely to have a direct impact on the identification of other phenomena. In Figure 9. it can also be seen that the problems with the schedule structure are also dependent on direction.

The clearest of all the 3D-graphs of punctuality is the cancellation-phenomena as can be seen in the Figure 10. The data of cancelled departures is shown in the graphs as missing points. The data had not been pre-processed at this stage, which would give the missing data some value for calculation. Had the pre-treatment been done, the more likely the cancellation-phenomena would still have been clearly noticeable. The choice of pre-treatment method is of great importance in this matter. In terms of cancellation detection, unprocessed data at the 3D-aggregation level produces the desired result. As such, the graph does not even need information about punctuality, but only information about whether the departures have been cancelled. However, it would be good to produce a separate indicator for cancellation, so that it could be observed from the most suitable graph for it.

Table 8. Functionality of 3D-level graph options with different phenomena.

		Selected aggregation level	
Time – Date – Value	**Location reduction**	**Partial location and time reduction**	**Removing Bias**
Rush	Partly distinguishable	Distinguishable	Distinguishable
Delay	Not indistinguishable	Partly distinguishable	Weakly distinguishable
Bunching	Not indistinguishable	Distinguishable	Partly distinguishable
Cancellation	Distinguishable	Not indistinguishable	Not indistinguishable
Combination	Partly distinguishable	Distinguishable, risk of saturation	Distinguishable

Other modelled phenomena were not observable without partial aggregation of the time-dimension. Dividing the route into segments also improves to locating the punctual problems. The removal of the bias clearly highlights the place of the problems in graphs. Therefore, it is recommended to consider the disturbances due to the schedule structure in the calculation of the indicator. It is identified that it causes significant effects on the results. The downside here is that the calculation of indicator is becoming more laborious and need more pre-treatment of the data. In Figure 10. there is a case where the bias has been taken into account and it is clear from it when there have been problems with punctuality. But for example, it is difficult to distinguish between rush and delay if they are both present at the same time. Table 8. collected estimates of the functionality of different 3D-aggregation levels to describe different phenomena as well as the aggregation level verified to work best.

For the 2D-aggregation, a simple line graph was created. In created graph the events of one day are grouped on a single coloured line. In Figure 11. is an example of a created graph in the case of a rush-phenomenon. The graph is separated into punctuation information for arrival and departure punctuality at the timing stop and at the stop delimiting the segment. For runtime planning, it may also be a good idea to create the aggregation level by segment and calculate intra-segment values, such as total segment punctuality, change in segment punctuality, and the effect of timing point on punctuality.

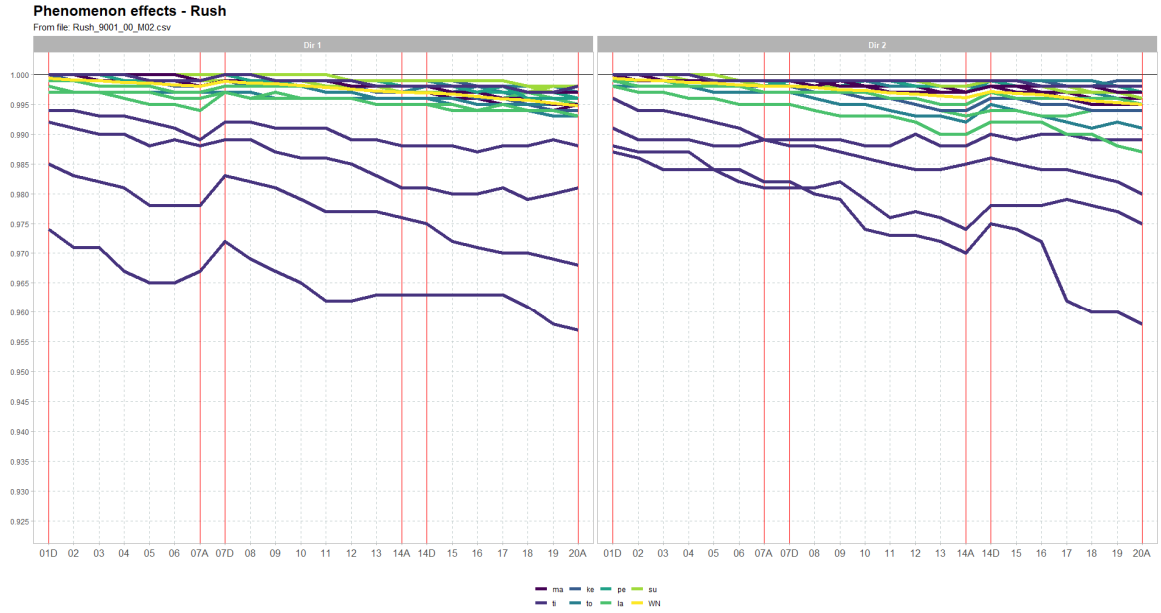


Figure 11. Punctuality location-value graph with timing point separation in each day type.

By distinguishing between arrival and departure punctuality, the effect of improving the punctuality of the timing point is clearly demonstrated at the aggregation level. In addition, the effect of the location of the timing point emerges in the first half, the timing point on the route has a more punctual-improving effect than in the second half. Thus, when calculating the average punctuality, the timing point has the most probable effect of raising the average. In the calculating of the average punctuality, it is necessary to consider how the effect of the timing points is taken into account, so that the punctuality does not give a wrong picture of the situation.

Identifying and distinguishing between different phenomena themselves is difficult at the 2D-aggregation level. The dates of events can be distinguished, but the reason behind them remains obscure. In some cases, different phenomena can interfere each other out and the result is a good punctuality value, even if there are serious problems. For example, in every day repeating bunching, saturation also occurs. As a result of which the cause of a poor punctuality result cannot be distinguished from a systematic error. This saturation is aided by data calculated on a bias, which can be computationally removed from all data or described on the same graph as the observations. Describing into the same graph provides information on what can be achieved with this schedule structure in case of punctuality.

4.2.3 Regularity

As with punctuality, the aim was to find two different levels of aggregation for the regularity indicator: 3D and 2D levels. The 3D-aggregation level of these included the time, location, and value-dimensions, and the 2D-aggregation level included the time and value-dimensions. At the 3D-aggregation level, three different aggregation levels were tested to present the results of the indicator. These levels were defined as the aggregation of the data-dimension over the month, week, and day type. During the analysis, it was found that the results of the regularity indicator and their changes are very small. This small size of values is due to the definition of the indicator itself. As a result of the small values of the indicator, there is a requirement for the graphs to be able to have higher accuracies. The accuracy requirement could be achieved, for example, by calculating the relative change or by fitting the colour scale to a smaller value range. In the calculation of the indicator, a deficiency was also observed in the headway calculated at the changeover point of the day. This shortcoming did not affect the outcome of the calculated indicator, when using the data used in this work, but it needs to be corrected in the calculation later.

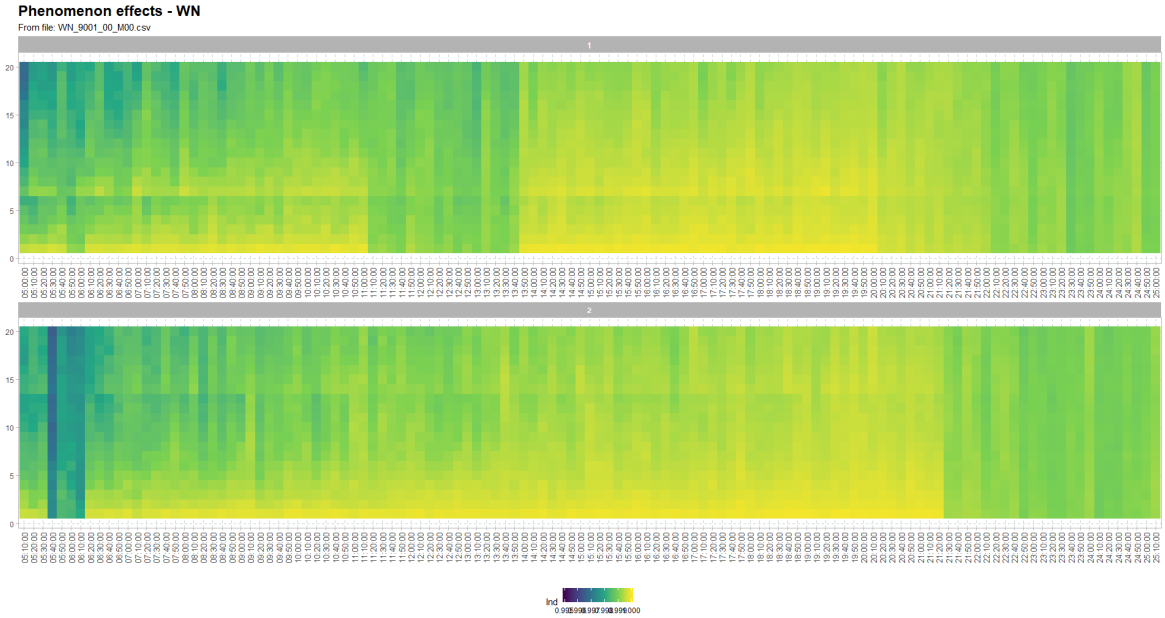


Figure 12. The effect of randomness in location-time graph.

Unlike in the case of punctuality, the value of the indicator is shown at all stops. And since the aggregation is done in terms of the date-dimension, calculating the headway is not a problem. The location of the timing points along the route was also found to have a similar effect on regularity as on punctuality. Although the levels of data aggregation and indicators are different, the impact is still very visible. When aggregating the location-dimension, the effect of the timing points should be taken into account, as it can distort the outcome and misrepresent the conclusions made to justify it. The calculation must also consider the headways that could, for example, variate between the line or between different days and different times. It would therefore be advisable to relate the observed headway to the scheduled value. In this case the results of the calculation can be used for comparison between different line and schedule structure. If the headway remains constant over the entire data, as in the data produced for this analysis, the proportioning headways not necessarily need to be used.

The aggregation of the regularity indicator shows the same problems due to the schedule structure as in punctuality. In Figure 12. the effect of randomness is presented in the monthly summary. For example, it can be seen from the figure that the regularity is poor before the start of the morning rush hour, especially in the case of direction two. This poor result of the indicator is exacerbated with different phenomena and makes it difficult to identify the problems caused by the phenomena. One general consideration of the results of the indicator's aggregate levels is that adding interference of randomness appears to improve regularity in random situations. This may be due to natural variation caused by randomness or a possible correlation between multiple phenomena.

There was already a problem with saturation of punctuality indicator's aggregation, but there was also a significant problem with regularity. For the aggregation levels of the regularity indicator, the risk of saturation, especially for the recurring problem, is very high. The result of saturation risk is an unreliability of the indicator results. In Figure 13. is an example of the disappearance of phenomena caused by saturation. The figure should clearly show the problem caused by bunching, for example. When the Figure 13. is compared with the Figure 12., a small difference can be observed between the figures, for example towards the end of rush time. But, for example, one of the worse indicator results during the noon cannot be said to be certain without reference data. For each level of aggregation, it would be a good idea to include information on irregularity due to mere randomness. This helps to compare what level of regularity is practically achievable at the operational level. It is also possible to correct the saturation if e.g. only one day of pure randomness or only scheduled events are included to the calculation.

When data is aggregated over a month, the effect of individual problematic days disappears, except in the case of bunching. In the case of a single day, it is easy to confuse the observation with the noise caused by randomness. As the number of event days increases, the indicator begins to detect changes caused by other phenomena as well. When aggregating data for the weekly level, one-week data was tested because the data, which modelling method produced, does not see differences between the weeks. The only difference that arises between different weeks is the proportion of phenomenon days out of all days. In a weekly aggregation, the effect of rush on a single day can be most clearly detected by comparing the situation to the situation which only contains randomness. Saturation also appeared to be lower in the weekly level than in the monthly level. The reason for this is the smaller amount of data and the calculation method of aggregation. In Table 9. the functionality of different 3D-aggregation levels to detect different phenomena has been collected. In general, it can be said that aggregation at the weekly level works very well and the indicator is very sensitive to individual anomalous observations. Therefore, it was chosen as the aggregate level of the regularity indicator. Of course, there are even more precise and appropriate levels of aggregation for the observation of individual phenomena.

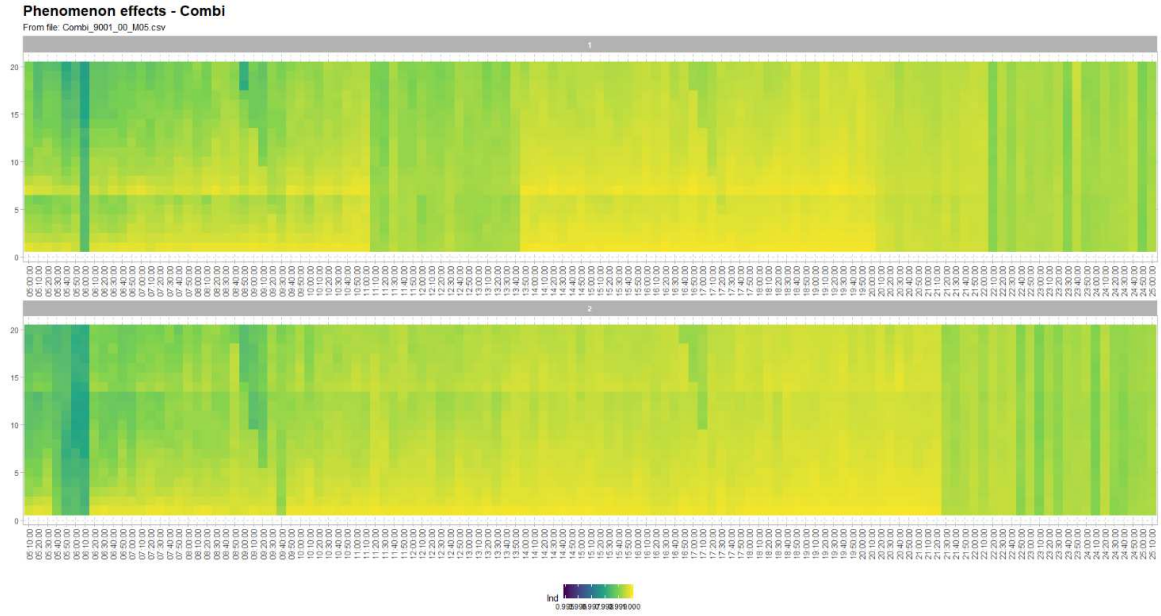


Figure 13. The combined effect of all phenomena in location-time graph for regularity.

In the analysis it was found that a single daily figure is not very useful for selected indicator. Sure, it tells you how one day has gone, but a comparison with other days gives a better picture of whether that day has been particularly problematic. About the delay, it was also noted that the phenomenon is not clearly visible in the regularity indicator or in its aggregates. Cancellation-phenomena is not displayed at the selected aggregate levels because aggregates view events over multiple days and a cancelled departure is some random departure from some random day. However, the visibility of cancellation can be affected by the right kind of data pre-processing. In Figure 14. a graph model made for the 2D-aggregate level is shown. The delay phenomenon is not clearly visible even in a 2D-aggregation. Since all departures are regularly approximately as late, the indicator does not see any problem in traffic. The indicator must be aggregated over another dimension for the delay to be visible. At the 2D-aggregate level, it is also difficult to detect individual days of phenomena. It is a good idea to add a day of mere randomness to the figure. This makes it possible to outline which result of the indicator can be achieved with the planned vehicle tasks.

As the regularity indicator was intended to be best detecting bunching, it was looked at a little more closely than other phenomena. Bunching occurred at all aggregation levels as a clear ripple, and the phenomenon was already noticeable as a one-day problem when compiled over the entire month. This was due to the chaining modelling method, in which a constant proportion of outputs were shifted to late departure. In reality, bunching is not so evident unless there are problems with one and the same outputs. It must also be borne in mind that, as a recurring event, saturation becomes a problem in the aggregation. Active traffic control could also have been modelled for the bunching phenomenon. Traffic control can affect the headway and its evenness. Therefore, it would have been interesting to know whether the indicator can see the impact of the guidance.

Table 9. Functionality of 3D-level graph options with different phenomena.

		Selected aggregation level	
Seq – Time – Value	**Month**	**Week**	**Day type**
Rush	Distinguishable	Distinguishable, risk of saturation	Partly distinguishable, risk of saturation
Delay	Partly distinguishable, risk of saturation	Distinguishable, risk of saturation	Not indistinguishable
Bunching	Distinguishable, risk of saturation	Distinguishable, risk of saturation	Partly distinguishable, risk of saturation
Cancellation	Not indistinguishable	Not indistinguishable	Not indistinguishable
Combination	Distinguishable, risk of saturation	Distinguishable, risk of saturation	Not indistinguishable

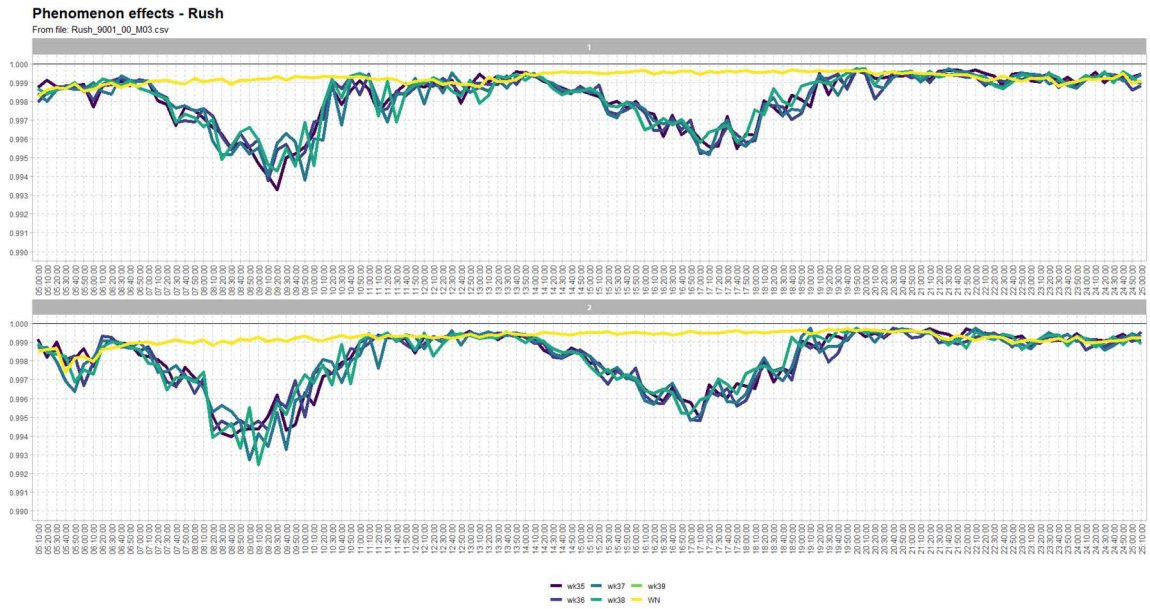


Figure 14. Time-value graph for the regularity indicator in case of rush with randomness (WN).

