

# Master thesis:

Develop a demand-oriented train service plan  
by using the railway passenger demand pattern

Graduation project  
By Xinran Meng

# Master thesis:

Develop a demand-oriented train service plan  
by using the railway passenger demand pattern

by

By Xinran Meng

in partial fulfilment of the requirements for the degree of

**Master of science**

in Transport & Planning

at the Delft University of Technology

to be defended publicly on July 20th 2023.

Student number: 5467284

TU Delft, chair: TU	R.M.P. (Rob) Goverde
Delft, supervisor: TU	N. (Niels) van Oort
Delft, supervisor:	R.J.H. (Renate) van der Knaap
NS, external supervisor:	Menno de Bruyn

# Preface

*Sometimes it's like we're having one dream after another.*

This is what I feel whenever one phase is over. From middle school, along to my master period, I was always in a different city, with different people. I feel so lucky to travel through those well designed segments, and all views among them motivate me to step forward and explore. I guess that's why hiking is one of my favourite activities — if you keep walking, there is always something unexpected, that can satisfy your curiosity.

I never thought I would come to Delft, to a European country like Holland, to study. Even though everything is so different here, I'm glad I didn't try to run away from all the discomfort. It's so wonderful to interact with people from other cultures, especially ones that are so different from Chinese culture. It's also wonderful to see that the sun go down around 12 o'clock.

This thesis, is the final project for obtaining my master's degree in Transport & Planning at Delft University of Technology. I would like to thank all my committee members here. I could not have completed the project so quickly without your help. I would like to thank the chair of the committee, Rob, who gave a lot of pertinent guidance on the logic and structure of the paper. I would like to thank Niels for reminding me of making good plans and scheduling everything well. As you told me when in our first meeting: A good plan is half of success. I would like to thank Renate. Even though you usually say you're a first time supervisor, I think you did a pretty good job. Sometimes when I felt more or less discouraged, I can always get some motivation from the communications with you. Those conversations with you about hiking are always enjoyable! I would like to thank Menno, who would always encourage me whether or not I am making a progress. I've also benefited from a lot of your feedback about the program.

I also want to give a special thanks to Ivan and Grisha, who are like my big brothers. We had a lot of fun in the gym these past two years, and exercise always makes me less stressed. You guys always went out of your way to teach me all sorts of things and spurred me on to workout, even though I'm a lazy guy.

This should be the official end of my master's program in TU Delft. I really appreciate all the opportunities during my stay in Delft.

I will miss this dream for always:)

*By Xinran Meng  
Delft, July 2023*

# Summary

The railway undertakings in the Netherlands, particularly NS (Nederlandse Spoorwegen), have traditionally provided a cyclic railway timetable, offering fixed departure times and regular interval services throughout the day. While this periodic schedule has been successful in meeting peak hour passenger needs, it does not fully match the characteristics of off-peak passenger demand, which often differs in terms of volume and origin-destination combinations. However, recent developments such as increased train frequencies and the availability of the NS mobile app have reduced the reliance on periodic schedules. For instance, the addition of high-frequency service lines with trains departing every ten minutes between major cities (Rotterdam-Schiphol, Schiphol-Arnhem and Rotterdam-Dordrecht) has provided more flexibility. These changes have led to a decreased desire for periodic timetables and opened the door for implementing a more flexible train schedule.

The current train service in the Netherlands primarily caters to peak-hour passenger demand, which may result in underutilized capacity during off-peak periods. While the periodic schedule effectively serves off-peak demand, it lacks insight into passenger travel patterns, leading to unnecessary redundancy in train services. This redundancy incurs additional costs for rolling stock usage and maintenance. By gaining a clear understanding of passenger flow characteristics over different timeframes, railway undertakings can optimize train schedules accordingly. This optimization would benefit railway companies by reducing costs and improving the overall passenger travel experience, including shorter travel times. This study contributes to the existing literature by exploring the methodology to develop a demand-oriented train service plan by analyzing the railway passenger demand pattern. This study aims to answer the following research question:

- How to develop a demand-oriented train service plan by using the railway passenger demand pattern?

To achieve this objective, the study began with a literature review to explore existing methods for analyzing flow demand patterns in public transport. The concept of the base demand from NS gives a good start to find the first type of demand pattern: the base demand. Base demand refers to the consistent and representative passenger demand that exists throughout the day or week, as opposed to peak hour flows. Designing a train service based on the base demand aims to meet the needs of the majority of passengers. The concept recognizes that a significant percentage of railway travelers use trains during off-peak periods. By focusing on the base demand, train services can be optimized to better serve passengers and improve overall efficiency. Defining the base demand involves identifying suitable periods with minimal flow fluctuations, ensuring that the train service can adequately meet passenger requirements. Besides, various approaches, including the use of smart card data, visualization techniques, statistical indicators, and clustering methods, were reviewed. The literature provided valuable insights into understanding commuter patterns and identifying base demand periods.

Based on the literature review, the study focused on utilizing clustering methods to characterize passenger demand patterns. Four commonly used clustering methods, namely K-means, bisecting K-means, DBSCAN, and Hierarchical clustering, were compared. Hierarchical clustering, with its average-link measure, is able to overcome chain effects (which is the defect owned by all other methods). Besides, no initial values such as the number of clusters are needed to be set. It was found to be the most suitable method for identifying base demand. The silhouette coefficient was utilized to ensure optimal clustering results.

The methodology was applied to a specific sub-network of the Dutch railway network using a one week of demand data from 2022. This one week data represents the typical autumn/winter demand pattern in the network. For each pair of ODs in the network, the clustering method is applied to its week of data, where the cluster with the most data is considered to be the cluster representing the base demand. This is because this feature is consistent with the definition of base demand: it floats less in the corresponding period and is always distributed in off-peak periods. For each period, the frequency they are considered as the base demand period is counted. The clustering method was once again applied to the frequency results, where a cluster containing values that all have high frequencies was considered to represent the base demand period, since for the periods in this cluster, almost all ODs considered the period to be the base demand period. The clustering results also helped to delineate 15 periods of the week that might contain different demand patterns. Each of these 15 periods is represented by a matrix containing demand. By comparing the similarity between the matrices, which is performed by calculating the Manhattan distance between the matrices, those similar matrices are combined into the same matrix. The 15 matrices were finally merged into 8 matrices.

The analysis revealed eight distinct periods representing different demand patterns, including the base demand. The comparison between those eight periods and the existing train schedules suggests that the peak service on weekdays (Monday to Thursday) can be optimized further. For morning/afternoon peak hours, the Monday and Wednesday shows similarity in both demand structure and demand volume, which also works for Tuesday and Thursday. Specifically, the aggregate demand volume between 3:00 p.m. and 4:30 or 5:00 p.m. when the evening peak begins, will increase compared with the base demand. For the base demand period, the time period it represents is spread throughout the week, mostly during off-peak hours on weekdays and throughout the weekend, taking up more than half of the week. From the aggregate level, the demand volume is much smaller compared with the peak hour demand.

The analysis of the demand patterns serves as the basis for developing a demand-oriented train schedule. For the case study, a compact railway line consisting of 6 stations was chosen. Two distinct train service plans were formulated, each operating on an hourly basis. The first plan followed conventional rules and practices for developing train schedules, while the second plan was specifically designed based on the concept of base demand. These scenarios were evaluated and compared in terms of railway undertakings (RU) and passenger benefits, providing valuable insights into the effectiveness of the base demand approach.

The approach of making the line plan in scenario 2 ensured a better match between the train plan and the passenger demand pattern, avoiding unnecessary services and meeting passenger needs effectively. The study concluded that a demand-oriented train service plan, which

---

considers the base demand and add-on/subtracted services accordingly, can provide a more satisfactory solution. The results highlight the differences between developing train schedules based on base demand compared to the existing approach. The development of service plans becomes more intricate during peak periods, with more adaptations and adjustments required than before. If the train plan is designed around base demand, additional considerations need to be taken into account, including the preferences of other stakeholders during peak periods. In terms of cost, there is a potential reduction since the basic demand during off-peak periods is relatively low compared to peak periods, resulting in lower-cost services tailored to meet the base demand. For passengers, the total travel doesn't change little compared with the existing timetable. More resources, such as improved train equipment, can be allocated to serve these passengers, ensuring a more comfortable journey.

The limitations of this research and recommendations for future research were provided. This research acknowledges several limitations regarding the project's scope and methodologies used. To improve accuracy, it is recommended to classify the origin-destination (OD) mix before determining the base demand, considering irregular demand variations. Analyzing passenger flow attributes based on station locations and grouping OD combinations with different demand sizes would provide valuable insights. Additionally, choosing a reasonable capacity level is crucial to strike a balance between passenger satisfaction and operating costs. It is essential to consider the interests of various stakeholders, such as freight forwarders and the government. Although the final set of studies simplified the train service design process, a more comprehensive approach is needed to quantify costs and benefits accurately.

Recommendations for future research include expanding the case study to the entire network, using more comprehensive data for realistic results. Incorporating stakeholders' preferences and employing a realistic simulation process will provide a clearer understanding of the impact of service design on various aspects. Utilizing data from different periods or using predicted data for upcoming years would enhance the study's validity. For NS, adjusting the service plan based on the eight different demand patterns is recommended. For lines experiencing capacity saturation, reducing off-peak services and focusing on improving base demand services with amenities like higher speeds and passenger facilities would be beneficial. During peak hours, prioritizing capacity over quality, such as providing more seats, and offering personalized additional services tailored to specific needs would optimize the passenger experience.

Overall, the research successfully developed a demand-oriented train service plan by using the railway passenger demand pattern. The findings contribute to a better understanding of commuter patterns and provide insights for optimizing train services based on demand characteristics.

# Contents

<b>Preface</b>	<b>i</b>
<b>Summary</b>	<b>ii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	2
1.2 Problem Description . . . . .	3
1.3 Research Questions . . . . .	3
1.4 Literature Review . . . . .	4
1.4.1 Base demand . . . . .	4
1.4.2 Flow characteristics . . . . .	5
1.4.3 Impact analysis . . . . .	6
1.4.4 Conclusion . . . . .	7
1.5 Thesis Outline . . . . .	7
<b>2 Methodology</b>	<b>8</b>
2.1 Review of clustering methods . . . . .	9
2.2 Method Selection . . . . .	12
2.3 Silhouette coefficient . . . . .	13
2.4 Definition of base demand . . . . .	14
2.5 Conclusion . . . . .	16
<b>3 Case Study</b>	<b>17</b>
3.1 Case Study . . . . .	18
3.2 Data preparation . . . . .	19
3.3 Conclusion . . . . .	21
<b>4 Division of the periods with different demand characteristics</b>	<b>23</b>
4.1 Methodology Testing . . . . .	24
4.2 Clustering constraints . . . . .	27
4.3 Methodology application . . . . .	27
4.4 Base demand matrix . . . . .	29
4.5 Matrices merge . . . . .	32
4.6 Comparison with periods used in practice . . . . .	35
4.7 Conclusion . . . . .	35
<b>5 Line plan design and comparison</b>	<b>37</b>
5.1 Flow chart of designing the line plan . . . . .	38
5.2 Demand analysis . . . . .	39
5.3 Train schedule design . . . . .	43
5.3.1 Scenario 1 . . . . .	43
5.3.2 Scenario 2 . . . . .	44
5.3.3 Evaluation for two scenarios . . . . .	47

5.4	Results analysis . . . . .	48
5.4.1	Impact on the RU aspect . . . . .	48
5.4.2	Impact on the passenger aspect . . . . .	48
5.5	Conclusion . . . . .	49
<b>6</b>	<b>Conclusions and Discussions</b>	<b>50</b>
6.1	Conclusions . . . . .	51
6.2	Contributions . . . . .	53
6.2.1	Comparison with literature . . . . .	54
6.3	Discussions . . . . .	55
6.3.1	Reflections for the methodology . . . . .	55
6.3.2	Reflections for the case study . . . . .	55
6.3.3	Recommendations . . . . .	56
<b>A</b>	<b>Appendix</b>	<b>61</b>
A.1	Clustering process of the OD demand . . . . .	61
A.2	Clustering process of the frequency . . . . .	65
A.3	The demand matrix acquisition and consolidation . . . . .	67
A.4	Similarity calculation of matrices . . . . .	71
A.5	Acquire segment demand . . . . .	72



# List of Figures

1.1	The so-called “ten-minute train” between Rotterdam, Schiphol and Arnhem. Via:NS . . . . .	2
2.1	Chain effect phenomenon . . . . .	12
2.2	Clustering result example . . . . .	15
3.1	Sub-network geographical topology . . . . .	19
3.2	Total passenger flow of the whole Dutch railway network in 2022 . . . . .	20
4.1	The clustering result of station A - station B . . . . .	25
4.2	Histograms of the number of clusters of all ODs . . . . .	25
4.3	The clustering result of station C - station D . . . . .	26
4.4	The clustering result of station E - station F . . . . .	26
4.5	Base demand period frequency . . . . .	28
4.6	Example for selecting the value . . . . .	31
4.7	The change of total demand of designed service . . . . .	32
4.8	Manhattan distance . . . . .	33
4.9	The process of delineating periods with different demand characteristics . . . . .	36
5.1	The process of designing the demand-oriented line plans based on the base demand . . . . .	38
5.2	NS train types (Nederlandse Spoorwegen, 2022) . . . . .	42
5.3	Segment demand between big stations for matrix H (from 1 to 6) . . . . .	42
5.4	Line plan based on the peak hour demand . . . . .	43
5.5	Line plan based the base demand . . . . .	44
5.6	Line plan A . . . . .	44
5.7	Line plan B . . . . .	45
5.8	Line plan C . . . . .	45
5.9	Line plan D . . . . .	45
5.10	Line plan E . . . . .	46
5.11	Line plan F . . . . .	46
5.12	Line plan G . . . . .	46
6.1	The project process . . . . .	53

# List of Tables

2.1	Pros and cons for four methods . . . . .	13
3.1	The data format of the original dataset . . . . .	19
4.1	The data format of the new week's data . . . . .	24
4.2	15 periods over a week . . . . .	29
4.3	Covered base demand periods . . . . .	31
4.4	The sum of the values in each matrix . . . . .	34
5.1	Matrix A: The Monday/Wednesday morning peak . . . . .	39
5.2	Matrix B: Tuesday/Thursday morning peak . . . . .	39
5.3	Matrix C: Pre afternoon period from Monday to Thursday . . . . .	39
5.4	Matrix D: Monday/Wednesday afternoon peak . . . . .	40
5.5	Matrix E: Tuesday/Thursday afternoon peak . . . . .	40
5.6	Matrix F: Friday morning peak . . . . .	40
5.7	Matrix G: Friday afternoon . . . . .	40
5.8	Matrix H: Base demand periods . . . . .	40
5.9	The results of two scenarios . . . . .	47

# 1

## Introduction

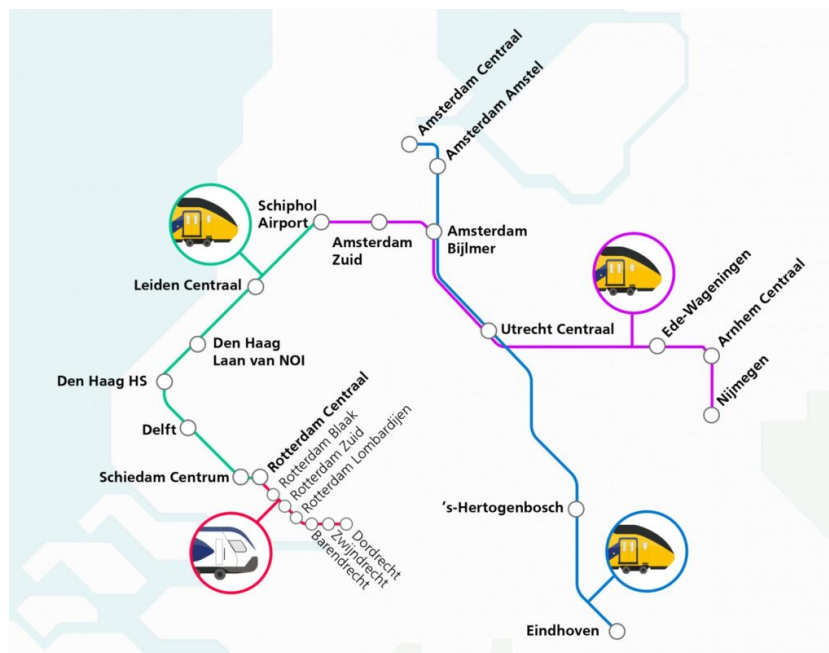
Chapter 1 introduces the graduation project. First of all, the timetable of the current Dutch railways and some possible directions for development are introduced. Based on this, the research question is proposed: How to develop a demand-oriented train service plan by using the railway passenger demand pattern? To answer this research question, the literature review is done to find the gap in current research. Finally, the thesis outline is given at the end of this chapter.

## 1.1. Background

The railway undertakings in the Netherlands provide a cyclic railway timetable for passengers. NS (Nederlandse Spoorwegen), as the principal passenger railway operator in the Netherlands, has been using the periodic timetable to provide service for railway passenger since 1970. The periodic schedule provides a regular interval service throughout the day, and thus a fixed departure time during each time unit cycle (such as per hour). The fixed departure time and the fact that trains are equally spaced over the hour provide convenience for most passengers, which makes the train service attractive.

However, although the periodic train service takes into account the needs of passengers during peak periods very well, it doesn't fully match the off-peak passenger demand. The off-peak passenger demand usually shows different characteristics, such as different volumes, as well as different OD pairs (origin and destination combinations). For example, people may not travel primarily for commuting purposes during off-peak periods; more people who travel for other purposes, such as entertainment, shopping, visiting relatives, etc., will take the train during those time periods.

As NS provides increasingly convenient train services, the need for periodic schedules is decreasing in recent years. With the increasing frequency of passenger trains, people usually don't need to wait long at the station. In 2022, three high frequency service lines were added, which are: Rotterdam-Schiphol, Schiphol-Arnhem and Rotterdam-Dordrecht (de Bruyn & Mestrum, 2021). In the figure 1.1, there will be a train departing every ten minutes between those big cities. Besides, the NS app on mobile brings a lot of convenience. People who have the demand for traveling can easily find suitable departing time and route through the app. Those changes decrease the demand for periodic timetable service to some degree, and provide a reasonable start of applying a flexible train schedule.



**Figure 1.1:** The so-called “ten-minute train” between Rotterdam, Schiphol and Arnhem. Via:NS

The existing train service in the Netherlands is mainly designed for peak-hour passenger demand, which possibly causes a waste of capacity in off-peak periods. While the periodic schedule also works well for off-peak demand and fully satisfies it, the lack of clarity about passenger travel patterns might lead to redundant train service. The redundant train service will cause unnecessary costs of using rolling stock and corresponding maintenance. Further, if the railway undertakings can acquire a clear perspective of passenger flow characteristics over a year or a month, or even a day, the railway undertakings can better schedule the train service according to those characteristics. The optimization of service will definitely bring benefits for railway companies, and the passenger travel experience can also be improved, such as less travel time.

## 1.2. Problem Description

To better identify passenger demand characteristics, i.e., how the passenger demand is distributed in different time periods, the method such as finding the homogeneous demand periods can be a good choice. There is already a lot of research on developing the method to find the homogeneous periods in the public transport field. Van der Knaap et al. (2022) developed a method to find periods of homogeneous railway passenger demand, and both the structure (the origin-destination combinations) and the volume of the passenger demand are considered. Mishalani et al. (2011) used a probability flow matrix to identify the homogeneous bus route passenger demand OD matrices throughout the day. Mahmoudzadeh & Wang (2020) used a clustering method, which considers both high student loads and graduate/undergraduate behaviors, to find the homogeneous periods of demand. The research helps improve the service of the shuttle bus in the university.

However, most research does tell us their ideas and methods about in which periods the demand pattern differs, but doesn't show how the demand changes in a day or a week. Based on the found homogeneous periods, further analysis can be done. For example, exploring how the passenger flow characteristics change between adjacent delineated homogeneous periods (how fast the flow changes, how the structure changes, etc.)

With the observation of the demand pattern change, a clearer perspective about how to adjust the train service can be acquired. Some operating strategies such as Short-turn (The train ends its run early at a station instead of continuing to the terminal) and Skip-stop (Trains skip stops at certain stations and travel directly from one station to another) can be used respectively to better match the passenger demand in different periods. For example, in the Netherlands, the Sprinter might skip some stops (in the existing schedule it stops at every stop along the corridor where it provides service), or the intercity train will be assigned more stops. In general, the train will have different operating schedule and will stop at different stops or stop at extra/fewer stops to adjust the service.

## 1.3. Research Questions

Based on the problem description, the research question is:

- How to develop a demand-oriented train service plan by using the railway passenger demand pattern?

According to the main research question, four sub questions can be generated:

- What is already known in the literature regarding analyzing the flow demand pattern in public transport demand?
- What method can be used to characterize the passenger demand patterns?
- What passenger demand patterns can be found when applying this method to concrete regions?
- How can the result of the analysis of the demand pattern be used for developing a demand-oriented train schedule?

## 1.4. Literature Review

In order to answer the main research question, the first thing is to explore methods that can characterize passenger flow; Secondly, some quantitative methods are needed to assess the accuracy of the observed flow demand characteristics. In this section, the base demand is introduced. Then two aspects are explored in current literature. The first one is characterizing the passenger flow demand; the second part is the way to evaluate the impact of both railway undertakings and passengers by the changed train service.

### 1.4.1. Base demand

An idea about designing the new train service on “base demand” is given by Bruijn et al. (2019). This base demand indicates the demand that exists all over the day or the week, instead of the peak hour passenger flow demand. The train service is then designed to satisfy the base demand, and the peak hour demand is met by additional or add-on peak hour services. One fact that supports this idea is that around 60%-70% of railway passenger take the trip during off-peak periods (Bruijn et al., 2019). If those passengers are seen as the base demand and the train service is designed based on it, some improvement in train service can be expected for most passengers.

