

VRIJE UNIVERSITEIT AMSTERDAM

MASTER THESIS BUSINESS ANALYTICS

---

# Improving the probabilistic passenger forecast distribution

---

*Author:*  
Sebastiaan Groeneveld

*Graduation supervisor:*  
Dr. Paulo Jorge de Andrade  
Serra

*Second reader:*  
Prof. dr. Mark Hoogendoorn

*External supervisors:*  
Bert de Vries  
Simone Griffioen  
Tjebbe Hepkema



September 2021



---

## Preface

Before you lies the Master Thesis “Improving the probabilistic passenger forecast distribution”. The research on how to improve the prediction of the probability distribution of the expected number of passengers was carried out at NS at Team Reizigersprognoses. This thesis was written as part of my graduation from the Business Analytics programme at VU University Amsterdam and on behalf of the NS company. From February 2021 to September 2021, I have been working on this research and writing this thesis.

I would like to thank my NS supervisors Simone Griffioen, Bert de Vries and Tjebbe Hepkema for their excellent guidance and support during this process. The weekly Teams meetings were incredibly helpful and it was also very pleasant that I could go to the office once a week despite the corona pandemic. This certainly improved my internship experience. To my other colleagues at NS: I would like to thank you for your wonderful cooperation. In particular, I would like to thank Mischa van der Haar for her guidance at NS. And thank you to Erik Ramaker for recommending the internship.

From the VU University I would like to thank Dr. Paulo Jorge de Andrade Serra. You were always very helpful in the meetings and you had a clear answer for everything. It was also really pleasant that you always responded to my e-mails very quickly and had feedback on the chapters of my report ready promptly. I would also like to thank Prof. dr. Mark Hoogendoorn for being my second reader.

I hope you enjoy reading my thesis.

Sebastiaan Groeneveld

Utrecht, September 2021



---

## Abstract

The Dutch passenger railway operator NS (Netherlands Railways) aims to offer the best possible train schedule to their passengers. To achieve this, NS uses passenger number forecasts for each train that is deployed. However, these forecasts are uncertain as these numbers are rarely fully correct. In order to capture this uncertainty, NS has developed two prediction methods that determine the probability distribution of the expected number of passengers. Yet, NS thinks that these prediction methods can be improved. In this study we investigate how the probabilistic passenger forecast distribution can be improved.

We first discuss when we consider a probabilistic prediction good and we suggest evaluation methods for this. Because good predictions for higher passenger numbers and/or route travel time are important factors to NS, these are included in the evaluation. In addition to the two prediction methods used by NS, the current method and the 2.0-prognoseverdeling method, we propose two new prediction methods: the 3.0-prognoseverdeling method and quantile regression.

The results show that the 2.0-prognoseverdeling method, the 3.0-prognoseverdeling method and quantile regression perform significantly better than the current method. Considering the importance of higher passenger numbers and/or route travel times, quantile regression achieves the best prediction of the probabilistic distribution. Therefore we can conclude that quantile regression is the best prediction method of the examined methods.

This study shows that there is room for improvement in the current prediction of the probabilistic passenger forecast distribution. Therefore, we recommend to further investigate the researched methods in this study, predominantly quantile regression since it shows the most promising results.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Data Analysis</b>	<b>5</b>
2.1	Datasets . . . . .	5
2.1.1	SOFA dataset . . . . .	5
2.1.2	Passenger forecast dataset . . . . .	6
2.1.3	Additional datasets . . . . .	8
2.2	Data transformations . . . . .	9
2.3	Data split . . . . .	11
<b>3</b>	<b>Methodology</b>	<b>12</b>
3.1	Evaluation method . . . . .	12
3.1.1	Percentages of stop-stop prognoses . . . . .	12
3.1.2	WMAPE( $\tau$ ) values . . . . .	13
3.1.3	WMAPE <sub>p<sub>50</sub> × t</sub> ( $\tau$ ) values . . . . .	14
3.2	Current method of NS . . . . .	15
3.3	2.0-prognoseverdeling method . . . . .	15
3.4	3.0-prognoseverdeling method . . . . .	16
3.5	Quantile Regression . . . . .	19
3.6	Quantile Regression with weights . . . . .	23
<b>4</b>	<b>Results</b>	<b>25</b>
4.1	Evaluation method: percentages . . . . .	25
4.2	Evaluation method: WMAPE( $\tau$ ) values . . . . .	27
4.2.1	Without observation weights . . . . .	28
4.2.2	With observation weights . . . . .	29
4.2.3	Results test set . . . . .	30
4.3	Error contribution per factor level . . . . .	33
4.3.1	Weekday . . . . .	33
4.3.2	Daypart . . . . .	35
4.3.3	Area . . . . .	36
<b>5</b>	<b>Discussion</b>	<b>38</b>
5.1	Key findings . . . . .	38
5.2	Interpretation of results . . . . .	38
5.3	Limitations of the research . . . . .	40

<b>6</b>	<b>Conclusions</b>	<b>41</b>
6.1	Main conclusions . . . . .	41
6.2	Recommendations for future research . . . . .	42
<b>A</b>	<b>Numbers of missing data</b>	<b>44</b>
<b>B</b>	<b>Details of variable Area</b>	<b>45</b>
<b>C</b>	<b>Step-up steps for quantile regression without observation weights</b>	<b>46</b>
<b>D</b>	<b>Step-up steps for quantile regression with observation weights</b>	<b>49</b>
<b>E</b>	<b>Example of differences in minimums</b>	<b>52</b>



# Chapter 1

## Introduction

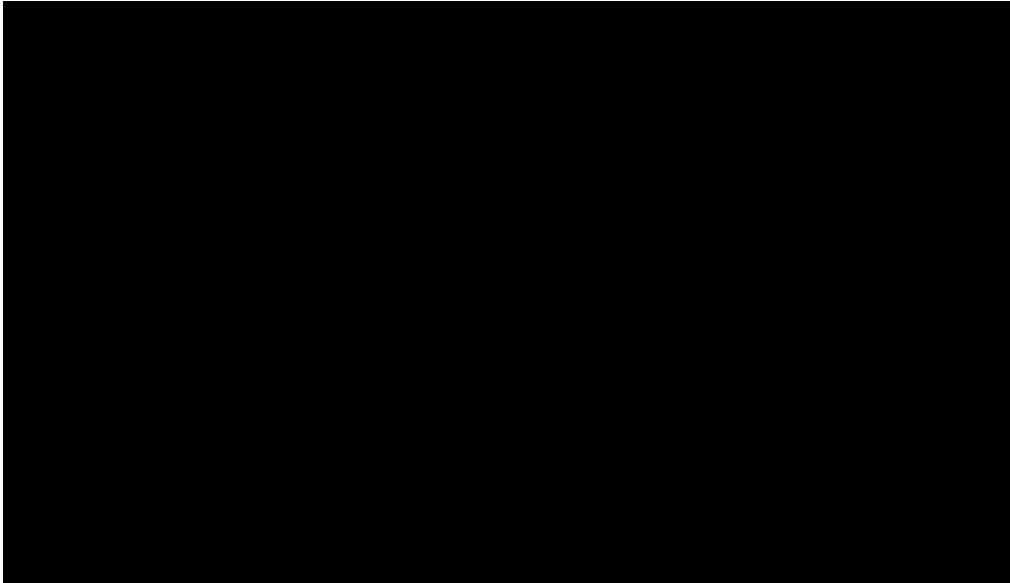
NS (Netherlands Railways) is the biggest passenger railway operator in the Netherlands. NS runs about 4,800 scheduled domestic trains a day, serving over one million passengers a day. Besides being the largest railway operator in The Netherlands, NS also enables traveling abroad to many European countries via NS International. The goal of NS is to transport passengers to their destination, as fast and as comfortable as possible, whilst being as sustainable as possible for the environment. To realise this, a timetable is constructed with a daily schedule of employed trains across the country. The timetable is updated annually and in the timetable the rolling stock plan is often assessed and optimised. This includes changes due to maintenance, disruptions or other issues.

Generally speaking, timetable planning at NS works as follows: a timetable is drawn up a year in advance for an entire year, followed by rolling stock planning and staff planning. After this, a “Basisdagen” (BD) plan is made about ten months in advance and sent to ProRail (the railway manager of the Netherlands) for approval. Once ProRail has determined the timetable, transport operators and ProRail can still make changes. This is done by means of the “Basisdagen Update” (BDu) six months in advance. The last planned update of the rolling stock plan is done by means of a plan for specific days (SD), five weeks in advance. The standard week is converted to specific dates. On those days, all kinds of changes can be defined, such as incidental maintenance, extra trains due to events and trains for ad hoc transport. The whole timeline for all NS passenger forecast types is shown in Figure 1. This plot shows that NS also makes long term and medium-long term forecasts, which are made even earlier than a year in advance. This study focuses on the passenger forecasts made six months in advance, the BDu.



Figure 1: NS passenger forecast types including a timeline.

In order to make the best timetable and rolling stock deployment plan for a BDU, accurate predictions of the number of passengers are required. NS has plenty of data on the daily offer of train rides in the national train network of the Netherlands including day and time of train rides, the amount of seats and the amount of passengers. The path from data to passenger forecasts is shown in Figure 2.



**Figure 2:** The path from data to passenger forecasts.

NS uses mainly check-in-check-out (CiCo) data to draw up Origin-Destination matrices per day, which represent the past number of passengers between each pair of train stations for that day. A factor for the expected growth in passenger numbers for the following year is applied to this. The passengers are then allocated to all trains in the schedule of that day using the traffic planning software ‘VISUM’, which is the only professional traffic planning software that provides a highly detailed representation of all modes of public transport according to PTV Group (2021). The result is a matrix containing a number of expected passengers for each combination of train number, day of the week and stop-stop (which is a part of the route of a train, described by an origin and a destination station) for each day in the BDU. So if a BDU covers two months, a specific combination of train number, day of the week and stop-stop will have the same train length for about eight days, and therefore the same passenger forecast is used for these days.

In this study, we will call a combination of train number, day of the week and stop-stop a “stop-stop prognose”. The prediction of the number of passengers for a stop-stop prognose in a BDU will be called the “ $P_{50}$ ”, which is the median.

The prediction of the number of passengers for each stop-stop prognose is given by a probability distribution, which is created in two steps: first, a prediction of the number of passengers is made (point 8 in Figure 2), which is the  $P_{50}$ , and then the distribution of this prediction is determined (point 9 in Figure 2). The reason why

the prediction of the number of passengers is determined separately is because of the new timetable for the to be predicted BDU period. It is likely that there will be new combinations of train number, day of the week and route that did not exist in the past. On top of that, there are also no properties that allow you to use data on similar stop-stop prognoses in the past. After all, it does not depend on the properties of that stop-stop prognose, but on the overall coherence of the timetable. As a result, the data at that level can no longer be used, making it necessary to use an allocation model for the Origin-Destination passengers. It may not be statistically correct to use a different model for the distribution, but this ensures that it is very robust to timetable changes.

This research will not focus on predicting the  $P_{50}$  of the number of passengers, but on its distribution, represented by percentiles  $P_{\tau}$  with  $\tau \in (0, 100)$ . In order to create the best train schedule for the passengers, it is not only important to know the mean or median expected number of passengers, but also the other possible numbers of passengers and the probability of those. A train schedule that accurately fits passenger demand is very important to NS: if too few rolling stock units are used, passengers would have to travel without a seat, possibly resulting in complaints. Yet, if too many rolling stock units are deployed, this will lead to unnecessarily high costs. Of course, both are undesirable.

NS has provided two methods for determining the percentiles for the  $P_{50}$  of a stop-stop prognose: the “current method” and the “2.0-prognoseverdeling method” (2.0-forecast distribution method). The current method to determine the desired percentiles for a given  $P_{50}$  does not take the uncertainty in this  $P_{50}$  into account and does not perform well enough as we will see in this study, resulting in inaccurate passenger forecasts. The 2.0-prognoseverdeling method was developed later than the period of the data we use and does take the uncertainty in the  $P_{50}$  into account. It determines the percentiles in a different way than the current method and will also be compared with the current method and the newly introduced methods. All of the above motivates the research question of this thesis: “How can the estimation of certain percentile values in the probability distribution of a predicted number of passengers for a combination of train number, day of the week and route in a BDU be improved?”