4.2.4 Wait time

The aim was to produce 3D and 2D-aggregate levels for the waiting time indicator. The 3D-aggregation level included the date, location, and value-dimensions, and the 2D-aggregation level contained the date and value-dimensions, respectively. For the 3D-aggregation level, three different options were defined as compiling the entire time-dimension, dividing the time-dimension into hour groups, and dividing the time-dimension into hour groups by filtering out by punctuality. Scheduled headway is a constraint on hourly groupings. There

should not be too few observations in one group, as this will affect the calculation. In the third 3D-aggregation level option, the punctuality filtering sought to highlight problematic days and their impact on the value of the indicator.

The indicator can be characterized as function based on aggregates, i.e. it combines a larger amount of data. Thus, problems in produced aggregate levels do not easily arise. The aggregate level of the indicator would also need average number to aid interpretation. The result of the indicator itself only describes the variance. In the analysis was found, that the number of observations used in the calculation is a very useful additional information. The small number of observations affects the result of the indicator, making it unreliable, because the quantiles used in the calculation must be approximated with a small number of observations. It is therefore worth paying attention to the amount of data used. One clear difference between wait time and previous, the punctuality and regularity, aggregations is the lack of effect of the schedule structure on the indicator aggregations. It is perfectly logical, as the aggregation of the wait time is done over one day, at which point the schedules of the day and the problems caused by its structure disappear when it is done. There is still a clear variability due to randomness between individual days.

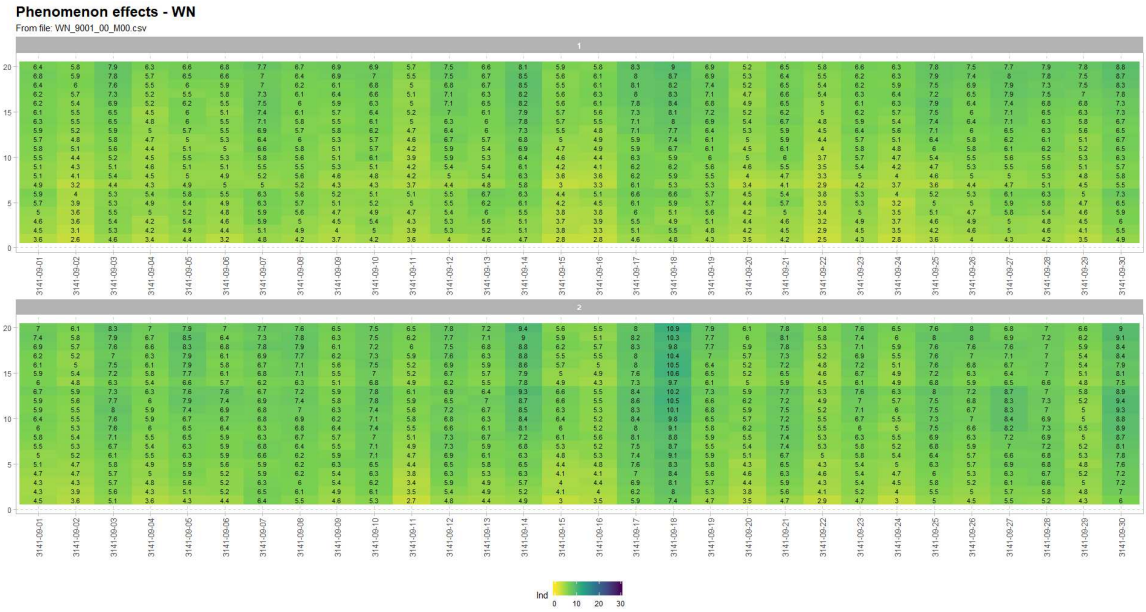


Figure 15. The effect of randomness in the location-date graph with numerical value.

The analysis of the wait time indicator in the aggregate levels revealed the need to relate the value of the indicator to the headway. Then the comparison of alternating lines with the indicator would be successful. Since the indicator is a compilation of data, it can have a problem with headway and headway ratio. Either all observations or the result of the indicator should be related to the average scheduled headway. Proportioning could also be done to the observed waiting time.

A recurring phenomenon may saturate the result of the aggregate, as in the case of regularity. To reduce interpretability and saturation, a numerical indicator value was added to the graphs to represent. In Figure 15. the effect of mere randomness on the result of the indicator is

presented, which includes the value of an individual observation as a numeric value. The values produced by the aggregates for the indicator are in the order of about five minutes. This is half the headway and corresponds approximately to what the average waiting time produces as a result. Another way to reduce saturation is to add one day, which includes only randomness in the data. In this case, the comparative information is included in the aggregation and no other figure, such as the average waiting time, is required. The data produced by randomness were found to be really useful additional information for all indicators. The usefulness and production of such additional information should therefore be considered at previous levels of analysis.

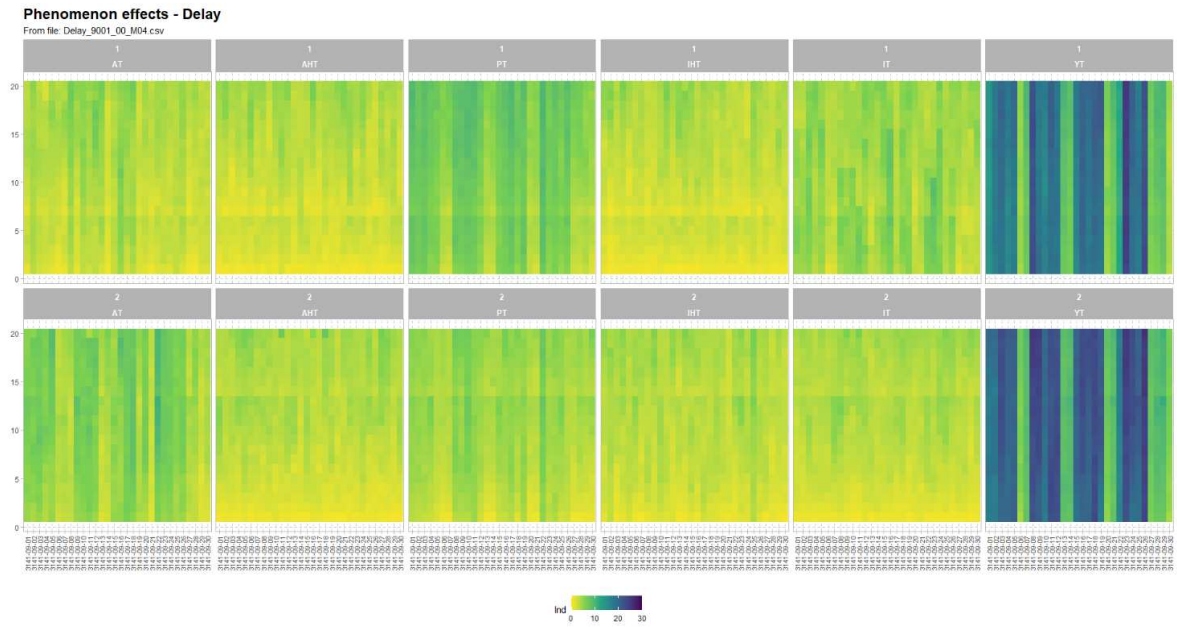


Figure 16. The effect of delay on the wait time indicator result in location-date graph with hour groups.

Aggregating the date-level also works well for a single stop. On the other hand, for example, dividing a day into rush time and day traffic can provide additional information about the average wait time for a day and the differences within that day for each stop. For example, the disturbance that occurs on all days is reflected in the hourly division. This is because in the morning hour group the phenomenon is not yet so intense, and the phenomena effects had not cumulated yet. In other words, the morning hour group serves as a comparison group. This supports the idea that the interpretation of the results of the indicator should always be done on some reference data. Without it, it is still very difficult to distinguish between individual phenomena with a mere daily aggregation.

The grouping of daily information into hours was done by dividing the events into morning (AT), morning rush (AHT), day (PT), evening rush (IHT), evening (IT), and night (YT) groups. In Figure 16. an example of an aggregation made for this hourly division is given. When aggregating, it was found that presenting one month in hourly groups at the aggregate level is the upper limit for the number of days to present. With larger numbers of days, the graph can become too confusing and therefore require additional aggregation or filtering over the days. By dividing the hours into groups e.g., the effect of rush on the aggregation

can be revealed. Disturbances due to the structure of the schedule are also shown, as in Figure 16. can be seen especially in the last hour group of the night.

In the third aggregation option for wait time in 3D-aggregation level, filtering was used to separate the data according to punctuality. The aim of the filtering was to reveal the effect of different punctualities on the indicator. It is also possible to make the division according to regularity. For various phenomena, filtration did not yield a high-quality aggregate and therefore the aggregate level was found to be ineffective. The most effective aggregate level was the daily aggregate without hourly grouping, and it was chosen to be wait time indicator's aggregation level. This was decided because of the aggregating nature of the indicator itself, but also to ensure that the amount of input data needed in the calculation is enough. In Table 10. an evaluation of the functionality of different 3D-aggregation levels with different phenomena has been compiled.

Table 10. Functionality of 3D-level graph options for different phenomena.

	Selected aggregation level		
Seq – Date – Value	**Day**	**Hour group**	**Hour group, unpunctual**
Bunching	Partly distinguishable	Partly distinguishable	Not indistinguishable
Cancellation	Not indistinguishable	Not indistinguishable	Not indistinguishable
Combination	Distinguishable	Partly distinguishable, risk of saturation	Not indistinguishable
Delay	Distinguishable	Partly distinguishable, risk of saturation	Partly distinguishable, risk of saturation
Rush	Distinguishable	Distinguishable, risk of saturation	Distinguishable, risk of saturation

For individual phenomena, the rush appeared to increase the value of the indicator by about a headway based on the aggregates. This effect appears to be repeated regardless of the number of days of disruption. Correspondingly, the effect of the delay-phenomenon turned out to be slightly larger. Separating the rush and delay phenomena is tricky when using time-aggregation. The phenomena stood out, but their effect was very close to the values caused by mere randomness. In addition, in the hour group, the effect of the delay cannot be clearly distinguished without comparative data.

The effect of bunching on the aggregation appears to be small. The result of the indicator is likely to be affected by the quantiles used in the calculation. The indicator may not be correct to detect bunching. Like case of bunching phenomenon, cancellation is not clearly reflected in the aggregations of indicator. The need to produce a separate indicator for cancellation was identified, with which its effects on waiting time, for example, could be studied. The definition of this indicator is limited beyond this work. When making a 2D-aggregation, the feature that aggregates the information of the indicator also appeared as a problem. Often, in aggregations, phenomena cannot be clearly distinguished without the aggregate of other

dimensions. Therefore, the aggregation was done in part with both date and time-dimensions. In Figure 17. a partial additional aggregation over a time-dimension for a 2D-aggregation is shown. The value of the indicator, calculated using randomness alone, has also been added to the figure as reference material.

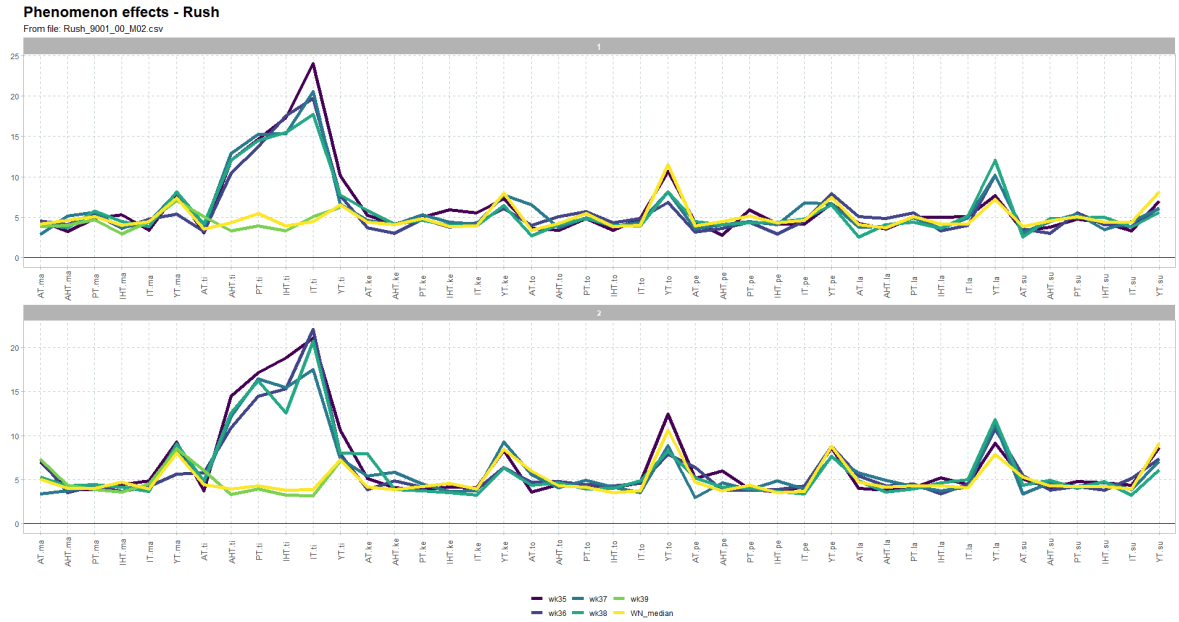


Figure 17. Date-value graph for the wait time indicator in case of rush-phenomena with randomness (WN).

4.3 Example case

The graphs of the punctuality indicator highlight identification problems at certain stops, as well as phenomena due to changes in identification logic. As shown in Figure 18., the combined line's punctuality problems focus on morning and evening rush hours. The reason is probably the large number of passengers, but other traffic (in rush time) can also affect. The reason can be deducted by the improvement in punctuality over time, observable the same figure, due to the reduction in traffic because of the current global situation. The unpunctuality remains visible and this is probably a structural unpunctuality. Spatially, the unpunctuality focused on the common section of lines as the indicator shows.

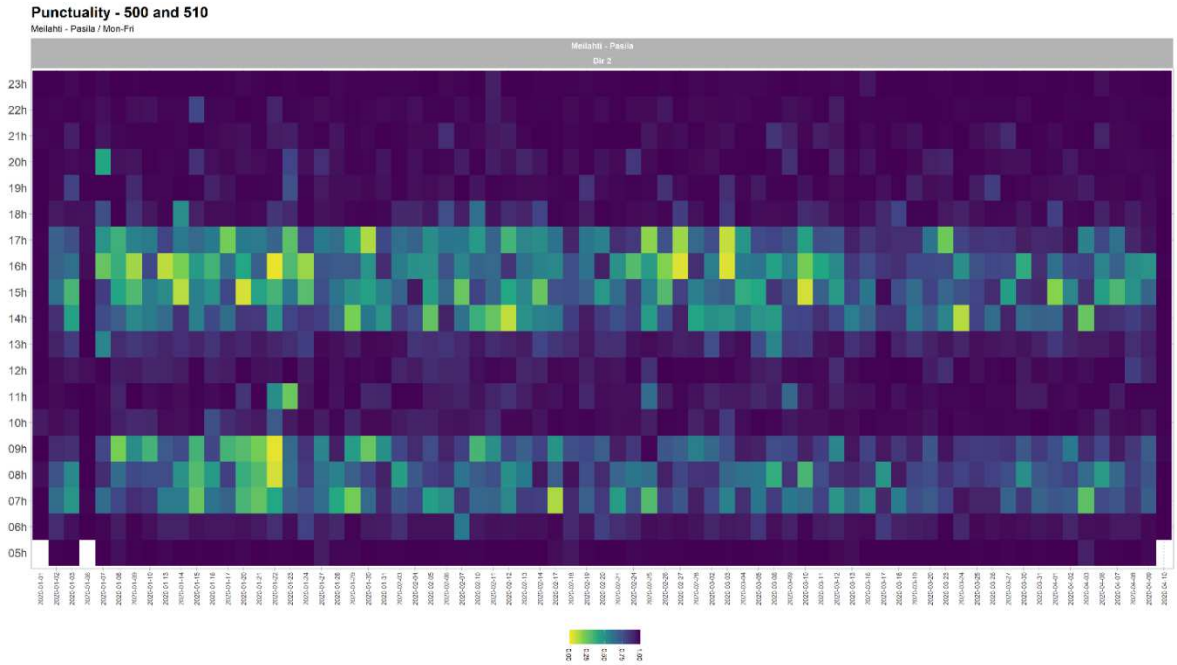


Figure 18. Punctuality of 500 and 510 in time-date graph.

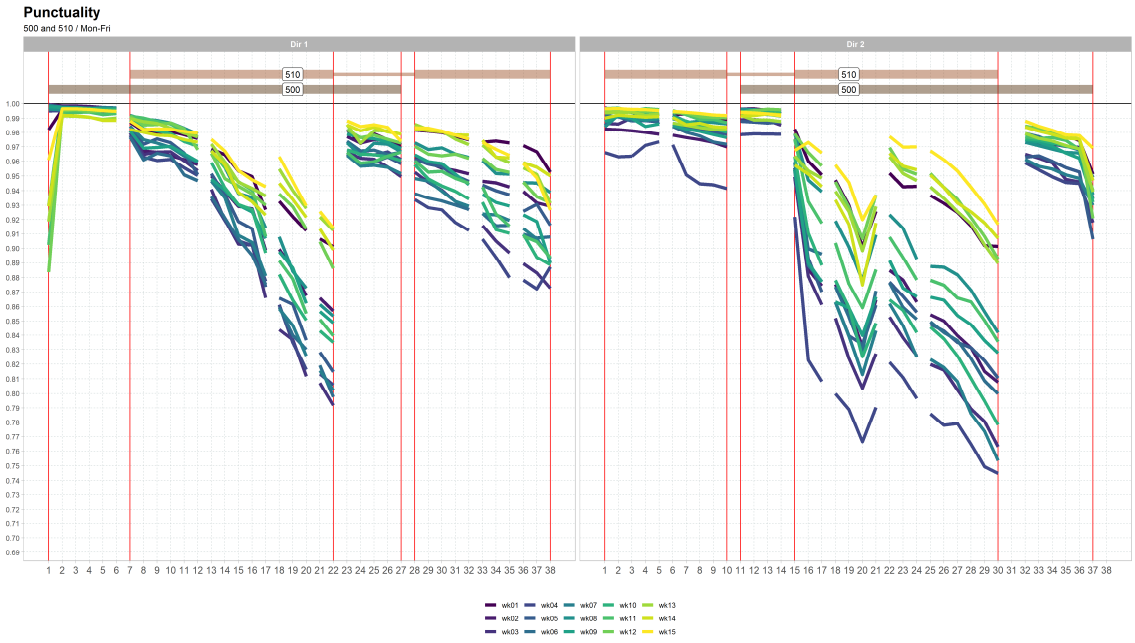


Figure 19. Common punctuality of lines 500 and 510 with segments in location-value graph.