However, how to reasonably define the base demand needs more discussions and considerations. If the lower bound of the base demand is too low, then most periods of the day or the week will require add-on service, which increases the complexity of designing the timetable; if the upper bound of the base demand is too high, then it comes back to the existing train service design thought, which is the peak hour demand based service. Therefore, a suitable base demand should enable the corresponding base train service to well serve passengers over the day or the week. At the same time, there will still be capacity for add-on train service to deal with the peak hour demand.

Van der Knaap et al. (2022) provides a good start for delineating the boundaries of different demand patterns. What the research has done is that they use hierarchical clustering to determine which periods are homogeneous over continuous time periods of a day or a week, based on the passenger flow volume and the structure (where passengers come from and where they finally go to). The homogeneous periods include different periods within a day and within a week. The result provides a base for deciding when the passenger demand pattern is different and when it is suitable to change the train service.

But to define the base demand, there still exists a problem: which period, or which periods are suitable for working as base demand periods? This base demand should be representative

and be able to satisfy what has been discussed about the base train service. In the meantime, within the based demand periods, the flow fluctuation should be as few as possible. To figure out which periods will be reasonable for the base demand, a further observation is needed to give a view about how demand differs in different time periods.

### 1.4.2. Flow characteristics

To have a better view regarding how passenger flow demand pattern be like over time periods, it is important to first transfer the flow data into useful information, which requires some approaches to directly indicate those patterns.

The record of passenger flow of different stations or different lines can generally describe the fluctuation of flow volume. Yu et al. (2019) use the smart card data of the metro passenger flow to perform the analysis of demand space-time characteristics. For time aspect, some indicators such as the whole day average passenger flow or the peak hour coefficient of passenger flow of different lines are used to show the fluctuations and variability. For the space or structure aspect, thermodynamic charts are used to show the flow change in different stations and lines within a day. Jinjing et al. (2020) used the accumulative variance volume to indicate the congestion degree combined with the flow fluctuation characteristics. The interquartile-range (IQR) method is used for setting the anomaly threshold so as to generate the corresponding warning level when the detected passenger flow changes and the calculated indicator exceeds the threshold. Zhao et al. (2017) use simple statistical indicators to show the passenger flow characteristics of Chongqing Rail transit system, such as the commuter identification, the commuter line OD flow, and the commuter station passenger flow.

Some indicators are used to distinguish the level or importance of the station. Limtanakool et al. (2007) use two indices to characterize the structure of the network. The directional dominance index is defined as the rate of interactions of one node to average interactions of all other nodes. This indicates whether one node is dominating in the network and how much this node interacts with other nodes. Besides, the relative strength is used to measure the proportion of the flow interactions of one link between two nodes among all interactions in the network. And the link symmetry is further used for considering the direction of the interaction on each link. Similarly, Zhang & Ng (2021) use the rich-club coefficient to indicate the characteristics of the network that includes the “rich” nodes (nodes that are highly connected), and further tell whether the subnetwork composed by those nodes is influential and whether those nodes intensely interact with one another.

Visualization is one popular way to show the flow characteristics of the network. Sun et al. (2016) use visualization methods to better show the change of passenger flow pattern in metro. Spatial distribution over the network and the passenger boarding distribution over the time are used for showing the passenger travel pattern over the space and the time. The directional imbalance, which is the rate of inbound flow over outbound flow of each station in the network, is used to further observe the passenger flow. Zhang & Ng (2021) chooses to use the box-whisker plot to indicate the passenger flow fluctuations over different periods. Besides, the inbound and outbound passenger flow of each station over different time periods are expressed in the geographic map.

Clustering methods are used a lot and they are mainly used to partition a dataset into classes or clusters according to a specific criterion (e.g. distance). Based on specific OD pairs, the volume data over a day or a week can also be clustered into different clusters with different characteristics, which contributes to finding the base demand periods. Hartigan et al. (1979) come up with the K-means algorithm, which is an iterative solving algorithm for cluster analysis. Distance is used as a similarity metric so that K classes in a given dataset are found and the center of each class is obtained from the mean of all values in the class. The Jenks Natural Breaks method was proposed by Jenks (1967), and it seeks to reduce intra-class variance and maximize inter-class variance. The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a method proposed by Ester et al. (1996). It is a density-based clustering nonparametric algorithm: for a set of nodes in a space, it will combine nodes that are closely stacked together and mark those that are individually located in low-density regions as anomalies.

### 1.4.3. Impact analysis

To explore what impacts can the demand-oriented method for designing the train service can bring, an important thing is to find an evaluation methodology, to analyse the impact for both passengers and railway undertakers. This can start from exploring some ways in current research to evaluate the train service.

Goverde & Hansen (2013) use some indicators to evaluate the timetable quality and performance. Five indicators are included for showing the performance of specific timetable, which are: infrastructure occupation, timetable stability, feasibility, robustness, and resilience. The infrastructure occupation is the percentage of time periods required for train operations on specific railway tracks during specific timetable pattern. The feasibility means the ability of all trains to follow their scheduled train path. The stability is the ability of a timetable to absorb delays so that delayed trains are able to return to their intended train paths. The robustness indicates the ability of the timetable to withstand design errors, parameter changes and changing operating conditions. The resilience indicates the flexibility to use scheduling to prevent or reduce secondary delays.

There are many intuitive indicators to simply reflect the performance of the designed schedule. Castillo et al. (2011) choose minimizing the sum of the relative travel times as the objective when dealing with the timetabling problem. Brännlund et al. (1998) use the maximum profit as the objective to design the timetable, and the schedule decisions are based on the fact that different types of train services will have different values. Mu & Dessouky (2011) set the objective to be minimum total train delay when considering scheduling issues. To improve the robustness of the timetable, Vansteenwegen & Van Oudheusden (2007) use the objective of minimizing the waiting cost of passengers to improve the timetable schedule. The waiting cost is measured by the waiting time, and different types of waiting time will correspond to different amount of waiting cost. For example, if the train running time is longer than the ideal running time and it caused delay, the waiting time for passengers on trains is weighted 1.0; for passengers who wait on the platform, the waiting time caused by the delayed by the train is weighted 2.5.



#### **1.4.4. Conclusion**

To summarize, the base demand must have the following characteristics: it is generally distributed during off-peak periods and represents the passenger flow that is always present in the network; the demand characteristics are similar during the time period in which the base demand is distributed, which includes the size and structure of the demand. Based on these characteristics, the clustering method is considered a suitable approach to find these flows with similarities. Clustering methods can effectively identify data with the same characteristics, and this method is also used by Van der Knaap et al. (2022) to identify periods with different demand characteristics in the railway area. Therefore, in the methodology part, suitable clustering methods will be further explored.

The evaluation of any train service can be done in terms of both the passengers and the Railway undertakers (RU). For passengers, the total travel time as well as the waiting time can be calculated and the results often represent the impact on passenger travel. For RU, train operating costs are a commonly used indicator, which can often be obtained from data such as frequency of departures, number of stops, etc. When exploring the impact of the demand-oriented service, these two aspects can be mainly discussed.

### **1.5. Thesis Outline**

In the following content, chapter 2 characterizes the methodology to be used for analyzing the passenger demand pattern. In chapter 3, the scope of the case study is mainly described, mainly including the study area and data scope, while some data preparation is done to provide input for the application of the methodology that follows. Chapter 4 section describes the application process of the methodology, delineating several periods characterized by different requirements. In chapter 5, a case study is performed for further data analysis. Finally, conclusions and discussions are discussed in Chapter 6.

# 2

## Methodology

In this chapter, we delve into methodologies for finding the flow demand pattern. Since the data is presented in the form of passenger volume for each OD pair in half-hourly intervals, clustering methods are considered to be suitable for this task. We present and compare four widely known methods, namely K-means, bisecting K-means, DBSCAN, and Hierarchical clustering. The silhouette coefficient is used to obtain the optimal clustering results without requiring any initial values to be set. In addition, we define the concept of base demand and explain it in the form of data.

## 2.1. Review of clustering methods

The analysis of passenger flow patterns starts with the passenger flow data contained in each OD pair in the network. The general idea is: For each OD pair, the passenger flow demand in each half hour interval will be seen as a data point. The points with similar values can be seen to have similar demand patterns. If a significant portion of points have similar values, then the periods they represent can be considered as base demand periods. The basis for such consideration is that a train service based on a base demand design should be able to cover as many time periods as possible.

Clustering method is considered to be suitable to achieve this objective. Clustering is to partition a data set into different classes or clusters according to a specific criterion (e.g. distance), so that the similarity of data objects within the same cluster is as large as possible, while the difference of data objects not in the same cluster is also as large as possible. With the help of clustering methods, the previously proposed ideas can be better implemented. In this case, each OD pair is able to get its own base demand through a week, and the base demand is different between each pair.

To perform the clustering process, a suitable method should be chosen. Several commonly used clustering methods are described below, and their performances are discussed.

### K-means

The first method for clustering is K-means, which is firstly proposed by Hartigan et al. (1979). Since the basis of clustering analysis is one-dimensional passenger flow data, whether for filtering the right week or finding the right base demand, the relatively simple and frequently used K-means algorithm can do the job. The logic of the algorithm is:

- Create  $k$  points as initial clustering centers, the  $k$  value usually has to be set manually.
- For each sample  $X_i$  in the dataset (where  $X_i \in X = [X_1, X_2, \dots, X_n]$ ), calculate its distance to  $k$  cluster centers and assign it to the center with the smallest distance.
- For each clusters  $A_i$  (where  $A_i \in A = [A_1, A_2, \dots, A_k]$ ), recalculate its clustering center, which is the center of mass of all samples belonging to the class. This is indicated by the average value of all sample points.
- The last two steps are repeated until the termination condition is satisfied. The termination conditions can be: no (or a minimum number of) points are reassigned to different clusters, no (or a minimum number of) cluster centers change again, or the error sum of squares is locally minimal.

The K-means algorithm has several advantages: the complexity of the algorithm is low and easy to understand and use. However, the value of  $k$  needs to be set manually and different values yield different results. That is, the K-means algorithm is sensitive to the initial cluster center, and different selections of  $k$  value will yield different results. Besides, the goal of the algorithm is a local optimum, which leads to that the result is not a global optimum.

### Bisecting K-means

A measure of clustering effectiveness is SSE (Sum of Squared Error), which indicates the sum of the squared distances between each point in a cluster and the cluster's center. Bisecting

K-means is an improved algorithm for the defect that K-means algorithm will fall into local optimum (Steinbach et al., 2000). The algorithm is based on the principle of SSE minimization. First, all the data points are considered as one cluster, and then the cluster is divided into two, after which one of the clusters is selected to continue the division, and the choice of which cluster to divide depends on whether the value of SSE can be minimized for its division. Assuming the data set to be  $A$ , and the logic of this algorithm is:

- Take all objects as a cluster  $A_p$ , and determine the number of dichotomous trials  $m$  at this time.
- Take the cluster with the largest SSE in the dataset and perform dichotomous trials  $m$  times: call the k-means clustering algorithm, take  $k = 2$ , divide the dataset into two clusters:  $A_{i1}$ ,  $A_{i2}$ , and get a total of  $m$  dichotomous results set  $B = [B_1, B_2, \dots, B_m]$ , where  $B_i = [A_{i1}, A_{i2}]$ , where  $A_{i1}$  and  $A_{i2}$  are two clusters obtained from dichotomous trials.
- Calculate the total SSE values of the two clusters obtained by each division method in the set  $B$  of dichotomous results in the previous step, select the result obtained by the dichotomous method with the smallest total SSE:  $B_j = [A_{j1}, A_{j2}]$ , and add clusters  $A_{j1}$  and  $A_{j2}$  to the set  $A$ , and remove  $A_p$  from  $A$ .
- Repeat last two steps until  $k$  clusters are obtained, i.e., there are  $k$  clusters in the set  $A$ .

Compared with the K-means algorithm, the bisecting K-means method mainly improves the problem of uncertainty of clustering results caused by the randomness of choosing the initial center of mass in the K-means algorithm. However, although the bisecting K-means can overcome the limitation that the K-means converge to a local minimum to some degree, it does not guarantee convergence to the global optimum.

## DBSCAN

Both K-means and bisecting K-means are algorithms that belong to the delineated clustering method. This type of algorithm requires specifying the number of cluster classes or cluster centers in advance, and iterating until the final goal of “points within a cluster are close enough and points between clusters are far enough” is achieved. However, for non-convex shaped data points, it is easy to make mistakes, and this is where the density-based clustering method is needed. Ester et al. (1996) proposed a data clustering algorithm called Density-based spatial clustering of applications with noise (DBSCAN).

DBSCAN is a density-based clustering algorithm that works on the assumption that clusters are dense regions in space, separated by low-density regions. The method requires the definition of two parameters  $E$  (epsilon) and  $M$  (minPoints) to denote the neighborhood radius of the density and the neighborhood density threshold, respectively. The logic of this algorithm is:

- Mark all data points as unvisited.
- Randomly select a point  $p$  among all unvisited points and mark it as visited.
- If there are at least  $M$  points in the neighborhood range  $E$  of the point  $p$ , a new cluster  $C$  is created and then  $p$  is placed in  $C$ . If not, then  $p$  is marked as a noise point.
- For each point within range  $E$ : if this point  $p_i$  is unvisited, mark it as visited; if there are at least  $M$  points in the neighborhood  $E$  of  $p_i$ , add these points into the cluster  $C$ .

- Repeat last three steps until all points are visited.

In general, for DBSCAN, it doesn't need to set the number of clusters in advance. Also, DBSCAN can distinguish noise points and is insensitive to outliers in the dataset. However, it does not work well for data aggregation with uneven density. In the DBSCAN algorithm, a uniform epsilon value is used. When the data density is not uniform, if a smaller epsilon value is set, the node density in the sparser clusters will be smaller than `minPoints` and will be considered as boundary points and not used for further expansion; if a larger epsilon value is set, the denser and closer clusters are easily classified as the same cluster.

## Hierarchical clustering

The idea of sum of squares of deviations (Ward's method) was proposed by Ward Jr (1963) in 1963 based on ANOVA, and the algorithm of hierarchical clustering was formed in 1967. Jain et al. (1999) provided a systematic explanation and summary of the clustering algorithm, where the development of hierarchical clustering method is introduced. Hierarchical clustering divides the dataset into clusters in one layer, and the clusters generated in the later layer are based on the results of the previous layer. Hierarchical clustering algorithms are generally divided into two categories:

- Agglomerative hierarchical clustering: also known as bottom-up hierarchical clustering, each object is a cluster at the beginning, and each time the two closest clusters are combined to generate a new cluster according to certain criteria, and so on, until eventually all objects belong to one same cluster.
- Divisive hierarchical clustering: Also known as top-down hierarchical clustering, all objects belong to one cluster at the beginning, and each time a cluster is divided into multiple clusters according to certain criteria, and so on, until each object is the same cluster.

Hierarchical clustering algorithms have been used and studied for decades, and agglomerative algorithms are used more than divisive algorithms. Here the logic of agglomerative clustering algorithm is introduced:

- At the beginning, each point is viewed as a cluster and the distance matrix  $D$  is calculated, where element  $D_{ij}$  is the distance between point  $i$  and point  $j$ .
- Iterate through the distance matrix  $D$  and find the minimum value (except for the elements on the diagonal). The two clusters corresponding to this minimum element are merged into a new cluster and the distance matrix  $D$  is updated according to the distance metric.
- Repeat last step until there is only one cluster left.

Then it is important to decide the criteria of the distance between clusters. The generally used distance measures, which are also called linkage measures, include three: single linkage, complete linkage and average linkage. A lot of comparative studies so far have been done to analyse those methods, and here we refer to Jarman (2020) and Sharma et al. (2019).

- Single-linkage defines the distance between two clusters as the distance between the two closest points between the two clusters. And because of this, it is susceptible to outliers and noise. At the same time, because the calculation of single link is too simple, it is

prone to chain reaction, i.e., a long chain is formed first, and then individual points are added gradually, so that all points are in the same cluster.

- Complete-linkage method calculates the distance between clusters as the distance between the two farthest data points from each cluster. This method is less sensitive to outliers and noise, but it can be influenced by local minimums. Similar to single linkage, complete linkage can also result in chain reactions.
- Average-linkage is a compromise between Single-linkage and Complete-linkage methods, which defines the distance between two clusters as the average of the distances of all points between the two clusters. This method considers all distances between data points, resulting in less sensitivity to outliers and noise, and less chain reactions. However, this method takes longer to calculate due to the need to consider all distances between data points, and it can also be influenced by noise and local minimums.

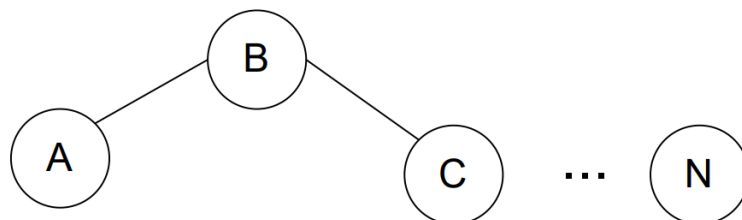
Based on the features of three linkage measures, the average-linkage is considered more appropriate to be used in this case, which help to solve the chain effect problem. Hierarchical clustering algorithm is a greedy algorithm, as each merge or division is based on some locally optimal choice.

It has many advantages: Compared to non-hierarchical clustering methods such as K-means clustering, hierarchical clustering does not require a pre-specified number of clusters, which makes it more flexible; Hierarchical clustering can generate a tree-like structure, which helps visualize and interpret the organizational relationships of the data. However, it has a high computational complexity, especially when the dataset is large, which can increase the computation time significantly; the result is influenced by the chosen distance metric, and different distance metrics may lead to different clustering results.

## 2.2. Method Selection

In order to select a suitable one of the methods presented above, some comparisons are needed to make a selection based on their characteristics.

The chain effect is further explained here. The methods introduced before (K-means, bisecting K-means and DBSACN) can indeed obtain good results when the dataset is not too complex. But there exists a chain effect phenomenon for all of them, for example: A is similar to B, B is similar to C, then it will be easier to cluster them together when clustering A, B, and C. But if A is not similar to C, it will cause a clustering error. In severe cases, this error can be passed on forever.



**Figure 2.1:** Chain effect phenomenon

**Table 2.1:** Pros and cons for four methods

	Pros	Cons
K-means	Low complexity	Initial clusters needed to be set manually; local optimum; sensitive to initial cluster center; chain effect
Bisecting K-means	Low complexity; no more sensitive	Still not global optimum; chain effect
DBSCAN	No need to set initial cluster number	Doesn't work well for uneven density data; chain effect
Hierarchical Clustering (average linkage)	Overcome the chain effect; no initial values needed	High complexity

In table 2.1, all pros and cons for four methods are concluded. After analyzing and comparing the above four methods, the hierarchical clustering methodology with the average linkage measure is considered as the most suitable one for the adoption of clustering. It overcomes the defects possessed by some other algorithms (such as chain effects) and is not influenced by any initial setting values. Of course, in the case of adopting hierarchical clustering, a method for finding the optimal clustering result is also needed. In the next subsection, more information regarding hierarchical clustering method and a detailed description of the method to evaluate the clustering result will be given.

### 2.3. Silhouette coefficient

The next problem to be solved is finding a suitable method to determine the optimal clustering result. Here we refer to Rousseeuw (1987), where the silhouette coefficient is considered to be a suitable method that can be used to achieve this goal. Silhouette coefficient was first introduced by Rousseeuw (1987) to evaluate the quality of clustering result. The logic of this method is:

- Assume all samples  $i \in I$ . Calculate the average distance  $a_i$  from sample  $i$  to other samples in the same cluster. The smaller  $a_i$  is, the more sample  $i$  should be clustered into that cluster.  $a_i$  is called as the intra-cluster dissimilarity of sample  $i$ .

Assume that there are already  $n$  clusters, and for sample  $i$ , it belongs to cluster  $C_{i'}$ . Let  $|C_{i'}|$  be the number of samples in cluster  $C_{i'}$ . The indicator can be calculated as:

$$a_i = \frac{1}{|C_{i'}| - 1} \sum_{j \in C_{i'}, i \neq j} d(i, j) \quad (2.1)$$

$d(i, j)$  here means the distance between sample  $i$  and  $j$  in the cluster  $C_{i'}$ .

- The average distance  $b_{ij}$  of sample  $i$  to all samples of some other cluster  $C_j$  is calculated and is called the dissimilarity of sample  $i$  to cluster  $C_j$ .

Let sample  $i \in C_{i'}$ , and  $C_k$  indicates the  $k^{\text{th}}$  cluster in all clusters. Define the inter-cluster dissimilarity  $b_i$  of sample  $i$ :

$$b_i = \min_{C_k: 1 \leq k \leq n, k \neq i'} \left\{ \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j) \right\} \quad (2.2)$$

- Based on the intra-cluster dissimilarity  $a_i$  and inter-cluster dissimilarity  $b_i$  of sample  $i$ , define the Silhouette coefficient of sample  $i$ :

$$s_i = \frac{b(i) - a(i)}{\max \{a(i), b(i)\}} \quad (2.3)$$

If  $s_i$  is close to 1, it means that sample  $i$  is clustered well; if  $s_i$  is close to -1, it means that sample  $i$  is more deserving of classification into another cluster; if  $s_i$  is close to 0, it means that sample  $i$  is on the boundary of two clusters. The mean value of  $s_i$  for all samples is called the Silhouette coefficient of the clustering result, which is a measure of whether the cluster is reasonable and valid. Let  $|I|$  be the number of samples that are clustered. Here the mean value is indicated as SC:

$$SC = \frac{\sum_{i \in I} s_i}{|I|} \quad (2.4)$$

Then the criteria to acquire a good clustering result will be the result with the highest SC value among all results. In this way, for each dataset (which will be the passenger flow data for each OD pair), one optimal clustering result can be found.