Two new methods are introduced in this study: the “3.0-prognoseverdeling method”, which is an extension of the 2.0-prognoseverdeling method, and “quantile regression”. Quantile regression is a method specifically developed for determining percentiles for some dependent variable and one or more independent variables as predictors (Koenker and Hallock, 2001).

The methods are evaluated in several ways. The first evaluation method is by determining the percentage of some test data that is less than or equal to the percentiles. The closer the percentages are to the percentile ranks, the better the method has estimated them. The percentile rank of the 25th percentile ( $\tau = 25$ ), for example, is 25 (Glen, 2021). NS has used this evaluation method before, but found that this does not capture well enough whether the predictions of the percentiles were correct as

this only says something about the average of the forecasts. If a prediction method performs well on average, it does not necessarily mean that the method predicts individual stop-stop prognoses well.

In order to have a better error measurement for the prediction methods, a more suitable evaluation method is also sought in this study. We look for an evaluation method that also takes into account the forecast errors for individual stop-stop prognoses. On top of that, for NS it is less important to predict the distribution perfectly for lower passenger numbers since the least number of seats a train can have is about 150 seats. Therefore, the planned train will be the same regardless of whether the distribution is predicted well or not. In addition to passenger numbers, route travel time is also an important factor, because for a passenger it is more annoying to have to stand for a longer amount of time. So we are also looking for an evaluation method that favours good predictions of higher passenger numbers and/or route travel times.

The remainder of the report is organised as follows. In Chapter 2 the datasets for this research are explained, along with some added variable transformations. The methods used to answer the research question are described in Chapter 3, with its results in Chapter 4. Lastly, the discussion of the results will be presented in Chapter 5, followed by the conclusion including recommendations for future research in Chapter 6.

# Chapter 2

## Data Analysis

In this section the datasets that are used for this research will be explained and explored, along with a number of descriptive graphs. After this, the data transformations that are used will be described, followed by how the data is split into a training set, validation set and test set.

### 2.1 Datasets

This research focuses on the data from the month of April in 2019 because starting from the BDU in that period (starting on 7 April), NS has carried out a model change to the model that determined the percentiles. As a result, the percentiles of the period before this BDU are not representative. The reason we do not choose a month in 2020 or 2021 is because of the coronavirus, as passenger numbers in this period are not representative of normal conditions. In order not to make the calculation time of the methods too long due to the limited time of the study, we decide to only look at the available data for the month of April 2019 (i.e., from 7 to 30 April). On top of that, we only considered weekdays, not weekend days. We did this because the distribution for these days is the most important for NS to predict. The variables for the methods are extracted from different datasets, which will be explained individually below.

#### 2.1.1 SOFA dataset

The first dataset is called the “SOFA dataset”, containing 619,687 stop-stop prognoses for the month of April. For each of these stop-stop prognoses, the dataset contains the exact departure and arrival times and the realised number of passengers. This realised number of passengers is the best estimate NS has at the moment. For the purpose of this study, the SOFA data are taken as truth.

We extract two additional variables from the date: the “Weekday” and the “Daypart”. The variable Weekday has five levels: Monday to Friday. The variable Daypart has three levels: morning rush hour, off-peak hours and evening rush hour. Morning

rush hour is from 7:00 to 9:00, evening rush hour from 16:00 to 18:00, and off-peak hours are all remaining hours of a day.

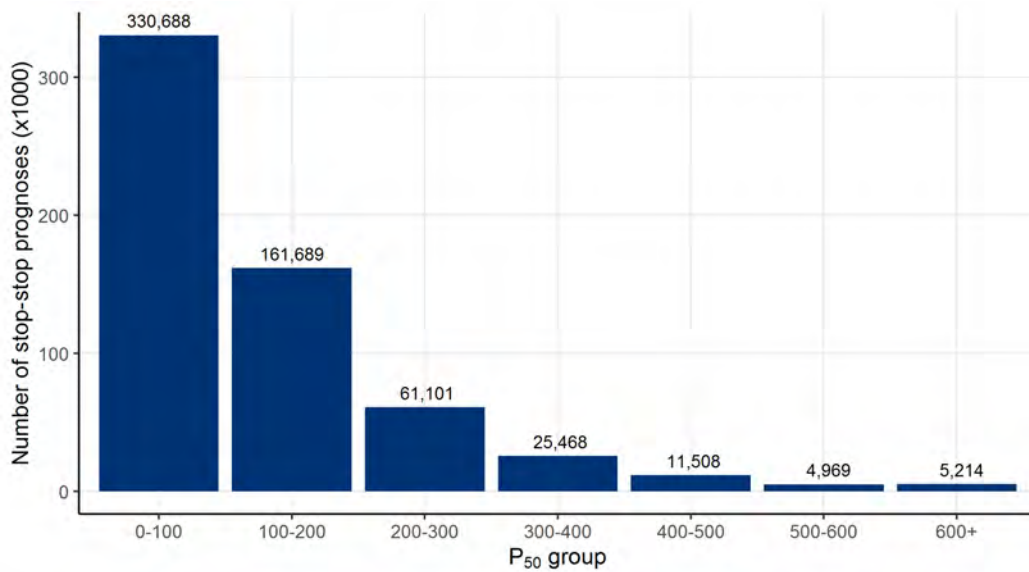
Instead of dividing a day into dayparts, making it a factor variable, we could also use the exact departure time of a train, making it a temporal variable. The reason is that in the off-peak hours, the trains with departure times just before or after the rush hours are probably more similar to the rush hours than to the off-peak hours themselves. It might also be that the peaks in passenger numbers could be more accurately captured when using the exact times. To do this, the departure times have to be transformed in order to become useful to include in the methods. This transformation is explained in detail in Section 2.2.

### 2.1.2 Passenger forecast dataset

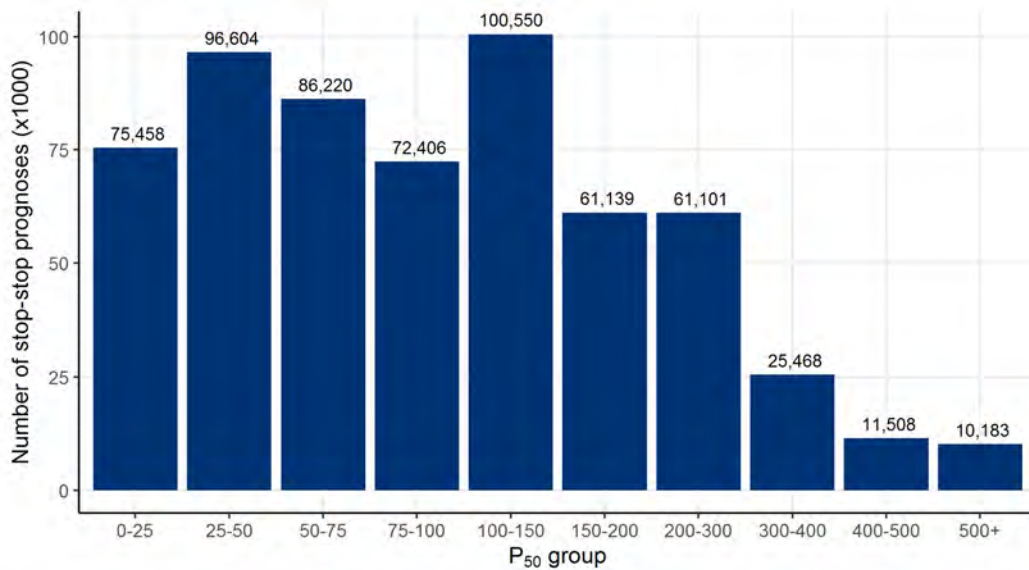
We merged a second dataset with the SOFA dataset containing all desired percentiles of the passenger forecast for 641,151 stop-stop prognoses. This dataset also contains the passenger predictions per stop-stop prognose, the  $P_{50}$ . The number of observations in this dataset and the SOFA dataset may differ, as it is possible that a planned train will not run for some reason. In this case, there is data from one or more stop-stop prognoses in the passenger forecast dataset missing in the SOFA dataset. It could also be that a train has to run a different route due to a timetable change. In this case, there is data on realised passengers for which there was no prediction. The exact differences in numbers of stop-stop prognoses can be seen in Appendix A. The datasets are linked in such a way that both the percentiles of the passenger forecasts as well as the realised number of passengers are present for all stop-stop prognoses, resulting in a dataset containing 600,637 stop-stop prognoses. So, for this study, stop-stop prognoses where either the realised number of passengers or the predictions for the percentiles are missing have been discarded.

We created two new variables using the  $P_{50}$  forecast, dividing the data into different  $P_{50}$  groups. This is done to change the  $P_{50}$  forecast from a numerical variable to a factor variable, which is needed for two of the prediction methods. On top of that, we make the groups to aggregate similar stop-stop prognoses. Because for an individual stop-stop prognose, there are only about eight observations for a BDU of, say, two months, which is far too few observations to give a reliable prediction of the distribution. The first  $P_{50}$  group is used in the 2.0-prognoseverdeling method and divides the  $P_{50}$  forecasts into groups of 100 passengers up to 600 passengers and one group of 600+ passengers, i.e.  $[100(n-1), 100n)$  for  $n \in \{1, \dots, 6\}$  or  $[600, \infty)$ . This grouping variable will be called the Static size  $P_{50}$  group. Figure 3 shows the number of stop-stop prognoses in each of these groups.

Figure 3 shows that the distribution of the number of stop-stop prognoses between the groups is very positively skewed. The ratio of the highest to the lowest number of stop-stop prognoses is over 90:1. It might be more useful to make  $P_{50}$  groups with an approximately equal number of stop-stop prognoses per group in order to have a sufficient amount of stop-stop prognoses in each group to make reliable predictions. Figure 4 shows a plot of more evenly distributed  $P_{50}$  groups.



**Figure 3:** Number of stop-stop prognoses per static size  $P_{50}$  group.



**Figure 4:** Number of stop-stop prognoses per variable size  $P_{50}$  group.

Figure 4 shows that using these variable size  $P_{50}$  groups, the number of stop-stop prognoses per group is much more evenly distributed. The ratio of the highest- to the lowest number of stop-stop prognoses in these groups is about 14:1. Having a more equal number of stop-stop prognoses per group and the fact that this variable has more groups (ten instead of seven) means that we are likely to have more accurate predictions for most groups. The variable Variable size  $P_{50}$  group is used in the 3.0-prognoseverdeling method.

### 2.1.3 Additional datasets

We merged three additional datasets with the passenger forecast dataset to add additional predictive variables. These datasets are discussed separately below.

The first additional dataset contains the train type per train number. It describes three train types: Intercity trains (IC), Sprinter trains (SPR) and international trains (e.g., ICE). As the number of realised passengers in 2019 is not reliable enough for international trains, these trains are not considered in this study. Adding the train types could be useful in determining the percentiles as it could be that the distribution of Intercity trains is different from that of Sprinter trains. From now on we will refer to the train type as Rolling stock.

The second additional dataset divides all NS train stations in the Netherlands into three global areas: “Randstad”, “invloedsgebied” and “periferie”. Randstad consists of the largest cities in the Netherlands, located in the mid-west of the Netherlands. Invloedsgebied refers to the area of influence of the Randstad as introduced by Govers (2011). It denotes the area immediately surrounding the Randstad. The remainder of the Netherlands is classified as periferie. The exact distribution of the three areas is shown in Figure 5. Adding the global areas could be useful as well when determining the percentiles since the distribution of passengers for Randstad, invloedsgebied and periferie could differ from each other.