In the interpretation, it must be considered that the headway is used as a divisor in the calculation of the punctuality indicator, and this affects the varying headways of sections. Direction two appears to be a direction with more punctuality problems. The line 510 has its own section in this direction, with construction sites causing disruption. This is precisely why the unpunctuality caused by the disturbance are highlighted by the common section. There is a clear linearity in the stop-specific change in punctuality in certain segments. This

linearity can be seen in Figure 19., which also shows the average stop-specific accuracy for different weeks. In the Figure 19., the precision lines are divided into segments, which would make the figure useful in planning schedules of lines as well.

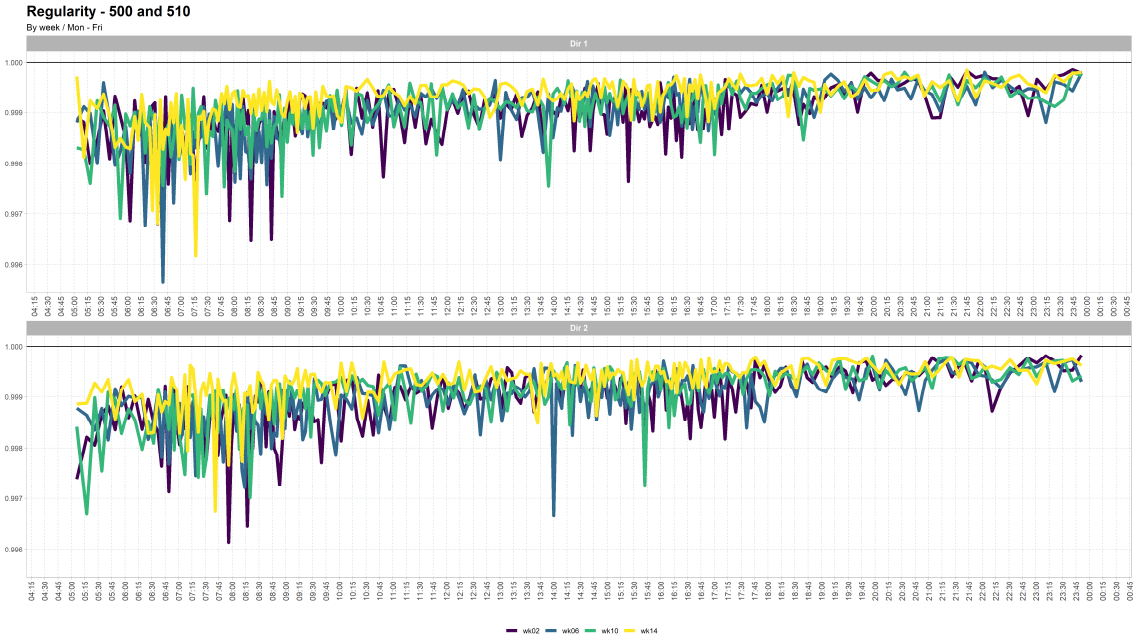


Figure 20. Common regularity of lines 500 and 510 in time-value graph.

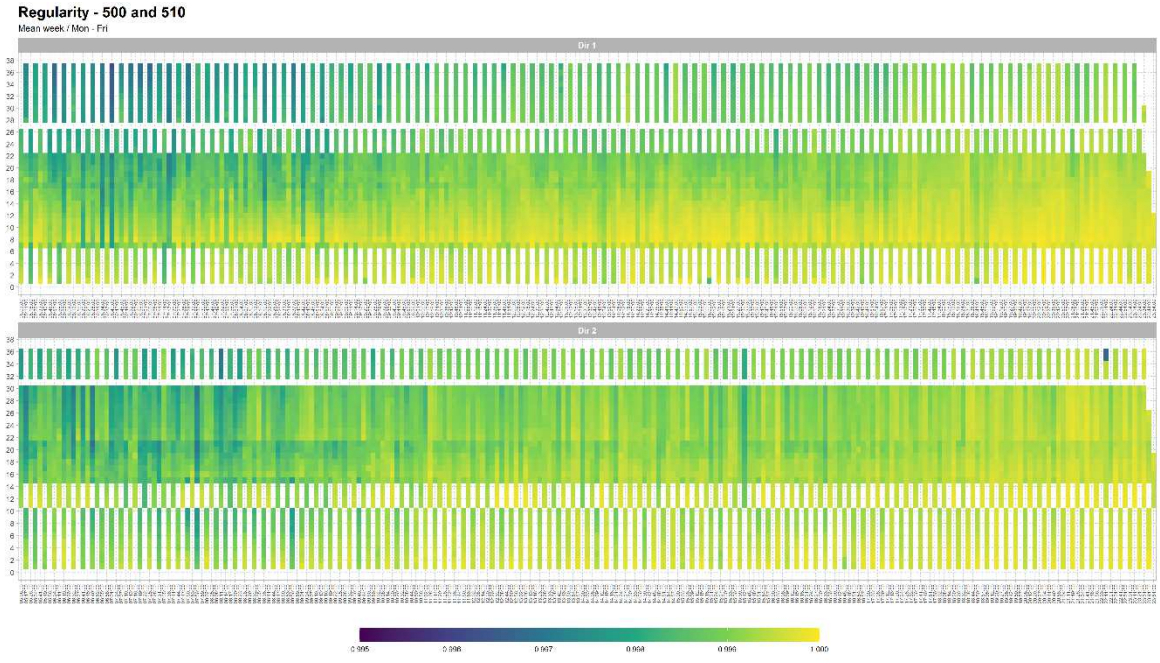


Figure 21. Regularity of lines 500 and 510 in location-time graph.

The graphs of regularity clearly show the bunching on the lines and this is strongest in direction two. The reason for this is the same as for punctuality problems in this direction: a longer own section for line 510 with construction sites. In Figure 20. it can also be seen that the regularity is worse at the beginning of the traffic than in the evening traffic, for example. The reason for the poor regularity at the very beginning of the traffic may be in the operation of the operator and this should be examined in more detail. Bunching also seems to occur in general, as the average regularity over the entire data shows the same bouncing behavior of the indicator. In Figure 21. the common regularity of the lines of the average weekday is presented. Bunching also appears to occur in the same manner as modeled in spatio-temporal analysis.

In Figure 22. it can be seen that the waiting time indicator is, on average, at its best in the early morning, although the least dispersion between different weeks is in daytime traffic. The figure can also be used to outline waiting times develop during the day in direction two: until the end of the evening rush hour, the time increases steadily, after which it starts to decrease. In direction one, a similar uniform change is not visible from the graph. There is not much difference in the normal situation between different weekdays, but as the situation changes, the spatial behaviour of the waiting time also changes, as Figure 22. shows. Starting a second line along the route of the first line affects the waiting time by reducing it. Unlike the problems of regularity and punctuality, the waiting time seems to be worse in the direction of one. From Figure 23. it can be seen how the waiting time increases steadily towards the end of the direction and under normal circumstances even the timing stop does not affect the result of the indicator.

Figure 22. Common wait time in date-value graph.

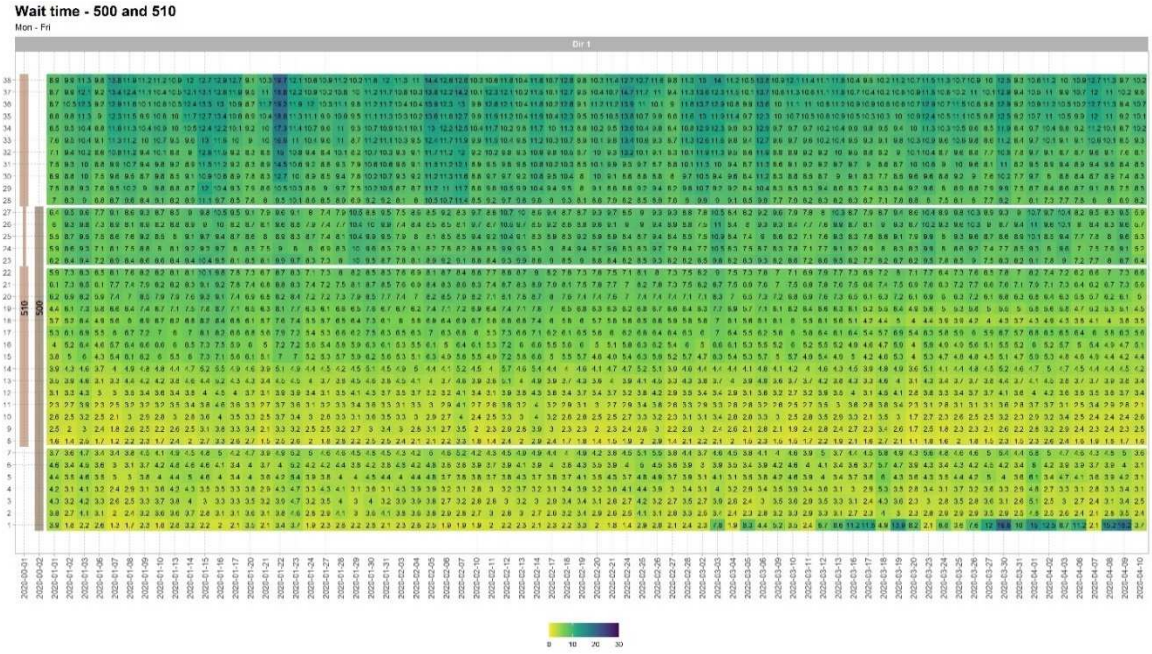


Figure 23. Wait time of the lines 500 and 510 in location-date graph.

In addition to the exemplar analysis of the lines, it is a good idea to include line-specific information in order to distinguish the line that is causing the problems. On the other hand, when lines are planned together, it is also justified to study them together. One point to be also highlight is the need to standardize the headway in different parts of the route into one and the same number, especially in terms of punctuality. The same headway between different parts of the route facilitates and clarifies the interpretation of the indicator. The effect of the timing stop is visible in all indicator graphs, and the magnitude of its effect can also be found out. As clearly seen in the modeling, the strength of the effect depends on where on the route the stop is located. The case study also highlighted the need for other aggregation levels for indicators in order to better understand the causes of the various problems. Thus, it can be said that spatio-temporal analysis provided information on the required aggregation levels. When working with graphs for the case study, additional needs for graphs also arose. These additional needs were applied in the analysis and can be seen, for example, by comparing the graphs outlined in section 4.2 Spatio-temporal analysis with the graphs made for the exemplar study.

5 Discussion and conclusion

5.1 Summary of results

Overall, the outlined performance metric selection process worked as expected. There is a lot of minor correction in the details of the process and the process needs some clarification in some points. The use of the literature to determine alternative indicators provided a good basis for the early stages of the process. The literature review was also used as thought-provoking and to find the limitations of alternative indicators. In determining the components of reliability, the literature was a good helpful tool.

The AHP used to create evaluation framework and evaluate the indicators worked as planned and can be easily extended to other similar needs in the selection process. It was a good thing to separate technical requirements from user needs, and requirements and create an objective and subjective evaluation framework. With the consistency ratio, the evaluators were able to follow the success of subjective evaluation clearly and without affecting the outcome of the evaluation itself. The AHP is also good tool to define criteria for evaluation. Use of participants to define those criteria shows that different participant groups have their own opinion what is good and useful indicator.

Based on the feedback from the workshop's participants, the workshop itself and the involvement of participants were considered successful. The feedback also provides development ideas for workshop. For example, evaluation form should be simplified, and indicator alternatives should be explored in advance. The way in which the indicators were evaluated led to a partial failure in the workshop. A comparison of the criteria proved to be the main reason for the failure of the workshop and the evaluation. That is a reason why result of the workshop is not as reliable as it should be. And therefore, the results of the workshop were not used in this work.

A stepwise model of spatio-temporal analysis achieved a functional description of the analysis process. The analysis revealed the multidimensionality of the used input data. Within the framework of the analysis, it was also possible to find out the main dimensions of the data, based on which aggregates can be made as possible graphs. The dimension reduction, in turn, revealed the constraints imposed by the indicator on the calculation. The spatio-temporal analysis also noted clearly how the systematic error in the results of the indicators created by the planning of the schedules can be seen. As the example analysis show, the systematic error can be considered in two different way. Error could be removed from indicators data by modelling it or displaying error as a benchmark graph.

Spatio-temporal analysis shows how important is define the pre-process of the input data. In the pre-processing phase, certain phenomena can be taken into account or can be removed from data. In this work, cancellation was that kind of phenomena. Input data could be reformatted in the way that indicators are observed clearly, or it could be completely removed. If it was removed from data, it requires its own indicator.

In the example case, the results of spatio-temporal analysis were put into a test. The example case shows that outlines of indicator graphs work fine. The graphs need refinements for some indicators, but overall the graphs can be used in further analysis of reliability. Example case also shows that all individual indicator needs all graph options that was defined in spatio-temporal analysis but was excluded from the scope of this work.

5.2 Discussion of results

The literature review shows that there are several options for indicators. In some cases, there are several different options, but only a few in other cases. On the other hand, those suggested options are very similar and often there are only minor differences between the alternatives. This limited number of suggested options can be seen in two ways: either those few options are most suitable for the reliability indicators or the reliability component has not been the focus of the studies. Either way, there should be a few options for each component that look at it from different perspectives. By offering indicators which have different point of view, the needs and requirements of the user can be met more precisely. Different alternatives for indicators can be achieved also in workshop by using participant knowledge. Defining options can be own separate part in workshop where participant can create own indicators alternatives.

A minor thought that rise from the workshop and reviewed literature not pointing out clearly is how suitable an indicator is for different modes. In general, different modes are planned in different way and they operate in their own way in day-to-day level. These differences create needs that the selected indicator should meet and can limit where the selected indicator is useful. If the purpose of the indicator is showing a general overview of transit network, different modes create major problems and limitations in results interpretation. To avoid the problem, mode sensitive indicators can be chosen, but it can increase number of selected indicators. How the different modes are handled should be decided in early on.

Other considerations, that emerged during defining the indicator alternatives, are the scaling of the indicator results to a standard scale. This does not mean that the different indicators should have same results value, but it means that the best possible result's, for example, valued to one. Also, not all indicators are limited to values between 0 and 1. In terms of comparability of different indicators and comprehensibility of results, indicators using the same scale are better. An exception to the need for scaling are indicators that use time for their results, for which scaling to a constant scale does not make sense. If all indicators use the same scale, in users' point of view it is easier to make interpretation of the results in case of different indicators.

There are multiple issues in workshop that had to be paid attentions. At first, evaluation scale and method should be as simple as possible for the workshop. Evaluating just one indicator in each criterion alone and assess how well the indicator fits within the framework of the criterion. The evaluation is done on a simple scale, for example from one to ten. And the result of evaluation should be reformatted into an original form of AHP-method. This way the evaluation and evaluation form can be simplified. There may be some limitation in reformatting phase, and it need further investigation and thoughts.

Familiarization with the indicator alternatives is likely to require more thought and redesign. Earlier user involvement solves some of this problem, but evaluators still need more time and background material to do the evaluation. If participants is involved in earlier stage in indicator's defining process, they can also provide knowledge and ideas for indicator alternatives. Involving participant in the alternative definition phase opens new opportunities. As the participant defines the alternatives themselves, other alternatives found in the literature

can be discussed at the same time. In this case, the alternatives from the literature will also become more familiar.

Some attention should be paid to how many alternative indicators are chosen in evaluation and how much earlier involvement may affect subsequent evaluation. A fewer alternatives makes it easier to evaluate and familiarize with them. Presenting a smaller number of indicators in the evaluation phase requires dividing the alternatives into different categories, such as qualitative and quantitative indicators. In this case, one or two options for each category are selected to proceed into evaluation phase. Each indicator selected for evaluation would be associated with one category and highlight more strongly the differences between the indicator alternatives. Other small thing, which was observed during the workshop, is how much other participants can affect the opinion and thus the outcome of the evaluation. To avoid the effect, it is better use different participants in different phases of workshop if possible.

The spatio-temporal analysis itself proved to be a teaching method. With the analysis, it is easy to find limitation of data in indicator's results calculations. It is very rare that data is not corrupted, distorted or part of it is missing. So, it is good thing that analysis can point out how the calculation is limited by the data. The limitations of the indicators could also be found in same time as data limitations. By defining primary dimensions, it is clear to find which graphs are possible and in which cases further aggregations are needed. It can be said that spatio-temporal analysis is one way to get acquainted with indicators and data.

In this work, spatio-temporal analysis show that there are multiple places in analysis process where certain things are made repeatedly. Repeating same work in multiple times doesn´t necessarily add new value to the analysis. The analysis step model should therefore be refined and supplemented to eliminate possible recurrences and speed up the analysis process itself. Some of the repeating cases can be avoided by defining and using standards. In other cases, the analysis should be done a few times for additional information and learning. With additional knowledge of the analysis, some correlation between indicators, indicators limitations and initial setups of the analysis can be found. Assuming that correlation happens in same way, some analysis phase can be omitted.

The phenomena used in the spatio-temporal analysis could also be made available to the evaluators in order to illustrate the reactivity of the indicator alternatives to different phenomena. It is also a good idea to define modelled phenomena with the users. With this way the most important phenomena of users' point of view can be modelled. Also, most sensitive indicators and aggregation levels can be found for those phenomena. The need of the separate indicator for phenomena can be determined with help of spatio-temporal analysis. If defining phenomena is include in workshop, the workshop should be divided in two parts: workshop for defining initial setups and evaluation workshop.