## 2.4. Definition of base demand

The hierarchical clustering method described previously, as well as the silhouette coefficients, are able to cluster the dataset and come up with an optimal clustering result. At the beginning of section 2.1 we say if a significant portion of points have similar values, then the periods they represent can be seen as base demand periods. Among all clusters, we can define that the cluster that owns the largest number of points can be considered as the base demand periods cluster. For this cluster, the largest value within the cluster will be considered as the base demand for this OD pair. However, the result is still on OD level, which still cannot provide a base for designing the base demand service for the network. Due to different characteristics of flow patterns of different OD pairs, the generated clusters can also show differences.

The ideal situation is that most defined base demand periods of different OD pairs will overlap, then it is reasonably to say those are base demand periods for the whole network. But if the number of periods is too few, here we call them as “common base demand periods”, then it is not a suitable result.

So now, the problems fall at finding a suitable threshold, which indicates what is a reasonable number of common base demand periods. Since the concept of base demand itself is relatively new, there is no research that can help to delineate a reasonable range in a day, as far as we know. However, from the Railway undertakings' point of view, if a service based on base demand can serve more than half portion of a certain time period, it can be considered a successfully



designed service. We assume there is a certain period range with a threshold, then there will be two situations:

- The common base demand periods are enough, which means the number of common base demand periods is larger than the threshold, then it will be considered a good result. We can say those are the base demand period for the sub-network.
- The common base demand periods are not enough and the number is smaller than the threshold. Due to different characteristics of flow patterns, the clustering result will also be different between OD pairs. If the first ideal scenario doesn't show up, then the next step is to consider transferring more periods into common periods, to come up with at least enough base demand periods for the network.

Here figure 2.2 is used to indicate this situation. Assume that there are six OD pairs and six time periods. The colored cells are the found base demand periods for each OD pair. Then the base demand period for the whole network is only the second time period. Clearly it is not enough, the base demand service should not just serve one period, leading to too complex an add-on or subtract service in other periods.

OD pair	Time period					
	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						

**Figure 2.2:** Clustering result example

Obviously there should be a plan to deal with the second situation. The idea is to convert some periods into common base demand periods, to satisfy at least the threshold number of base demand periods for the network. This process will bring some costs, and this is how we evaluate which period is suitable to be converted.

Take the first period for an example. If this period has to be converted into the common base demand period, then for OD pair 2, 5 and 6, they all have to be converted. Take the OD pair 2 for an instance, if the base demand are 1000, and the first period demand of this OD pair is 500, then to cover the first period passenger travel demand, there will be a 500 passenger capacity waste. If the first period demand is 1500, then to cover the this period, the base demand should at least be 1500, which leads to  $(1500-1000) * 3 = 1500$  passenger capacity waste (for period 2, 3 and 4). The capacity waste can be seen as a certain type of cost.

In this way, the cost for OD pair 5 and 6 can also be calculated, which means the cost for converting period 1 to common base demand period can be got. Then we know for each period, what is the cost to convert then into common base demand periods. The minimum value of cost will be chosen, until the number of common base demand periods attains the threshold.

## 2.5. Conclusion

This chapter gives the main methods to characterize the passenger flow in the sub-network. By means of hierarchical clustering and silhouette coefficient methods and the definition of base demand, it is possible to obtain the base demand time periods for each OD pair and, consequently, for the whole sub-network.

However, several problems still exist before performing the application. The first one is to define a study case. A representative study case should be chosen, for example the scope of the sub-network. The range of time should also be considered carefully. Although the passenger flow demand shows similarity for every week, there are still a lot of demand fluctuations between weeks, because of unexpected issues such as strikes, holidays and Covid-19. So it is necessary to filter some outliers and select the representative time period.

In the next section, the study case and the data preparation steps will be introduced. The format of the dataset will be explained in detail so as to describe how the methodologies will be applied.

# 3

## Case Study

In this chapter, the case study is introduced, including both the study region and the data format. Some analysis regarding the network flow pattern is given and the process of the data preparation is introduced.

### 3.1. Case Study

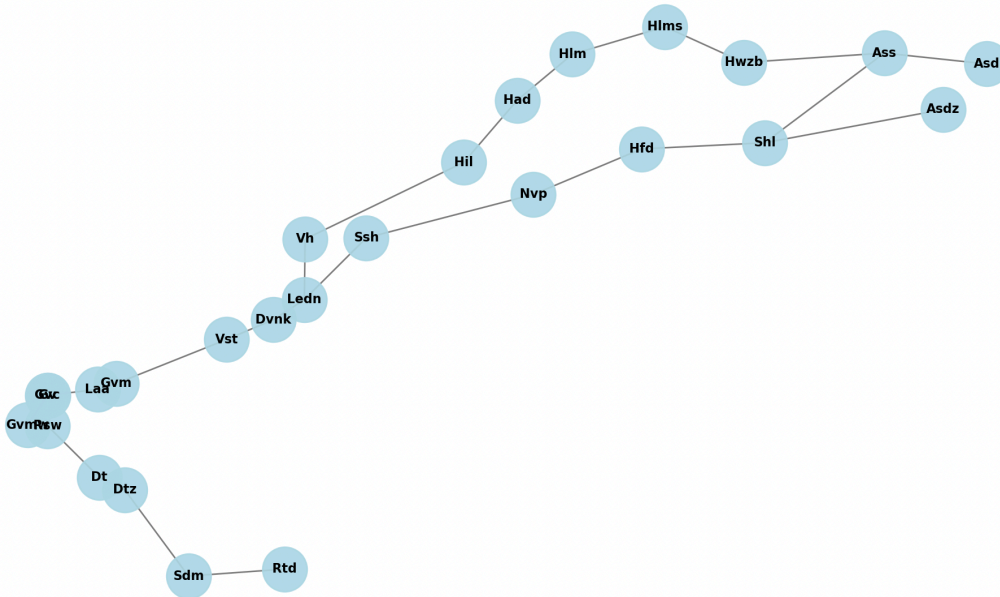
In order to observe the distribution of passenger flow patterns in the network, a study region should first be selected. Since the data contained in the entire Dutch rail network would be too complex, it is necessary to select a representative sub-network to perform the research. This sub-network should have a certain complexity, containing several lines in it, and the passenger flow patterns over different OD pairs should be diverse. These characteristics enable the methodology applied to the sub-network and the conclusions drawn to be also applicable to the network as a whole, which ensures the significance of the study.

A study region is chosen and it is a sub-network of the Netherlands railway, which contains 26 stations in total:

Number	Station	Abbreviation
1	Schiphol Airport	Shl
2	Hoofddorp	Hfd
3	Nieuw Vennep	Nvp
4	Amsterdam Sloterdijk	Ass
5	Amsterdam Zuid	Asdz
6	Amsterdam Centraal	Asd
7	Haarlem	Hlm
8	Heemstede-Aerdenhout	Had
9	Leiden Centraal	Ledn
10	De Vink	Dvnk
11	Voorschoten	Vst
12	Den Haag HS	Gv
13	Den Haag Centraal	Gvc
14	Den Haag Laan van NOI	Laa
15	Den Haag Mariahoeve	Gvm
16	Rijswijk	Rsw
17	Delft	Dt
18	Delft Zuid	Dtz
19	Schiedam Centrum	Sdm
20	Rotterdam Centraal	Rtd
21	Den Haag Moerwijk	Gvmw
22	Halfweg-Zwanenburg	Hwzb
23	Haarlem Spaarnwoude	Hlms
24	Hillegom	Hil
25	Voorhout	Vh
26	Sassenheim	Ssh

This region includes different types of stations. It has some big stations like Amsterdam Centraal and Rotterdam Centraal, some small stations like Delft and Delft Zuid (Delft campus). Schiphol Airport provides a rather different flow patterns compared with other stations, because the airport there will continually provide and attract continuing flow. At the same time, this sub-network is a busy region within the whole railway network. These complexities guarantee this area to be representative.

The geographical topology of this sub-network is shown: each station is represented by a point, and stations with rail connections between them are connected by straight lines.



**Figure 3.1:** Sub-network geographical topology

The raw data contains the travel information for each recorded passenger (anonymized) in 2022. Each data entry is in the form of: a record of the number of commuters per OD pair, in half-hour increments. The format of the dataset is shown in table 3.1:

**Table 3.1:** The data format of the original dataset

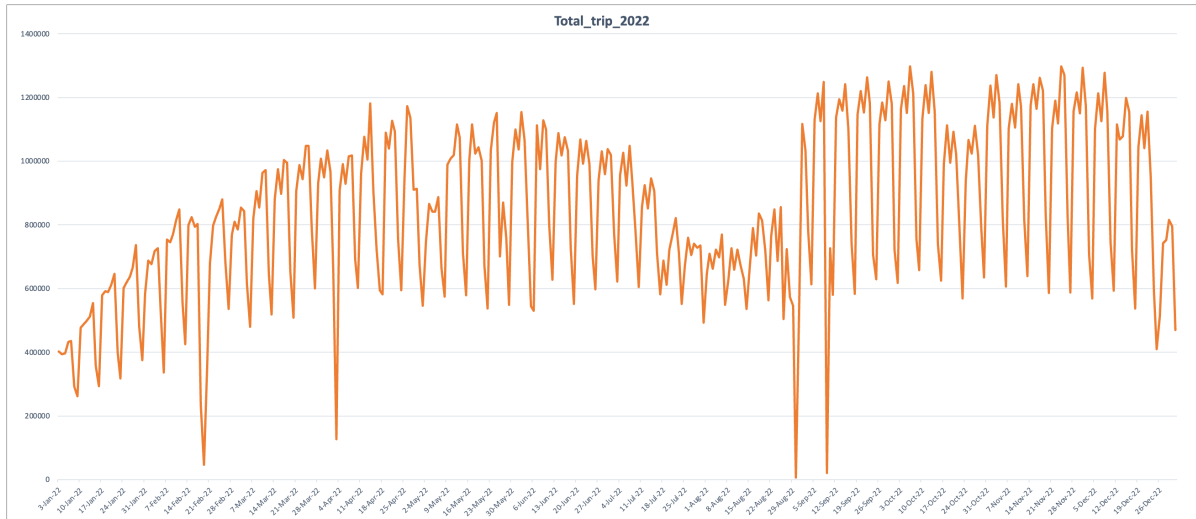
Date	09062022
Day of the week	5
Time of the day	07:00
Class level	2
Origin code	230
Destination code	239
The number of passenger	47.78622

## 3.2. Data preparation

For the study data used as input, the inclusion of sub-network passenger flow data within a week is a relatively ideal scope for the study. This is because the weekly traffic patterns are similar over a long period of time (e.g., a year). The current schedules are all designed on a week-by-week basis as well. For each week, the flow pattern differs as time goes by, due to different reasons. Thus, the representative passenger flow data over a week should be decided, before applying any methods to the data. This preparation aims at excluding the effect of outliers, for example, the rather low value when there are big strikes, or the relatively low value because of Covid-19.

The data used as a case study for this research includes passenger flow data for the Dutch rail network for the whole year 2022. The forms of data described in the previous section were recorded by smart cards. The data is provided by the Dutch railway undertaking NS. NS carries much fewer passengers in 2020 and 2021 than in previous years. Ridership begins to pick up again in 2022 due to the end of the Covid-19 pandemic (Nederlandse Spoorwegen, 2022).

The step of data preparation aims at selecting a representative week to perform the research. To do this, we can start by looking at the total network flow for the whole year 2022 in figure 3.2.



**Figure 3.2:** Total passenger flow of the whole Dutch railway network in 2022

The figure was generated by counting the total number of passenger trips (the trip indicates one person travels from one station to another station by train) between all stations in the network for each day in 2022. As can be seen, there is a certain similarity in flow patterns within each week. Weekday flow throughout the network often peaks on Tuesdays and Thursdays, while Monday, Wednesday and Friday traffic is lower in comparison, creating an "M" shaped traffic trend. Weekend traffic is lower on both days compared to all weekdays.

The figure shows a slow trend of increasing passenger flow in the early three months of 2022, i.e. January, February and March. This is due to the end of the pandemic and the return of people's lives to their usual patterns, with a gradual pick-up in travel volume as a result. There were four extremely low flow days throughout the year, in February, April, August and September, due to large strikes. At the same time, there are obvious changes in the passenger flow patterns of some holidays. On the one hand, there is a certain decrease in overall traffic; on the other hand, the traffic pattern during the week no longer shows a clear "M" shape. Some typical holidays in the Netherlands include Easter (April 18), King's Day (April 27), summer holidays (July and August), autumn holidays (two weeks starting October 17), and Christmas (starting December 25). Furthermore, passenger flow also shows different characteristics between the summer and fall/winter seasons. Although the weekly traffic fluctuations are very similar, the fall and winter seasons will have greater volume overall compared to the summer.

We plan to get a week's flow information that will give a more representative picture of the usual flow pattern performance of the network of the fall and winter seasons. In this way, the beginning filtering steps include:

1. Months before fall will be excluded, and the stop point will be the end of summer holiday, which is 2022/09/04.
2. The Christmas holidays are excluded. Consider the flow pattern fluctuates from the 12<sup>th</sup>, then the period 2022/12/12 – 2022/12/31 will be excluded. (The week from 12<sup>th</sup> -19<sup>th</sup> is not Christmas yet, but there is already an obvious change in the demand pattern of this week, so it is excluded also).
3. The fall holiday is excluded: from 2022/10/17 to 2022/10/30.
4. A big strike day is excluded, which is 2022/09/09. In order to reach a new collective labor agreement, trade unions FNV, CNV and VVMC announced that on this day NS workers will go on strike and that there will be no NS train service throughout the Netherlands all day due to industrial action by NS employees.
5. The night time is excluded. Because most passengers choose to travel during the daytime, so only the period 6:00 to 22:00 of every day is included. Then for every day, there will be 8 hours excluded.
6. The number of first and second class tickets are summed up, to count all people in that time period.

So finally, the chosen period will be: 2022/09/05 to 2022/09/08, 2022/09/10 to 2022/10/16, 2022/10/31 to 2022/12/11. The days included in this period are considered representative.

The next step is to find a week that can be representative during this period. Now we know that, the observation will be based on the whole sub-network level (where 26 stations are included), and the analysis will be based on OD-level flow data. So for each OD pair and each time period (half an hour), the median value of the flow demand can be a representative value to indicate the typical demand pattern.

To illustrate how this step is performed, let us look at the Monday. First we select all Monday flow data in the research period. We assume that there are  $k$  Mondays in total. Then for each time interval, such as the period 6:00-6:30, there will also be  $k$  corresponding periods. Among the flow values in these intervals, we choose the median as the representative value, creating a typical demand pattern for each time interval within a day. By combining these typical patterns across all time intervals, we obtain a daily demand pattern that reflects the network's general demand. The resulting patterns for all origin-destination pairs are then used to develop a weekly demand flow pattern suitable for exploring the base demand period.

### 3.3. Conclusion

This chapter describes a specific case study, including the scope of the network and the form of the data. Some data processing steps are done to prepare the data for the methodology application. In this way, a typical autumn/winter flow characteristics data can be obtained, which contains a week's passenger flow of the network.

In the next chapter, the application steps of the methodology are described, as well as the result obtained.



# 4

## Division of the periods with different demand characteristics

In this section, the methodology is applied to the dataset. Several constraints are added to address several shortcomings of the methodology. According to the different demand pattern shown by the clustering result, 15 demand matrices representing different demand patterns are initially divided. By comparing the observations of the 15 matrices divided, 8 periods representing different demand patterns during the week were finally merged. Comparing the existing train service plan to the new demand pattern periods, some observations and analysis were made.

## 4.1. Methodology Testing

Before formally applying the methodology, some tests should be completed to check whether the methodology logic is sound. Based on this idea, the hierarchical clustering methodology with the average linkage measure is applied to the new week data obtained from the data preparation.

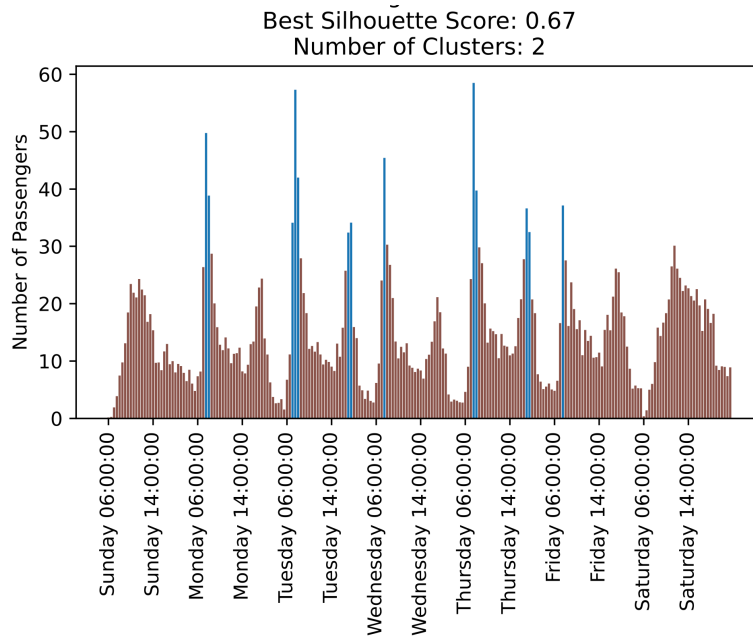
**Table 4.1:** The data format of the new week's data

Day of week	Time	Origin Code	Destination Code	Number of passengers
1	6:00	230	231	1.855272035
1	6:00	230	232	0.129766938
1	6:00	230	234	4.383570109
1	6:00	230	237	3.280628065
1	6:00	230	239	33.69149754
..	...	...	...	...

Table 4.1 indicates the data format of the new week's dataset. The dataset includes five columns, which records the information of the day of the week, the time of the day, the origin station and destination station code, and the number of passengers. Totally there 130340 data entries, and if there are no passengers travelling between some OD pair in some period, it will be recorded as 0. Every entry of data is seen as a data point.

The algorithm will go through situations of different number of clusters, and the range for the number of clusters is from 2 to the number of points of this OD pair. The Silhouette Score will be calculated for each result with different clusters. The optimal clustering result is selected according to the SC: the value that is close to 1 represents the optimal clustering result.

The output will be the clustering result for each OD pair. The results of clustering are represented by colored bar graphs, where different colors represent different clusters. One clustering result is shown here to illustrate the output. Because of the confidentiality of the data, different letters are used here to indicate different station names. Figure 4.1 indicates the clustering result of the OD pair: station A - station B. The blue bar represents one cluster and the brown bar represents another cluster.

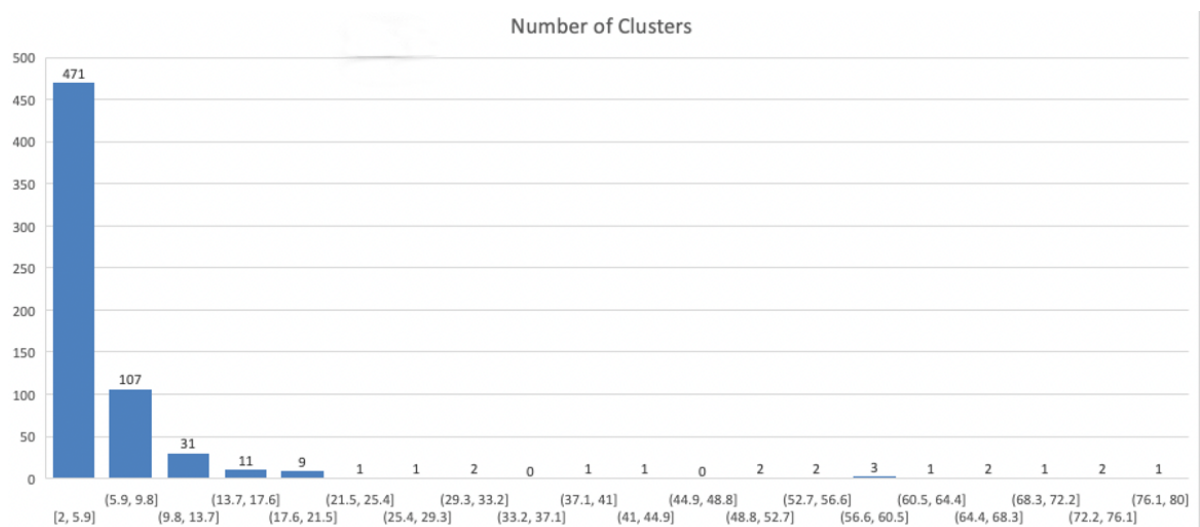


**Figure 4.1:** The clustering result of station A - station B

Totally there should be 650 results, because there are 26 stations in the sub-network, which brings  $26 * (26 - 1) = 650$  combinations. But the result only show 649 OD pairs, which is due to the one OD pair has only one period of data.

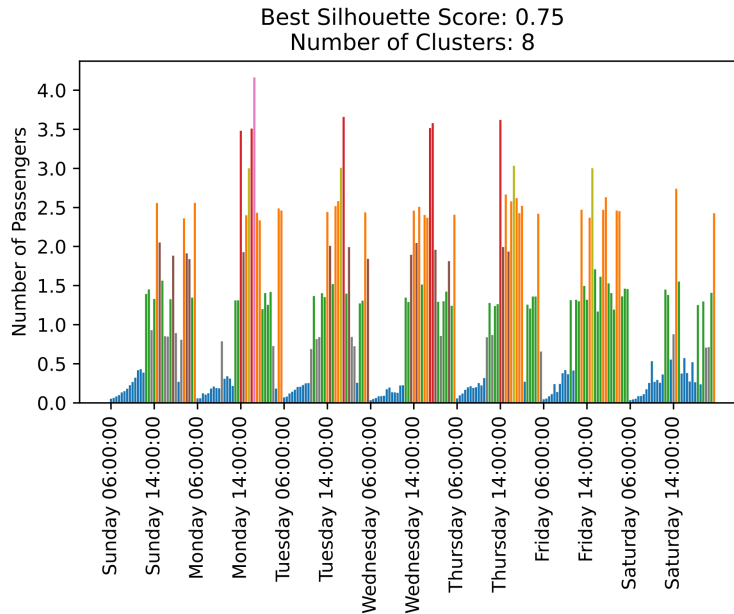
Different OD pairs will have different clustering results, and for each OD, their optimal clustering results are recorded, where the main parameters include the SC and the number of clusters. In figure 4.2, a histogram is used to show the number of clusters for all OD pairs.