**Figure 5:** The distribution of areas in RIP, where areas with a Roman numeral are Randstad, areas with a letter are invloedsgebied and areas with a digit are periferie. (Source: Govers (2011))

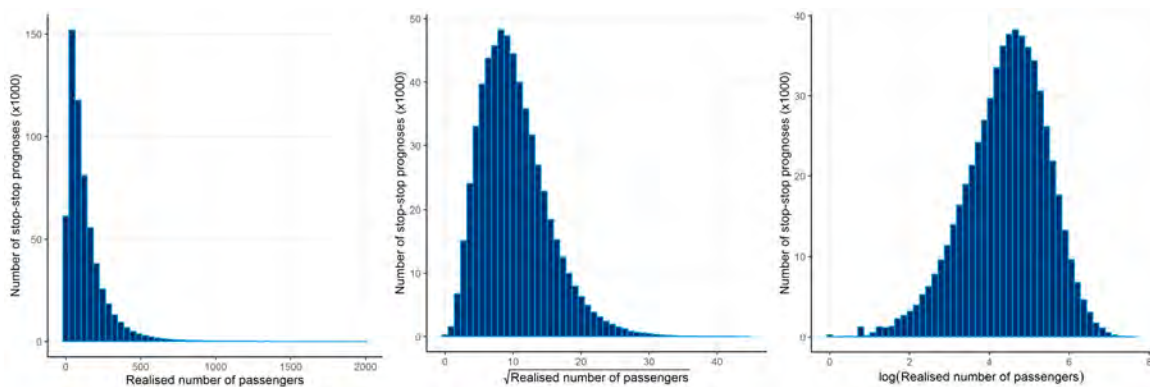
The last additional dataset also contains areas in which the train stations are divided. This dataset contains 27 areas that are more specific than the areas of the



previous dataset, dividing the Netherlands on a provincial or big city level. Some provinces have fewer or no major cities and therefore fall into a provincial level. Other provinces do contain large cities that are assigned their own level, such as Amsterdam in North Holland. All details concerning this dataset can be found in Appendix B. It could be that the more specific areas than the three global areas are better areas to add in terms of properly estimating the percentiles.

## 2.2 Data transformations

The variable for which we want to predict the distribution is the realised number of passengers for each stop-stop prognose. The fact that it is per stop-stop prognose entails a number of things. For instance, many passengers travel several stop-stop prognoses on the same train route, which means that the same passenger is counted for several stop-stop prognoses and that the stop-stop prognoses of a train route therefore depend on each other. It also means that Sprinter trains count more heavily than Intercity trains because they generally have more stop-stop prognoses per complete train run, and therefore more data points. The left histogram in Figure 6 shows the distribution of the number of stop-stop prognoses per realised number of passengers.



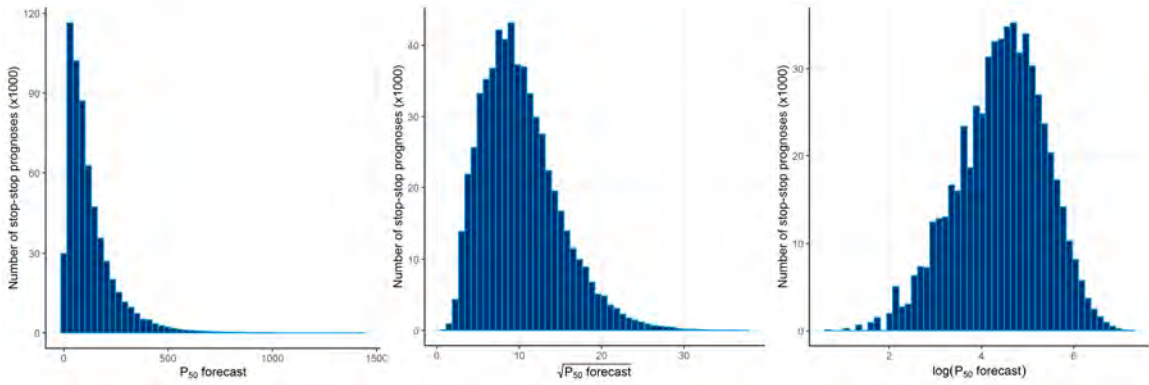
**Figure 6:** Histograms of the number of stop-stop prognoses per realised number of passengers (left), its square root (middle) and its logarithm (right).

The left histogram in Figure 6 shows that about 80% of the data of the response variable is between 0 and 200 passengers and is therefore not very spread out. This can make it difficult to fit a regression line. Because there is such a large number of stop-stop prognoses with a low realised number of passengers, this will weigh heavily in determining the best regression line. However, for NS it is less important to predict the distribution perfectly for lower passenger numbers since the least number of seats a train can have is about 150 seats. Therefore, the planned train will be the same regardless of whether the distribution is predicted well or not.

To make the variance of the realised number of passengers no longer dependent on the number of passengers itself, we use a variance stabilising transformation (Foi,

2009). Using a transformation of the data also spreads out the data, which makes creating a linear model easier. Two transformations are applied to the realised number of passengers and are added to the dataset: the square root and the logarithm, respectively the middle and right histogram in Figure 6.

The  $P_{50}$  forecast variable shows approximately the same distribution as the number of realised passengers, where about 80% of the data is between 0 and 200 passengers. For this variable, the square root and logarithm transformations are added to the dataset as well. Histograms of the  $P_{50}$  forecast, its square root and its logarithm are shown in Figure 7.



**Figure 7:** Histograms of the number of stop-stop prognoses per  $P_{50}$  forecast (left), its square root (middle) and its logarithm (right).

The independent temporal variable concerning the exact departure time of the trains requires some attention as this is a cyclic variable. To include the departure time as a numerical variable, we choose to represent it as the total number of seconds after midnight  $s$ , as is done by London (2016). As a result, the distance between the seconds is correctly represented. However, the times 00.00 (0 seconds after midnight) and 23.59.59 (86,399 seconds after midnight) are very close in terms of time, but are as far apart as can be in this numerical representation. In order to make variable  $s$  cyclical, it is split into two variables, namely:

$$s_{\sin} = \sin\left(2\pi\frac{s}{S}\right), \quad (2.1)$$

and

$$s_{\cos} = \cos\left(2\pi\frac{s}{S}\right), \quad (2.2)$$

where  $S = 86,400$  seconds is the maximum number of seconds in a day (24 hours  $\times$  60 minutes  $\times$  60 seconds = 86,400 seconds). Using this cyclic transformation, the distance between all seconds is correctly represented.

## 2.3 Data split

The data is split at random into a training set and a test set. This is done with an 80/20 ratio at train number level to keep all routes in a complete train run in the same set. This is important because passenger numbers in consecutive routes of the same train run are dependent on each other, as many passengers travel several routes in one journey. The distribution of all levels within all factor variables is kept approximately equal. For the numerical variables, the distribution of low and high values is kept roughly equal as well.

To assess the generalisation capability of predictive models and to prevent overfitting,  $k$ -fold cross-validation is performed on the training set (Berrar, 2019). This is used to determine the precision of the prediction methods. In  $k$ -fold cross-validation, the training set is partitioned into  $k$  subsets of approximately equal size (Berrar, 2019). The methods are trained using  $k - 1$  subsets, which represent the training set. Then the method is evaluated using the  $WMAPE(\tau)$  values (which will be explained in detail in Section 3.1.2 in the Methodology), on the remaining subset, which represents the validation set. This procedure is repeated  $k$  times until each subset has served as validation set. To measure the robustness of a method, the mean of the  $WMAPE(\tau)$  values of the  $k$  repetitions and corresponding standard deviation are then calculated. The smaller this deviation is, the more robust the method is.

We decided to use  $k = 5$  subsets (dividing the training set into five subsets, each containing approximately 20% of the training set data) in order to keep enough repetitions to get an estimate of the robustness of the methods, but at the same time not to increase the execution time too much.

# Chapter 3

## Methodology

In this section, the evaluation methods and the methods used to predict the percentiles are presented. First, the evaluation methods are described in detail. After this, the two NS methods are described: the current method and the 2.0-prognoseverdeling method. Lastly, the two newly introduced methods are explained: the 3.0-prognoseverdeling method and quantile regression. Quantile regression is divided into two versions, one is trained without observation weights and one is trained with observation weights.

### 3.1 Evaluation method

Comparing the methods is a difficult issue as we are looking for a single measure for thousands of stop-stop prognoses, each with its own distribution of passenger numbers. Three ways to compare methods are explained in this section. The reason why we first explain the evaluation methods and only then the prediction methods is because we first want to answer the question of when we consider a forecast good. We need to know this before we can create models to make good predictions of the percentiles.

#### 3.1.1 Percentages of stop-stop prognoses

The first evaluation method we use is a very basic method previously used by NS. For each percentile separately, we determine the percentage of stop-stop prognoses in a given data set that is less than or equal to its percentile prediction. We then compare this percentage to the percentile rank of the specific  $\tau \in \{0.01, 5, 10, 25, 75, 90, 95, 99.99\}$ . For  $\tau = 25$ , for example, the closer the percentage of stop-stop prognoses is to 25, the better the method has predicted this percentile of the realised number of passengers on average.

In addition to the direct comparison between the percentages and percentile ranks, we also look at the factor (percentage – percentile rank) / percentile rank. This allows us to take a better look at the relative difference between percentage and

percentile rank. The closer this factor is to zero, the better the method has determined the percentile on average.

This method provides a first impression of how well the method performs on average. However, its disadvantage is that it only says something about the average forecast, and not about how good the stop-stop prognose predictions are individually. For example, if a method overestimates the percentiles for one stop-stop prognose and underestimates for another stop-stop prognose, then the prediction is good on average, but bad for both stop-stop prognoses individually. This means that if a prediction method performs well according to this evaluation method, it does not necessarily mean that it is a good predictor per stop-stop prognose. However, if a method does perform poorly according to this evaluation method, then it performs already poorly on average, let alone per stop-stop prognose.

### 3.1.2 WMAPE( $\tau$ ) values

The main method used in this study to evaluate the prediction methods is an extension of the Weighted Mean Absolute Percentage Error (WMAPE). According to Chockalingam (2007), WMAPE is the sum of absolute errors divided by the sum of the actuals as follows:

$$WMAPE = \frac{\sum_{t=1}^n |A_t - F_t|}{\sum_{t=1}^n A_t}, \quad (3.1)$$

where  $n$  is the number of observations,  $A_t$  is the actual value and  $F_t$  is the predicted value. This measure is volume weighted, meaning that it is not skewed by very small passenger numbers (Kolassa et al., 2007). So in essence, in this measure, a 10% deviation for 1,000 passengers has a much larger impact on the total error than a 10% deviation for 10 passengers. In order to also take the percentile ranks into account, we modify WMAPE to

$$WMAPE(\tau) = \frac{\sum_{t=1}^n \rho_\tau(A_t - F_t)}{\sum_{t=1}^n A_t}. \quad (3.2)$$

In this function  $\rho_\tau$  has been added, which is a function that assigns asymmetric weights to the error depending on the quantile  $\tau$  and the overall sign of the error according to Koenker and Bassett Jr (1978). It has the following form:

$$\rho_\tau(u) = u(\tau - \mathbb{1}(u < 0)). \quad (3.3)$$

In this function,  $\mathbb{1}$  is the indicator function, which is one if the condition is true and zero if it is false. For example, for the  $P_{90}$ , this function multiplies positive errors by 0.9 and negative errors by  $-0.1$ . This also eliminates the need for the absolute values in Equation 3.1 as the result of this multiplication will always be positive. This evaluation method allows the total error of a method for a certain percentile

rank to be represented very specifically, while also being weighted by the total size of passenger numbers.

In addition to the  $WMAPE(\tau)$  values, we also used a modified version of the local goodness of fit measure that Koenker and Machado (1999) introduced, the  $R^1(\tau)$  value. Koenker and Machado (1999) state that this measure is motivated by the  $R^2$  goodness of fit for classical least squares regression. Unlike  $R^2$ , which provides a global measure of goodness of fit,  $R^1(\tau)$  constitutes a local measure of goodness of fit for a particular quantile  $\tau$ . This implies that the  $R^1(\tau)$  values cannot be compared between different  $\tau$ s. The equation for  $R^1(\tau)$  is

$$R^1(\tau) = 1 - \frac{\hat{V}(\tau)}{\tilde{V}(\tau)}, \quad (3.4)$$

where  $\hat{V}(\tau)$  is the error of the method to be tested and  $\tilde{V}(\tau)$  is the error of some baseline method. If the method to be tested obtained a lower error than the baseline method,  $R^1(\tau)$  lies between zero and one. The closer  $R^1(\tau)$  is to one, the greater the difference between the methods. However, the version we will use is

$$R^1(\tau) = \frac{\hat{V}(\tau)}{\tilde{V}(\tau)} - 1, \quad (3.5)$$

which is the opposite of Equation 3.4. We do this because in this study we are not maximising something, but minimising the total error. So in this case, it makes more sense to flip the sign. The closer this  $R^1(\tau)$  value is to  $-1$ , the better the method performs compared to the baseline method.