One crucial factor influencing the value of the indicator, which was not demonstrated in the literature review and which emerged during the spatio-temporal analysis, is the systematic errors caused by the planning process. Because even different modes can have different planning processes, it should be included at least in indicator results' calculation phase. As previous chapter mentions there are two ways to consider this: by removing it from calculation or showing effects of planning as a benchmark. It is recommended that effects of planning in indicators are modelled, and if needed, some adjustments are made. It is also good thing

to monitor systematic errors from planning as monitoring reliability of produced transport. Monitoring the planning results may need their own performance indicators.

5.3 Future development

There are few main developments in the selection process of performance indicators. First, the question is whether there will be participants in determining the indicator alternatives and defining of modelled phenomena. These two things define the final structure of workshop and if the workshop should be kept in two completely separated parts. Also reformatting evaluation form into more user-oriented mode is needed before next workshop. Evaluation framework works as intended, and it doesn´t need any revisiting at this point. In spatio-temporal analysis there is some minor development to make it more easily accomplished. Process description of key performance indicators' selection has to be made, before the selection process is used with different participant group.

References

Amin-Naseri, M. R., & Baradaran, V. (2015). Accurate estimation of average waiting time in public transportation systems. *Transportation Science*, 49(2), 213-222.

Barabino, B., Di Francesco, M., & Mozzoni, S. (2015). Rethinking bus punctuality by integrating Automatic Vehicle Location data and passenger patterns. *Transportation Research Part A: Policy and Practice*, 75, 84-95.

Barabino, B., Di Francesco, M., & Mozzoni, S. (2016). An offline framework for the diagnosis of time reliability by automatic vehicle location data. *IEEE Transactions on Intelligent Transportation Systems*, *18*(3), 583-594.

Cats, O. (2014). Regularity-driven bus operation: Principles, implementation and business models. *Transport Policy*, *36*, 223-230.

Cham, L. C. (2006). *Understanding bus service reliability: a practical framework using AVL/APC data* (Doctoral dissertation, Massachusetts Institute of Technology).

Chen, X., Yu, L., Zhang, Y., & Guo, J. (2009). Analyzing urban bus service reliability at the stop, route, and network levels. *Transportation research part A: policy and practice*, *43*(8), 722-734.

Diana, M., & Daraio, C. (2010). Performance indicators for urban public transport systems with a focus on transport policy effectiveness issues. In *World Conference on Transport Research,(WCTR, Lisboa 2010). Paper. Lisboa, World Conference on Transport Research*.

Dziekan, K., & Kottenhoff, K. (2007). Dynamic at-stop real-time information displays for public transport: effects on customers. *Transportation Research Part A: Policy and Practice*, *41*(6), 489-501.

Firew, T. (2016). Analysis of service reliability of public transportation in the Helsinki capital region: The case of bus line 550. Master thesis. Aalto-university. Espoo. 107.

Furth, P. G., & Muller, T. H. (2006). Service reliability and hidden waiting time: Insights from automatic vehicle location data. *Transportation Research Record*, *1955*(1), 79-87.

Golshani, F. (1983). System regularity and overtaking rules in bus services. *Journal of the Operational Research Society*, *34*(7), 591-597.

Henderson, G., Kwong, P., & Adkins, H. (1991). Regularity indices for evaluating transit performance. *Transportation Research Record*, *1297*, 3-9.

Hess, D. B., Brown, J., & Shoup, D. (2004). Waiting for the bus. *Journal of Public Transportation*, *7*(4), 4.

HSL (Helsingin seudun liikenne) (2020). *HSL Reittiloki*. Viewed 2020-04-17. https://reittiloki.hsl.fi.

Kho, S. Y., Park, J. S., Kim, Y. H., & Kim, E. H. (2005). A development of punctuality index for bus operation. *Journal of the Eastern Asia Society for Transportation Studies*, *6*, 492-504.

KFH Group. (2013). Transit capacity and quality of service manual.

Ruan, M. (2009). Probability-based bus headway regularity measure. *IET intelligent transport systems*, *3*(4), 400-408.

Liu, R., & Sinha, S. (2007). Modelling urban bus service and passenger reliability.

McLeod, F. (2007). Estimating bus passenger waiting times from incomplete bus arrivals data. *Journal of the Operational Research Society*, *58*(11), 1518-1525.

Mishalani, R. G., McCord, M. M., & Wirtz, J. (2006). Passenger wait time perceptions at bus stops: Empirical results and impact on evaluating real-time bus arrival information. *Journal of Public Transportation*, *9*(2), 5.

Mladenovic, M. N., Mangaroska, K., & Abbas, M. M. (2017). Decision support system for planning traffic operations assets. *Journal of Infrastructure Systems*, *23*(3), 05017001.

Saaty, T. L., & Vargas, L. G. (2012). *Models, methods, concepts & applications of the analytic hierarchy process* (Vol. 175). Springer Science & Business Media.

Saberi, M., Zockaie, A. K., Feng, W., & El-Geneidy, A. (2013). Definition and properties of alternative bus service reliability measures at the stop level. *Journal of Public Transportation*.

Ap. Sorratini, J., Liu, R., & Sinha, S. (2008). Assessing bus transport reliability using microsimulation. *Transportation Planning and Technology*, *31*(3), 303-324.

Strathman, J. G., Dueker, K., Kimpel, T. J., Gerhart, R., Turner, K., Taylor, P., ... & Griffin, D. (1999). Automated bus dispatching, operations control, and service reliability: the initial Tri-Met experience.

Trompet, M., Liu, X., & Graham, D. J. (2011). Development of key performance indicator to compare regularity of service between urban bus operators. *Transportation research record*, *2216*(1), 33-41.

van Oort, N., Boterman, J. W., & 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., Sparing, D., Brands, T., & Goverde, R. M. (2015). Data driven improvements in public transport: the Dutch example. *Public transport*, *7*(3), 369-389.

van Oort, N., & van Nes, R. (2004). Service regularity analysis for urban transit network design. In *Proceedings of 83rd Annual Meeting of Transportation Research Board* (pp. 1-26).

Vincent, M. P., & Hamilton, B. A. (2008). *Measurement valuation of public transport reliability*. Wellington: Land Transport New Zealand.

Vuchic, V. R. (2017). *Urban transit: operations, planning, and economics*. John Wiley & Sons.

Weckström, C., Kujala, R., Mladenović, M. N., & Saramäki, J. (2019). Assessment of large-scale transitions in public transport networks using open timetable data: case of Helsinki metro extension. *Journal of Transport Geography, 79*, 102470.

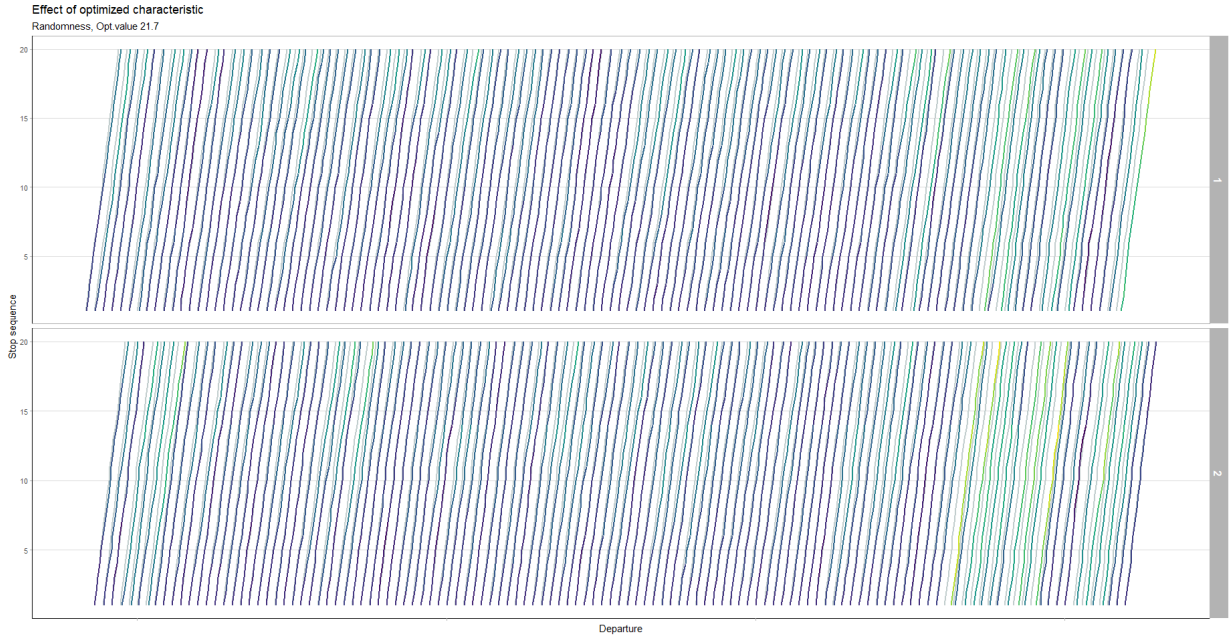
Yaakub, N., & Napiah, M. (2011). Public Transport: Punctuality Index for Bus Operation. *World Academy of Science, Engineering and Technology*, *60*, 857-862.

Appendix

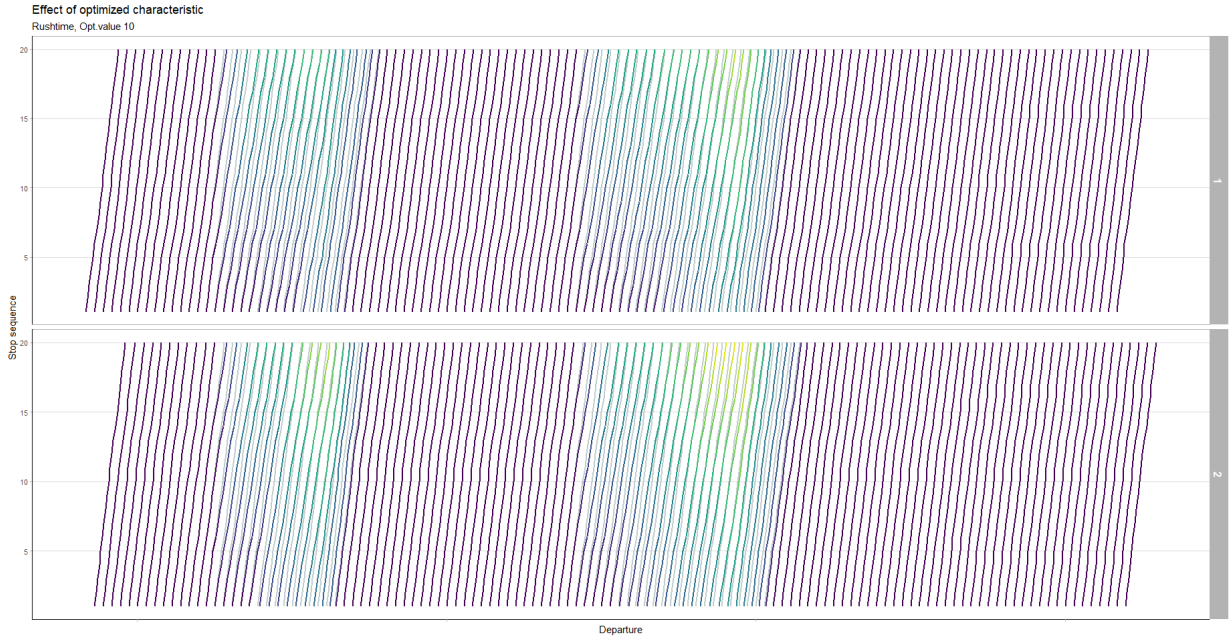
Appendix 1. The effects of randomness and phenomena on schedule structure. 4 pages.
Appendix 2. Exemplar study graphs. 12 pages.

Appendix 1. The effects of randomness and phenomena on schedule structure

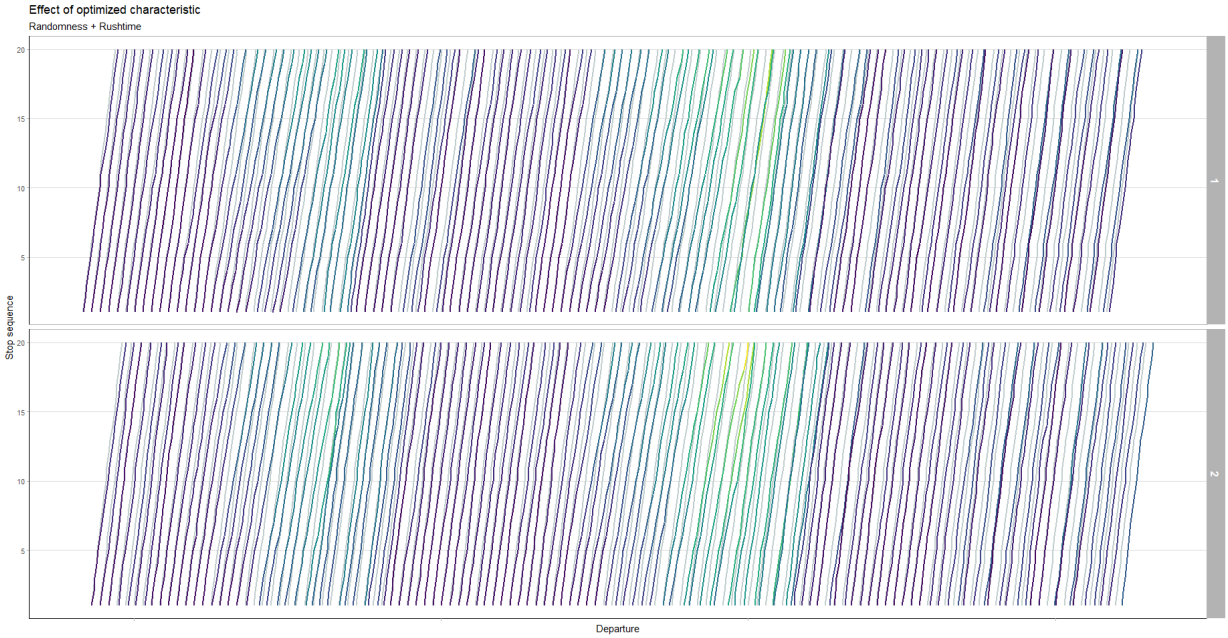
Randomness



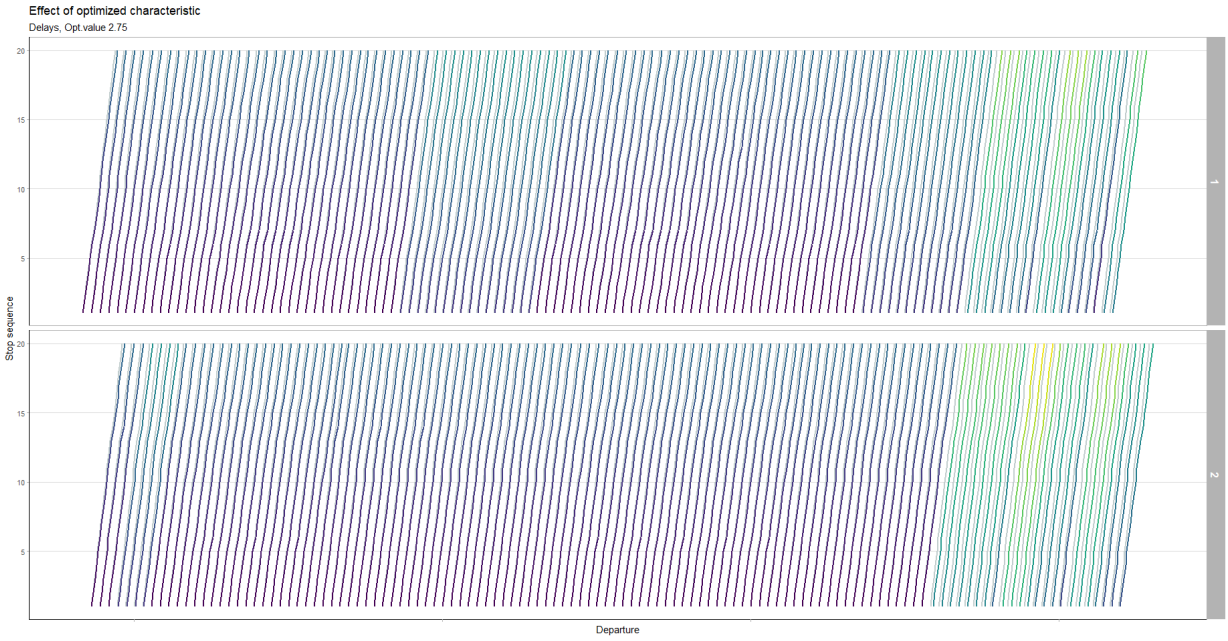
Increased driving time in rush hour



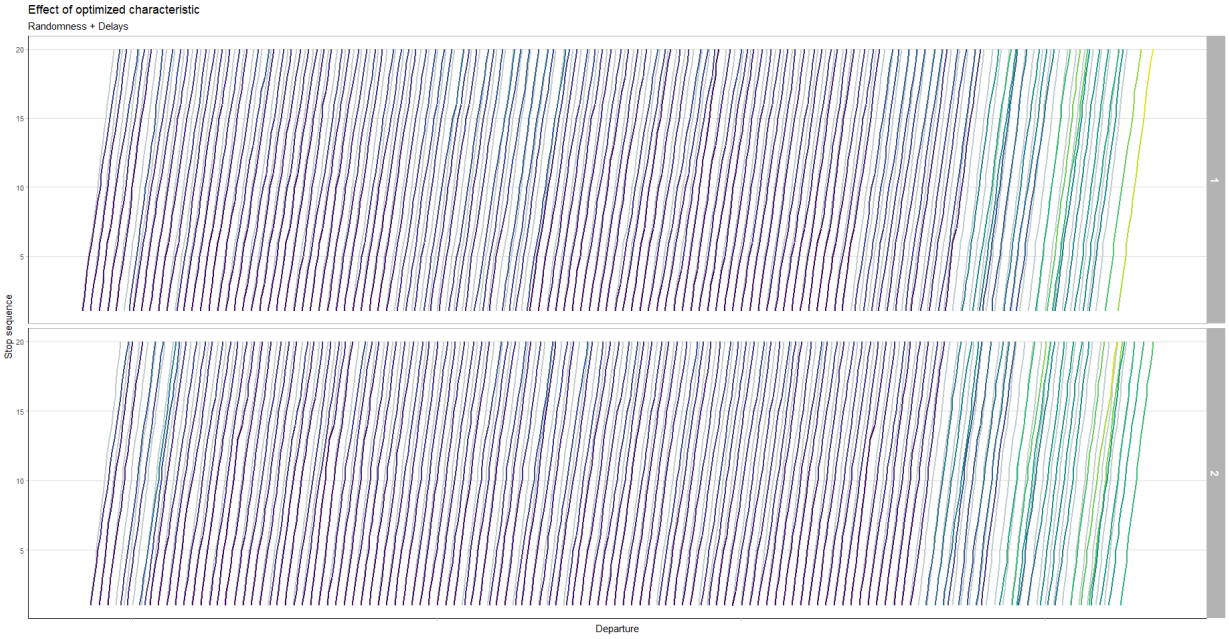
Increased driving time in rush hour with randomness