As figure 4.2 indicates, most pairs are in the range of  $[2, 6)$ , which is ideal. Imagine that RU wants to design a timetable for the Dutch railways, it would be possible to implement a total of up to 5 train schedules for different time periods during the week.



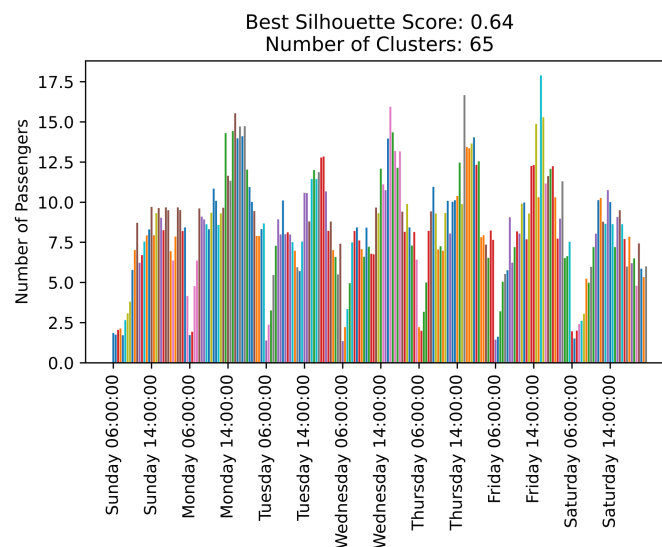
**Figure 4.2:** Histograms of the number of clusters of all ODs

There are 107 results in the range of [6,10). The flow levels of OD pairs in this interval are generally between 1 and 10, meaning that they are all fairly low demand OD pairs. Figure 4.3 is a typical result in this range, and it is the clustering result between station C and station D. Despite the low flow demand during the week, there still 8 clusters and the points within the clusters have little difference.



**Figure 4.3:** The clustering result of station C - station D

Some clustering results have more than 10 clusters, and few of them have even more than 70 clusters. These results are seen as undesirable. Although the original purpose of the search for the base demand period was to design a train service that better matched the passenger demand, delineating dozens of periods with different passenger flow characteristics make the analysis too complicated. As the figure 4.4 shows, with 65 clusters, the results show a great complexity and cannot be used for further analysis.



**Figure 4.4:** The clustering result of station E - station F

## Conclusion

From the results of the test, there are two areas of the original methodology that deserve reconsideration. The first is that it is necessary to set a range of clusters in the clustering algorithm to ensure that the results obtained are not too complex to be used. Of course, it is also important to consider how to define this range.

Secondly, in the process of defining the base demand period of the whole network, it seems unreasonable to use the common base demand period. Some OD pairs exhibit very distinctive traffic characteristics. One example is that for many OD combinations with Schiphol as origin station, their demand characteristics do not show significant morning and evening peaks during the week, but rather vary irregularly during the week. In this case, counting the number of times that each period is determined to be base demand period is one option.

## 4.2. Clustering constraints

Following the previous analysis, we first add a limit to the number of clusters. A simple way to do this is to add a constraint to the algorithm that only allows the algorithm to traverse the number of clusters in a range. In the original setting, the lower limit of this range is 2. Here, the upper limit is set to 5. This is done considering that in the previous clustering results, most of the optimal results were in this interval. For the range [6, 10), the OD pairs in this range basically have very low traffic and are combinations of small stations. Adhering to the idea that train services are designed to serve the majority of traffic, it would be slightly redundant to divide this part of OD pairs with low volume into 6-10 different demand characteristic periods.

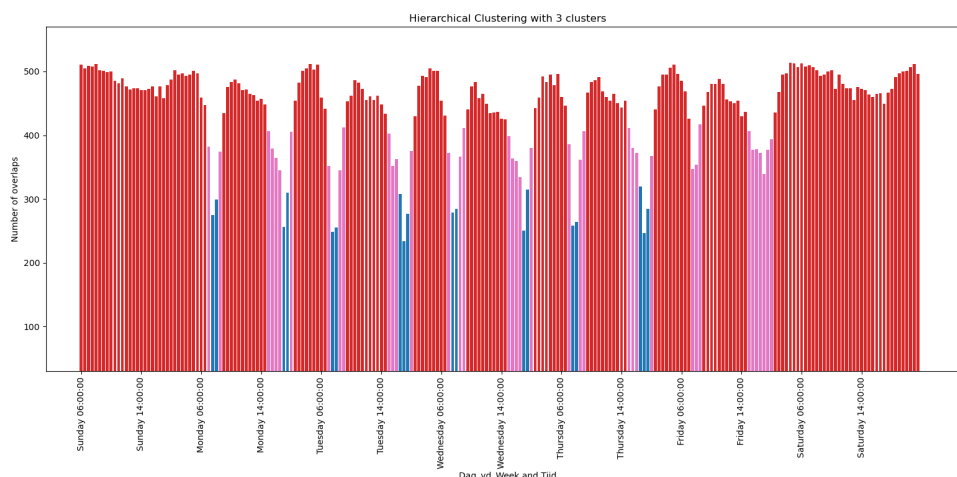
Instead of finding common base demand period, to find a period that is common for most OD pairs is a better idea. This means we record the number of times that each period is selected as the base demand period. We use frequency to represent this number in following text. Again, the clustering method will be applied to this set of data to make a reasonable classification. The higher the frequency, the more reasonable this time period is considered to be the base demand period of the whole network. In this way, we can classify time periods with different traffic characteristics according to the magnitude of the frequency.

## 4.3. Methodology application

The new methodology is applied to the data set and for each time period the frequencies are shown in figure 4.5.

According to our previous discussion, the frequency represents the number of times each time period is selected as the base demand period. For periods with similar high frequencies, we can say that the demand characteristics, which is the volume of these periods are similar throughout the network. Therefore, the hierarchical clustering approach is applied to the data again and the highest SC is used to get the optimal results. The results of the clustering are shown in figure 4.5, with three clusters.

The red cluster contains periods with high frequencies, and they are defined as the base traffic periods of the network. Most of the blue clusters are concentrated in the morning peak and evening peak hours, and both are relatively small. This mirrors the reality: the peak periods



**Figure 4.5:** Base demand period frequency

always occur at fixed times of the day and are relatively short. The pink clusters are concentrated in the part before and after the blue clusters, representing the time period before and after the peak. In the morning, pink clusters make up a much smaller portion, usually in the half hour before and after the morning peak; in the afternoon, they occupy the hours before the evening peak, and usually start at three o'clock.

Friday's demand is a special case. There are no blue clusters during the day, which indicates that there is no very clear peak traffic characteristic of Friday, either in the morning or in the evening. For the weekend, Saturday and Sunday are classified in the red cluster, indicating that traffic fluctuations on these two days are low and similar to the off-peak periods during the working days.

Although the clustering results have three clusters, we cannot arbitrarily decide that the demand pattern is divided into three categories based on the characteristics. The frequency is obtained from the demand volume of each OD pair, so the clustering results only tell us that the demand pattern begins to change at the boundaries between different clusters. For the time period represented within the red cluster, since most ODs are concentrated within this cluster, we can assume that this cluster can represent a demand pattern. For both blue and pink clusters, it is not reasonable to decide them as the same pattern if they are not adjacent.

### Period division

Based on the clustering results, 15 time periods were first delineated:

**Table 4.2:** 15 periods over a week

No.	Day of week	Time						
1	Monday	07:00	07:30	08:00	08:30			
2	Tuesday	07:00	07:30	08:00	08:30	09:00		
3	Wednesday	07:00	07:30	08:00	08:30	09:00		
4	Thursday	07:00	07:30	08:00	08:30	09:00		
5	Friday	07:00	07:30	08:00				
6	Monday	15:00	15:30	16:00	16:30			
7	Tuesday	15:00	15:30	16:00				
8	Wednesday	15:00	15:30	16:00	16:30			
9	Thursday	15:00	15:30	16:00				
10	Friday	15:00	15:30	16:00	16:30	17:00	17:30	18:00
11	Monday	17:00	17:30	18:00				
12	Tuesday	16:30	17:00	17:30	18:00			
13	Wednesday	17:00	17:30	18:00				
14	Thursday	16:30	17:00	17:30	18:00			
15	All other periods							

In table 4.2, 07:00 represents the period from 07:00 to 07:30. The pink and the blue clusters around the same time are combined. This is because when designing different patterns, the time periods within different patterns need to be adjacent and not too broken. Here, one hour is viewed as the minimum period. Imagine in reality, when developing train services, it is not desirable to switch one set of services every half hour. For the two clusters in the morning peak period, merge them into one. Also for the clusters after the afternoon peak, they are merged into the afternoon peak period cluster. For the red cluster, it is considered as one pattern, which is the base demand pattern of the network.

The next step is to get the corresponding flow OD matrix for each of these 15 periods. The idea is that for each OD, a reasonable demand value during the time period in which it is located will be chosen, as the representative of the demand for this OD. In this way, 15 matrices of 26\*26 scale will be generated.

## 4.4. Base demand matrix

This section illustrates how the base demand matrix is generated. Although the data cannot be shown here for confidentiality reasons, some observations can be given. In all matrices, the demand level between most OD pairs is low, which corresponds to most of the small sites (which are also the majority of all stations) in the network. Sites in larger cities generally have higher demand levels, with some typical OD pairs being: Schiphol - Amsterdam Centraal, Leiden Centraal - Den Haag Centraal and Amsterdam Centraal - Haarlem.

One of the observations is that if we compare the base demand matrix with the group 6 to 10 (that is, the hours before the afternoon peak of each working day; for Friday, it is from 15:00 until 18:29), that the base demand matrix always has a higher demand. This observation is contrary to reality. Usually, on weekdays, the demand in the network rises gradually from

around three o'clock in the afternoon, rather than decrease.

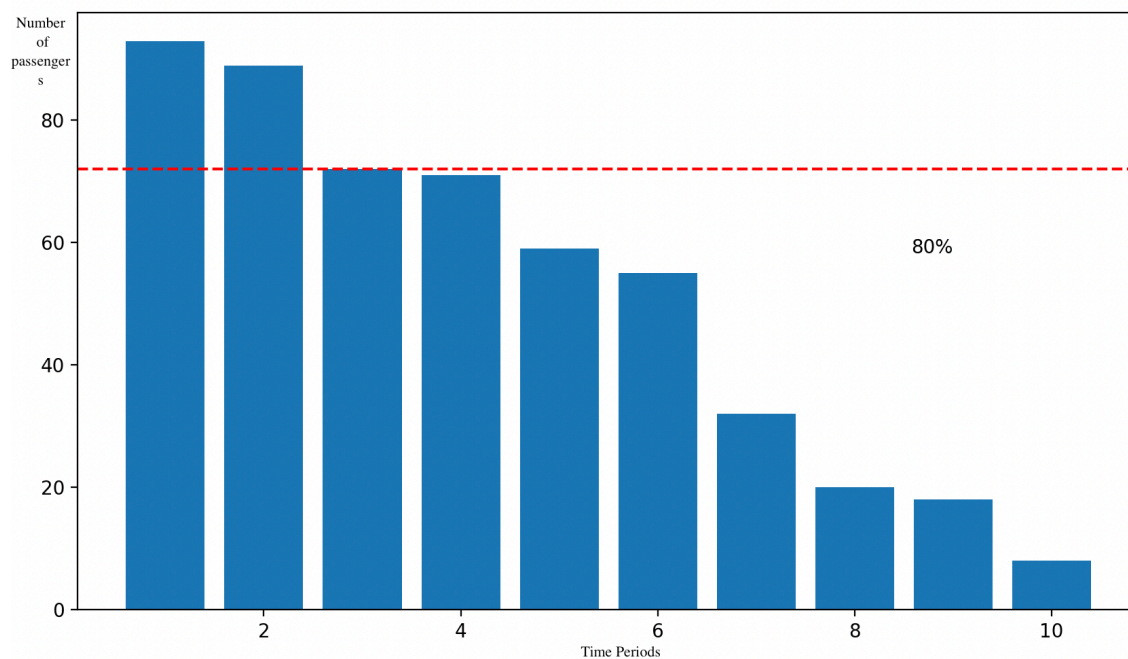
This phenomenon illustrates the unreasonable method of using the maximum value to determine the base demand matrix. Because the base demand period covers a long period of time, outliers in any period (e.g., a certain half-hour having a higher flow relative to the whole period) will have a great impact on the final results obtained. For all periods, because of the consistency of train service, three scenarios will occur for each period: demand is met and capacity exceeds demand for that period; demand is met and capacity is exactly equal to demand for that period; and demand is not met and capacity is less than demand for that period. For the case where the maximum demand is used to determine the service, although all demands will be satisfied, a considerable amount of service redundancy will appear, which is why the obtained base demand matrix will have more traffic than matrices 6-10. Thus, it is important to select a reasonable value for each OD pair to represent the base demand of it.

A simple comparison method is used to select this value. For each OD pair, the demand for all its periods is counted. Suppose we need to satisfy a certain number of periods such that the demand contained in these periods can be completely satisfied. In this way, based on the total amount of demand and the total amount of demand that is satisfied, we are able to derive how much demand can be satisfied for that OD pair. 80% is chosen as a start to explore this value. Two indicators are generated, which includes the percentage of satisfied demand and the total demand of the designed service. The latter is calculated by summing all values in the generated base demand matrix.

In figure 4.6, one example is used to illustrate how to find a fit value. Assume we are going to serve 80% periods, then the corresponding value (which is the red dashed line) is chosen. This value is assumed to be the capacity value that will be provided for each period. Then for all demands larger than this value, the percentage of the demand that is not satisfied can be calculated, and accordingly we can get the percentage of all demands that are satisfied.

For each OD, one assumed capacity value can be obtained, and the sum of all those values indicates the total capacity provided for every period in the sub-network. The statistics of this value allow us to see the level of service that needs to be provided when satisfying the travel needs of different percentages of passengers.





**Figure 4.6:** Example for selecting the value

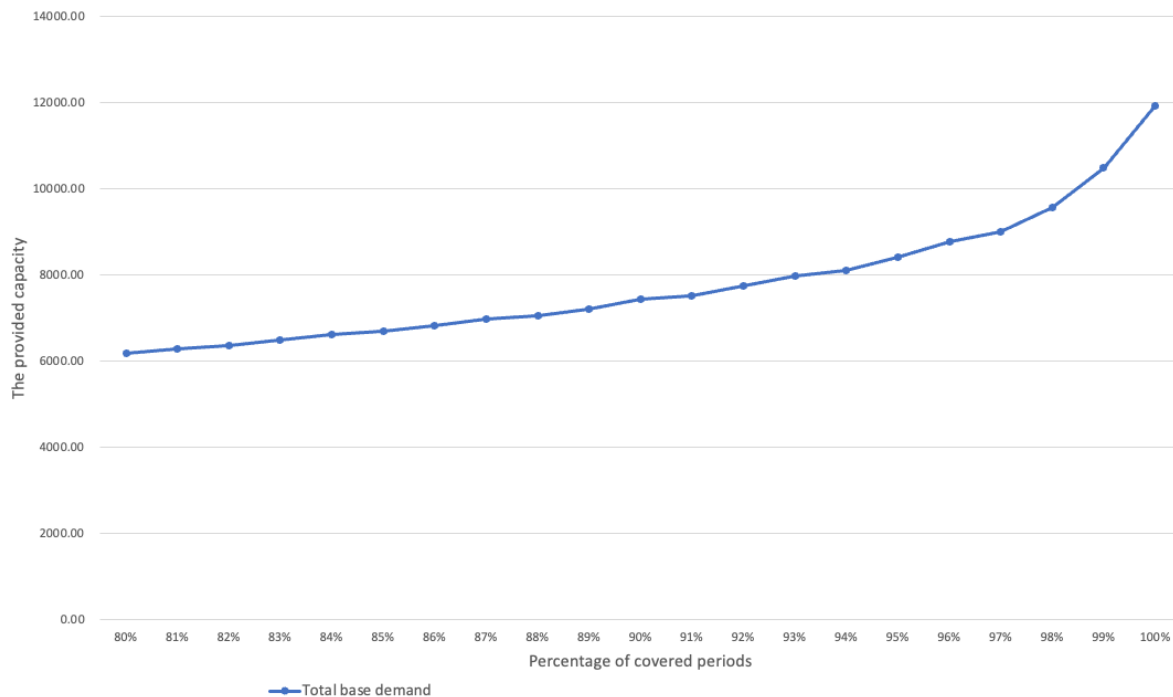
From 80% to 100%, all indicators are calculated, and the result is shown in table 4.3:

**Table 4.3:** Covered base demand periods

Cover periods	Satisfied demand	The provided capacity
80%	92.978%	6175.19
81%	93.451%	6281.78
82%	93.739%	6350.61
83%	94.234%	6473.50
84%	94.741%	6610.05
85%	94.979%	6682.12
86%	95.463%	6827.69
87%	95.865%	6960.26
88%	96.083%	7038.57
89%	96.539%	7210.45
90%	97.043%	7420.58
91%	97.263%	7521.02
92%	97.695%	7740.63
93%	98.112%	7984.09
94%	98.302%	8112.88
95%	98.677%	8403.03
96%	99.051%	8766.87
97%	99.233%	8994.51
98%	99.583%	9571.52
99%	99.881%	10494.42
100%	100.000%	11924.50

In table 4.3, one obvious point is that the provided capacity has nearly doubled from satisfying 92% demand at the beginning to satisfying 100% demand at the end. This shows the amount of wasted capacity. Because the service design cannot perfectly match passenger demand at all times, a more reasonable percentage should be chosen from these values. To better observe the change, a line graph is created (which is figure 4.7), where the x-axis is the covered periods and the y-axis is the provided capacity.

It can be seen that the provided capacity increases according to the increase in the percentage of the period being satisfied. But the relationship between them is not completely linear. From 95% to 96%, the total demand will gain a significant increase. And each subsequent percentage point increase brings an exponential increase in total demand. From a cost-saving perspective, 95% is a good point to stop and choose.



**Figure 4.7:** The change of total demand of designed service

So far, a total of 15 matrices representing different demand characteristics during a week were obtained. However, for service design, having 15 services in a week is too much. Therefore, further observation and merging of the matrices is the next step to be taken.

## 4.5. Matrices merge

In order to merge any two or more matrices, we must first make observations about the characteristics of the matrices. Since the 15 matrices have the same dimension, it is only necessary to compare the corresponding values within the matrices. Since the comparison is between the differences of corresponding values in the matrices, Manhattan distance (Black, 2019) is used here to compare the differences between matrices. Suppose the size of matrix A is  $m \times n$  and the size of matrix B is also  $m \times n$ . The difference between matrix A and matrix B is calculated as:

$$d(A, B) = \sum_{i=1}^m \sum_{j=1}^n |A_{ij} - B_{ij}| \quad (4.1)$$

$A_{ij}$  denotes the element in row  $i$  and column  $j$  of matrix  $A$ . Manhattan distance is a measure of the distance between two vectors and is suitable for calculating the difference between elements in corresponding positions in a matrix. The Manhattan distance is intuitive and easy to understand. It measures the degree of difference by calculating the sum of the absolute values of the differences between the elements of the two matrices at the corresponding positions. This method ignores the positive and negative relationships between the elements and focuses only on the numerical differences between them. In the process of comparing the matrices, if the difference between the corresponding positions is smaller, it proves that the matrices representing the two periods are more similar, both in terms of demand level and demand structure. Therefore Manhattan distance is considered as a simple and suitable method for comparison purposes. In figure 4.8, the comparison result is shown:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0														
2	2556	0													
3	1625	3288	0												
4	1979	1433	2680	0											
5	4121	6373	3444	5602	0										
6	8898	10770	8350	10069	6457	0									
7	9342	11293	8667	10546	6369	1563	0								
8	8791	10686	8159	9950	6110	1147	1289	0							
9	8953	10966	8272	10197	6013	1671	1041	1337	0						
10	7959	9655	7470	9034	6003	2652	3175	2457	2938	0					
11	9021	10214	8608	9823	7844	4442	5666	4798	5544	3992	0				
12	9107	10166	8786	9810	8253	5702	6780	5913	6671	4846	1749	0			
13	8491	9764	8056	9358	7228	3955	4934	4063	4854	3142	1488	2213	0		
14	8544	9696	8226	9317	7746	5202	6280	5411	6143	4043	1571	1417	1815	0	
15	8846	11028	8216	10285	5499	3615	3018	3342	2760	4181	6867	8040	6200	7413	0

**Figure 4.8:** Manhattan distance

In order to explain the results in the matrix more clearly, an interpretation of the period of the week represented by each group is made. As defined before, 1 to 5 represent the morning rush from Monday to Friday. 6 to 9 represent the period before afternoon rush. 10 represent the the majority period for Friday which is from 15:00 to 18:30. 11 to 14 represent the afternoon rush period from Monday to Thursday. 15 represent the base demand period.

The value in the first column represents the difference between group 1 and all other groups, where a larger value indicates a larger difference between the matrices. It can be seen that the combinations between group 1 and groups 2 to 5, that are, (1, 2), (1, 3), (1, 4), and (1, 5) have the values 2556, 1625, 1979, and 4121, respectively. And the differences with other groups were generally around 8000. We can find that for the first five columns of data, they all share this pattern: the difference values between groups 1 to 5 (the position of the red-brown color in the matrix) are generally small, but the differences between any one of them and the other combinations are large. So one message is that there is a large difference between the matrices divided in the morning and afternoon, whether they are distributed on the same day or not. This is because the traffic pattern is symmetrical between a significant portion of the sites, with a portion of people leaving from site A to B in the morning and returning from B to A in the

afternoon. Therefore, in the network, AM and PM will usually have different demand patterns.