### 3.1.3 $WMAPE_{P_{50} \times t}(\tau)$ values

NS can only use complete rolling stock units with a fixed number of seats. It is not possible to use half a unit, for example. The smallest train that NS can deploy has about 150 seats. This means that the forecast for lower passenger numbers is less important. For example, if the distribution around 60 passengers is slightly inaccurate, it will not affect which train is scheduled. It will matter for higher  $P_{50}$ s, where perhaps other rolling stock units are scheduled if the distribution is more accurate.

In addition to the number of passengers, the route travel time of a stop-stop prognose is also an important factor. After all, it is more annoying for a passenger to have to stand for a long period of time than for a short period of time. We also do not want passengers who travel several shorter stop-stops to be weighted more heavily than passengers who travel few longer stop-stops. For NS, it is therefore more important to predict the distribution of passenger numbers at a stop-stop prognose with higher passenger numbers and/or longer route travel times more accurately than at other stop-stop prognoses. To achieve this, we add observation weights to Equation 3.2 to get a value weighted  $WMAPE(\tau)$ :

$$WMAPE_{P_{50} \times t}(\tau) = \frac{\sum_{t=1}^n w_t \rho_{\tau}(A_t - F_t)}{\sum_{t=1}^n w_t A_t}, \quad (3.6)$$

where  $w_t$  are the weights for stop-stop prognose  $t$  in  $P_{50}$  forecast times stop-stop prognose route travel time (in minutes). Also for this method, the  $R^1(\tau)$  values of Equation 3.5 can be used to compare the prediction methods with a baseline method.

## 3.2 Current method of NS

THIS SECTION HAS BEEN REMOVED BECAUSE IT CONTAINS CONFIDENTIAL INFORMATION.

## 3.3 2.0-prognoseverdeling method

The current method uses only the realised number of passengers of the past data to predict the desired percentiles and not the  $P_{50}$  forecast itself as well. As a result, it does not take into account the uncertainty in the  $P_{50}$  forecast itself, but assumes it to be 100% correct. NS has developed a second method for determining the desired percentiles that does take this uncertainty into account: the 2.0-prognoseverdeling method. This method calculates a new distribution based on the medians of the passenger forecast and the realised number of passengers. By doing this, the method also takes the uncertainty in the realised  $P_{50}$  into account.

For this method, we start by computing the multiplication factor between the realised number of passengers  $r_i$  and the passenger forecast  $(P_{50})_i$  for all  $i$  stop-stop prognoses in a specific reference year and BDU combination as follows:

$$\alpha_i = \frac{r_i}{(P_{50})_i}. \quad (3.7)$$

In this equation,  $\alpha_i$  is the multiplication factor that is needed to get from the  $P_{50}$  forecast to the actual number of passengers for the past reference data. In this method we try to estimate the percentiles of the multiplication factors for grouped parts of the data. The reference data is grouped as follows:

- Weekday (Monday - Friday)
- Daypart (morning rush hour, evening rush hour, off-peak hours)
- Static size  $P_{50}$  group ( $[100(n-1), 100n)$  for  $n \in \{1, \dots, 6\}$  or  $[600, \infty)$ )

The percentiles that NS wants to calculate for the  $P_{50}$ s of the stop-stop prognoses in each of these groups are  $P_\tau$ , with  $\tau \in \{\text{min}, 5, 10, 25, 75, 90, 95, \text{max}\}$ . The multiplication factors for the percentiles other than  $P_{\text{min}}$  and  $P_{\text{max}}$  are calculated following Hyndman and Fan (1996). They define sample quantiles as follows:

$$Q(p) = (1 - \gamma)x[j] + \gamma x[j + 1], \quad (3.8)$$

where  $\frac{j+p-1}{n} \leq p < \frac{j+p}{n}$ ,  $x[j]$  is the  $j$ th order statistic,  $n$  is the sample size and the value of  $\gamma$  is  $p(n-1) - j + 1$ . The sample quantiles can then be obtained by linear interpolation between the points  $(p[k], x[k])$ , where  $p[k] = \frac{k-1}{n-1}$  and  $x[k]$  is the  $k$ th order statistic. The result is then used as multiplication factor  $\tilde{\alpha}_\tau$ .

The multiplication factors for  $P_{\text{min}}$  and  $P_{\text{max}}$  are calculated in a different way. For  $P_{\text{min}}$ , the multiplication factor  $\tilde{\alpha}_{\text{min}}$  of a given group is defined as the minimum factor in the group.

For the multiplication factors for percentiles between  $P_{95}$  and  $P_{\text{max}}$  NS assumes a linear interpolation. If we take the maximum for  $P_{\text{max}}$ , this can lead to problems as it is very sensitive to outliers. To avoid this, NS wants a lower  $P_{\text{max}}$  to make the linear interpolation somewhat reliable, using the following equation:

$$\tilde{\alpha}_{\text{max}} = 2\tilde{\alpha}_{99.5} - \tilde{\alpha}_{99}. \quad (3.9)$$

The estimation of the multiplication factor for  $P_{\text{max}}$  is thus a linear extrapolation of the 99th and 99.5th percentile. Using the approximation of the percentiles for the multiplication factors for all combinations of groups, the new percentiles of the expected number of passengers can be calculated with the following equation:

$$P_\tau = \tilde{\alpha}_\tau P_{50}. \quad (3.10)$$

### 3.4 3.0-prognoseverdeling method

The 3.0-prognoseverdeling method focuses on finding a better combination of variables to split the reference data than the combination of variables used in the 2.0-prognoseverdeling method. The combination of variables that is used in the 2.0-prognoseverdeling method for the reference data in a BDU are: Weekday, Daypart and Static size  $P_{50}$  group (groups of 100 passengers). For the variables Weekday and Daypart, the percentage of the dataset and the five-number summary for the realised number of passengers per weekday/daypart in April 2019 is shown in Table 1 and Table 2 respectively.

A number of characteristics of the distribution of passengers per weekday and daypart can be extracted from the five-number summary: the location, the spread and the shape. According to Heckert et al. (2002), the location is the expected value of the output being measured. The spread is the amount of variation associated with



the output. This tells us the range of possible values that we would expect to see. The shape shows how the variation is distributed about the location. For all variables in this dataset, the shape is very positively skewed and will therefore not be investigated further. The location will be indicated by the median and the spread by the inter-quartile range (IQR), which is the difference between the third and first quartile.

**Table 1:** The percentage of the dataset and the five-number summary of the realised number of passengers per weekday.

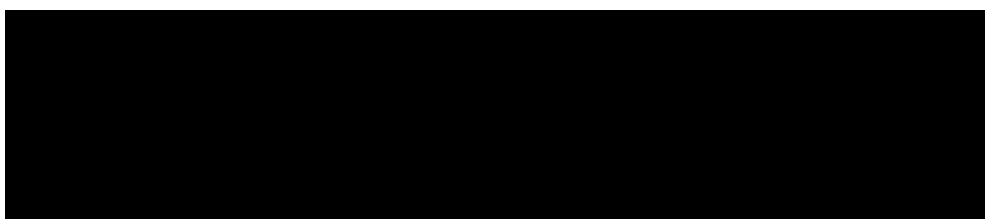
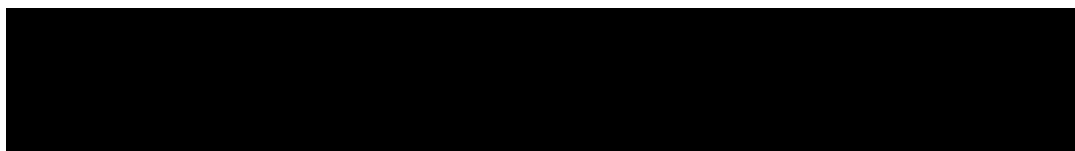
A large black rectangular box redacting the content of Table 1.

Table 1 shows that there is an approximately equal ratio of the number of stop-stop prognoses between the five weekdays in April 2019. It also shows that Monday has the lowest location and spread and Thursday the highest location and spread in terms of the number of realised passengers. Therefore, there are some differences in the distribution of the number of passengers between the weekdays and it is probably useful to split the data on this.

**Table 2:** The percentage of the dataset and the five-number summary of the realised number of passengers per daypart.

A large black rectangular box redacting the content of Table 2.

As shown in Table 2, the differences in number of passengers between dayparts are considerably larger than between weekdays. The off-peak hours differ greatly from the morning and evening rush hours as the location and spread of the number of passengers are about twice as small. There is also a difference in location and spread between the two rush hours, but this difference is much smaller.

Since both variables Weekday and Daypart show considerable differences between their factor levels, we choose to include both variables in the 3.0-prognoseverdeling method.

Regarding the Static size  $P_{50}$  groups of the 2.0-prognoseverdeling method, we looked at the ratio between the number of stop-stop prognoses in each  $P_{50}$  group. In these groups, we see that the ratio between the groups is very positively skewed, with

468,820 stop-stop prognoses in  $P_{50}$  group  $[0, 100]$  and only 5,130 stop-stop prognoses in  $P_{50}$  group  $(500, 600]$ . To make the ratio between the number of stop-stop prognoses in the  $P_{50}$  groups more equal, we choose to use the Variable size  $P_{50}$  groups in the 3.0-prognoseverdeling method. As explained in Section 2.1.2, this variable splits the  $P_{50}$  forecasts into groups of variable sizes to make the number of stop-stop prognoses in each group more equal.

The 2.0-prognoseverdeling method does not take the area in which a stop-stop prognose is located into account. Adding a variable concerning this could be of great value as there could be a difference in distribution of passengers between different parts of the Netherlands. Two variables about the location of a departure station of a stop-stop prognose are available: RIP and Area. The variable RIP divides the stop-stop prognoses into three regions: Randstad, invloedsgebied and periferie. The variable Area divides the stop-stop prognoses into 27 areas at province or large city level. To find out which of these variables is best for splitting the data, we look at all areas of variable Area within invloedsgebied of variable RIP. The full name, in which RIP area it is located, the percentage of the dataset and the five-number summaries of these areas is shown in Table 3.

**Table 3:** The full name, RIP, percentage of the dataset and five-number summary of the realised number of passengers per area in invloedsgebied.

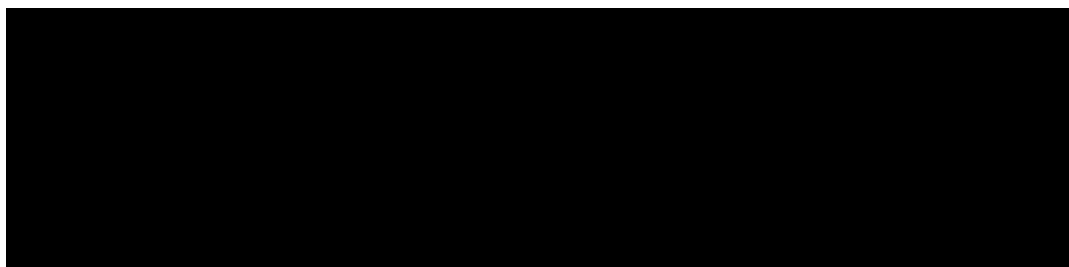


Table 3 shows that there is already a substantial difference between the areas in invloedsgebied itself, let alone between areas of different RIPs. The biggest difference in these areas can be seen between Area 9 and Area 23, where the median and the IQR are about twice as high for Area 23. Because the differences between the areas within one RIP area are already so large, we choose to include only the variable Area in the 3.0-prognoseverdeling method. The full name, in which RIP area it is located, the percentage of the dataset and the five-number summary of all 27 areas in variable Area can be found in Appendix B.

The last variable in the data set that might be interesting to split the data on is the Rolling stock. The percentage of the dataset and the five-number summary of Intercity trains and Sprinter trains is shown in Table 4.

Table 4 shows that about two-thirds of the data consists of Sprinter trains. It can also be seen that the median and IQR for Sprinter trains is about twice as small as

**Table 4:** The percentage of the dataset and the five-number summary of the realised number of passengers per rolling stock.

for Intercity trains. As a large number of stop-stop prognoses with low passenger numbers are Sprinter trains, it is probably helpful to the method to split the data on Rolling stock.

The combination of variables that will be used in the final 3.0-prognoseverdeling method is: Weekday, Daypart, Variable size P<sub>50</sub> group, Area and Rolling stock. With this combination of variables, it is possible that there are no stop-stop prognoses in a specific combination of factor levels in the training set, but that there are stop-stop prognoses of this combination in the test set. In this study, we choose to use the average of all other stop-stop prognoses as the prediction if this is the case.