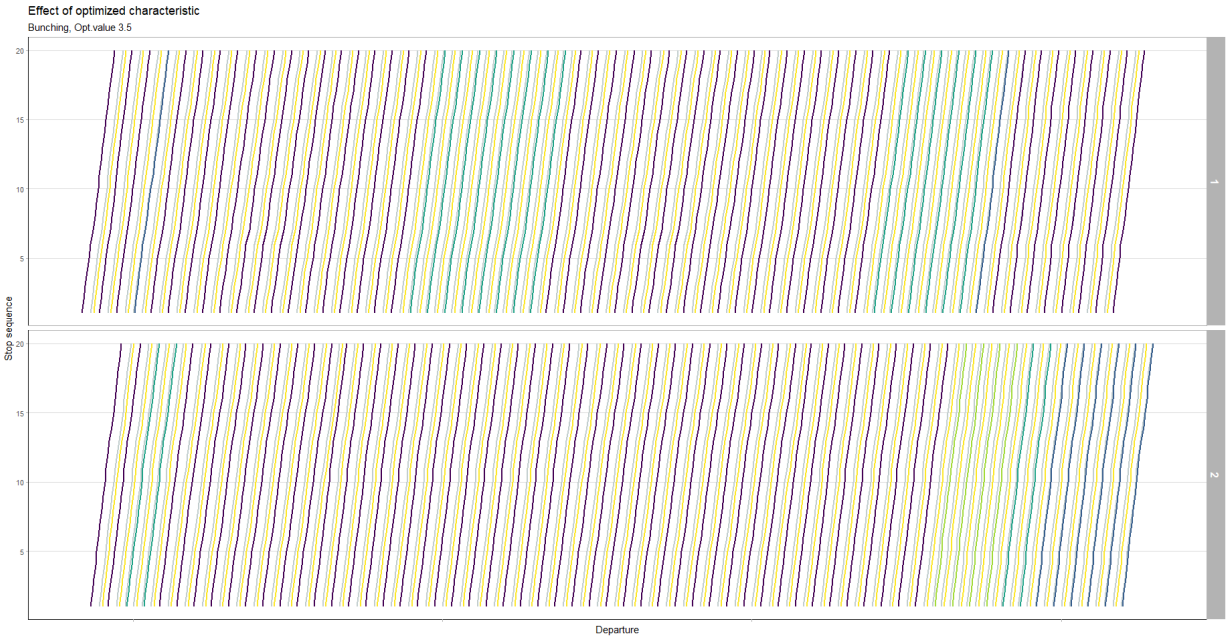
Increased driving time through day



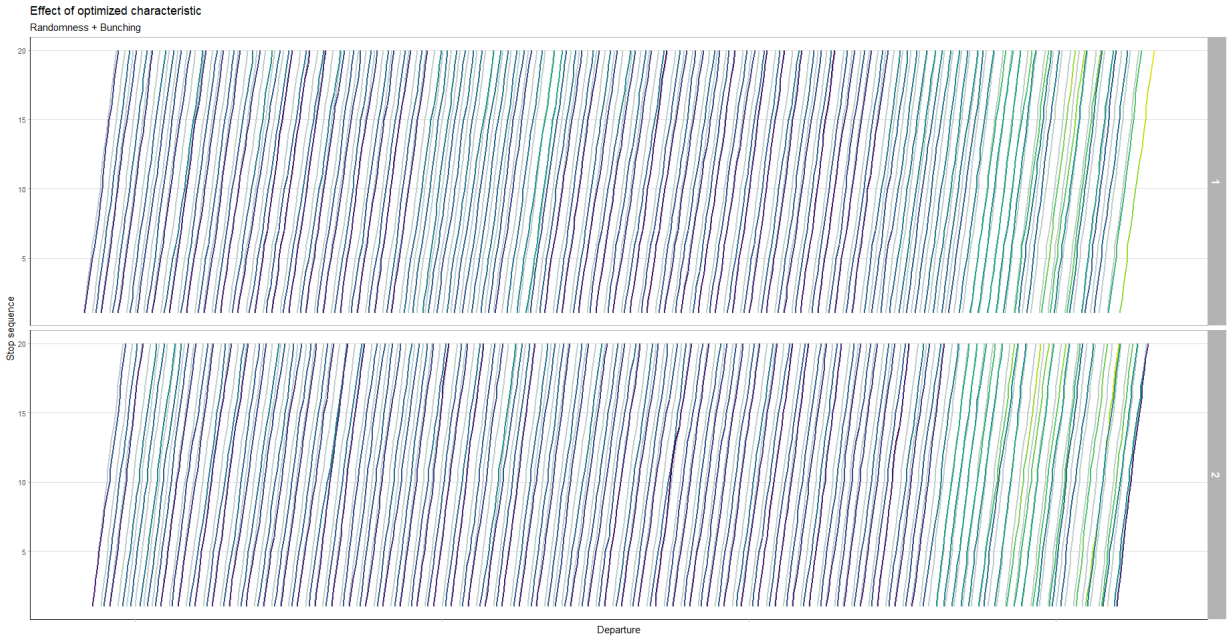
Increased driving time through day with randomness