Among the Monday through Friday morning rush, Monday and Wednesday, and Tuesday and Thursday show smaller differences (the deeper color area in 1-5 area). This phenomenon corresponds to the characteristics shown in Figure 3.2: Tuesdays and Thursdays always show similar higher level of demand. For all periods before the late peak, that is, groups 6 to 9, they were all less different from each other (values in the blue area). A special example is Friday, where both the morning and afternoon matrices are very different from any other matrix. For afternoon rush, Monday and Wednesday, and Tuesday and Thursday also show smaller differences (the deeper color area green area).

Putting those information together, one consolidation scenario would be to combine the morning peak on Monday and Wednesday, the morning peak on Tuesday and Thursday, all the periods before the afternoon peak from Monday to Thursday (that is, groups 6-9), the afternoon peak on Monday and Wednesday, and the afternoon peak on Tuesday and Thursday. Friday morning and afternoon both represent one pattern.

The rule for merging matrices is: select all merged matrices, and for each value in the corresponding position, select a maximum value and fill it into the corresponding position of the new matrix. Together with the base demand matrix, all 15 matrices are merged into 8 matrices representing different demand patterns over a week. These eight matrices are:

- Matrix A: Group 1 and 3, including Monday morning peak and Wednesday morning peak.
- Matrix B: Group 2 and 4, including Tuesday morning peak and Thursday morning peak.
- Matrix C: Group 6 to 9, including all periods before afternoon peak from Monday to Thursday.
- Matrix D: Group 11 and group 13, including Monday afternoon peak and Wednesday afternoon peak.
- Matrix E: Group 12 and 14, including Tuesday afternoon peak and Thursday afternoon peak.
- Matrix F: Group 5, the Friday morning peak.
- Matrix G: Group 10, the Friday afternoon.
- Matrix H: Group 15, the base demand matrix.

The sum of the values in each matrix is shown in the table below to indicate the sum of the capacity to be provided for each period, which also reflects to some extent the aggregate demand level for each period.

**Table 4.4:** The sum of the values in each matrix

A	B	C	D	E	F	G	H
15978	18214	10932	14656	16026	11530	11816	8113

Numerically, the total demand for Tuesday/Thursday is larger than the corresponding time for Monday/Wednesday for both the morning and evening peaks. For the hours before the

evening peak represented by C, the sum of demand will be slightly larger than the sum of the base demand. This is different from the morning peak, which tends to switch over from the base demand pattern at 19:00. The demand will slowly increase at 3:00 pm each day, first switching from the base demand pattern to the demand pattern before the evening peak, and then to the evening peak demand pattern. Friday morning's peak, as well as the afternoon hours (15:00-18:30), exhibit similar levels in aggregate. They are slightly above base demand levels but well below peak levels on several other working days.

## 4.6. Comparison with periods used in practice

In terms of the point when services are switched, the time periods delineated by the eight matrices are similar, but partially different, compared to the schedules used today. In the existing schedule, the services remain consistent from Monday to Thursday, with one set of services used on Friday and one set on the weekend. During the daytime hours, additional train service is added on some routes only from 7:00-9:00 and 16:00-18:00.

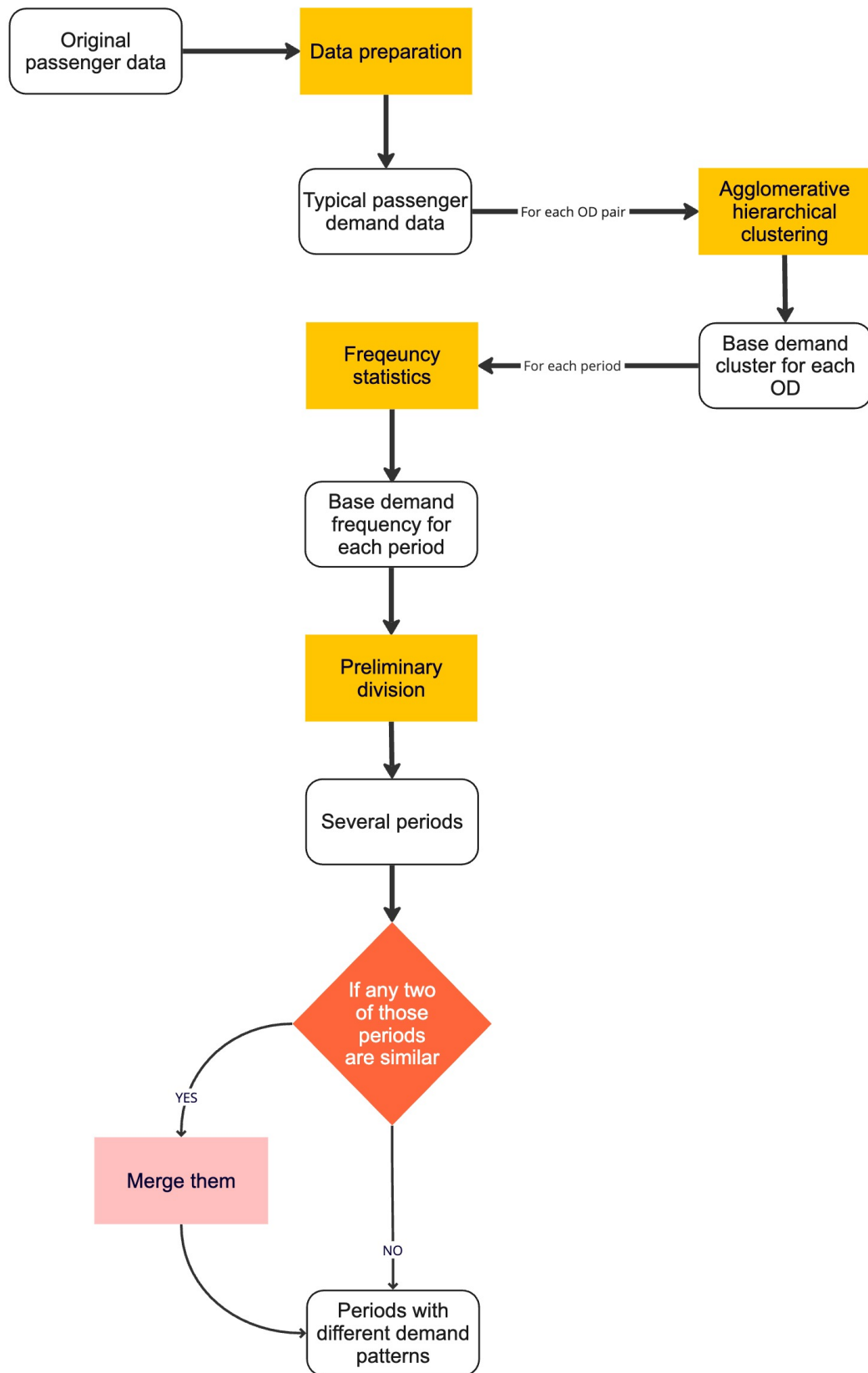
The results obtained in this chapter show that the peak service from Monday to Thursday can be further broken down. For the morning and evening peak, different levels of service could be considered for Monday/Wednesday and Tuesday/Thursday. For the period represented by the base demand, appropriate service subtractions should be considered to match service and demand even further. For Monday through Thursday from 3:00 p.m. until 4:30 or 5:00 p.m. when the evening peak begins, service fine-tuning, such as service subtraction based on peak service standards, could be considered.

For Friday, its morning and afternoon periods are divided into two patterns, plus the basic demand period, for a total of three periods, compared to the previous full day of consistent service.

## 4.7. Conclusion

In figure 4.9, the process of the whole methodology is shown in the flow chart. It indicates how the original passenger flow data is processed and how those periods with different demand patterns are delineated.

This chapter describes the application of the methodology. Eight matrices, representing eight different demand patterns are generated to indicate one week's typical demand pattern of the 26 stations sub-network. In the next chapter, we will look deeper into the demand pattern of several sites in this network.



miro

**Figure 4.9:** The process of delineating periods with different demand characteristics

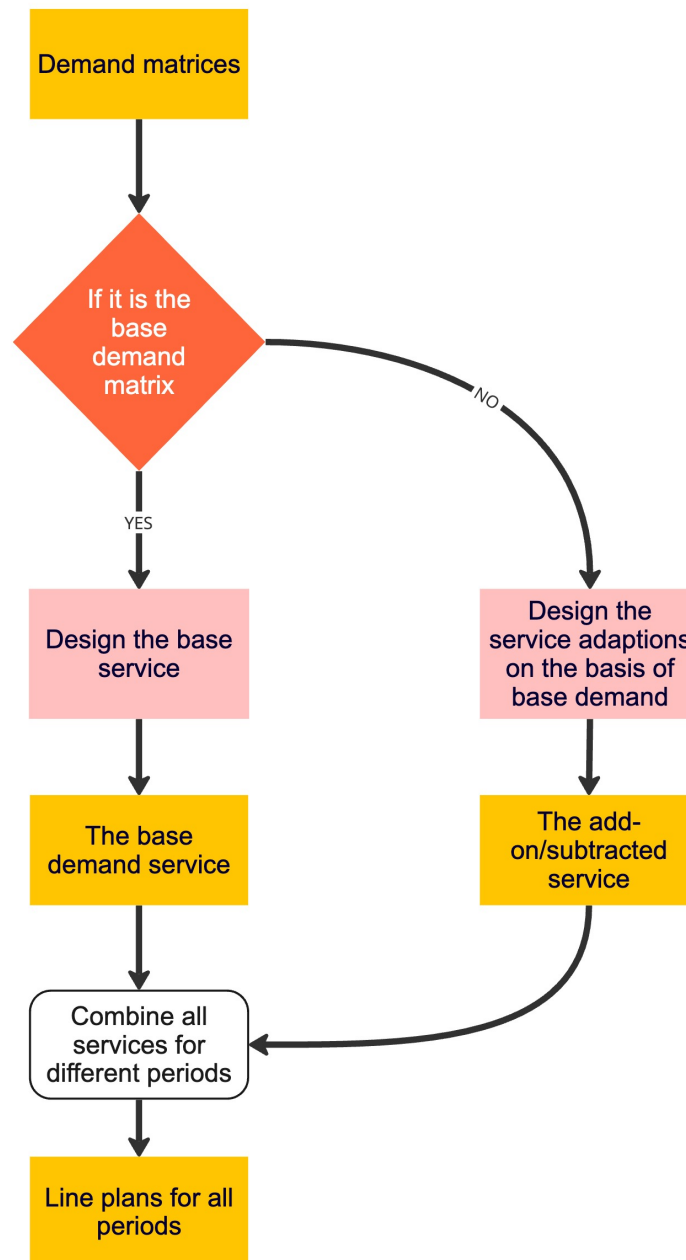
# 5

## Line plan design and comparison

In this chapter, a short line containing 6 stations was selected for the case. Two train service plans (based on an hourly basis) were developed based on demand during 8 different periods and some scenario assumptions are made. The first scenario is designed based on present rules for developing train schedules, while the second one was based on base demand. Two scenarios are compared in terms of RU and passengers.

## 5.1. Flow chart of designing the line plan

In figure 5.1, the process of using the different demand patterns to design the demand-oriented line plans is shown in a flow chart.



miro

**Figure 5.1:** The process of designing the demand-oriented line plans based on the base demand



## 5.2. Demand analysis

To get a deeper look at the distribution of demand, a line is selected from the network. In the process of developing a train schedule, the entire network is not selected as the case study. This is due to the multitude of factors that need to be taken into account when designing train plans, and the complexity increases with the number of stations involved. The focus of this project is primarily to showcase train schedules designed based on different demand patterns. Therefore, a specific portion of the network was chosen, which simplifies the design process without compromising the intended objectives. By selecting a subset of the network, we can streamline the design process and concentrate on creating schedules that align with the specific requirements, ensuring efficiency and effectiveness in meeting passenger demands.

This line contains a total of six stations from 1 to 6, where 1, 5, and 6 are large stations and 2, 3, and 4 are small stations. From matrix A to H, the demand of those six stations is used. The demand matrices including the six stations are shown in table 5.1 and table 5.8:

**Table 5.1:** Matrix A: The Monday/Wednesday morning peak

Matrix A	1	2	3	4	5	6
1	0	5	7	36	198	942
2	23	0	4	15	39	125
3	22	4	0	2	23	97
4	79	9	69	0	5	14
5	356	14	61	9	0	7
6	839	14	160	15	3	0

**Table 5.2:** Matrix B: Tuesday/Thursday morning peak

Matrix B	1	2	3	4	5	6
1	0	5	10	35	271	1166
2	24	0	2	11	44	166
3	26	2	0	2	33	132
4	93	6	67	0	5	17
5	394	15	62	10	0	9
6	846	18	164	22	7	0

**Table 5.3:** Matrix C: Pre afternoon period from Monday to Thursday

Matrix C	1	2	3	4	5	6
1	0	19	17	44	198	567
2	10	0	2	7	8	22
3	15	2	0	86	73	257
4	25	10	2	0	7	9
5	140	21	16	4	0	12
6	684	85	57	9	5	0

**Table 5.4:** Matrix D: Monday/Wednesday afternoon peak

Matrix D	1	2	3	4	5	6
1	0	18	18	74	337	872
2	7	0	1	2	7	16
3	10	0	0	10	15	44
4	38	7	3	0	9	17
5	204	32	16	2	0	6
6	1044	119	88	13	6	0

**Table 5.5:** Matrix E: Tuesday/Thursday afternoon peak

Matrix E	1	2	3	4	5	6
1	0	21	22	79	311	908
2	7	0	0	2	7	24
3	11	2	0	7	15	34
4	39	8	3	0	11	20
5	211	41	22	4	0	9
6	1209	144	106	14	7	0

**Table 5.6:** Matrix F: Friday morning peak

Matrix F	1	2	3	4	5	6
1	0	3	11	16	121	629
2	18	0	2	8	19	79
3	22	0	0	2	15	61
4	63	6	56	0	2	4
5	304	15	64	2	0	4
6	630	17	135	9	2	0

**Table 5.7:** Matrix G: Friday afternoon

Matrix F	1	2	3	4	5	6
1	0	21	19	58	205	650
2	10	0	0	7	10	24
3	14	2	0	69	69	236
4	24	6	2	0	4	7
5	125	14	9	2	0	6
6	632	71	55	7	3	0

**Table 5.8:** Matrix H: Base demand periods









Matrix H	1	2	3	4	5	6
1	0	10	12	27	101	426
2	12	0	1	4	10	33
3	10	2	0	2	5	27
4	30	3	2	0	2	6
5	126	7	5	2	0	5
6	391	33	29	5	3	0

To demonstrate the impact of the summarized 8 different matrices on train service schedule of different periods, three train schedules, based on one hour interval, is made to show the difference. Because developing a train plan is a complex matter, a simple scenario is set up for the next step in the analysis:

- The demand will be served by the specified train type.
- There are two types of train: Intercity train (IC) and Sprinter (SP). The capacity of IC is 500 and the capacity of SP is 263.
- The goal of the service is to meet all demand, with a load factor of no more than 100%.
- The travel time of IC will be 9 minutes between station 1 and 5, 3 minutes between station 5 and 6. The travel time of SP will be 3 minutes between each adjacent station. The stop time will be two minutes for each stations.
- The period is for a whole week, and for each day only the period 6:00 - 22:00 is included.
- When doing the service adaptations, if the load factor exceeds 100%, extra frequency or extra stop are needed to satisfy the extra demand.
- The passenger will either come to the platform randomly or with a plan.
- In existing schedule development, the government has established a minimum of two trains per hour for any passenger line. Here, to clearly see the impact of demand on service design, we remove this restriction.

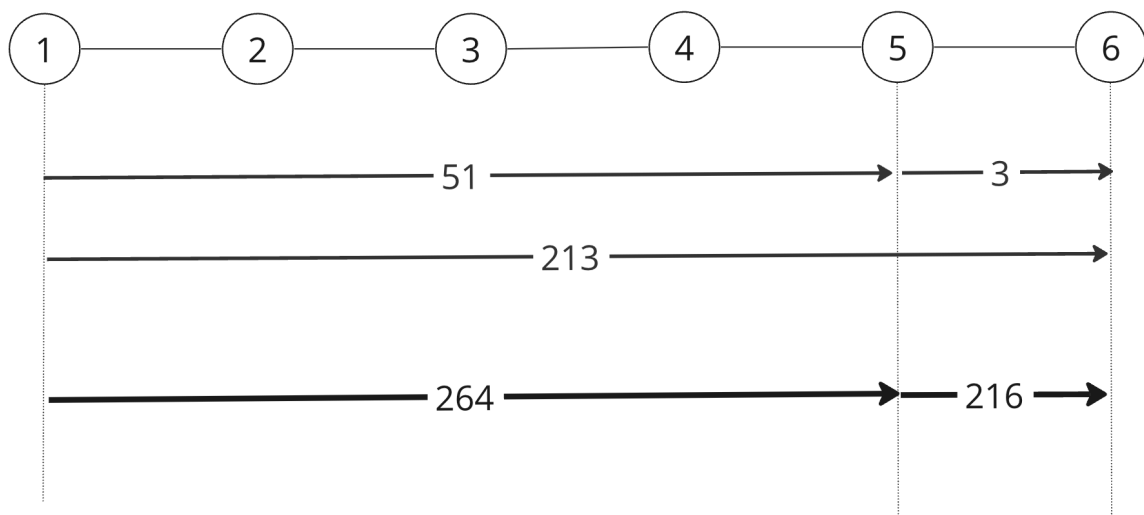
The assumptions for the two train capacities are motivated by the types of trains NS currently has. The capacity is calculated by dividing the seat capacity by the number of trains. For the Sprinter, the chosen type is SLT. For the IC, the chosen type is VIRM. The specific data are shown in Figure 5.2.

The assumptions for the way of how the passenger goes to the platform come from two situations by Van Oort (2011). If the passenger comes randomly to the platform, the average waiting time for each passenger will be half headway of the departing trains. If the passenger comes to the platform with the plan, which means he schedules first before going to the platform, the average waiting time is always 5 minutes for each passenger. Different waiting times due to two situations will be both calculated after designing the line plan.

	Number of trains as at 31 December 2021	Number of coaches/carriages	Seating capacity
<b>Sprinter trains:</b>			
 SLT	131	648	34.468
 SNG	171	588	30.843
 FLIRT (excl. TAG)	58	199	11.990
<b>Intercity trains:</b>			
 ICRmh	38	294	23.747
 VIRM	177	866	88.557
 ICMm	133	449	33.791
 DDZ	49	238	23.135
 ICNG	0	(National)	79 trains ordered
(Inflow starting in late 2022)	0	(International)	20 trains ordered

**Figure 5.2:** NS train types (Nederlandse Spoorwegen, 2022)

Among the six stations, the IC is used to serve passengers between station 1, 5 and 6, which are considered large stations. The rest of the passenger flow is considered to be served by SP. Figure 5.3 gives an example on how to get the segment demand from the OD matrix. It shows the segment demand with the direction from station 1 to station 6.



**Figure 5.3:** Segment demand between big stations for matrix H (from 1 to 6)

This figure indicates that one IC train is needed to satisfy the demand between those big stations, which should depart from station 1, stop at station 5 and finally stop at station 6.

### 5.3. Train schedule design

Based on what has been analysed before, assume now we have to make a train schedule for the whole week, during which there are eight different demand patterns. Two scenarios about designing the service are introduced here:

- Scenario 1: Make a schedule based on peak hour demand. This line plan is one hour based and will be applied to all the other hours in this week.
- Scenario 2: Make a schedule based on peak hour demand. For the other demand patterns, add service adaptations. The service adaptations will only be made when it is necessary: the load factor of some segment exceeds 100%. The service doesn't need to be symmetric.

Scenario 1 is based on the design process for train schedules for most lines in today's schedules. RU tends to design a one-hour train service based on the peak demand for this line and then replicates it for each hour throughout the day. Scenario 2 is a plan based on base demand: a train plan is developed for the base demand period, then services are added or removed depending on different demand patterns.

The rules for designing the line plan is: for on direction, we looking at all segments demand first. The capacity provided by the train must satisfy all demand for those segments demand. Take figure 5.3 for instance, from 1 to 5, the segment demand is 264; from 5 to 6, the segment demand is 216. Both of them are smaller than 500, so one IC train is enough for satisfying the demand. If the 1-5 segment demand exceeds 500, which is larger than the capacity of one IC train, the frequency will be 2. When train frequencies have to be increased, this does not mean that the new trains need to serve every station, but only stop at stations that need additional service. For instance, when the demand of 1-5 increases to over 500, but the demand of 5-6 stays same with before, then the added one train will only run between 1-5, instead of stopping at station 6.