### 3.5 Quantile Regression

Since we are looking for percentiles (quantiles), it makes sense to investigate the method of quantile regression. Quantile regression is a type of linear regression analysis, which is a statistical method to summarise the linear relationship between a dependent variable and a set of independent variables. Where standard linear regression techniques do this based on the conditional mean, quantile regression does this based on any quantile.

Quantile regression has some advantages over ordinary least squares (OLS) regression. One benefit of quantile regression is that it makes no assumptions about the distribution of the residuals. It also drops the assumption that the variance of the variable must remain constant. On top of that, quantile regression is much more robust against outliers in the dependent variable compared to OLS.

The equation of the quantile regression model for the  $\tau$ th percentile is described by the following equation:

$$y_i = \mathbf{x}_i^\top \mathbf{b} + e_i, \quad i = 1, \dots, n \quad (3.11)$$

where  $\mathbf{x}_i^\top$  represents a vector of predictors,  $\mathbf{b} = \beta(\tau)$  is the vector of coefficients associated with the  $\tau$ th percentile,  $n$  is the number of data points and  $e_i$  are the residuals of the model and are assumed to make the  $\tau$ th quantile equal to zero (Koenker and Hallock, 2001). Based on Equation 3.11, we also make the assumption that the  $\tau$ th percentile is given as a linear function of independent variables. To find the best estimate for percentile  $\tau$ ,  $\mathbf{b}$  is estimated by solving the minimisation function

$$\min_{\mathbf{b} \in \mathbb{R}^p} \sum_{i=1}^n \rho_{\tau}(y_i - \mathbf{x}_i^{\top} \mathbf{b}), \quad (3.12)$$

where  $p$  is the number of independent variables used in the model and  $\rho_{\tau}$  is the same function as in Equation 3.3. If a method underestimates, then the error is positive, which means that the error receives a penalty of  $\tau$ . If a method overestimates, then the error is negative, so the error receives a penalty of  $(1-\tau)$ . The  $\rho_{\tau}$  function ensures that the minimum in the minimisation function is exactly at the right percentile of the data, where the ratio of positive errors to negative errors is  $1-\tau : \tau$ . For example, for the  $P_{90}$  the function ensures that it searches for the line for which the ratio of positive errors to negative errors is exactly  $0.1 : 0.9$ .

Solving Equation 3.12 reduces to solving a linear program as this is a problem that seeks to optimise a linear function subject to linear constraints (Koenker and Hallock, 2001).

### Variable selection

We want to determine the distribution of the number of passengers for stop-stop prognoses. To spread out the observations more, making it easier for quantile regression, we try to predict the logarithm of the number of passengers. Then, to go from this prediction back to a prediction of the number of passengers, we take the exponential. To select the most important explanatory variables for this, the step-up method is used. In this method, we start with an empty model and step by step add the variable that yields the minimum error (calculated as in Equation 3.12, represented by  $\hat{\beta}(\tau)$ ). A restriction for adding the variable to the model is that the model contains only variables where at least one level is significantly different from zero. This is important, because if a variable is not significantly different from zero, it might as well not be included in the model. A second restriction is that we make sure that the model is a significant improvement on the previous model.

To determine whether or not there is a statistically significant difference between the model with an added variable and the previous model (which did not include the added variable), a partial  $F$ -test is used, following Pardoe et al. (2021). This tests whether the variable that is not in the full model is actually useful and should therefore be included. The null and alternative hypothesis for this test are as follows:

**Hypothesis  $H_0$ :** All coefficients removed from the full model are zero.

**Hypothesis  $H_1$ :** At least one of the coefficients removed from the full model is non-zero.

The significance level used is  $\alpha = 0.05$  and the  $F$  test-statistic that this test calculates is as follows:

$$F = \frac{\frac{\text{RSS}_{\text{reduced}} - \text{RSS}_{\text{full}}}{p}}{\frac{\text{RSS}_{\text{full}}}{n-k}}, \quad (3.13)$$

where  $RSS_{\text{reduced}}$  is the residual sum of squares of the reduced model,  $RSS_{\text{full}}$  is the residual sum of squares of the full model,  $p$  is the number of predictors removed from the full model,  $n$  is the total observations in the dataset and  $k$  is the number of coefficients (including the intercept) in the full model (Pardoe et al., 2021).

As the  $P_{90}$  is the most important percentile for NS, we choose this percentile for the variable selection for the quantile regression model. Table 5 shows the results of all quantile regression models with a single explanatory variable.

**Table 5:** The  $\hat{\beta}(0.9)$  values (in number of passengers) and significance of factor levels for a model with each explanatory variable separately.

Explanatory variable	$\hat{\beta}(0.9)$	All variables significant?
log( $P_{50}$ )	33,206.71	Yes
$P_{50}$ group (100)	43,176.39	Yes
$P_{50}$ group (variable)	34,979.62	Yes
Weekday	77,703.46	Yes
Daypart	71,782.92	Yes
Time (sin & cos)	76,969.30	Yes
RIP	76,178.70	Yes
Area	74,703.64	Yes
Material	72,127.63	Yes

As can be seen in Table 5, the  $P_{50}$  forecast seems to be the most important variable to predict the  $P_{90}$  of a stop-stop prognose as its errors are about half as small compared to the other variables. In this study, we have also created a quantile regression model without the  $P_{50}$  as a predictor. However, the error for the  $P_{90}$  that is then obtained is more than twice as much. Therefore, there is no need to present the results of this model if the difference is so obvious.

The logarithm of the  $P_{50}$  forecast achieves the lowest error value and therefore forms the basis of the model. Step by step, the variable that achieves the smallest error is then added to the model until no smaller error is achieved or there are variables that are not significantly different from zero. The interaction between the different factor variables is also included, which allows for even better results. The only variable for which no interaction is included, is Area. This is due to the fact that if the interaction is included, this will result in many variables that are not significantly different from zero.

The intermediate steps of the step-up method are shown in Appendix C. The best model found with all variables significantly different from zero is shown in Table 6.

**Table 6:** The  $\hat{\beta}(0.9)$  values (in passengers) and significance of factor levels for a model with  $\log(P_{50})$ , area, daypart and one remaining explanatory variable.

Explanatory variables	$\hat{\beta}(0.9)$	All variables significant?
$\log(P_{50})$ , Area, Daypart, Weekday	31,624.88	Yes
$\log(P_{50})$ , Area, Daypart, RIP	31,695.95	Yes
$\log(P_{50})$ , Area, Daypart, Rolling stock	31,683.04	Yes

As can be seen in Table 6, the model that obtains the lowest error for the  $P_{90}$  includes the  $\log(P_{50})$ , Area, Daypart, Weekday and the interaction between Weekday and Daypart. To see whether this model performs significantly better than the previous model (without Weekday), the  $F$ -test is performed, resulting in an  $F$  value of 123.18, with a corresponding  $p$ -value of  $< 10^{-15}$ . This  $p$ -value is much smaller than the chosen significance level of  $\alpha = 0.05$ , indicating that the null hypothesis can be rejected. This means that there is sufficient evidence that the variable Weekday is statistically significant to the model. Therefore, the best model for the  $P_{90}$  found is as follows:

$$\begin{aligned} \log(\#\text{passengers}_{P_{90}}) = & 0.74 + 0.87 \log(P_{50}) + 0.14A2 + 0.13A3 + 0.33A4 + 0.05A5 \\ & + 0.10A6 + 0.15A7 + 0.14A9 + 0.12A10 + 0.37A11 + 0.21A12 + 0.15A13 \\ & + 0.35A14 + 0.32A15 + 0.43A16 + 0.51A17 + 0.33A18 + 0.24A19 + 0.29A20 \\ & + 0.19A21 + 0.11A22 + 0.16A23 + 0.17A24 + 0.03A25 + 0.09A26 - 0.03O \\ & - 0.02M + 0.04Tu + 0.05We + 0.08Th - 0.01Fr - 0.05Tu/O - 0.03Tu/M \\ & - 0.08We/O - 0.03We/M - 0.09Th/O - 0.05Th/M + 0.04Fr/O - 0.12Fr/M, \end{aligned} \quad (3.14)$$

where  $A2, \dots, A26$  are the different areas,  $O$  and  $M$  are off-peak hours and morning rush hour respectively,  $Tu, \dots, Fr$  are Tuesday to Friday and  $Tu/O, \dots, Fr/M$  are the interaction terms between variables Weekday and Daypart.

To be able to provide some interpretation on how big the influence of the variables in this model is, we can use the natural exponential function on both sides of the equation, resulting in

$$\#\text{passengers}_{P_{90}} = e^{0.74} \times e^{0.87 \log(P_{50})} \times e^{0.14A2} \times \dots \times e^{0.04Fr/O} \div e^{0.12Fr/M}. \quad (3.15)$$

Since the factor levels can only be one or zero, the exponential functions for the factor levels depend only on the size of the coefficients. Most of these coefficients are smaller than 0.2, and can therefore be approximated by  $1+x$ , which is a linearisation of  $e^x$ . For  $x \leq 0.2$  the maximum error that can be made with this linearisation is  $e^{0.2} - (1 + 0.2) \approx 0.021$ .

Taking Area 2 as an example, we see that the influence on the prediction of the number of passengers by this variable is described by  $e^{0.14A2}$  and can be approximated by  $1 + 0.14 \times 1 = 1.14$ . This effectively means that in the multiplications in Equation 3.15, the prediction of the number of passengers is increased by approximately 14% if the stop-stop prognose is in Area 2.

### 3.6 Quantile Regression with weights

Due to the fact that the smallest train NS can deploy contains about 150 seats, a good prediction of the percentiles for smaller  $P_{50}$ s has less priority than higher  $P_{50}$ s. It might therefore be advantageous to assign weights to the stop-stop prognoses in order to indicate the importance of their percentile estimation. NS also considers it important to make a good prediction for stop-stop prognoses with longer route travel times. For this reason, we choose to take as observation weights the  $P_{50}$  forecast times the route travel time in minutes ( $P_{50} \times t$ ), which are the same as in the evaluation method with observation weights. Following Huang and Rat (2017), these weights can be added to the objective function in Equation 3.12 to obtain a weighted loss function. To find the best estimate for percentile  $P_\tau$ ,  $\mathbf{b}$  is estimated by solving the minimisation function

$$\min_{\mathbf{b} \in \mathbb{R}^p} \sum_{i=1}^n w_i \rho_\tau(y_i - \mathbf{x}_i^\top \mathbf{b}), \quad (3.16)$$

where  $w_i$  is the weight for stop-stop prognose  $i$ , defined as  $P_{50}$  forecast times route travel time in minutes. By adding weights to the stop-stop prognoses, we now have better control over where the method will perform better or worse. Despite the weights, the function will still look for the correct percentile. However, the method will make compromises somewhere in the stop-stop prognoses with lower weights to better predict the percentiles for stop-stop prognoses with larger weights.

The same step-up method as before is used to select the best explanatory variables for the quantile regression model trained with observation weights. We again search for the best combination of variables for the  $P_{90}$ . The results for all models with a single explanatory variable are shown in Table 7.

Again, the most important variable to predict the  $P_{90}$  seems to be some form of the  $P_{50}$  forecast, more specifically its logarithm. The intermediate steps can be found in Appendix D, with the final step shown in Table 8.

Table 8 shows that again variable Weekday achieves the lowest error value, with  $\hat{\beta}_{P_{50} \times t} = 32, 115.50$ . To examine whether the variable Weekday should be included in the final model, a partial  $F$ -test is performed. The hypotheses and significance level are as before. The test results in an  $F$  test-statistic of 63.98, with a corresponding  $p$ -value of  $< 10^{-15}$ , indicating that the null hypothesis can be rejected and that there is

**Table 7:** The  $\hat{\beta}_{P_{50} \times t}(0.9)$  (in passengers<sup>2</sup> minutes) and significance of factor levels for a model with each explanatory variable separately.

Explanatory variable	$\hat{\beta}_{P_{50} \times t}(0.9)$	All variables significant?
log(P <sub>50</sub> )	33,566.84	Yes
P <sub>50</sub> group (100)	43,678.33	Yes
P <sub>50</sub> group (variable)	35,063.00	Yes
Weekday	95,131.80	Yes
Daypart	82,977.94	Yes
Time (sin & cos)	97,751.11	Yes
RIP	93,300.24	Yes
Area	91,491.07	Yes
Material	82,296.87	Yes

**Table 8:** The  $\hat{\beta}_{P_{50} \times t}(0.9)$  (in passengers<sup>2</sup> minutes) and significance of factor levels for a model with log(P<sub>50</sub>), area, daypart and one remaining explanatory variable.