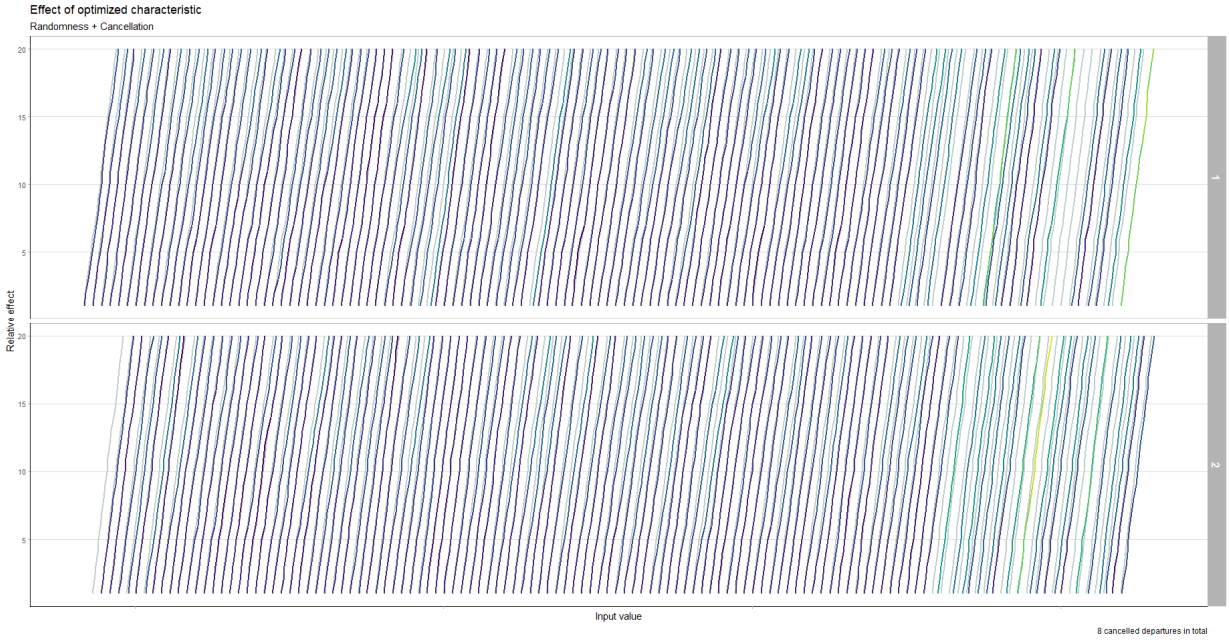
Bunching



Bunching with randomness



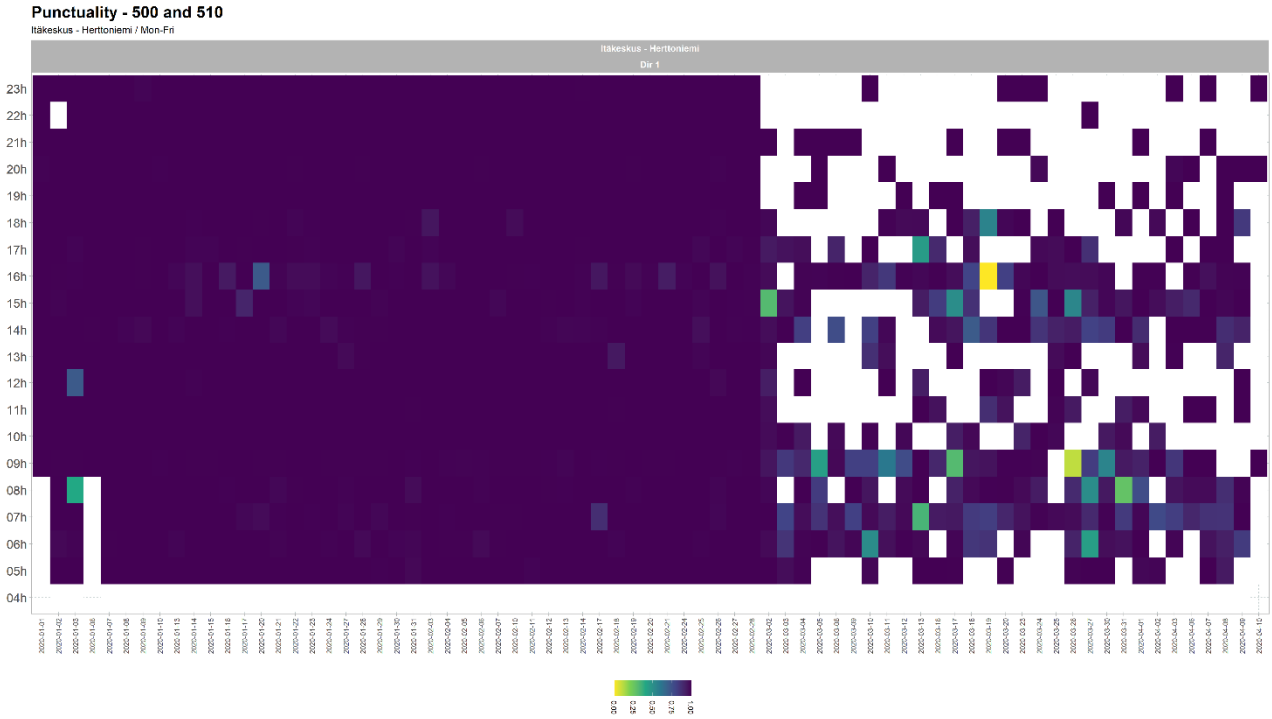
Cancellation with randomness



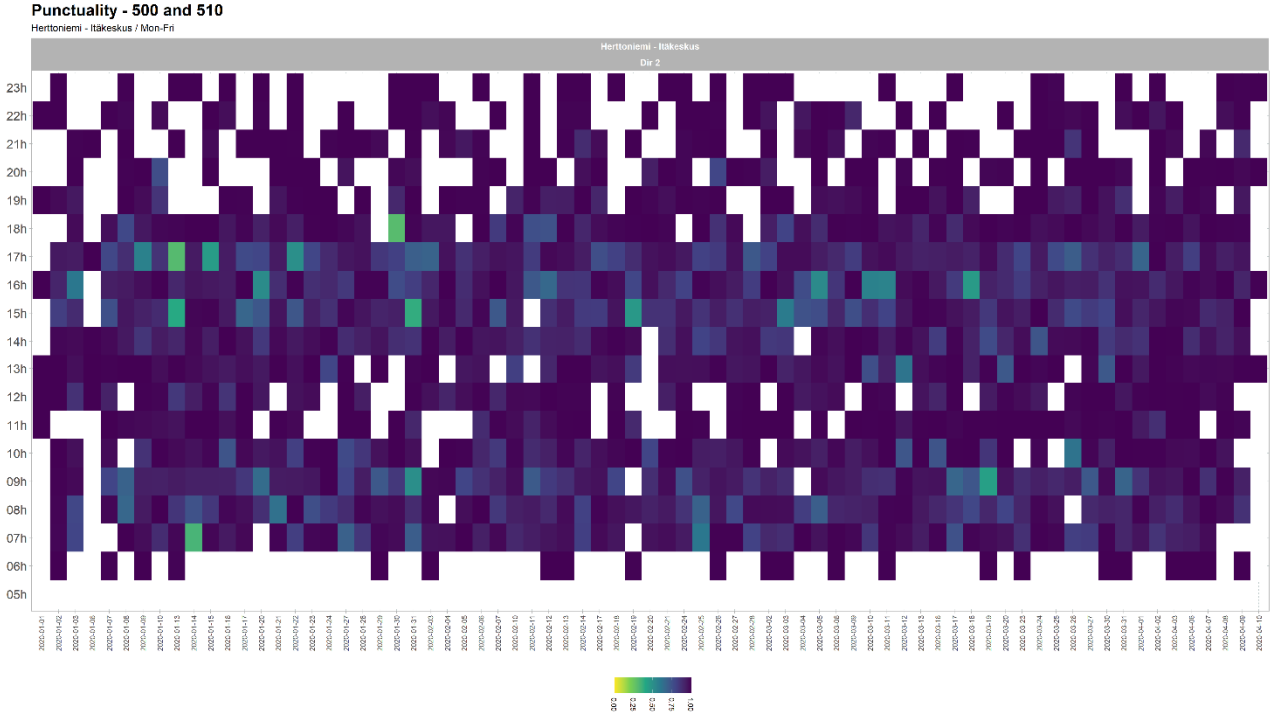
Appendix 2. Exemplar study graphs

Punctuality

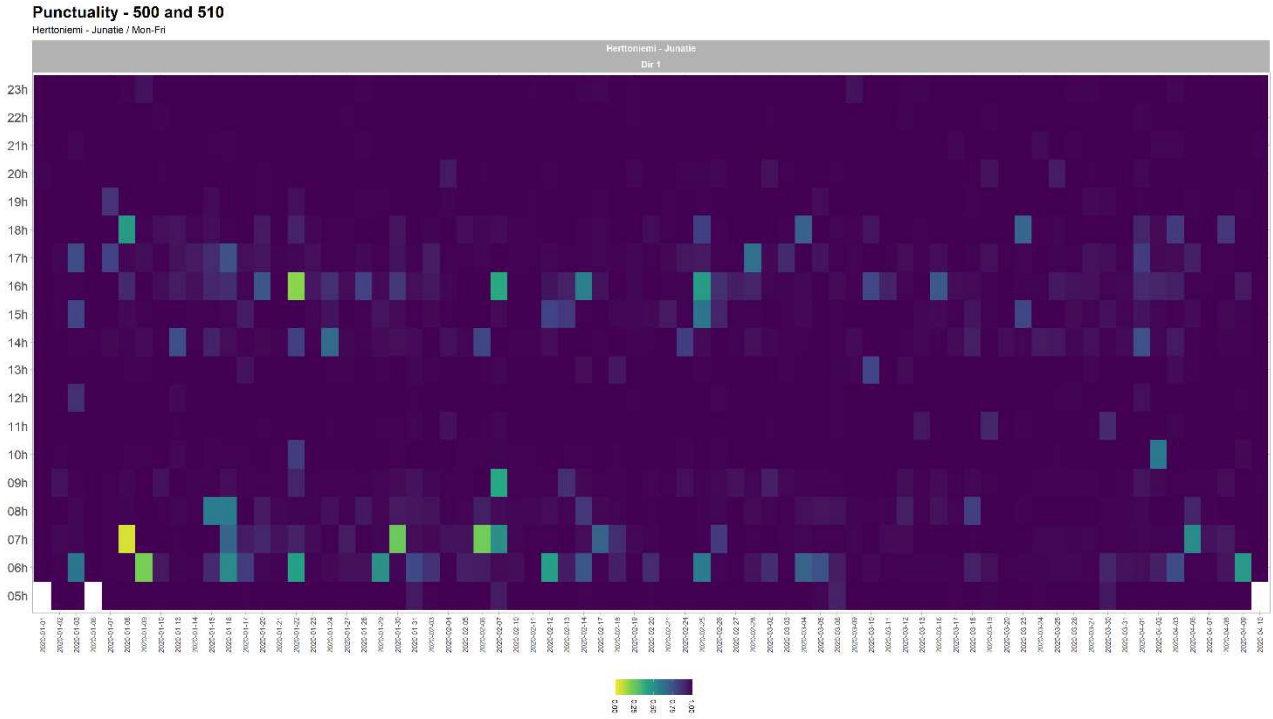
Itäkeskus – Herttoniemi



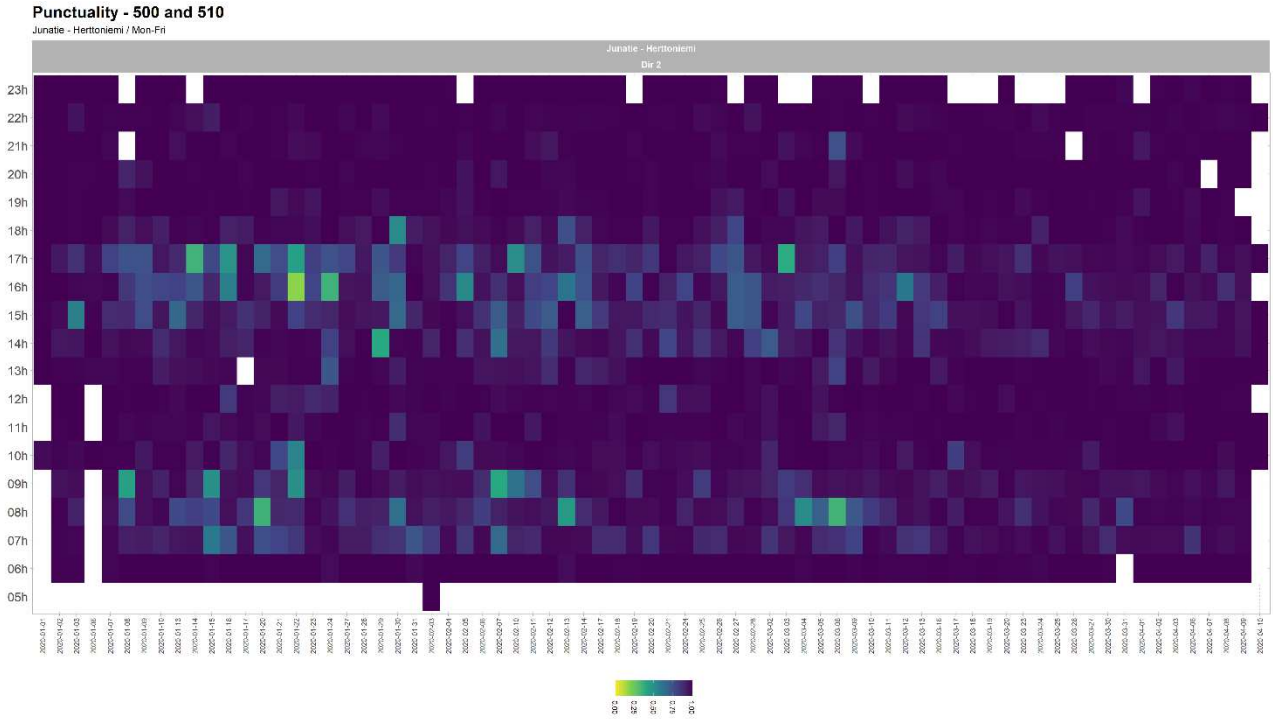
Herttoniemi – Itäkeskus



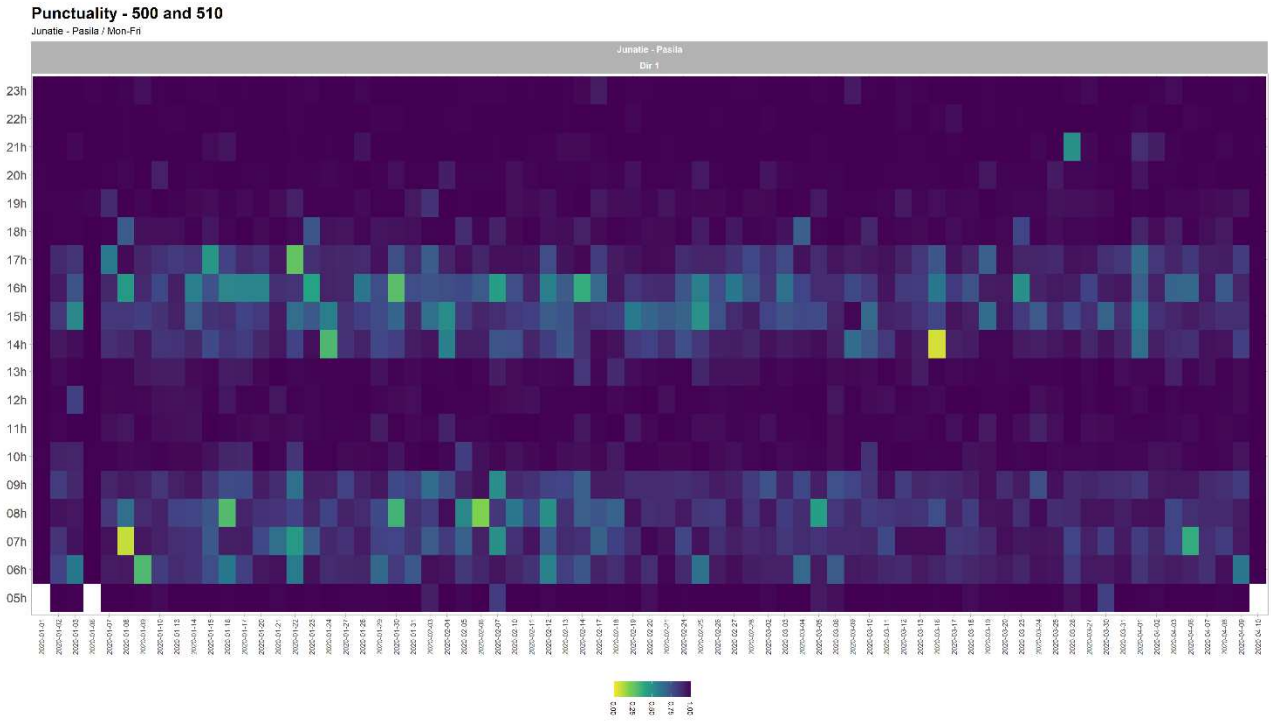
Herttoniemi – Junatie



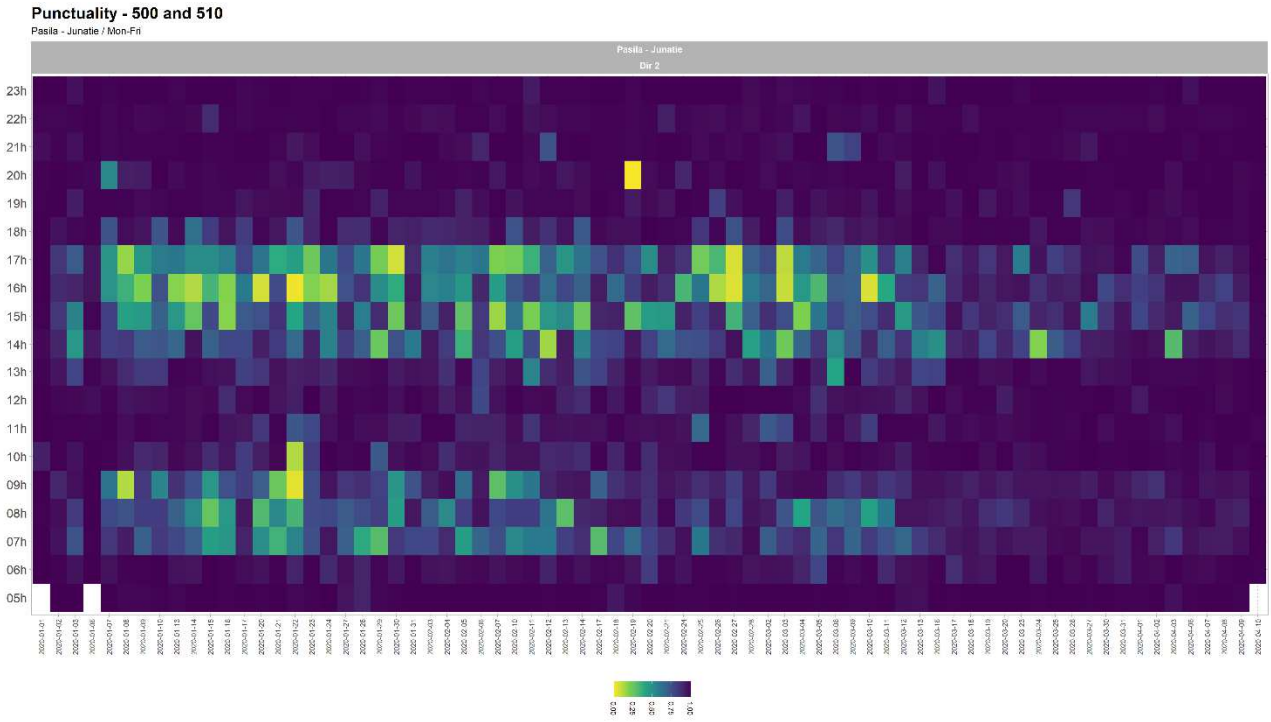
Junatie – Herttoniemi



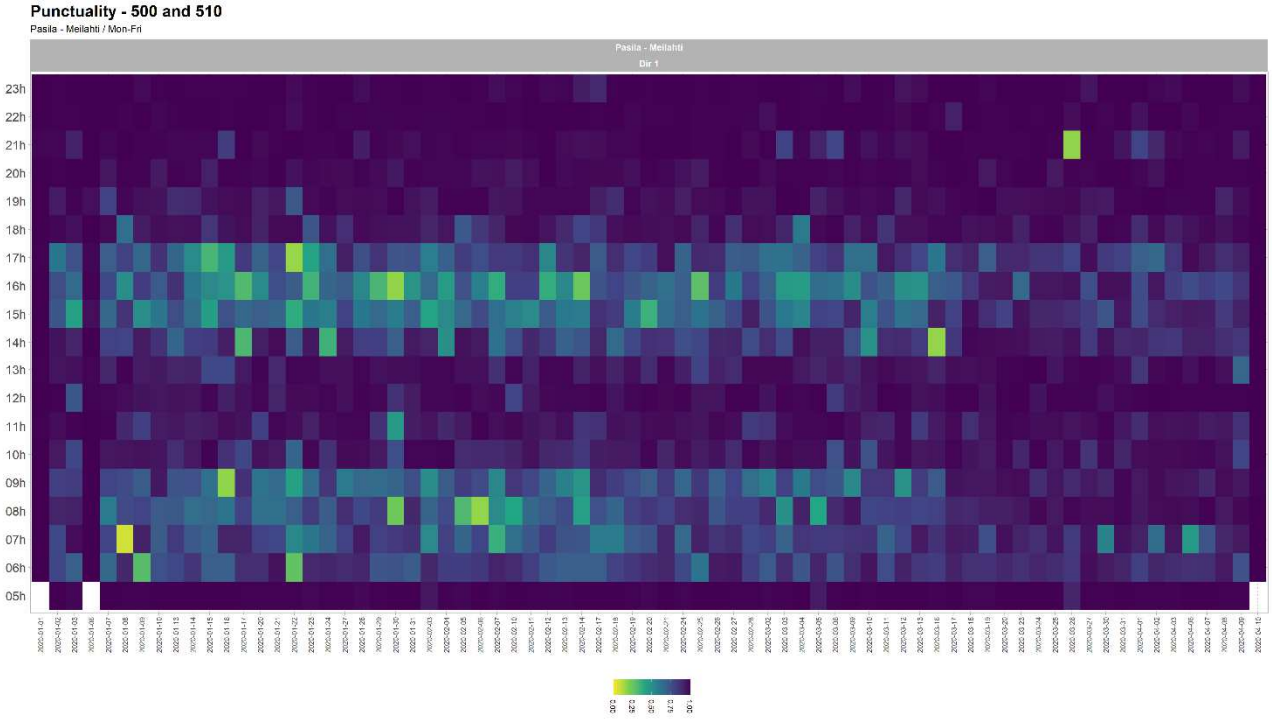
Junatie – Pasila



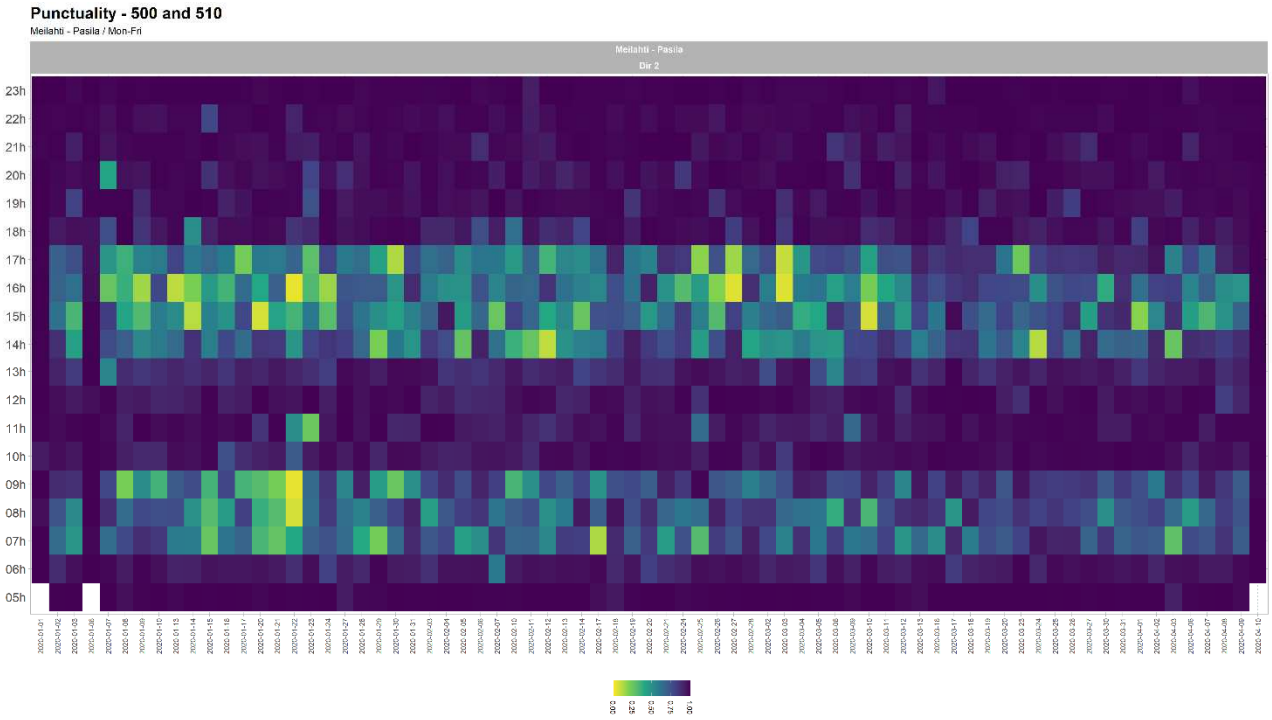