#### 5.3.1. Scenario 1

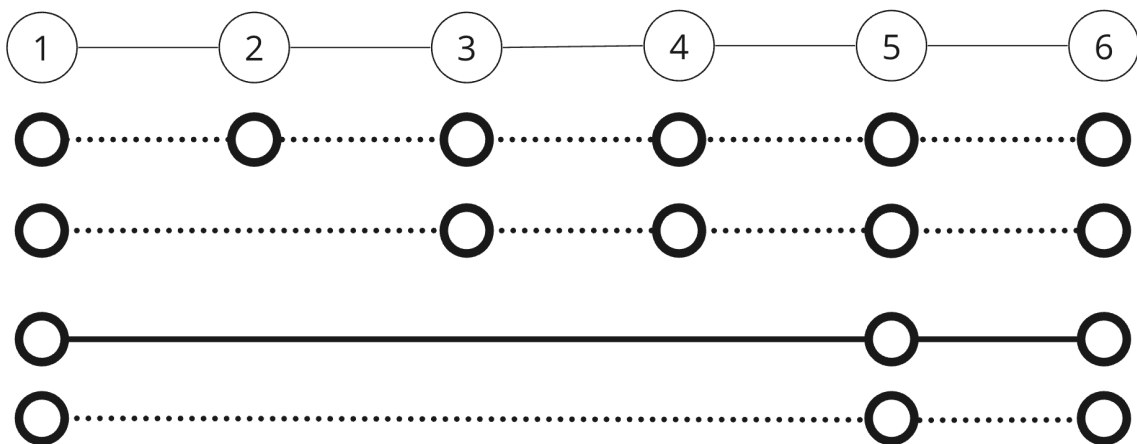


Figure 5.4: Line plan based on the peak hour demand

In figure 5.4, a line plan based on the peak hour demand is made. This demand represents the matrix B, which is the morning peak for Tuesday/Thursday. The type of line represents the frequency of trains, where the dotted line represents one train per hour and the solid line represents two trains per hour. The hollow circles mean that the train will stop at that station, only those stopping at 1, 5 and 6 are intercity trains, the rest are Sprinters. The line without arrows means that the service is symmetrical: the service is the same for both directions. The line with the right arrow means that the service is not symmetrical: it serves only the direction represented by the arrow. The line plan based on the demand of this period can well meet the demand of various periods of the week without any adjustment.

In figure 5.4, the Sprinter frequency is 2 per hour, and the IC frequency is 3 per hour. One of the Sprinters will not stop everywhere and it skips station 2.

### 5.3.2. Scenario 2

In this scenario, eight line plans are designed according to the corresponding demand. Figure 5.5 represents the line plan based on the base demand. From line plan A to G, all service adaptations are made on the basis of base line plan.

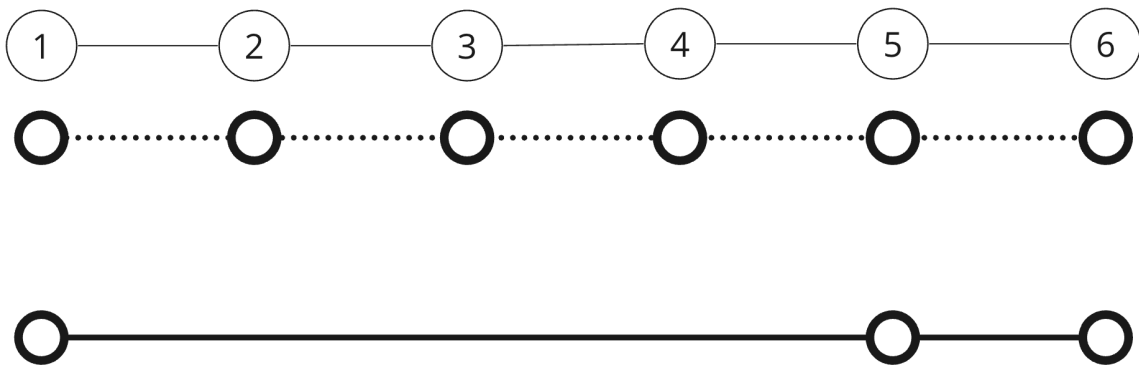


Figure 5.5: Line plan based the base demand

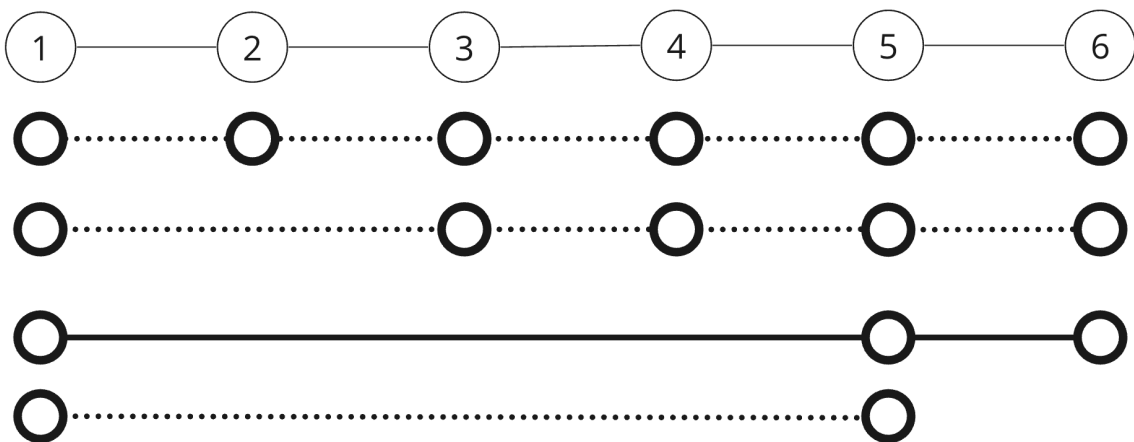


Figure 5.6: Line plan A

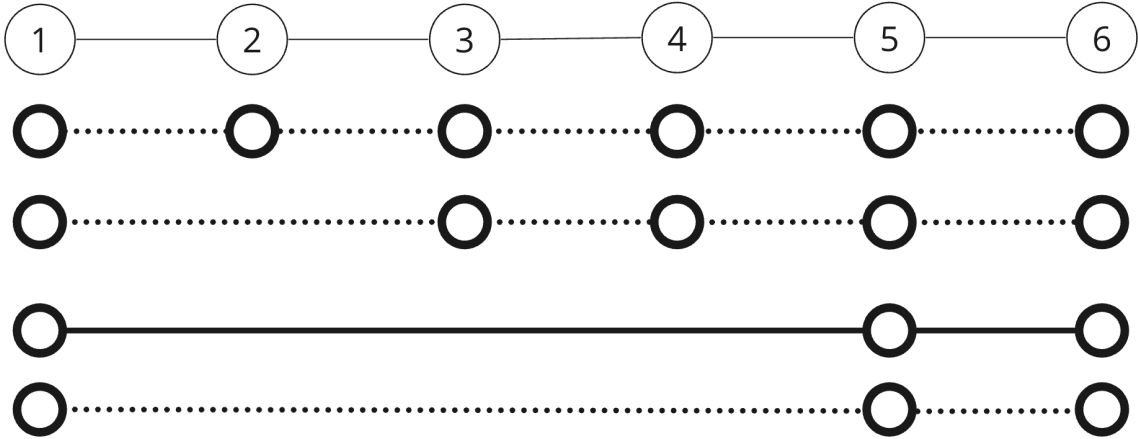


Figure 5.7: Line plan B

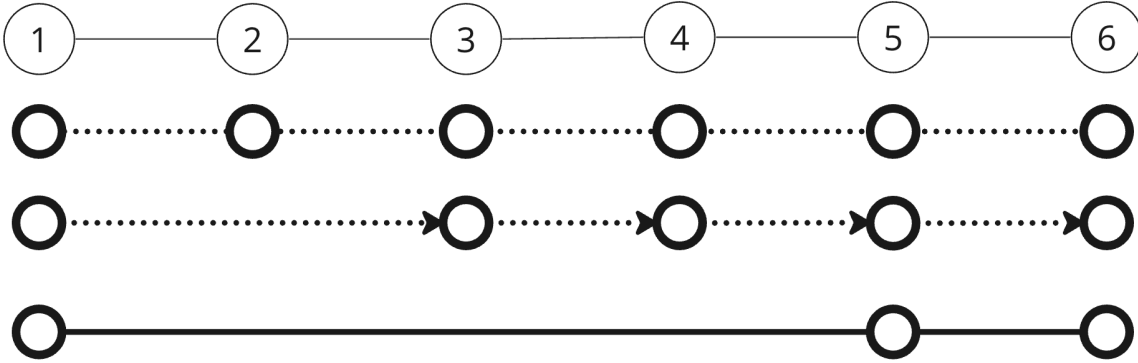


Figure 5.8: Line plan C

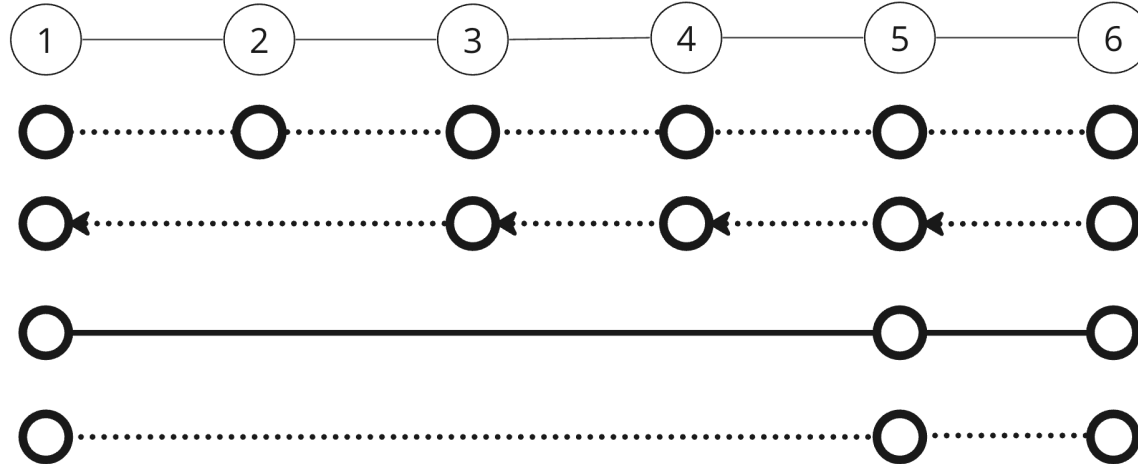


Figure 5.9: Line plan D

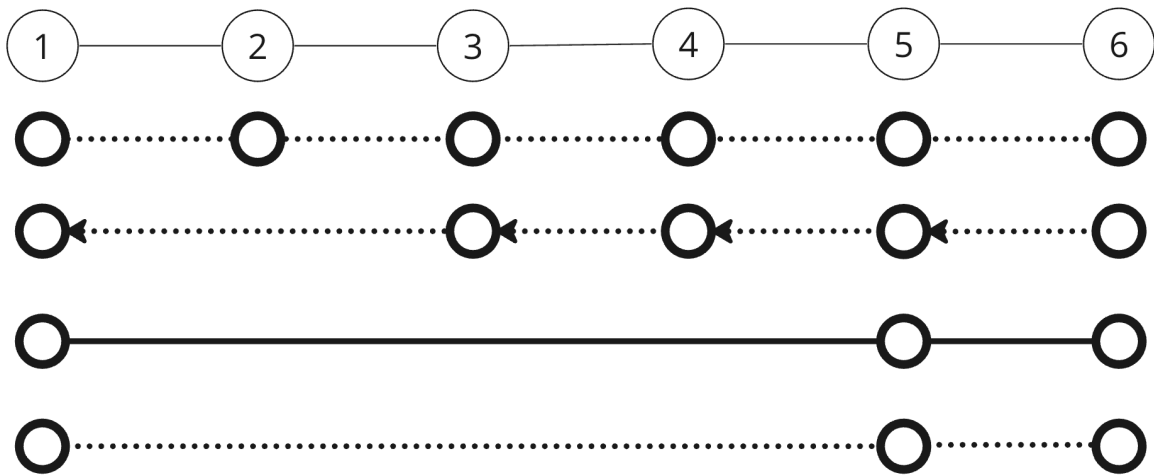


Figure 5.10: Line plan E

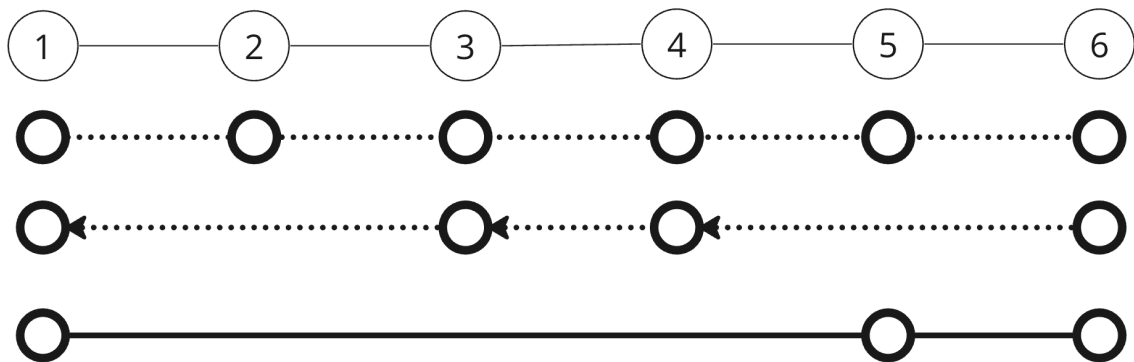


Figure 5.11: Line plan F

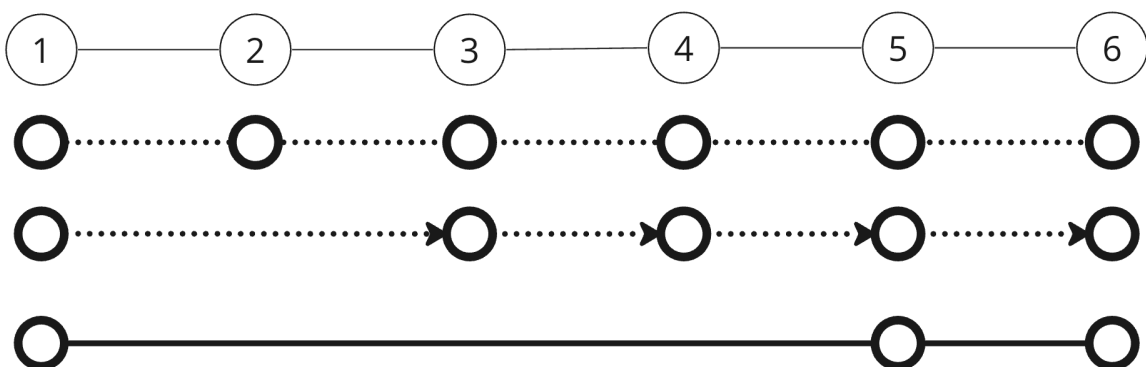


Figure 5.12: Line plan G

Compared with Scenario 1, Scenario 2 contains 6 different line plans (where D and E, C and G end up with the same line plan). For the base demand line plan, the train frequency is only three trains per hour: two ICs and one Sprinter. For the other service plans, morning and evening



peak, i.e. A, B, D and E, have the most additional services in terms of train frequency. There are seven line plans with frequencies ranging from three trains per hour to five trains per hour.

The additional services are diverse. For example, for period A, one additional IC does not need to stop at station 6, but only serves station 1 and station 5; for station 2, all additional Sprinter trains do not serve that station because demand is relatively low during the week in both adjacent segments of the station. Compared to the peak period, the line plan designed for pre afternoon period (line plan C), and Friday morning peak/afternoon period(line plan F/G) has less add-on service. Friday morning and evening traffic also shows some symmetry, as evidenced by the additional Sprinter trains (in figure 5.11 and figure 5.12: during the morning peak, the additional Sprinter only serves demand in the 6-1 direction; during the afternoon, the additional Sprinter only serves demand in the 1-6 direction. Although the additional services are not perfectly symmetrical, they serve mostly the same stations.

In addition to the observational analysis of line plan, some quantitative analysis tools can be used to compare the two types of plan development. In the next subsection, the specific metrics and results analyzed are presented.

### 5.3.3. Evaluation for two scenarios

Two scenarios are compared through two main aspects, the RU aspect and the passenger aspect. For RU, the indicators include the total number of stops, and the total train minutes. The total number of stops indicates the variable cost of train operating: Fuel costs, energy consumption, and maintenance and repair costs. The higher the number of stops, the higher the cost for train operation. The total train minutes indicates the variable cost of train crew revenue: the working time for all staff on the train.

For the passenger aspect, the total travel time will be calculated. This time includes two parts, which are the in-vehicle time and the waiting time. For waiting time, both situations of passenger come to the platform randomly or with plan will be calculated. Passengers will always prefer to travel on similar trains with shorter travel times. SP traffic will only be served by SP, and same for IC.

In table 5.9, all results are shown.

**Table 5.9:** The results of two scenarios

	Scenario 1	Scenario 2
Total number of stops	4480	2967
Total train minutes (h)	325	211
Total in-vehicle time (h)	47401	43693
Total waiting time (h) (come randomly)	79069	93283
Total waiting time (h) (come with plan)	32568	32568
Total travel time (h) (come randomly)	126470	136976
Total travel time (h) (come with plan)	79969	76261

## 5.4. Results analysis

The analysis of the results will be based on two aspects, including the RU aspect (total number of stops and total train minutes), and the passenger aspect (the total travel time).

### 5.4.1. Impact on the RU aspect

For scenario 1, the hourly train service designed for peak demand is replicated to all times of the week. This guarantees that the schedule will be easy to design and make. However, low design complexity brings redundant services and high costs. Compared with scenario 1, scenario 2 obtains a reduction of about one-third in both total number of stops and total train minutes, both of which reflect the variable cost of train operations.

Besides, for all the line plans developed under scenario 2, only the frequency of trains during the peak period from Monday to Thursday is consistent with the line plan of scenario 1. During the remaining periods of varying demand, the train frequency decreases - which results in the design of train services that use less track capacity. So another advantage of Scenario 2 is that services based on base demand do not require much “compromise”. An example of a “compromise” is: In 2017, NS wanted to run the Sprinter twice an hour between Roosendaal-Dordrecht. This conflicted with other wishes of NS and the wishes of some freight forwarders. The compromise reached was that the Sprinter between these two stations would run only once per hour. In reality, however, the freight forwarders’ freight trains do not run every hour of the day, so this Sprinter of NS is able to run twice per hour at certain times. The message from this example is that there is not a busy transportation demand on the rail network every hour. Therefore the compromise reached will remain in the schedule despite the fact that there are times when the transport demand does not exist, which is unnecessary.

Compared with scenario 1, scenario 2 brings less train operating costs. At the same time, this design idea frees up more capacity during off-peak hours, allowing the RU to be more flexible in scheduling all trains.

### 5.4.2. Impact on the passenger aspect

For passengers, the total time for scenario 2 is shorter in terms of in-vehicle time. This is because the service adaptation is added during the design of the service, depending on the demand. This is reflected in the added Sprinter and IC train services: they do not need to stop at every station, but serve the demand more precisely. For some passengers, their in-vehicle time will be shorter, because some stopping time is saved.

In terms of total waiting time, two different situations are considered: random arrival of passengers or planned arrival at the platform. This consideration was motivated by a study by de Bruyn et al. (2022): In 2007, most people considered a frequency of four trips per hour to be a “no schedule” situation: 64% of passengers said they would travel to the station without a plan. With six trips per hour, that percentage increased to 92%. Only a minority (33%) consider three trains per hour to be unscheduled. This research indicates that the higher of the train frequency, the more passenger will come to the platform randomly.

When all passengers arrive at the platform randomly, the total waiting time for passengers in-

creases for scenario 2 compared to scenario 1. This is because many periods of scenario 2, such as the base demand period of the line plan, have less train frequency, which leads to a larger waiting time according to half headway. As a result, the total passenger travel time is also greater for scenario 2. Compared to the change in cost, the change in the total travel time of passengers is not significant.

When all passengers are scheduled to arrive at the platform, the total waiting time is the same for both scenarios. Therefore, the total passenger travel time is less for Scenario 2, which is the desired outcome for both passengers and RU. Again, the change in total travel time for passengers is not significant compared to the change in cost.

Considering the reality that both kinds of people will be present in the passenger group, the correct value should fall in between the two cases. But in any case, the total travel time does not change much for the passengers.

## 5.5. Conclusion

This section describes a simple case study. The demand data of a short line including 6 stations is used to perform the analysis. Based on some assumptions, three scenarios of designing train schedules are generated. To compare the difference, the perspective of RU and passengers are both considered. The main considerations are the complexity of the service design and the cost of running the service for RU, and the total travel time is considered for passengers.

The result indicates the difference of making train schedules based on the base demand, compared to the existing train schedule design. Services based on base demand are well suited to meet the needs of passengers during the majority of time periods, and then services are added or subtracted at other times as necessary. This design idea avoids unnecessary compromises that would conflict with other transportation desires at the outset. It also allows for better utilization of rail capacity. In terms of service plan development, more service adaptations and adjustments occur during peak periods, which leads to more complex plan development during peak periods. After all, the existing plans based on peak hour demand are already the result of “compromise”. If the off-peak plan is based on base demand, the added/subtracted passenger services are needed to be considered together with the other stakeholders’ wishes for the peak period.

For the cost of running trains, there will be some reduction in cost because the basic demand during off-peak periods will always be relatively low compared to the higher demand during peak periods, and services designed to meet the base demand will have relatively less cost. For passengers, the in-vehicle time will be less because the train service is more matched with the demand. For the total travel time, there is no significant change.

# 6

## Conclusions and Discussions

In this final chapter, the conclusions are first given, where the main research question as well as the sub-research questions are answered based on the results of all the studies. Secondly, the contributions of this study are concluded. Finally, the discussion part includes some reflections for both the methodology and the case study, and recommendations.

## 6.1. Conclusions

The objective of this study is to develop a demand-oriented train service plan by using the railway passenger demand pattern. By summarizing the entire network traffic demand pattern over a period of time, for example, a week, it is possible to classify different periods based on demand characteristics. Such information can provide the basis for the development of demand-oriented train services. The main research question following from the goals of the study therefore was defined as:

- How to develop a demand-oriented train service plan by using the railway passenger demand pattern?

To address the main research question effectively, a sequence of sub-research questions was systematically devised, with each subsequent question building upon the insights gained from the previous one.