Explanatory variables	$\hat{\beta}_{P_{50} \times t}(0.9)$	All variables significant?
log(P <sub>50</sub> ), Area, Daypart, Weekday	32,115.50	Yes
log(P <sub>50</sub> ), Area, Daypart, RIP	32,167.54	Yes
log(P <sub>50</sub> ), Area, Daypart, Material	32,225.84	Yes

sufficient evidence that the variable Weekday is statistically significant to the model. The best found model for the weighted observations using quantile regression is as follows:

$$\begin{aligned}
\log(\#\text{passengers}_{P_{90}}) = & 0.68 + 0.90 \log(P_{50}) + 0.07A2 + 0.07A3 + 0.12A4 + 0.01A5 \\
& + 0.06A6 + 0.10A7 + 0.11A9 + 0.10A10 + 0.23A11 + 0.14A12 + 0.12A13 \\
& + 0.24A14 + 0.23A15 + 0.35A16 + 0.31A17 + 0.21A18 + 0.19A19 + 0.20A20 \\
& + 0.17A21 + 0.04A22 + 0.09A23 + 0.13A24 - 0.04A25 + 0.04A26 - 0.02O \\
& - 0.02M + 0.03Tu + 0.04We + 0.06Th - 0.04Fr - 0.03Tu/O - 0.001Tu/M \\
& - 0.07We/O - 0.01We/M - 0.08Th/O - 0.02Th/M + 0.003Fr/O - 0.06Fr/M.
\end{aligned}
\tag{3.17}$$

The difference between the quantile regression model trained with weights versus without weights is mainly reflected in the fact that the coefficient of the log(P<sub>50</sub>) is higher and the coefficients of all factor variables are lower. This implies that the P<sub>50</sub> plays an even higher role, compared to the other variables, in predicting the more important stop-stop prognoses in terms of P<sub>50</sub> forecast size and/or route travel time length.



# Chapter 4

## Results

This chapter shows the results. First, the methods are evaluated by comparing the percentages to the percentile ranks. Then, the robustness of the methods is tested using the  $WMAPE(\tau)$  values without and with observation weights. Afterwards, the methods are compared with each other by testing them on the test set. Lastly, the errors are broken down by factor variable to see which factor levels contribute most to the total error.

### 4.1 Evaluation method: percentages

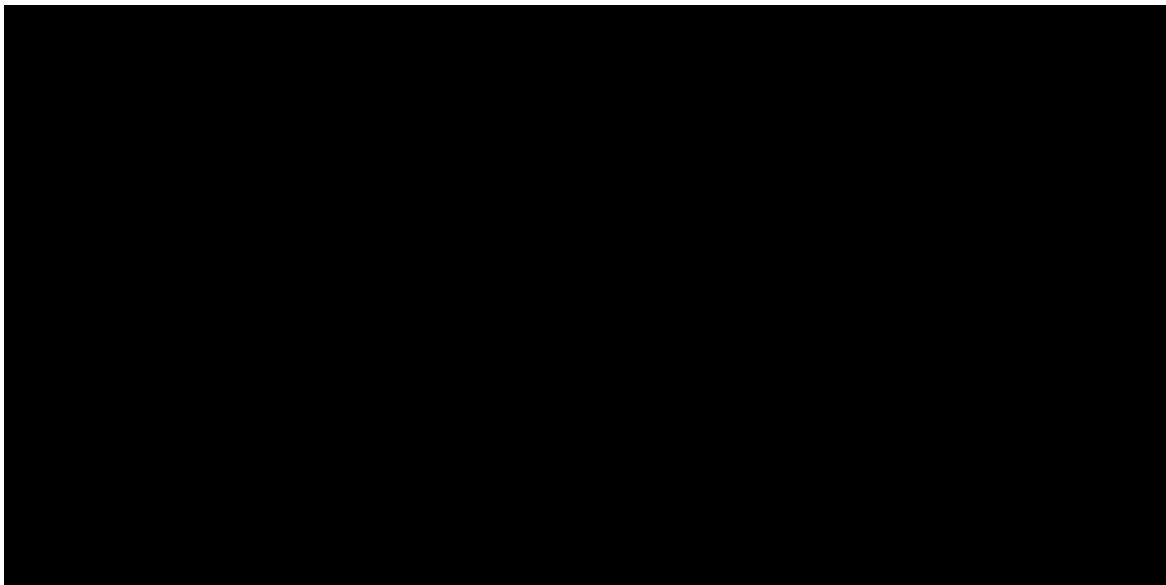
The first evaluation method of the various forecasting methods is to determine the percentage of stop-stop prognoses in the test set that are smaller than or equal to the percentiles. The closer these percentages are to the percentile ranks, the better the method reflects the distribution of the number of passengers per stop-stop prognose on average. A plot of this for the five methods (the current method, 2.0-prognoseverdeling, 3.0-prognoseverdeling, quantile regression trained without weights and quantile regression trained with weights) is shown in Figure 8.

Figure 8 shows that even on average the distribution of the realised number of passengers is not captured well by the current method. For the percentiles lower than  $P_{50}$ , the percentages are much higher than expected and for the percentiles higher than  $P_{50}$ , they are much lower than expected.

We can also see that for all methods other than the current method, the graphs form almost a straight line, indicating that all methods seem to perform well for all percentiles on average. To examine this in more detail, we determine the ratio (percentage of stop-stop prognoses where number of passengers  $\leq$  percentile – percentile rank) / percentile rank for all methods and percentile ranks. The result is shown in Figure 9.



**Figure 8:** The percentage of stop-stop prognoses where the number of realised passengers  $\leq P_\tau$  versus the percentile ranks for the five prediction methods.



**Figure 9:** A zoomed-in (left) and normal (right) plot of the ratios (percentage of stop-stop prognoses where the number of realised passengers  $\leq P_\tau$  - percentile rank) / percentile rank versus the percentile ranks for the five prediction methods.

As can be seen in Figure 9, the ratio does not have much meaning for the  $P_{0.01}$ , which makes sense since the percentages are divided by 0.01. So if the percentage of passengers at all stop-stop prognoses less than or equal to the  $P_{0.01}$  is for example 0.1%, then the ratio already blows up. The figure shows again that the current method does not perform well as for all percentile ranks the ratio is far from zero. It can

also be seen that the 2.0-prognoseverdeling method seems to outperform the other methods on average for the lower percentiles. For the higher percentiles, quantile regression seems to be closest to a ratio of zero on average, but the differences between the methods other than the current method are very small (i.e., a difference of less than 2%).

The figure also shows that for the percentiles lower than  $P_{50}$  quantile regression trained without observation weights achieves on average better results than trained with observation weights. However, for the higher percentiles this is the other way around. It could be that because of the observation weights, the focus for a good prediction lies more on higher  $P_{50}$  forecasts, so that the distribution is determined better there and worse for lower  $P_{50}$  forecasts. It could be that for the lower percentiles this change happens to be worse than quantile regression trained without observation weights and for the higher percentiles to be better. There probably are more factors that influence this, such as the route travel time (which is also part of the observation weights) that cannot be observed if only these percentages of stop-stop prognoses are considered.

The fact that a method predicts well on average does not necessarily mean that it predicts well for each individual stop-stop prognose. Therefore, this evaluation method is only used to filter out the prediction methods that do not perform well on average, which in this case is just the current method.

## 4.2 Evaluation method: WMAPE( $\tau$ ) values

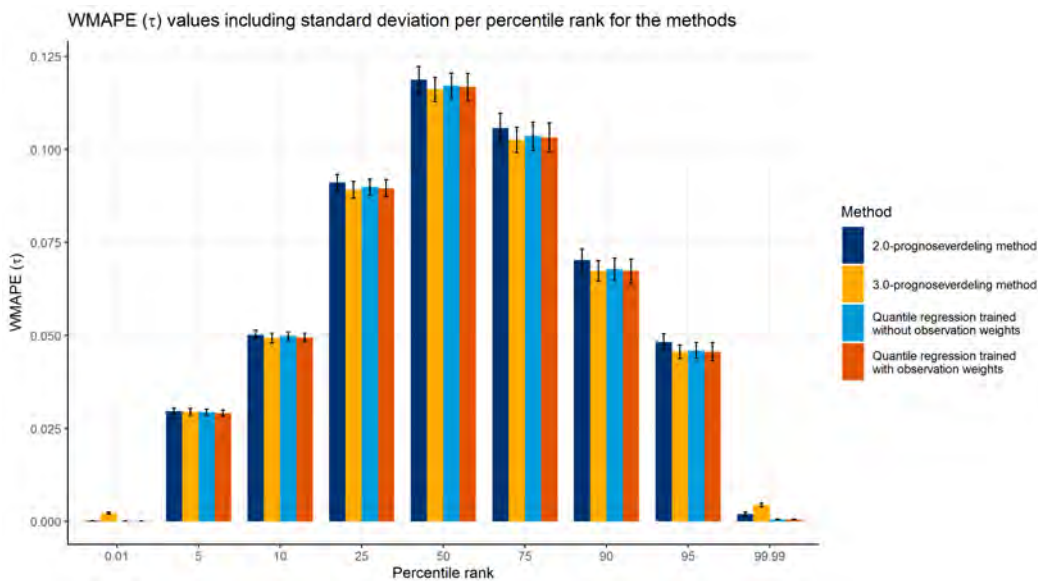
The main evaluation method uses the WMAPE( $\tau$ ) values for evaluating the prediction methods. This method assigns asymmetric weights to positive and negative errors for a certain percentile, allowing the total error of a method for a given percentile to be represented very specific for that percentile. It is also not skewed by very small passenger numbers as it provides a weighted mean of absolute percent errors, where it is weighted by the size of the passenger numbers (Kolassa et al., 2007). For this evaluation method, we do not include the current method as it already does not perform well on average. The 2.0-prognoseverdeling method will be used as the baseline method that the other methods will have to beat.

To assess the generalisation capability of predictive models and to prevent overfitting, 5-fold cross-validation was performed (Berrar, 2019). This is used to determine the precision of a method. The WMAPE( $\tau$ ) values are then determined for each fold, of which a mean with corresponding standard deviation per percentile is calculated. In addition, the  $R^1(\tau)$  values are calculated, allowing for better comparison of the 3.0-prognoseverdeling method and the quantile regression method with the 2.0-prognoseverdeling method.

After this, the methods are trained on the entire training set and tested on the test set. To compare the methods with each other, the WMAPE( $\tau$ ) values and  $R^1(\tau)$  values are determined again, but this time for the test set.

### 4.2.1 Without observation weights

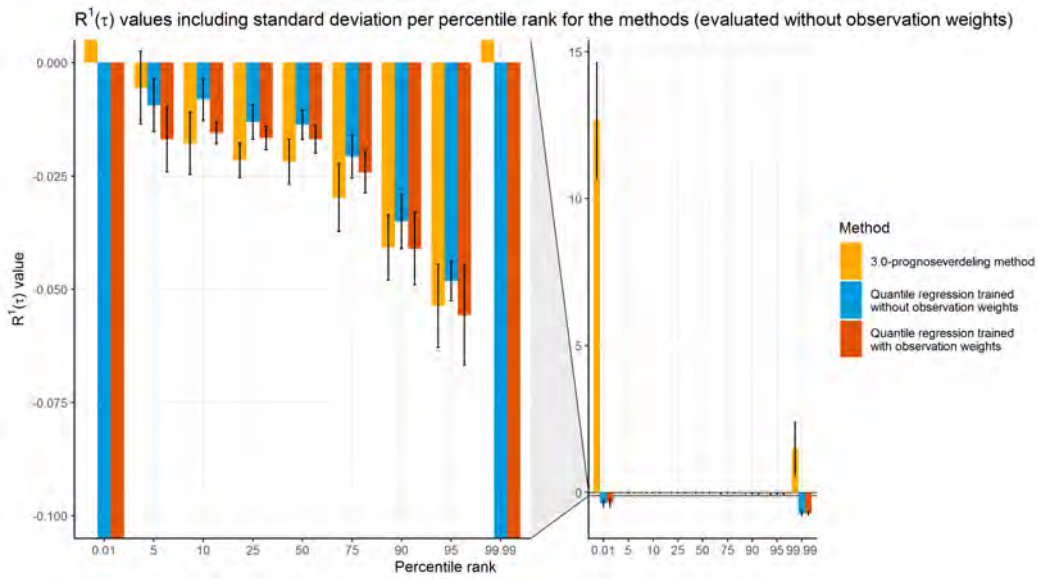
We first look at the  $WMAPE(\tau)$  values for the cross-validation without observation weights. These are shown in Figure 10. The figure shows that both the 3.0-prognoseverdeling method and quantile regression (both trained without and with observation weights) seem to achieve lower errors than the 2.0-prognoseverdeling method for all percentiles (except the 3.0-prognoseverdeling for  $P_{0,01}$  and  $P_{99,99}$ ). The percentiles cannot be compared with each other as each percentile has its own minimum. An example of this is shown in Appendix E. In addition to the differences in errors, it can be seen that the standard deviations of all methods for all percentiles are reasonably small, indicating that the methods are robust and can handle small changes in the data well.