Pasila – Junatie



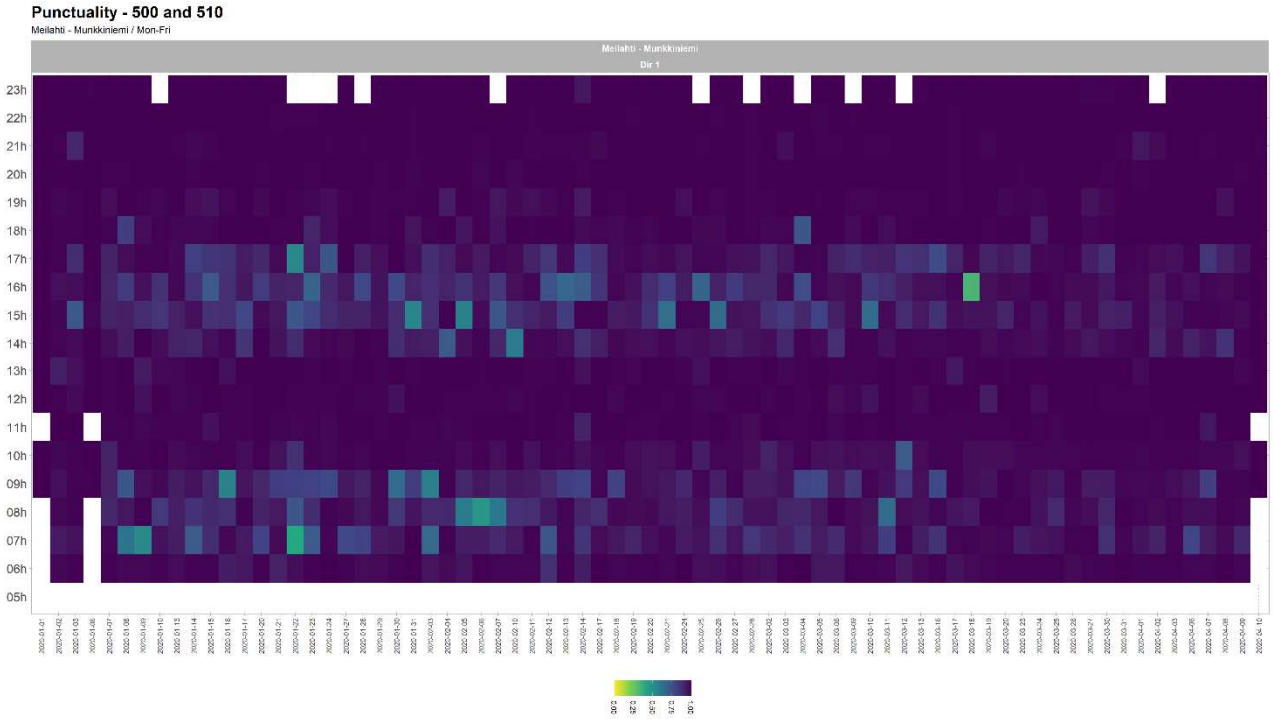
Pasila – Meilahti



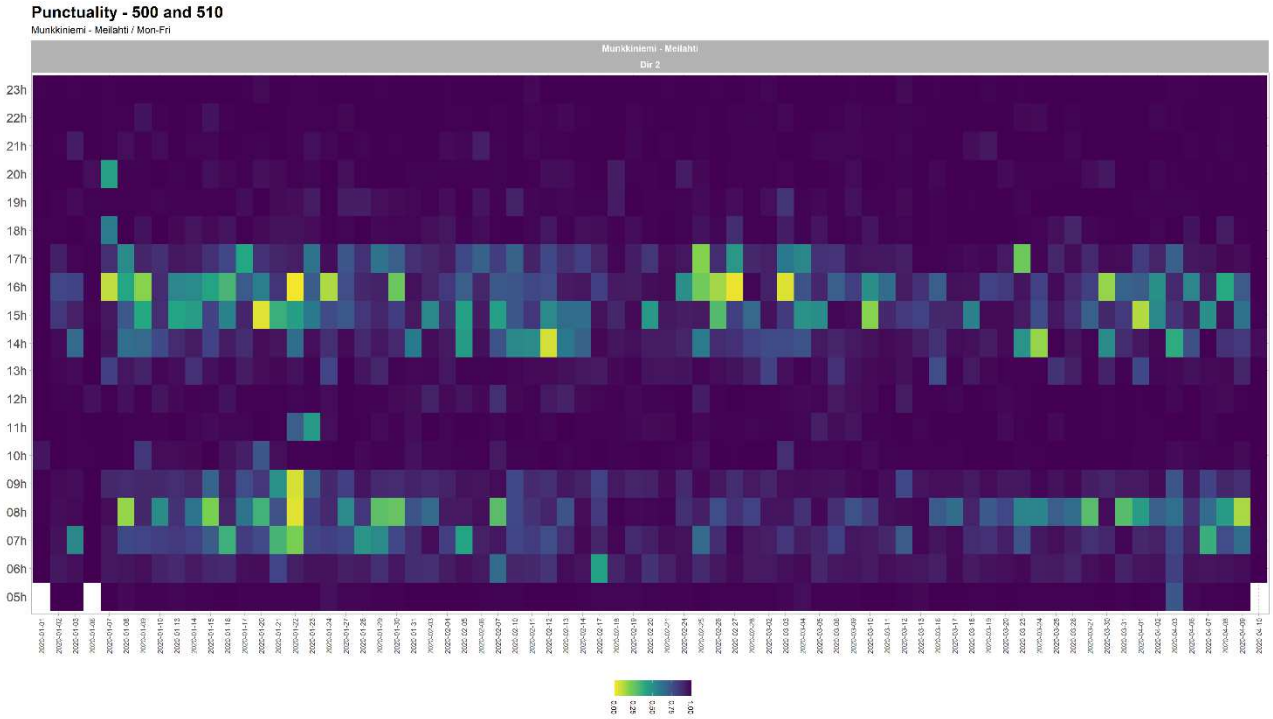
Meilahti – Pasila



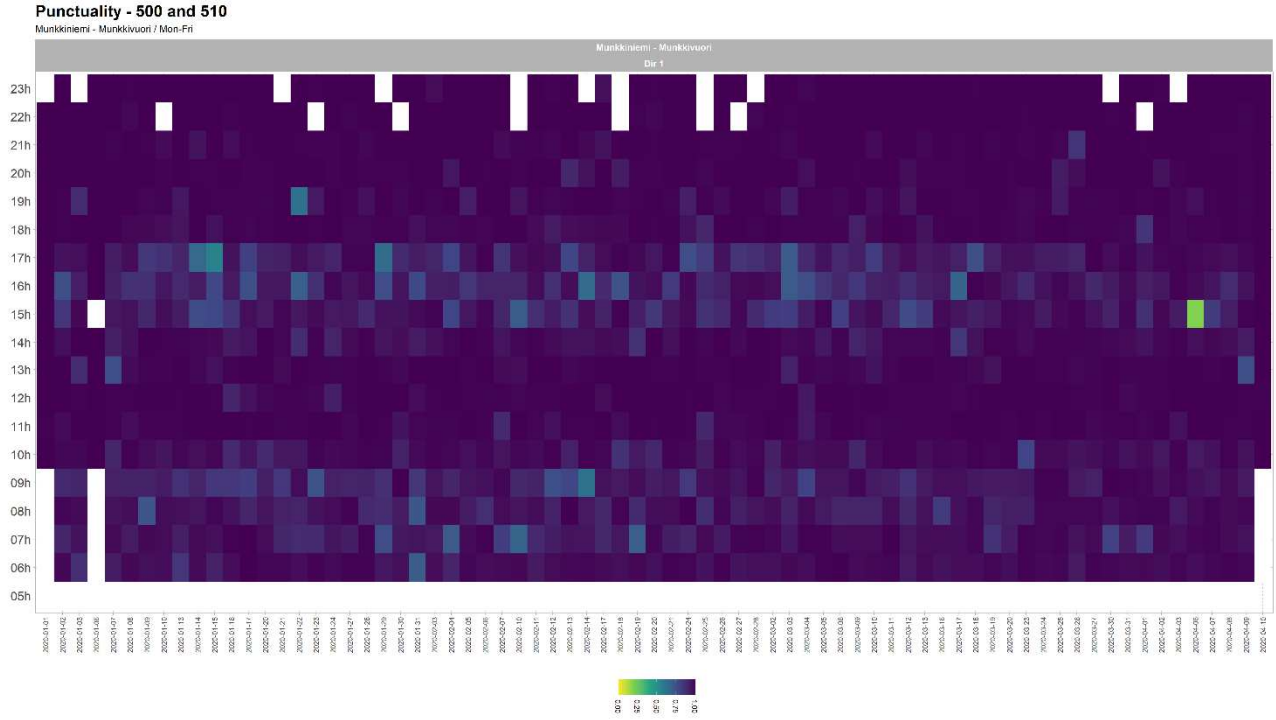
Meilahti – Munkkiniemi



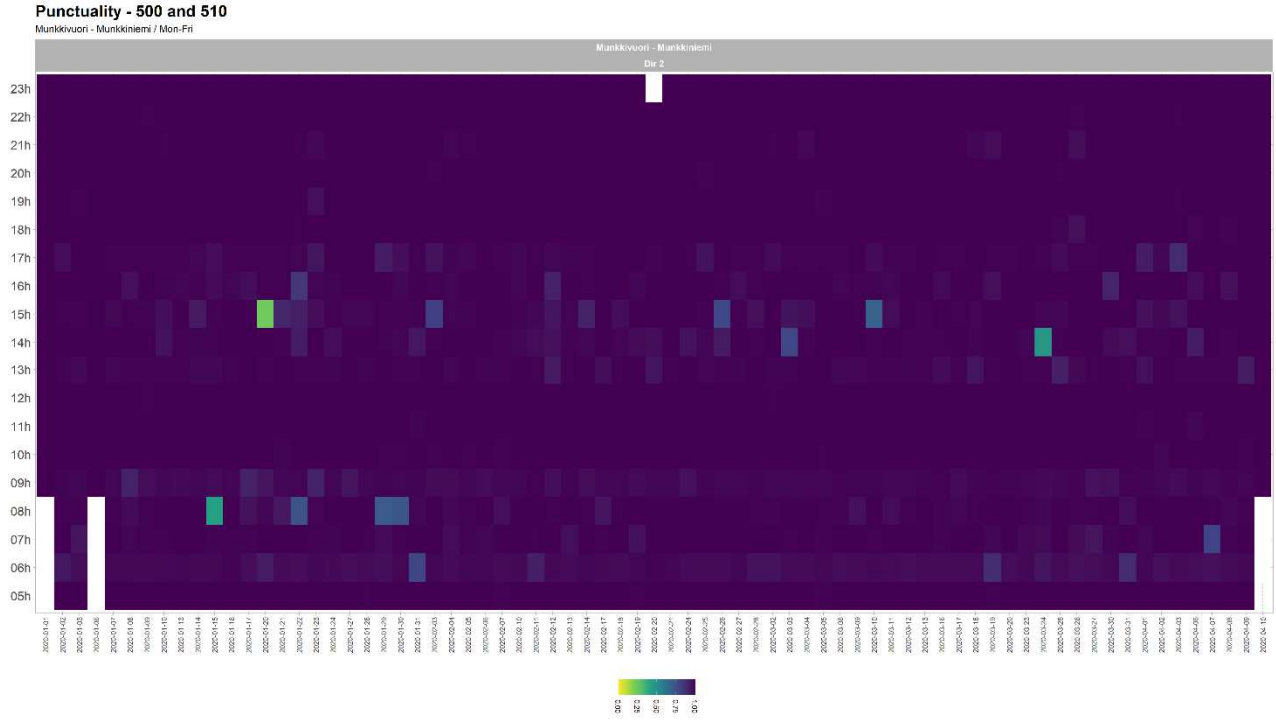
Munkkiniemi – Meilahti



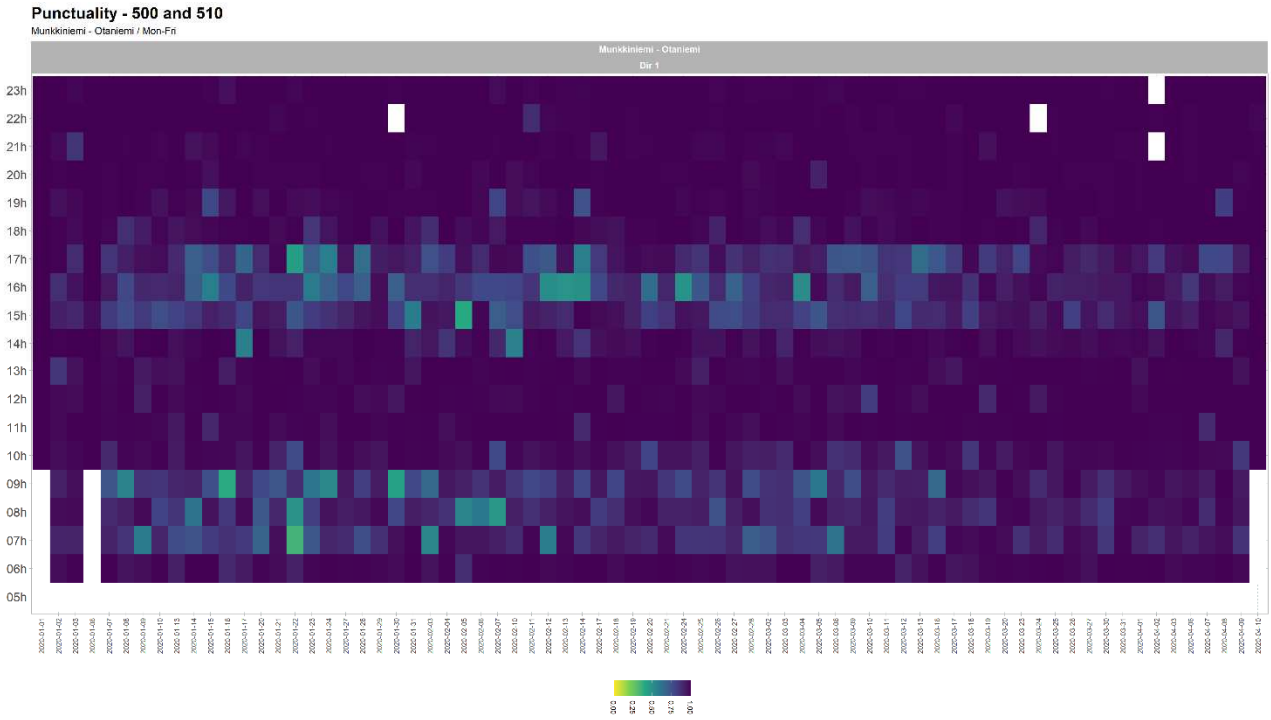
Munkkiniemi – Munkkivuori



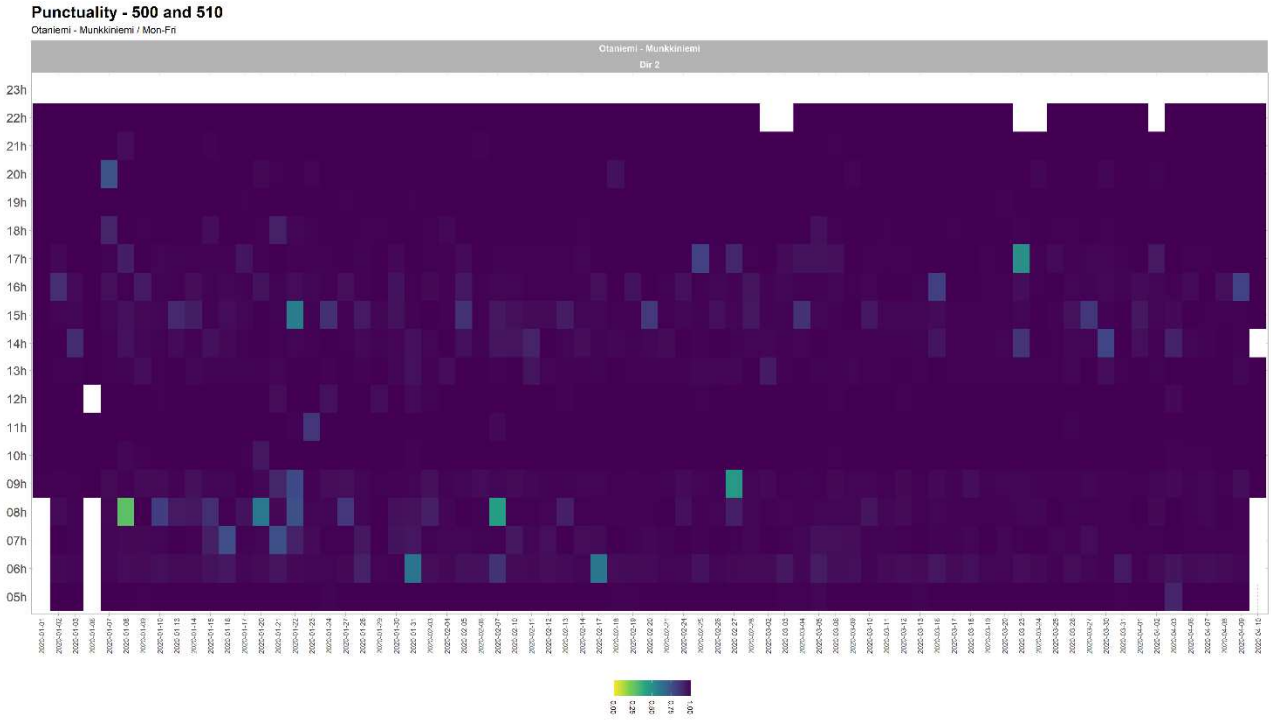
Munkkivuori - Munkkiniemi



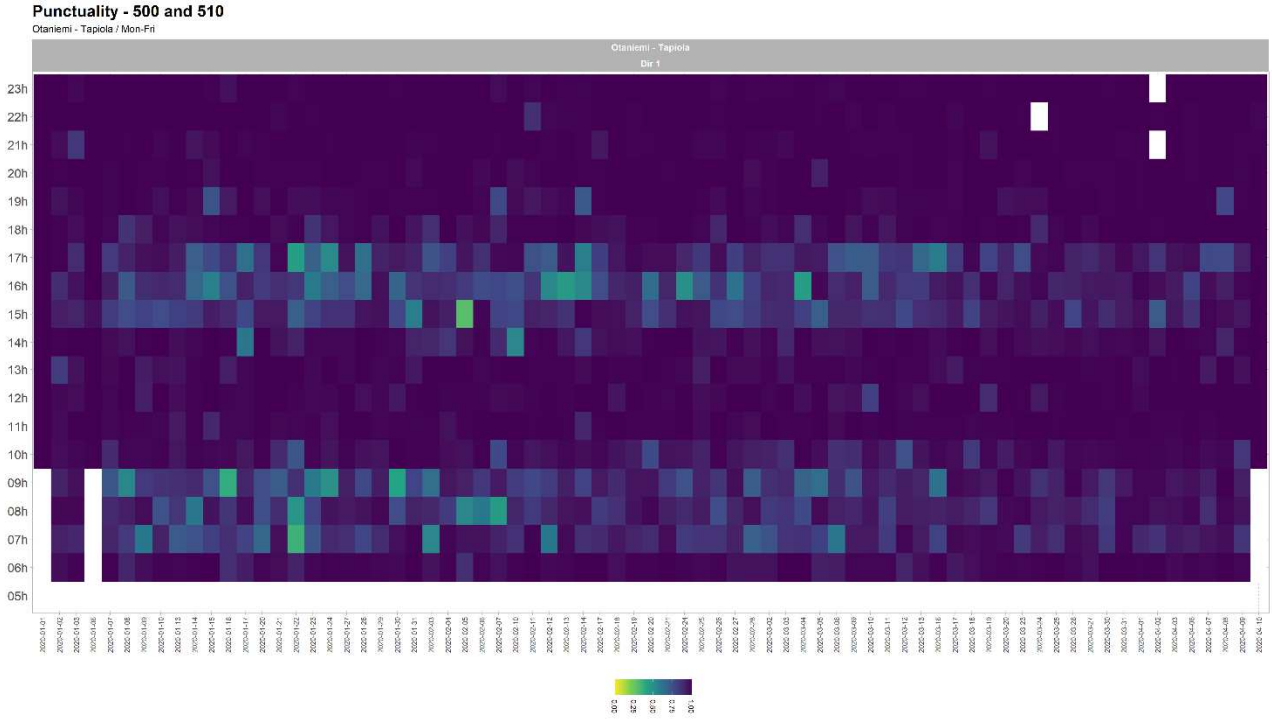
Munkkiniemi – Otaniemi



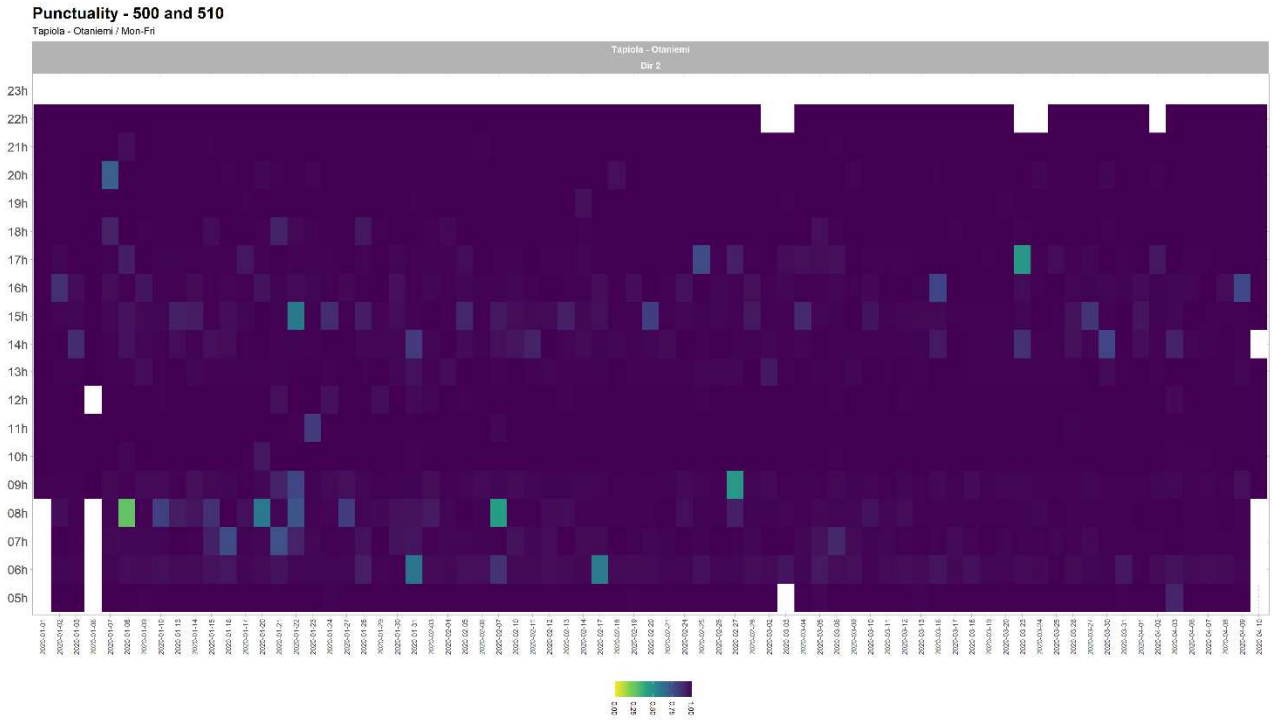
Otaniemi – Munkkiniemi



Otaniemi – Tapiola