### **SQ1:What is already known in the literature regarding analyzing the flow demand pattern in public transport demand?**

A literature review is done to explore what methods are already used in analyzing the flow demand pattern in public transport. Smart card data has been used to analyze space-time characteristics, employing indicators like average daily passenger flow and peak hour coefficients. Visualization methods, such as thermodynamic charts and geographic maps, have been utilized to illustrate flow changes across stations and lines. Statistical indicators and clustering methods have aided in understanding commuter patterns and identifying base demand periods. Overall, the literature offers valuable insights into analyzing public transport demand patterns using various approaches. The concept of base demand is also provided, which indicates to the demand that exist all the time in the network. Finally, clustering methods are chosen for identifying and analyzing base demand in public transport due to their effectiveness in grouping similar patterns and characteristics, enabling the recognition of consistent flow distributions during off-peak periods.

### **SQ2:What method can be used to characterize the passenger demand patterns?**

To characterize passenger demand patterns, the appropriate method to consider is clustering. We focus on methodologies for identifying the flow demand pattern using clustering methods, given that the data is presented as passenger volumes for each OD pair in half-hourly intervals. Four widely known methods, namely K-means, bisecting K-means, DBSCAN, and Hierarchical clustering, are presented and compared. The Hierarchical clustering is considered a suitable method to identify the base demand. With using the average-link measure, it can overcome the defect possessed by some other algorithms: chain effects. Besides, no initial values are needed to be set, such as the number of clusters. Furthermore, the silhouette coefficient is utilized to find optimal clustering results.

**SQ3:What passenger demand patterns can be found when applying this method to concrete regions?**

A sub-network of Dutch railway network with one week demand data in 2022, was selected to carry out the application of the methodology. A total of eight periods representing different demand patterns, including base demand, were delineated. The period representing the base demand covers the majority of the off-peak hours of the week, which includes weekends. During the peak hours, the morning peak shows a very different structure than the evening peak. This is because many riders are symmetrical throughout the day: from A to B in the morning and back to A from B in the afternoon. The evening peak differs from the morning peak in that the overall increase in demand already begins in the first few hours of the evening peak. On weekdays, the demand in the network starts to show a difference from the base demand starting at 15:00 each day. Comparing each day of the week, the peak hour demand patterns are closer on Mondays and Wednesdays, and on Tuesdays and Thursdays. Tuesdays and Thursdays have higher levels of demand. Friday's morning and evening peaks are not similar to the demand structure of any other weekday counterpart, but exhibit a lower level of demand size.

**SQ4:How can the result of the analysis of the demand pattern be used for developing a demand-oriented train schedule?**

Based on eight periods containing different demand patterns, a train service plan that can cover the base demand can be developed first. The service adaptations are made according to the demand characteristics of the other periods. Since the demand characteristics are already reflected in the corresponding matrix, which contains the demand size and demand structure, the added or subtracted services can better match the demand pattern.

The line plans generated based on the base demand can better match the passenger demand than the train plan based on the peak demand. More "compromises" can be avoided at the outset, some redundant services can be eliminated, and passenger needs can be well met. For the RU, the train operating cost will reduce; for passengers, the total travel time changes little. Of course, this will lead to some increase in design complexity. But in the future, when demand of all kinds is increasing and rail capacity is saturated, the idea of basic demand service can provide a more reasonable solution.

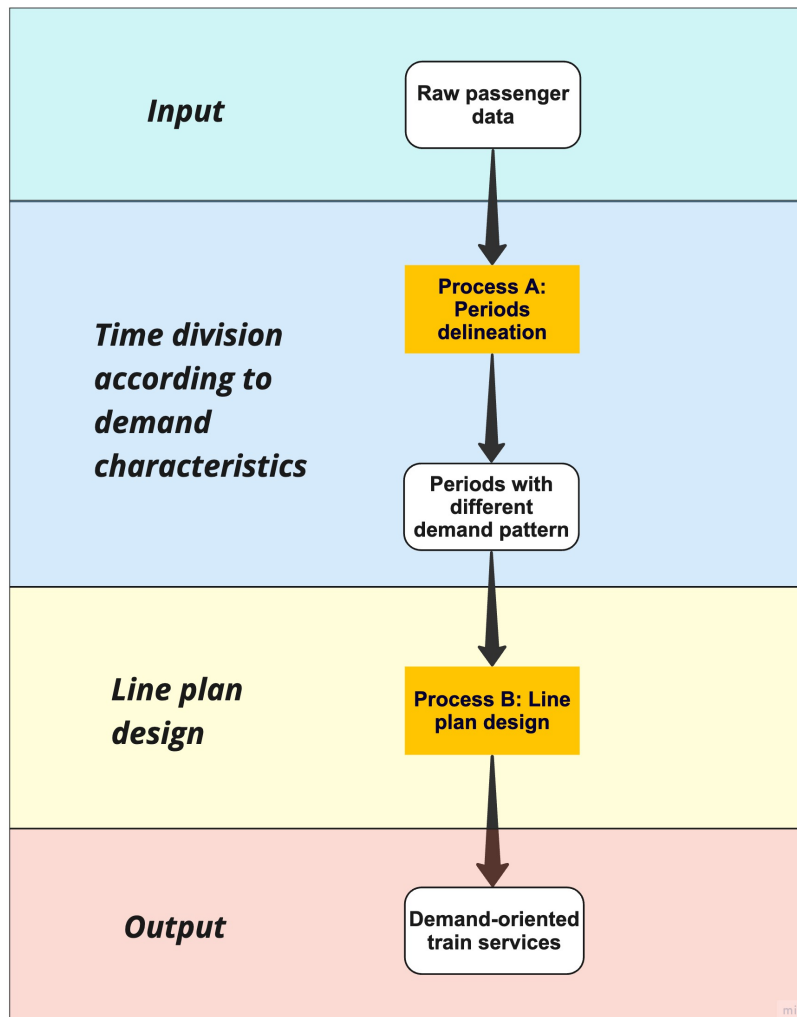
**Main Research Question:How to develop a demand-oriented train service plan by using the railway passenger demand pattern?**

To develop a demand-oriented train service plan, the railway passenger demand characteristics are needed to be analysed first. The analysis mainly starts with the volume and structure of the demand, and based on these characteristics, periods that have similar characteristics are classified into the same pattern. Thus a period of time, for example a week, can be divided into several periods with different demand patterns. One period is considered as the base demand period among all those period, and the demand contained in this period is always present during the week, its fluctuations are small and occupy the vast majority of the time except for the peak period.

Different from the existing method that the train service is designed according to the peak hour demand, the train plan will be first designed based on the base demand. This base service will exist all through the period (e.g. in one week) and provide service for all base demand service. For all other periods, the service adaption will be made and added to the base service. Those adaptations can be add-on or subtracted service, according to the different demand patterns. Those adaptations can adjust the service to match the changed demand over time.

## 6.2. Contributions

This project has developed a method of designing a demand-oriented train plan. In figure 6.1, the whole project process is shown. The detail of process A refers to figure 4.9, and the process of B refers to figure 5.1.



**Figure 6.1:** The project process

A method to identify periods with different demand characteristics is developed. This step of data preparation helps to filter the original data. After removing some outliers (e.g. due to strikes or holidays), pick a period length, e.g. one week. For the same time of the week (for each half hour) over an entire period, the median value of the traffic is selected to reorganize into a new week. This week is a good representation of the demand characteristics for this

entire period.

The clustering approach was applied to this week's data. For each OD combination, the demand value for each half hour is one data point. The clustering approach places these points into different clusters, and those clusters with the most data points are considered to be the base demand clusters. All the base demand periods are counted in terms of frequency, and then on the basis of cluster analysis of the results, the period that represents the base demand of the network is initially obtained: the one with the highest frequency. Cluster analysis of all frequency results also divides the week into several periods that may have different demand characteristics. After comparing these periods and merging those with similar demand characteristics, periods with different demand characteristics were obtained for the final demand representation.

The way of designing the line plan based on different demand pattern is developed. The base service is first developed based on the base demand, and then for other periods, the add-on/subtracted train service will be added according to the demand pattern. A number of metrics were used to analyze the methodology, compared to existing line plan designed for peak demand. The results showed significant savings in variable costs in terms of operating costs, as reflected in two areas: total number of train stops and train minutes. For passengers, the total travel time changes little, which means a little impact. The results of the analysis give a positive evaluation: this method is worthy of further explored and refinement.

### 6.2.1. Comparison with literature

In this section the results of this study are compared with previous studies. The study is unique in designing a method for finding periods with different demand characteristics and defining which period is the base demand period. The study summarizes this by analyzing the characteristics of the underlying demand: distributed in off-peak periods, with low fluctuations and similar traffic levels, always present in the network. The base demand period is defined according to those features. To the best of the author's knowledge, this base demand period has not been defined before. Previous studies however do give the idea and concept regarding about the base demand. Bruijn et al. (2019) indicates that the base demand is the demand that always exists during the period. Based on the base demand, there should be a schedule with no or little compromise. In terms of overall transportation volume, the basic demand should cover 60%-80% of all demands in the whole network, but at which level is not decided.

Van der Knaap et al. (2022) used the clustering to divided the moments of each day from Monday to Friday according to the characteristics of the demand (both structure and volume). For each day, from 6:00 to 00:00, the daytime is divided into 9 or 10 clusters. Based on the results of these clusters, the cluster-to-cluster boundaries exhibit similarities to the results of this studies: The cluster represented by the morning peak always ends around 9:00; the pre afternoon cluster always starts at 3:00 and then switches to another cluster at 4:30 or 5:00. The point in time when these clusters are converted to another cluster often implies a suitable point for converting different services. Different from Van der Knaap et al. (2022), this study defines one of them as the base demand cluster. Also, the idea of designing a line plan based on these periods with different demand pattern is given.



## 6.3. Discussions

In this part, some reflections based on the methodology and the case study are concluded. Recommendations are given based on those reflections for both the research and NS.

### 6.3.1. Reflections for the methodology

In terms of the demand data covered by the entire sub-network, it would be better to classify the OD mix prior to determining the base demand to avoid the impact on judgment of some ODs that contain irregular demand variations. For example, the demand to and from Schiphol tends to be irregular because there are no very distinct peak periods for flights at the airport. When the method summarized in this study is used to identify the base demand period, it is likely to deviate from the actual results. Since only passenger flow data provided by NS are available for this study, these stations with unusual demand can only be reflected from the observation of the data. This observation becomes difficult when the data base becomes larger, for example, when the number of stations increases. Therefore, in the future, it is a worthwhile direction to analyze and categorize the passenger flow attributes carried by the stations, taking into account the city where each station is located. It is also worthwhile to group OD combinations with different demand sizes, as it is often the high traffic combinations that are more worthy of consideration by RU. A possible approach is to assign different groups with corresponding weights according to the traffic levels they represent.

When comparing the similarity of 15 demand matrices, the Manhattan distance may not be the best method. Judging the similarity of two matrices from relatively low values needs further improvement. Imagine how much larger each matrix becomes (over 300 stations) when applying the method to a much larger network, such as the entire Dutch rail network. Some thresholds should be set based on the size of the network and the intention of the RU, to determine exactly when two matrices are similar enough to be merged.

In determining the demand matrix for each period, one improvement is the approach to choose a reasonable capacity that is provided. When deciding the value in demand matrices, if the demand to be satisfied rises, the number of services to be provided increases; because the services provided cannot change during the period (for example, several hours in a row, all within the basic demand period), the number of redundant services added also increases. Excessive waste of capacity is undesirable in the opinion of the RU. A point of balance between the interests of passengers and the operating costs of the RU deserves to be found. Of course, this is only partially in the interest of the RU and passengers. More stakeholders should be considered, e.g. the wishes of freight forwarders, the wishes of the government. Because this is not the main research purpose of this project and given the time constraints, not much research has been conducted in this area.

### 6.3.2. Reflections for the case study

When comparing the line plans of two scenarios, simple assumptions may not fully reflect the strengths and weaknesses of the services designed based on the underlying demand. For example, each switch of service, change of service frequency, and other measures can increase operating costs. These costs, including benefits to passengers, should be quantified and then considered in aggregate. However, because of the confidentiality of the data, it is difficult to

measure those costs or benefits. A simple example is train minutes: this often represents the revenue pay for all crew members. In a well-defined line plan, there is often more than one type of train used, and each type of train has different fixed costs, such as the cost of purchasing the train, and the cost of equipping the train. These data are more or less confidential, so it is not easy to fully consider the various factors involved in the cost.

For passengers, it is worthwhile to properly assess their total travel time, especially the waiting time. If the frequency increases, more people will arrive randomly. Conversely, if the frequency goes down, fewer people will arrive randomly. Although de Bruyn et al. (2022) indicates that when the train frequency is over six per hour, most passengers will arrive at the platform without a plan, the result may not apply when the new timetable is used in reality. If the train service is no longer the same throughout the day, the percentage of those passengers might also change.

The size of the case study brings limitations: In the whole network, many lines often have crossover and overlapping sections, which can lead to a more complex distribution of demand. The complexity might lead to different evaluation results (such as the total travel time). In the future, the study case can be performed on the basis of a complicated scenario: A simulation can be implemented. The study region includes all stations in the Netherlands, and develop the line plan based on all the factors that need to be taken into account.

### **6.3.3. Recommendations**

#### **Recommendations for future research**

The research demonstrates the feasibility of the approach of formulating the base service module based on the base demand and adding or subtracting service modules during other periods of different demand patterns. A possible direction for future research is to select the entire network as a case study, so as to produce more realistic results based on more comprehensive data. At the same time, more considerations such as more stakeholders' wishes and a more realistic simulation process can be added when developing services based on basic needs. This will allow a clearer view of the impact of this service design thinking on various aspects. In terms of data, the data selected for this case study are typical for the fall/winter period of 2022. Therefore, a good practice is to try more data from those unused periods for the study, or to use the predicted data for the future.

#### **Recommendations for NS in the future**

For NS, the first thing worth doing is to apply the methodology of the study to find periods with different demand characteristics. For example, using the projected demand data beyond 2023 as input, observe how the demand analyzed based on these data is distributed in the network. Although schedules may have already been produced for the next few years, it is still rewarding to compare them with those different periods: this helps NS to understand which areas deserve improvement. For example, some Sprinters do not need to stop at every stations, some IC trans can stop at more stations. Besides, the service does not necessarily need to be symmetrical. The train program can be more personalized for different groups. For example, additional train services are provided to target groups of students according to school hours. Depending on holiday schedules, increase the number of trains that can carry bicycles on board during holidays.

Explore how many “compromises” can be reduced: Compared with the line plan based on the peak period, the service designed according to the basic demand largely reduces the frequency and number of train stops. For the line in the case study, the new line plan reduces the amount of track capacity needed to meet passenger demand on the line. This is conceivably good news for the adjacent lines and the potential growth in demand. So another thing that NS can analyze and evaluate is: if the demand-oriented approach is adopted for designing the new timetable in the future, how much redundant service will be reduced and how much capacity will be freed up and more fully utilized.

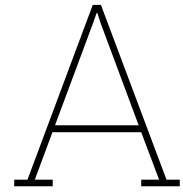
Consolidation and reallocation of resources: In terms of train equipment, NS can consider differentiating train equipment according to the different needs of the service. Because the basic service is designed based on the basic demand, the off-peak service can be made more attractive. For example, try to equip the base demand service with better train services, such as higher speed, and toilets, wheelchairs, etc. More first-class compartments may also be considered. And for all add-on services in the peak period, focus more on the capacity rather than the quality of travel. Instead of offering the first class compartment, there could be all second class compartments on the train, which provides with more seats and rooms for the passenger. Besides, the equipment for bicycles can also be removed to save more space. This provides a good separation between the functions of basic and service services: the former focuses on providing comfortable travel for all the basic needs that are always present, while the latter focuses on meeting the large additional growth in demand during peak periods.

# References

- Black, P. E. (2019). *Manhattan distance*. Dictionary of Algorithms and Data Structures [online]. Retrieved from <https://www.nist.gov/dads/HTML/manhattanDistance.html> (Accessed 2023)
- Bruijn, A., Hogenberg, J., & Guis, N. (2019). *Een betere dienstregeling door de grootste gemene deler*. Retrieved from [https://www.cvs-congres.nl/e2/site/cvs/custom/site/upload/file/cvs\\_2019/sessie\\_c/c3/cvs\\_83\\_een\\_betere\\_dienstregeling\\_door\\_de\\_grootste\\_gemene\\_deler\\_1\\_2019.pdf](https://www.cvs-congres.nl/e2/site/cvs/custom/site/upload/file/cvs_2019/sessie_c/c3/cvs_83_een_betere_dienstregeling_door_de_grootste_gemene_deler_1_2019.pdf)
- Brännlund, U., Lindberg, P. O., Nöu, A., & Nilsson, J.-E. (1998). Railway timetabling using lagrangian relaxation. *Transportation Science*, 32(4), 358–369. doi: 10.1287/trsc.32.4.358
- Castillo, E., Gallego, I., Ureña, J. M., & Coronado, J. M. (2011). Timetabling optimization of a mixed double- and single-tracked railway network. *Applied Mathematical Modelling*, 35(2), 859-878. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0307904X10002908> doi: <https://doi.org/10.1016/j.apm.2010.07.041>
- de Bruyn, M., Guis, N., & Banninga, J. (2022). *Vaker een trein, vaker op reis? de betekenis van frequentie voor het treinsysteem*. Retrieved from [https://cvs-congres.nl/e2/site/cvs/custom/site/upload/file/cvs\\_2022/150.pdf](https://cvs-congres.nl/e2/site/cvs/custom/site/upload/file/cvs_2022/150.pdf)
- de Bruyn, M., & Mestrum, D. (2021). *Een studie naar een vraagafhankelijke dienstregeling voor ns*. Retrieved from [https://cvs-congres.nl/e2/site/cvs/custom/site/upload/file/cvs\\_2021/sessie\\_a/a4/cvs\\_34\\_een\\_studie\\_naar\\_een\\_vraagafhankelijke\\_dienstregeling\\_voor\\_ns\\_1\\_2021.pdf](https://cvs-congres.nl/e2/site/cvs/custom/site/upload/file/cvs_2021/sessie_a/a4/cvs_34_een_studie_naar_een_vraagafhankelijke_dienstregeling_voor_ns_1_2021.pdf)
- Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, pp. 226–231).
- Goverde, R. M. P., & Hansen, I. A. (2013). Performance indicators for railway timetables. In *2013 IEEE International Conference on Intelligent Rail Transportation Proceedings* (p. 301-306). doi: 10.1109/ICIRT.2013.6696312
- Hartigan, J. A., Wong, M. A., et al. (1979). A k-means clustering algorithm. *Applied statistics*, 28(1), 100–108.
- Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3), 264–323.
- Jarman, A. M. (2020). Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method. *Georgia Southern University*.
- Jenks, G. F. (1967). The data model concept in statistical mapping. *International yearbook of cartography*, 7, 186–190.

- Jinjing, G., Jiang, Z., Fan, W., Wu, J., & Chen, J. (2020). Real-time passenger flow anomaly detection considering typical time series clustered characteristics at metro stations. *Journal of Transportation Engineering, Part A: Systems*, 146, 04020015. doi: 10.1061/JTEPBS.0000333
- Limtanakool, N., Dijst, M., & Schwanen, T. (2007). A theoretical framework and methodology for characterising national urban systems on the basis of flows of people: empirical evidence for france and germany. *Urban Studies*, 44(11), 2123–2145.
- Mahmoudzadeh, A., & Wang, X. B. (2020). Cluster based methodology for scheduling a university shuttle system. *Transportation Research Record*, 2674(1), 236-248. Retrieved from <https://doi.org/10.1177/0361198119900636> doi: 10.1177/0361198119900636
- Mishalani, R. G., Ji, Y., & McCord, M. R. (2011). Effect of onboard survey sample size on estimation of transit bus route passenger origin–destination flow matrix using automatic passenger counter data. *Transportation Research Record*, 2246(1), 64-73. Retrieved from <https://doi.org/10.3141/2246-09> doi: 10.3141/2246-09
- Mu, S., & Dessouky, M. (2011). Scheduling freight trains traveling on complex networks. *Transportation Research Part B: Methodological*, 45(7), 1103-1123. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0191261511000713> doi: <https://doi.org/10.1016/j.trb.2011.05.021>
- Nederlandse Spoorwegen. (2022). *Nederlandse spoorwegen, 2022 annual report*. (Available online at: <https://www.nsannualreport.nl/annual-report-2022/about-ns/the-profile-of-ns>)
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65. Retrieved from <https://www.sciencedirect.com/science/article/pii/0377042787901257> doi: [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- Sharma, S., Batra, N., et al. (2019). Comparative study of single linkage, complete linkage, and ward method of agglomerative clustering. In *2019 international conference on machine learning, big data, cloud and parallel computing (comitcon)* (pp. 568–573).
- Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques.. Retrieved from <https://hdl.handle.net/11299/215421>
- Sun, Y., Shi, J., & Schonfeld, P. M. (2016). Identifying passenger flow characteristics and evaluating travel time reliability by visualizing afc data: A case study of shanghai metro. *Public Transport*, 8(3), 341–363. doi: 10.1007/s12469-016-0137-8
- Van der Knaap, R. J. H., de Bruyn, M., Van Oort, N., Huisman, D., & Goverde, R. M. P. (2022). Extracting railway passenger demand patterns from origin-destination data for developing demand-oriented service plans.
- Van Oort, N. (2011). *Service reliability and urban public transport design* (Unpublished doctoral dissertation). Netherlands TRAIL Research School.