**Figure 10:** Mean  $WMAPE(\tau)$  values with corresponding standard deviation of the 5-fold cross-validation per prediction method.

Because the differences between the methods are quite close, we also make the comparison in a more detailed way, using an adaptation of Koenker and Machado (1999)'s goodness-of-fit value, the  $R^1(\tau)$  value. When calculating this value, we take the 2.0-prognoseverdeling as the baseline method and compare the other methods to it. The  $R^1(\tau)$  values can be seen in Figure 11.

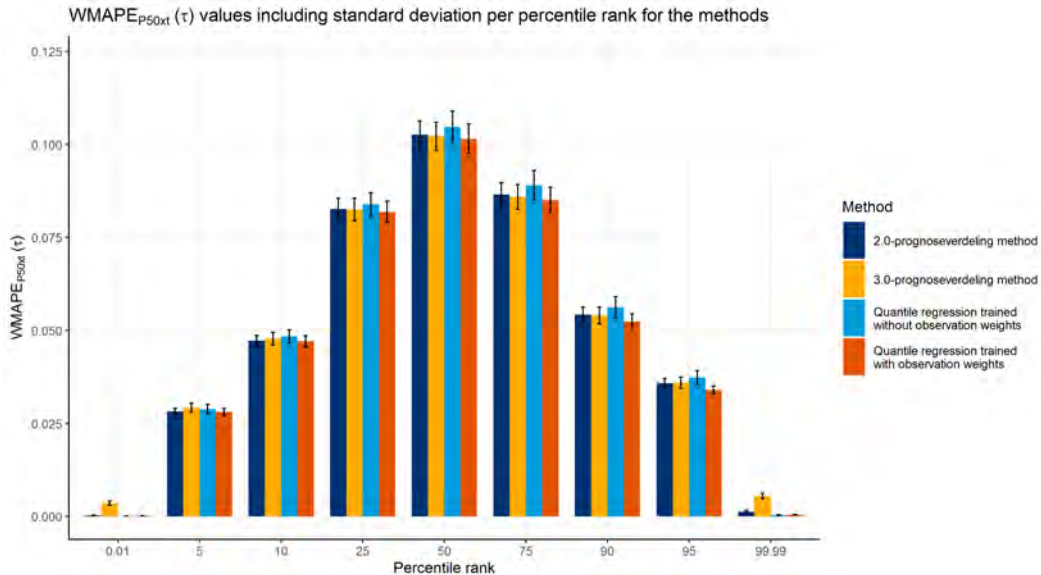
In Figure 11, the closer the  $R^1(\tau)$  value is to  $-1$ , the greater the difference in error value compared to the 2.0-prognoseverdeling method. The figure shows that for most percentiles the 3.0-prognoseverdeling method performs best compared to the 2.0-prognoseverdeling method. For some percentiles quantile regression trained with observation weights performs best. We can also see that for none of the percentiles quantile regression trained without observation weights performs best. We can see that the lower percentiles can be improved by about 2% and the higher percentiles by about 4 to 5% when using the 3.0-prognoseverdeling method.



**Figure 11:** Zoomed-in (left) and normal (right) plot of the mean  $R^1(\tau)$  values with corresponding standard deviation of the 5-fold cross-validation per prediction method.

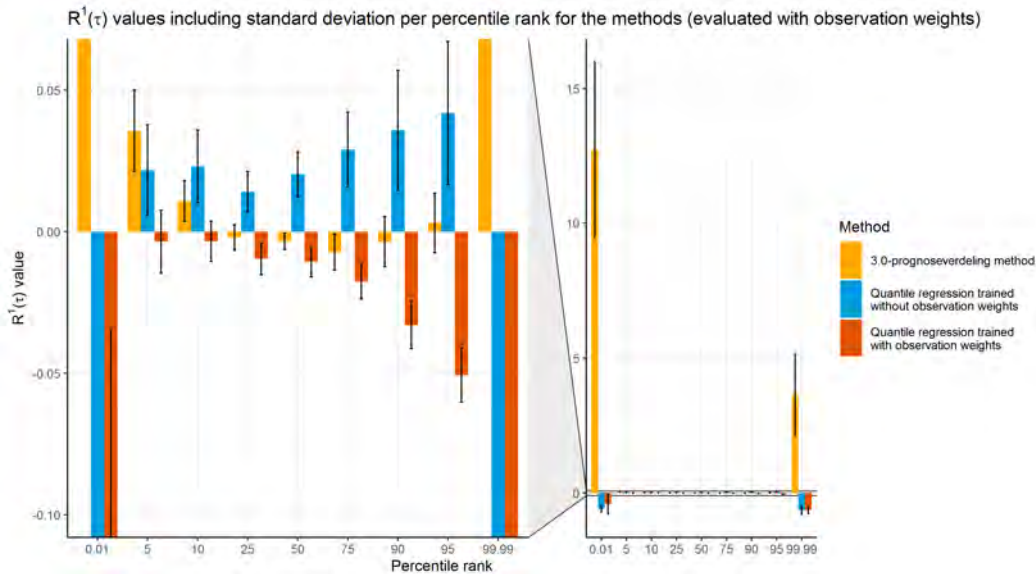
### 4.2.2 With observation weights

Because NS finds it more important to achieve better results for stop-stop prognoses with higher  $P_{50}$  forecasts and/or route travel times, we add the observation weights to the evaluation. The  $WMAPE_{P_{50} \times t}(\tau)$  values of the different prediction methods are shown in Figure 12.



**Figure 12:** Mean  $WMAPE_{P_{50} \times t}(\tau)$  values with corresponding standard deviation of the 5-fold cross-validation per prediction method.

Figure 12 shows that the methods again seem to be very robust, as evidenced by the small standard deviations. Also, the  $WMAPE_{p_{50} \times t}(\tau)$  values of the methods are very similar between the 3.0-prognoseverdeling method and quantile regression trained with observation weights. Quantile regression trained without observation weights seems to not outperform the 2.0-prognoseverdeling method. To better examine the differences, the  $R^1(\tau)$  values are shown in Figure 13.



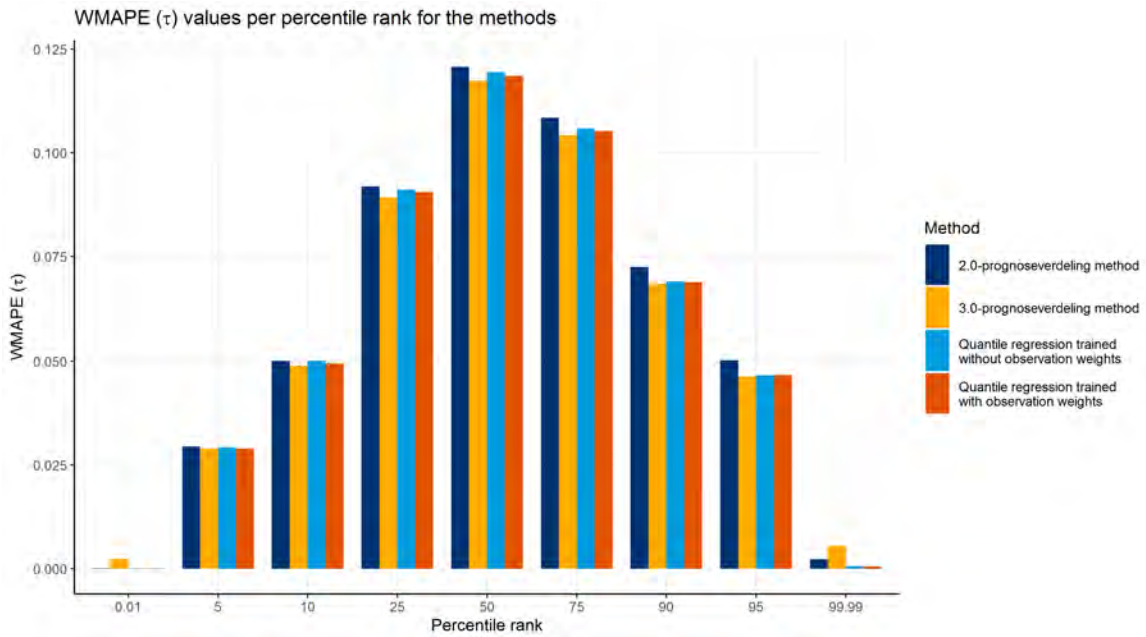
**Figure 13:** Zoomed-in (left) and normal (right) plot of the mean  $R^1(\tau)$  values with corresponding standard deviation of the 5-fold cross-validation per prediction method.

In this case, quantile regression trained with observation weights performs best for all percentiles. This is only a small reduction in average error of about 1% for the lower percentiles compared to the 2.0-prognoseverdeling method, but rises to about 4% to 5% for the higher percentiles (excluding the  $P_{0.01}$  and  $P_{99.99}$ , where the reduction in error is substantially higher).

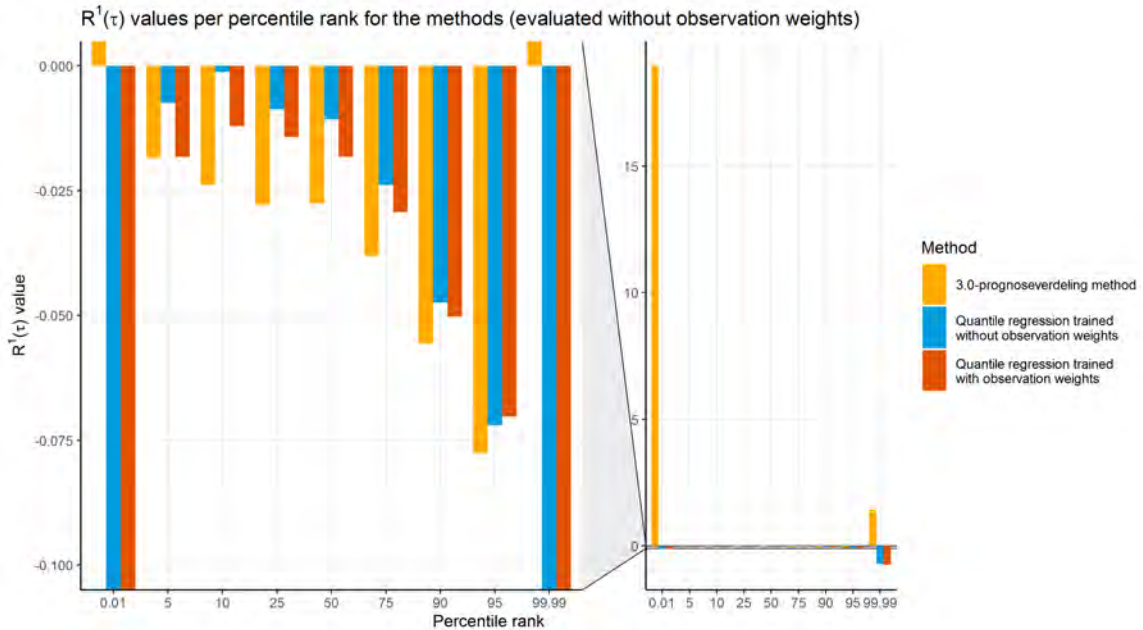
### 4.2.3 Results test set

After cross-validation, the methods are trained on the entire training set and tested on the, so far unseen, test set. The  $WMAPE(\tau)$  values of the methods for all percentiles are shown in Figure 14.