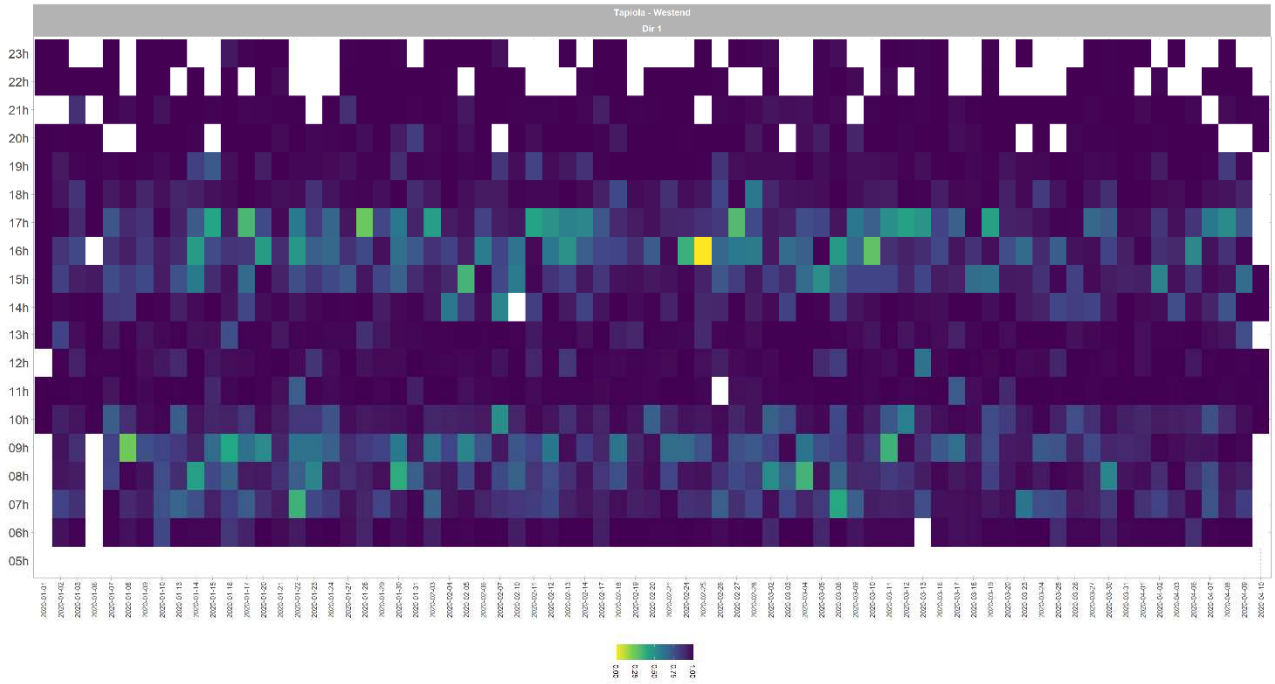
Tapiola – Otaniemi



Tapiola – Westend

Punctuality - 500 and 510

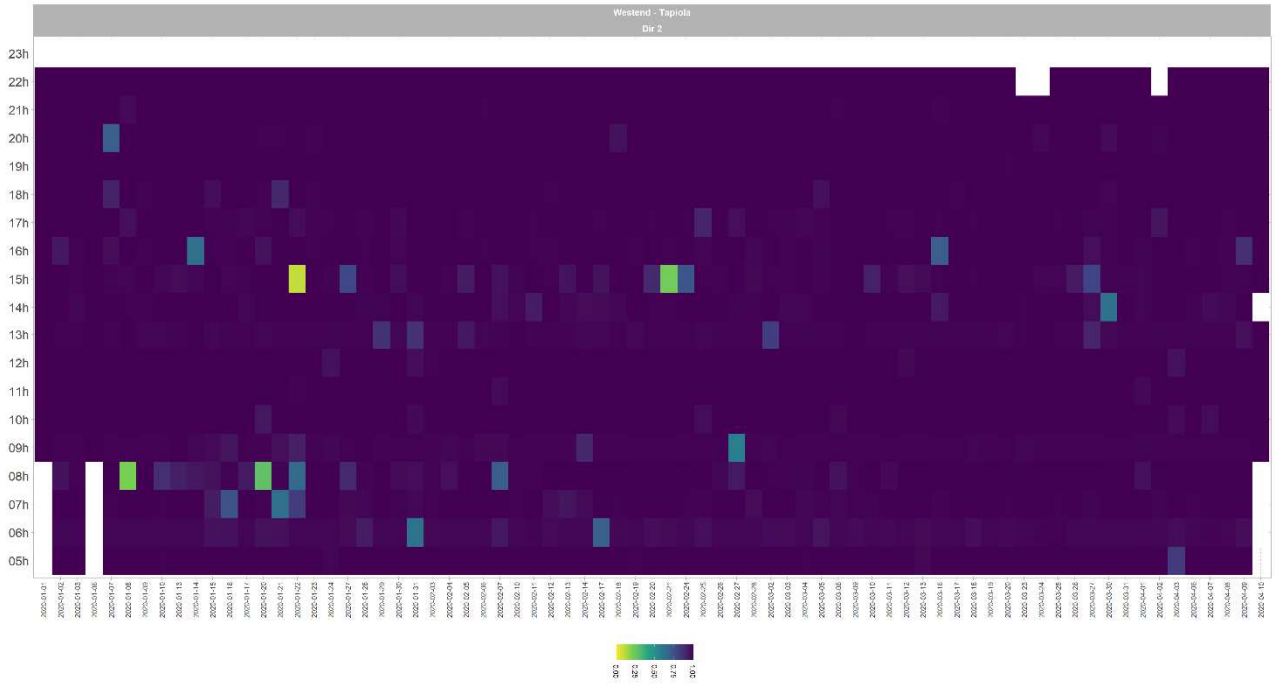
Tapiola - Westend / Mon-Fri



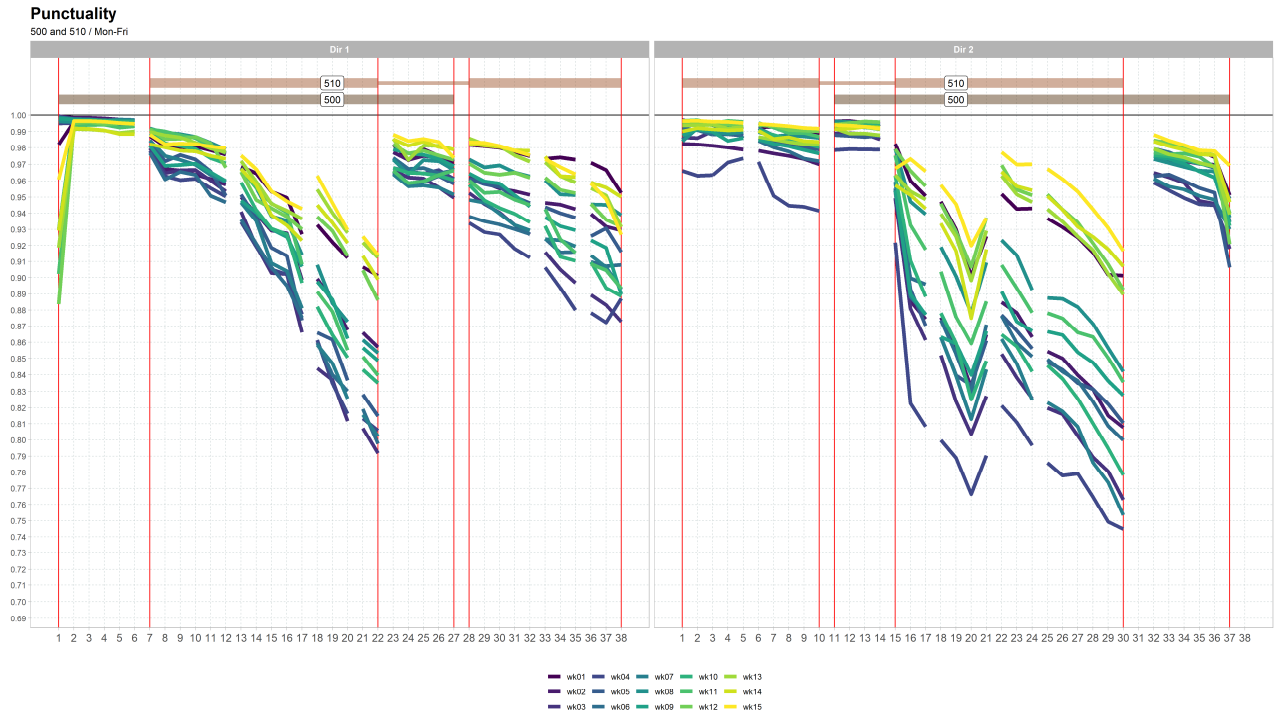
Westend – Tapiola

Punctuality - 500 and 510

Westend - Tapiola / Mon-Fri

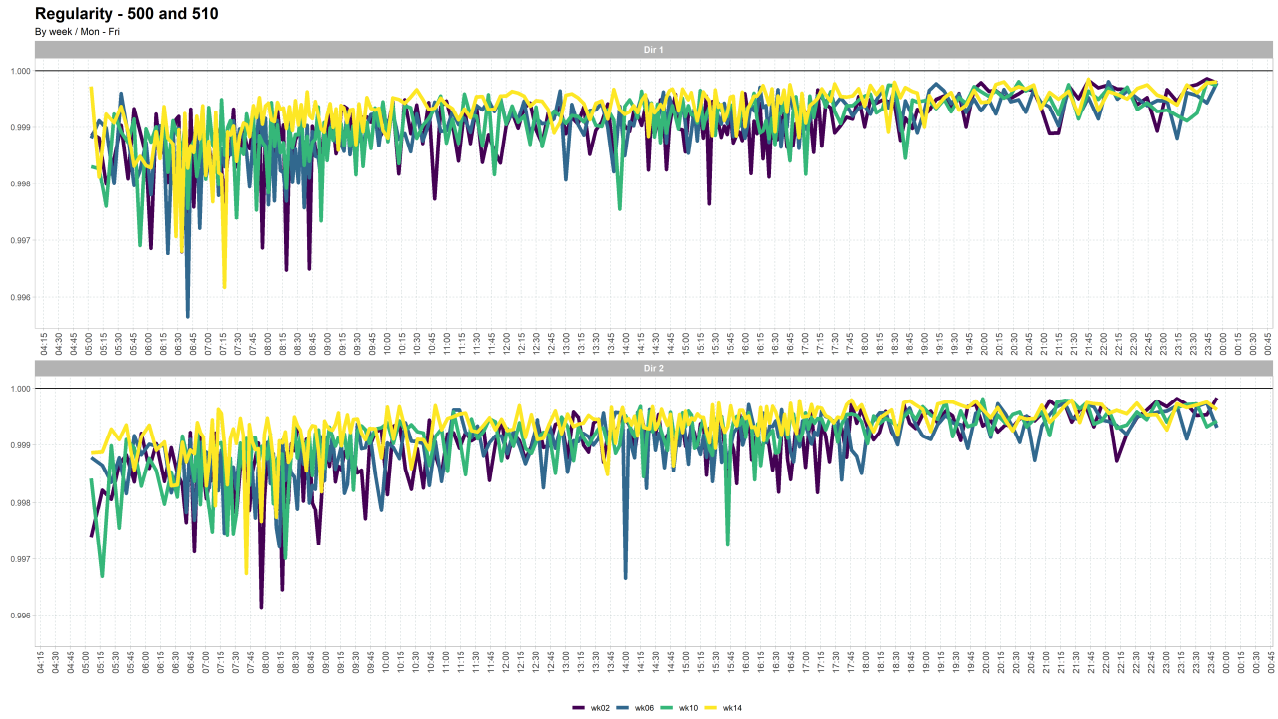


Punctuality per stop

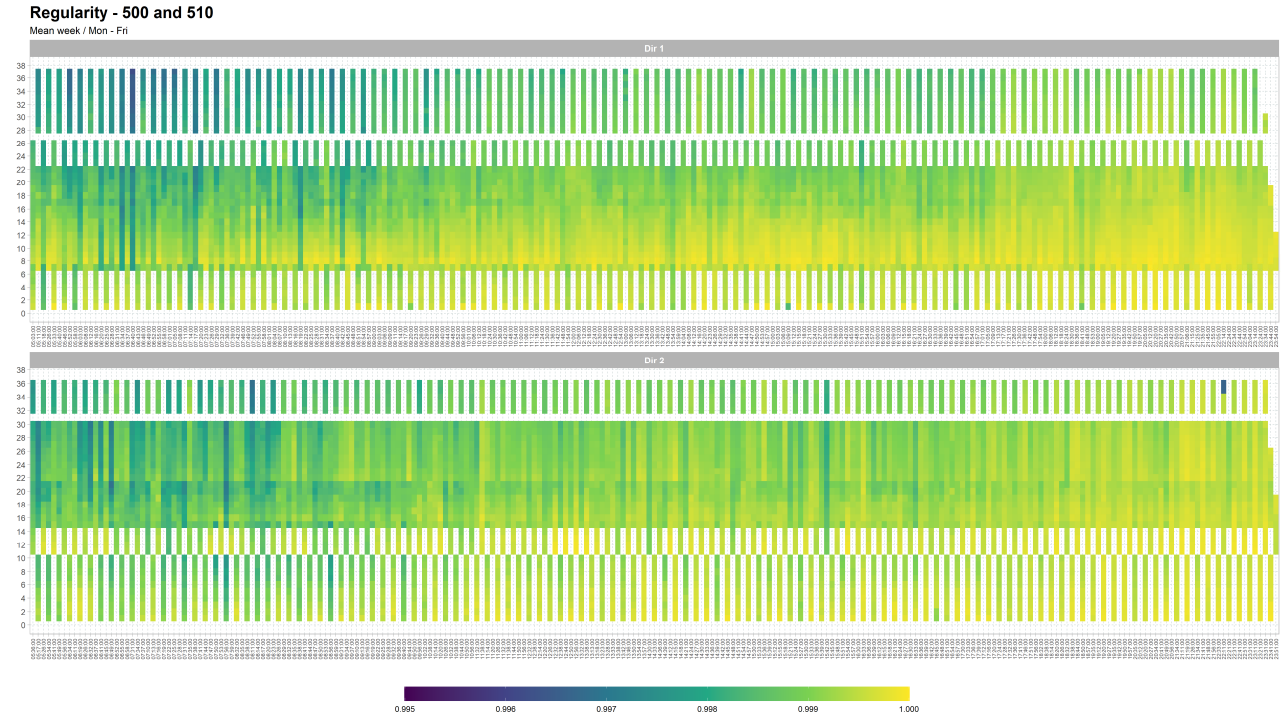


Regularity

Regularity per departure



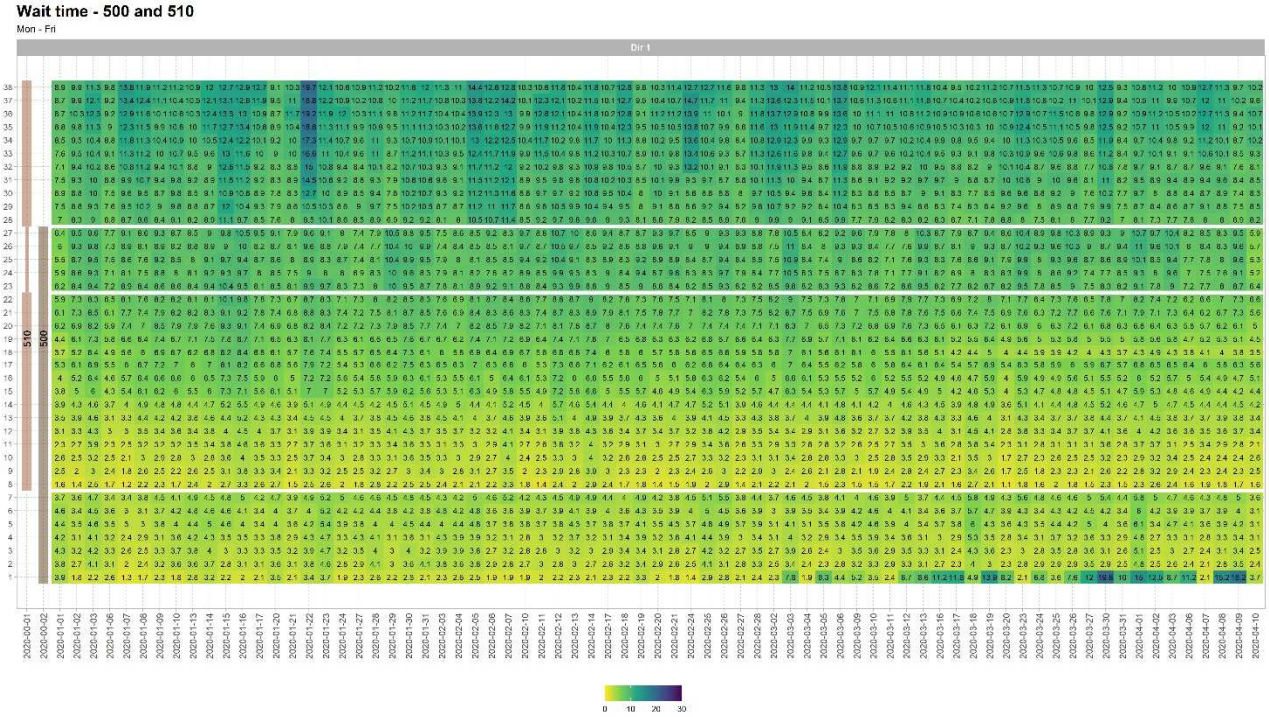
Mean weekday regularity



Wait time

Waiting time per weekday in hour group

Wait time in direction one



Wait time in direction two