- Vansteenwegen, P., & Van Oudheusden, D. (2007). Decreasing the passenger waiting time for an intercity rail network. *Transportation Research Part B: Methodological*, 41(4), 478-492. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0191261506001044> doi: <https://doi.org/10.1016/j.trb.2006.06.006>
- Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301), 236–244.
- Yu, W., Bai, H., Chen, J., & Yan, X. (2019). Analysis of space-time variation of passenger flow and commuting characteristics of residents using smart card data of nanjing metro. *Sustainability*, 11(18). Retrieved from <https://www.mdpi.com/2071-1050/11/18/4989> doi: 10.3390/su11184989
- Zhang, Y., & Ng, S. T. (2021). Unveiling the rich-club phenomenon in urban mobility networks through the spatiotemporal characteristics of passenger flow. *Physica A: Statistical Mechanics and its Applications*, 584, 126377. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0378437121006506> doi: <https://doi.org/10.1016/j.physa.2021.126377>
- Zhao, R., Zhang, Z., Cheng, F., & Tang, H. (2017). Characteristics of urban rail transit passenger flow in chongqing. *DEStech Transactions on Computer Science and Engineering(cii)*. doi: 10.12783/dtcse/cii2017/17288



# Appendix

## A.1. Clustering process of the OD demand

This code mainly performs the clustering process. The input should be already processed and represent the demand data within the network for a week. It should contain five types of data: the day of the week, the time of the day, the origin station code, the destination station code, and the corresponding demand volume.

The output of this code contains three categories: the first is for each OD clustering result in the form of a visual bar chart with different color labels for different clusters. The second is an excel table storing for each OD, its corresponding OD code, the number of clusters for the optimal clustering result and the silhouette score value. The third is also an excel spreadsheet, storing the frequency of each time period selected as the base demand period over a week.

```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import os
4 from sklearn.cluster import AgglomerativeClustering
5 from sklearn.metrics import silhouette_score
6 from openpyxl import Workbook
7
8 # Read in Excel file
9 df = pd.read_excel('all stations.xlsx')
10 # df = pd.read_excel('Final data.xlsx')
11
12 # Extract the required property columns
13 data = df[['dag_vd_week', 'tijd', 'herkomst_transcode', '
14           bestemming_transcode', 'Calculation']]
15 # Week Name Dictionary
16 weekdays = {
17     1: 'Sunday',
18     2: 'Monday',
19     3: 'Tuesday',
20     4: 'Wednesday',
21     5: 'Thursday',
22     6: 'Friday',
23     7: 'Saturday'
```

```
23 }
24 # Replace the number in the 'dag_vd_week' column with the corresponding
    week name
25 data['dag_vd_week'] = data['dag_vd_week'].replace(weekdays)
26
27 # Store the final base demand period statistics
28 statistics = {}
29
30 # Save the optimal results and parameters for each combination
31 results = {}
32 count = 0
33
34 # Iterate through all combinations
35 for combo in data[['herkomst_transcode', 'bestemming_transcode']].
    drop_duplicates().values:
36     # Counting
37     # count = (count + 1)
38     # print(count)
39     subset = data[(data['herkomst_transcode'] == combo[0]) & (data['
        bestemming_transcode'] == combo[1])]
40
41     if len(subset) <= 1:
42         continue
43
44     # Determine if the average value of the data in 'Calculation' is less
        than 10
45     # print('Average value:',subset['Calculation'].mean())
46     # if subset['Calculation'].mean() < 10:
47     #     continue
48
49     combo_results = {}
50     print(' OD',combo[0],'-',combo[1])
51     print('Length of the data',len(subset))
52
53     for n_clusters in range(2, len(subset)):
54         if n_clusters == 6:
55             break
56         clustering = AgglomerativeClustering(n_clusters=n_clusters,
            linkage='average')
57         clustering.fit(subset[['Calculation']])
58         labels = clustering.labels_
59         if len(set(labels)) <= 1:
60             continue
61         silhouette_avg = silhouette_score(subset[['Calculation']], labels)
62         print('The',count,' iterationthe SC score is',silhouette_avg)
63         best_silhouette_avg = combo_results.get('best_silhouette_avg', -1)
64         if silhouette_avg > best_silhouette_avg:
65             combo_results['best_silhouette_avg'] = silhouette_avg
66             combo_results['best_labels'] = labels
67             combo_results['best_n_clusters'] = n_clusters
68     # print('The optimal SC and cluster number results are',combo_results
        ['best_silhouette_avg'],combo_results['best_n_clusters'])
69     if not combo_results: # No optimal results were found for this
        combination
70         continue
71     results[tuple(combo)] = combo_results
```



```

72     ...
73
74     subset['tijd'] = pd.to_datetime(subset['tijd'], format='%H:%M:%S').dt.
       time.astype(str)
75     x = (subset['dag_vd_week'] + ' ' + subset['tijd'])
76
77     # Get the label of the cluster containing the most points
78     best_labels = combo_results['best_labels']
79     max_cluster_label = max(set(best_labels), key=best_labels.tolist().
       count)
80
81     # Get the corresponding time period
82     max_cluster_times = x[best_labels == max_cluster_label]
83
84     # Count the number of times each time period is tagged
85     for time in max_cluster_times:
86         if time in statistics:
87             statistics[time] += 1
88         else:
89             statistics[time] = 1
90
91     # Visualize and save results images
92     plt.bar(x, subset['Calculation'], width=0.8, color=plt.cm.tab10(
       combo_results['best_labels'] / combo_results['best_n_clusters']),
       edgecolor='black', linewidth=0)
93     # plt.xlabel('Dag_vd_Week and Tijd')
94     plt.ylabel('Number of Passengers')
95     plt.xticks(range(0, len(x), 16), x[::16], rotation=90) # Display
       every 16 scales and set the scale to string type x
96     plt.title('Hierarchical Clustering for Combination {}-{}\nBest
       Silhouette Score: {:.2f}\nNumber of Clusters: {}'.format(combo[0],
       combo[1], combo_results['best_silhouette_avg'], combo_results['
       best_n_clusters']))
97     plt.subplots_adjust(bottom=0.35)
98     plt.show()
99     plt.savefig(os.path.join('results', '{}-{}.pdf'.format(combo[0], combo
       [1])))
100
101     plt.clf() # Clear the current canvas
102
103 # Create a table to record the clustering results for each OD store,
       including OD pair names, SCs and number of clusters
104 new_data = pd.DataFrame(columns=["Combination", "Best Silhouette Score", "
       Number of Clusters"]).fillna(0)
105 for combo, combo_results in results.items():
106     new_row = {"Combination": combo, "Best Silhouette Score":
       combo_results['best_silhouette_avg'], "Number of Clusters":
       combo_results['best_n_clusters']}
107     new_data = new_data.append(new_row, ignore_index=True)
108 # write the DataFrame to an Excel file
109 new_data.to_excel("clustering result.xlsx", index=False)
110 #
111 # Create workbook and table objects to store the number of times each time
       period is determined to be a base demand period
112 wb = Workbook()
113 ws = wb.active

```

```
114 ws.append(['tijd', 'statistics'])
115 # Writing statistics to an Excel sheet
116 for time, count in statistics.items():
117     ws.append([time, count])
118 # Saving Excel files
119 wb.save('statistics.xlsx')
```

## A.2. Clustering process of the frequency

For this code, the results of further clustering analysis of the frequencies output from the previous step are mainly implemented. Where the input is a table containing the data for each time point, and the corresponding frequency values.

The output is a visualization of the clustering results, with different clusters having different frequency values. Where different clusters are represented by different colors, representing the periods that might have different demand characteristics.

```

1 import pandas as pd
2 import matplotlib.pyplot as plt
3 from sklearn.cluster import AgglomerativeClustering
4 from sklearn.metrics import silhouette_score
5
6 # Read in Excel file
7 df = pd.read_excel('base demand period.xlsx')
8
9 # Extracts the required property columns
10 data = df[['tijd', 'statistics']]
11 length = len(data['statistics'])
12 best_silhouette_avg = -1 # For saving the optimal Silhouette coefficients
13 best_labels = None # Used to save the clustering results corresponding to
    the optimal Silhouette coefficients
14 best_n_clusters = None # For saving the optimal number of clusters
15 best_distance_threshold = None # Used to save the optimal distance
    threshold
16
17 # Try different number of clusters, calculate Silhouette coefficients and
    save the optimal clustering results
18 for n_clusters in range(2,10):
19     clustering = AgglomerativeClustering(n_clusters=n_clusters,
        distance_threshold=None, linkage='average')
20     clustering.fit(data[['statistics']])
21     labels = clustering.labels_
22     silhouette_avg = silhouette_score(data[['statistics']], labels)
23     if silhouette_avg > best_silhouette_avg:
24         best_silhouette_avg = silhouette_avg
25         best_labels = labels
26         best_n_clusters = n_clusters
27         best_distance_threshold = None
28
29 # Add the optimal clustering results to the data frame
30 data['cluster_label'] = best_labels
31 data['n_clusters'] = best_n_clusters # Fix the variable named
    best_n_clusters
32
33 # Visualization display
34 plt.bar(data['tijd'], data['statistics'], width=0.8, color=plt.cm.tab10(
    data['cluster_label'] / best_n_clusters), edgecolor='black', linewidth
    =0) # cluster_labelbest_n_clusters
35 plt.xlabel('Dag_vd_Week and Tijd')
36 plt.ylabel('Number of overlaps')
37 plt.xticks(range(0, len(data['tijd']), 16), data['tijd'][:16], rotation

```

```
    =90)
38 plt.title('Hierarchical Clustering with {} clusters'.format(
    best_n_clusters)) # Use string formatting to fix title settings
39 plt.subplots_adjust(bottom=0.35)
40 plt.show()
```

### A.3. The demand matrix acquisition and consolidation

The output from last code enable the preliminary division of time periods, which will work as the input for this code. The input will be the content in the “groups” (the 48<sup>th</sup> line of the code).

The output includes two parts: The preliminary divided demand matrices (which are 15 in this study) and the merged 8 demand matrices. All results will be saved in excel.

```
1 import pandas as pd
2 import numpy as np
3 from openpyxl import Workbook
4
5 # Read the "all stations" file
6 df_stations = pd.read_excel("all_stations.xlsx")
7 # df_stations = pd.read_excel("Final data.xlsx")
8
9 # Read the "station_list" file
10 df_station_list = pd.read_excel("station_list.xlsx", usecols=[2])
11 station_codes = df_station_list["Code"].tolist()
12 def percentage(per,data):
13     # Sorting data in ascending order
14     sorted_data = np.sort(data)
15     # Calculate the index corresponding to 90% of the data
16     index = int(per * len(sorted_data))
17     # Select the value at the calculated index
18     selected_value = sorted_data[index]
19     # Find the number larger than the index value and calculate the sum of
20     # the difference with the flat index
21     diff_sum = sum([num - selected_value for num in data if num >
22     selected_value])
23     # 1 - Percentage of the sum of differences to the sum of all data
24     percentage = (1 - diff_sum / sum(data)) * 100
25     return percentage
26 def select_value_above_per_percent(per,data):
27     # Sorting data in ascending order
28     sorted_data = np.sort(data)
29
30     # Calculate the index corresponding to per% of the data
31     index = int(per * len(sorted_data))
32
33     # Select the value at the calculated index
34     selected_value = sorted_data[index]
35     return selected_value
36
37 #Merge matrix function
38 def merge_matrices(matrix1, matrix2):
39     merged_matrix = np.maximum(matrix1, matrix2)
40     return merged_matrix
41
42 # Merge the specified matrix numbers
43 def merge_selected_matrices(matrices, selected_indices):
44     merged_matrix = matrices[selected_indices[0]]
45     for index in selected_indices[1:]:
46         merged_matrix = merge_matrices(merged_matrix, matrices[index])
47     return merged_matrix
```

```

46
47 # Defining combinations and groupings
48 groups = {
49     1: ["2", "07:00:00", "07:30:00", "08:00:00", "08:30:00"],
50     2: ["3", "07:00:00", "07:30:00", "08:00:00", "08:30:00", "09:00:00"],
51     3: ["4", "07:00:00", "07:30:00", "08:00:00", "08:30:00", "09:00:00"],
52     4: ["5", "07:00:00", "07:30:00", "08:00:00", "08:30:00", "09:00:00"],
53     5: ["6", "07:30:00", "08:00:00", "08:30:00"],
54     6: ["2", "15:00:00", "15:30:00", "16:00:00", "16:30:00"],
55     7: ["3", "15:00:00", "15:30:00", "16:00:00"],
56     8: ["4", "15:00:00", "15:30:00", "16:00:00", "16:30:00"],
57     9: ["5", "15:00:00", "15:30:00", "16:00:00"],
58     10: ["6", "15:00:00", "15:30:00", "16:00:00", "16:30:00", "17:00:00",
59         "17:30:00", "18:00:00"],
60     11: ["2", "17:00:00", "17:30:00", "18:00:00"],
61     12: ["3", "16:30:00", "17:00:00", "17:30:00", "18:00:00"],
62     13: ["4", "17:00:00", "17:30:00", "18:00:00"],
63     14: ["5", "16:30:00", "17:00:00", "17:30:00", "18:00:00"],
64     # 15: ["1", "2", "3", "4", "5", "6", "7", "06:00:00", "06:30:00", "07:00:00",
65         "07:30:00", "08:00:00", "08:30:00", "09:00:00",
66         "09:30:00", "10:00:00", "10:30:00", "11:00:00", "11:30:00",
67         "12:00:00", "12:30:00", "13:00:00", "13:30:00", "14:00:00",
68         "14:30:00", "15:00:00", "15:30:00", "16:00:00", "16:30:00",
69         "17:00:00", "17:30:00", "18:00:00", "18:30:00", "19:00:00",
70         "19:30:00", "20:00:00", "20:30:00", "21:00:00", "21:30:00"]
71 }
72
73 # Initialize 15 matrices
74 matrices = []
75 for _ in range(15):
76     matrix = np.zeros((26, 26))
77     matrices.append(matrix)
78
79 matrices[14] = [[[] for _ in range(26)] for _ in range(26)]
80 # Record the combinations that have been used
81 used_combinations = set()
82 count = 0
83 # Putting data into the matrix
84 for _, row in df_stations.iterrows():
85     dag_vd_week = str(row["dag_vd_week"])
86     tijd = str(row["tijd"])
87     herkomst_transcode = row["herkomst_transcode"]
88     bestemming_transcode = row["bestemming_transcode"]
89     calculation = row["Calculation"]
90
91     # Find the group to which the combination belongs
92     group_id = None
93     for gid, group_values in groups.items():
94         if dag_vd_week in group_values and tijd in group_values:
95             group_id = gid
96             break
97     if group_id is None:
98         group_id = 15
99     # count = (count+1)
100    # print(group_id)
101    # print(count)

```

```

98
99     if group_id is not None:
100         if group_id == 15:
101             matrix = matrices[group_id - 1]
102             herkomst_index = station_codes.index(herkomst_transcode)
103             bestemming_index = station_codes.index(bestemming_transcode)
104             # Store all values of each position of the base demand matrix
105             # for subsequent comparison
106             matrix[herkomst_index][bestemming_index].append(calculation)
107         else:
108             matrix = matrices[group_id - 1]
109             herkomst_index = station_codes.index(herkomst_transcode)
110             bestemming_index = station_codes.index(bestemming_transcode)
111             matrix[herkomst_index][bestemming_index] = max(matrix[
112                 herkomst_index][bestemming_index], calculation)
113
114 matrix = matrices[14]
115 s = 0
116 volume = 0
117 for i in range(26):
118     for j in range(26):
119         s = s + sum(matrix[i][j])
120
121 for i in range(26):
122     for j in range(26):
123         a = matrix[i][j]
124         if a:
125             # try z from 0.8 to 0.99, where z is the percentage of periods
126             # that will be satisfied
127             z = 0.94
128             c = percentage(z, a)
129             volume = volume + c * 0.01 * sum(matrix[i][j])
130             matrix[i][j] = select_value_above_per_percent(z, a)
131         else:
132             matrix[i][j] = 0
133
134 per = volume/s
135
136 k = 0
137 for i in range(26):
138     for j in range(26):
139         k = k + matrix[i][j]
140
141 # Creating an Excel Workbook
142 workbook = Workbook()
143
144 # Write 15 matrices to an Excel sheet
145 for group_id, matrix in enumerate(matrices):
146     group_sheet = workbook.create_sheet(title=f"Group {group_id + 1}")
147
148     # Write table header
149     group_sheet.cell(row=1, column=1, value="")
150     for i, code in enumerate(station_codes):
151         group_sheet.cell(row=1, column=i + 2, value=code)
152         group_sheet.cell(row=i + 2, column=1, value=code)

```

```
151     # Write to matrix data
152     for i in range(26):
153         for j in range(26):
154             group_sheet.cell(row=i + 2, column=j + 2, value=matrix[i][j])
155
156 workbook.remove(workbook["Sheet"])
157 workbook.save("output.xlsx")
158
159
160 # Merge Matrix
161 merged_matrix1 = merge_selected_matrices(matrices, [0, 2])
162 merged_matrix2 = merge_selected_matrices(matrices, [1, 3])
163 merged_matrix3 = merge_selected_matrices(matrices, [5, 6, 7, 8])
164 merged_matrix4 = merge_selected_matrices(matrices, [10, 12])
165 merged_matrix5 = merge_selected_matrices(matrices, [11, 13])
166
167 workbook = Workbook()
168
169 # Write the merged matrix to an Excel sheet
170 sheets = [merged_matrix1, merged_matrix2, merged_matrix3, merged_matrix4,
171           merged_matrix5]
172 for i, sheet_data in enumerate(sheets):
173     sheet = workbook.create_sheet(title=f"Sheet {i+1}")
174     for row in sheet_data:
175         sheet.append(list(row)) # Converting numpy arrays to lists
176
177 # Save Excel file
178 workbook.save("output_new.xlsx")
```



## A.4. Similarity calculation of matrices

This code mainly calculation the similarity between matrices, where the Manhattan Distance is used as the measure approach. The input will be different demand matrices and they have the same size. The output will be the Manhattan Distance value between each matrix.

```
1 import pandas as pd
2 import numpy as np
3
4 # Read the Excel file
5 excel_file = pd.ExcelFile('output.xlsx')
6
7 # Get the sheet names
8 sheet_names = excel_file.sheet_names
9
10 # Initialize a list to store the matrices
11 matrices = []
12
13 # Iterate over each sheet and extract the matrix data
14 for sheet_name in sheet_names:
15     # Read the sheet into a DataFrame
16     df = excel_file.parse(sheet_name)
17
18     # Extract the matrix data from the DataFrame, excluding the header row
19     matrix = df.values[1:, :]
20
21     # Append the matrix to the list
22     matrices.append(matrix)
23
24 # Calculate the Manhattan distance between each pair of matrices
25 num_matrices = len(matrices)
26 manhattan_distances = np.zeros((num_matrices, num_matrices))
27
28 for i in range(num_matrices):
29     for j in range(i + 1, num_matrices):
30         manhattan_distance = np.sum(np.abs(matrices[i].flatten() -
31         matrices[j].flatten()))
32         manhattan_distances[i, j] = manhattan_distance
33
34 # Print the Manhattan distances
35 for i in range(num_matrices):
36     for j in range(i + 1, num_matrices):
37         if manhattan_distances[i, j] != 0:
38             print(f"Manhattan distance between matrix {i+1} and matrix {j
39             +1}: {manhattan_distances[i, j]}")
```

## A.5. Acquire segment demand

This code transform the demand OD matrices into the segment flow demand. The input will be the OD matrix for each period, and the output will be the segment demand value for each segment.

```

1 import pandas as pd
2
3 # Read Excel files
4 excel_file = pd.ExcelFile('Schedule.xlsx')
5
6 # Create a new Excel sheet
7 output_file = pd.ExcelWriter('line_flow.xlsx')
8
9 # Iterate through each sheet
10 for sheet_name in excel_file.sheet_names:
11     # Retrieve data from the current sheet
12     df = excel_file.parse(sheet_name)
13     print(df)
14
15     # Segment passenger flow dictionary
16     segment_flow = {}
17
18     # Calculation of passenger flow for the InterCity train
19     segment_flow['1-5'] = df.loc[0, 5] + df.loc[0,6]
20     segment_flow['5-6'] = df.loc[4, 6] + df.loc[0,6]
21     segment_flow['6-5'] = df.loc[5, 5] + df.loc[5,1]
22     segment_flow['5-1'] = df.loc[4, 1] + df.loc[5,1]
23
24     # # Calculation of passenger flow for the Sprinter
25     # segment_flow['1-2'] = df.loc[0,2] + df.loc[0,3] + df.loc[0,4]
26     # segment_flow['2-3'] = df.loc[0,3] + df.loc[0,4] + df.loc[1,3] + df.
27         loc[1,4] + df.loc[1,5] + df.loc[1,6]
28     # segment_flow['3-4'] = df.loc[0,4] + df.loc[1,4] + df.loc[1,5] + df.
29         loc[1,6] + df.loc[2,4] + df.loc[2,5] + df.loc[2,6]
30     # segment_flow['4-5'] = df.loc[1,5] + df.loc[1,6] + df.loc[2,5] + df.
31         loc[2,6] + df.loc[3,5] + df.loc[3,6]
32     # segment_flow['5-6'] = df.loc[1,6] + df.loc[2,6] + df.loc[3,6]
33     # segment_flow['2-1'] = df.loc[3,1] + df.loc[2,1] + df.loc[1,1]
34     # segment_flow['3-2'] = df.loc[5,2] + df.loc[4,2] + df.loc[3,2] + df.
35         loc[3,1] + df.loc[2,2] + df.loc[2,1]
36     # segment_flow['4-3'] = df.loc[5,3] + df.loc[5,2] + df.loc[4,3] + df.
37         loc[4,2] + df.loc[3,3] + df.loc[3,2] + df.loc[3,1]
38     # segment_flow['5-4'] = df.loc[4,4] + df.loc[4,3] + df.loc[4,2] + df.
39         loc[5,4] + df.loc[5,3] + df.loc[5,2]
40     # segment_flow['6-5'] = df.loc[5,4] + df.loc[5,3] + df.loc[5,2]
41
42     # Creating a Resulting DataFrame
43     result_df = pd.DataFrame({'Segment': list(segment_flow.keys()),
44                             'Flow': list(segment_flow.values())})
45
46     # Write the result to the Excel table corresponding to the current
47         sheet
48     result_df.to_excel(output_file, sheet_name=sheet_name, index=False)

```

```
43 # Save and close the Excel file
44 output_file.save()
```