Figure 14 shows that, for the test set, the 3.0-prognoseverdeling method and both quantile regression methods outperform the 2.0-prognoseverdeling method for all percentiles (except for the 3.0-prognoseverdeling method for  $P_{0.01}$  and  $P_{99.99}$ ). Since the  $WMAPE(\tau)$  values of the methods are very close to each other, we also look at the  $R^1(\tau)$  values, where we compare the 3.0-prognoseverdeling method and quantile regression methods to the 2.0-prognoseverdeling. These  $R^1(\tau)$  values are shown in Figure 15.



**Figure 14:**  $WMAPE(\tau)$  values for the prediction methods, tested on the test set and evaluated without observation weights.

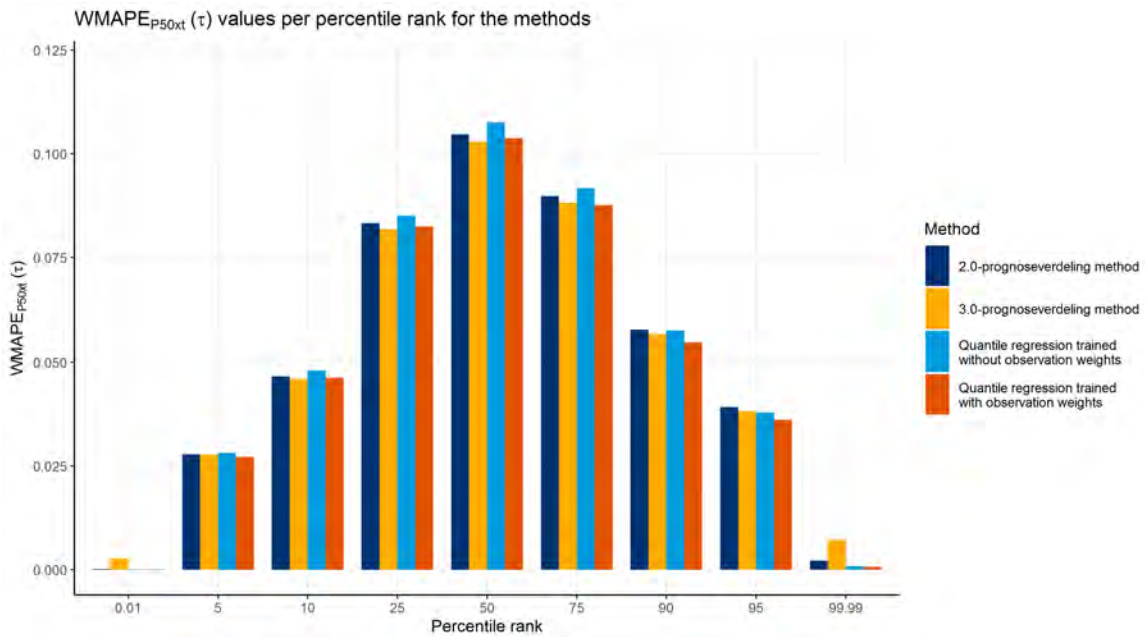


**Figure 15:** Zoomed-in (left) and normal (right) plot of the  $R^1(\tau)$  values for the prediction methods, evaluated without observation weights.

Figure 15 shows that the 3.0-prognoseverdeling method again achieves the best results for all percentiles other than  $P_{0.01}$  and  $P_{99.99}$ . For the  $P_5$ , both the 3.0-prognoseverdeling method and quantile regression trained with observation weights perform about equally well. We also see that quantile regression seems to perform

better for all percentiles except  $P_{95}$  when trained with observation weights instead of trained without observation weights.

If we look at the  $WMAPE_{P_{50 \times t}}(\tau)$  values of the methods with the weighted observations included, we see something different. These values are shown in Figure 16.

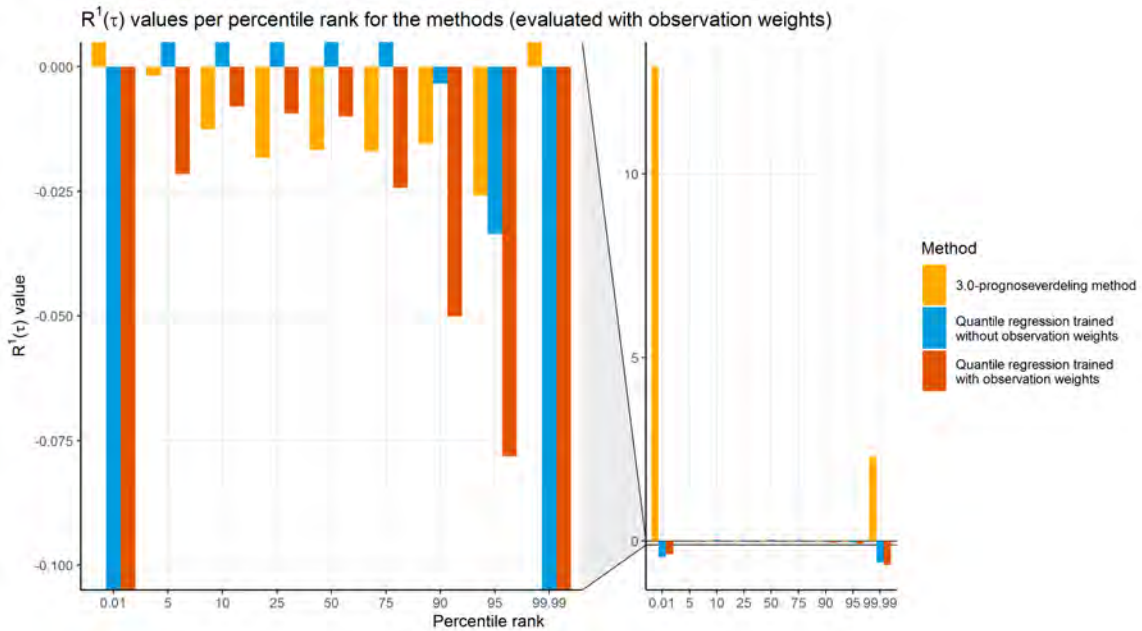


**Figure 16:**  $WMAPE_{P_{50 \times t}}(\tau)$  values for the prediction methods, tested on the test set and evaluated with observation weights.

This time we see that the quantile regression trained with observation weights seems to perform best on all desired percentiles other than  $P_{10}$ ,  $P_{25}$  and  $P_{50}$ , and thus outperforms the 3.0-prognoseverdeling method on the higher percentiles. We can also see that quantile regression trained without observation weights performs even worse than the 2.0-prognoseverdeling method for most percentiles. The  $R^1(\tau)$  values are shown in Figure 17 for a more detailed comparison.

Figure 17 shows that quantile regression trained with observation weights outperforms the other methods for the higher percentiles. It does not outperform the 3.0-prognoseverdeling method for the  $P_{10}$ ,  $P_{25}$  and  $P_{50}$ , which was the case for the 5-fold cross-validation. But for the other percentiles, the biggest gain seems to be available in the higher percentiles with about 5% to 7.5% improvement for the  $P_{90}$  and  $P_{95}$  respectively compared to the 2.0-prognoseverdeling method.





**Figure 17:** Zoomed-in (left) and normal (right) plot of the  $R^1(\tau)$  values for the prediction methods, evaluated with observation weights.

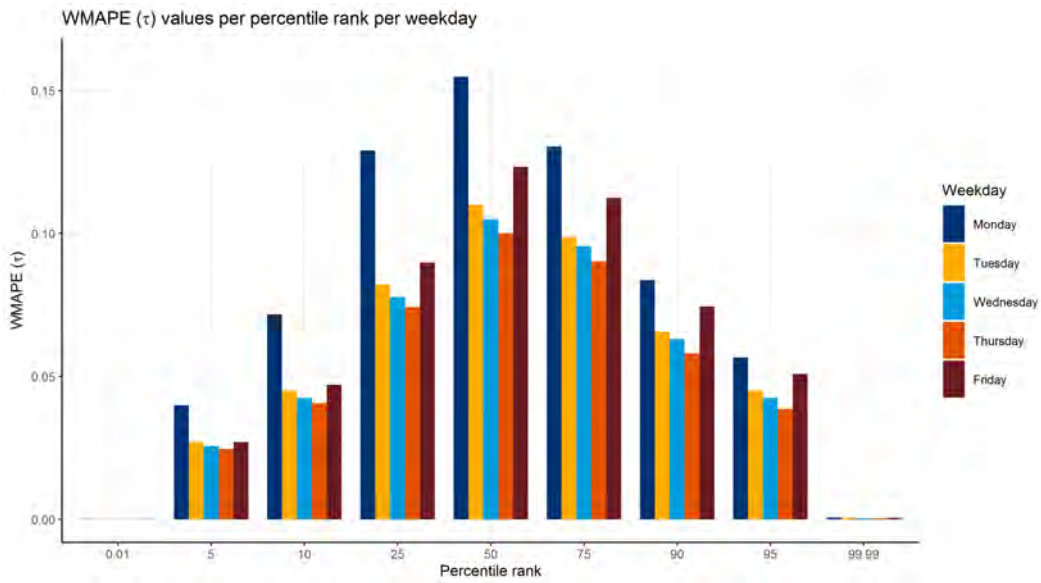
### 4.3 Error contribution per factor level

This section will zoom in on the difference in contribution to the error per level of each of the factor variables. These comparisons between factor levels are done using quantile regression trained with observation weights as predictor for the  $P_\tau$ , but the same patterns can be found when using the other methods. In the final quantile regression model, both without and with observation weights, there are three variables for which this applies: Weekday, Daypart and Area. The difference between the levels in each of these factor variables will be discussed separately below.

#### 4.3.1 Weekday

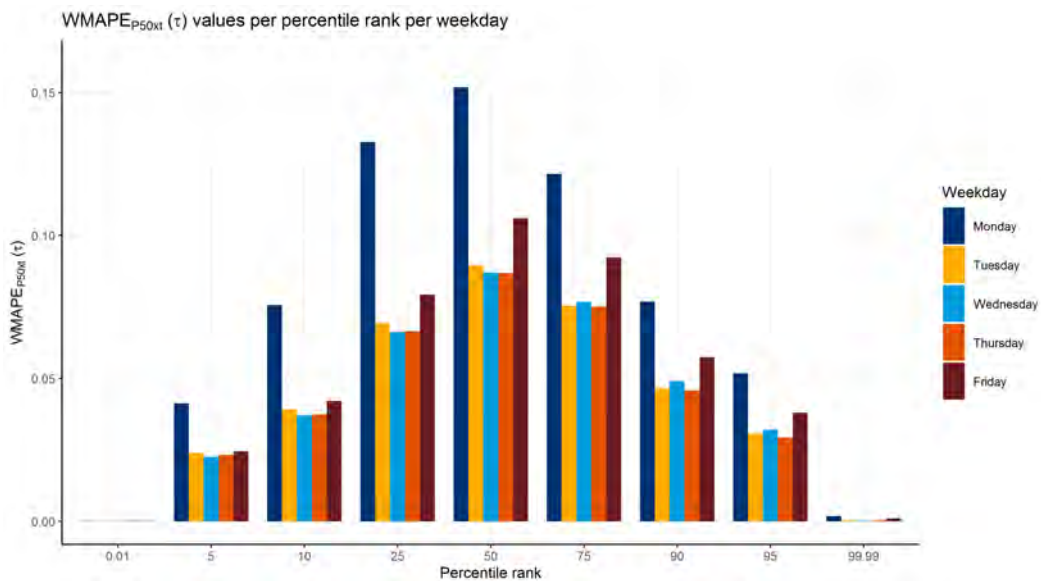
The ratio between the number of stop-stop prognoses for each weekday in the data set is: 21.9% Monday, 23.9% Tuesday, 18.1% Wednesday, 18.0% Thursday and 18.2% Friday, so the number of stop-stop prognoses per level is approximately equal. Figure 18 shows the  $WMAPE(\tau)$  values per weekday for quantile regression trained with observation weights.

Figure 18 shows that Monday, followed by Friday, contributes most to the total error, as it has the highest average over- or underestimation of true values for all percentiles. The peak of this is at  $P_{50}$ , where the average over- or underestimation for Monday is about 16% and for the other percentiles about 10%. It can also be seen that, of all other weekdays, Thursday has the lowest average over- or underestimation for all percentiles, closely followed by Wednesday and Tuesday.



**Figure 18:** WMAPE( $\tau$ ) values per percentile rank for the weekdays using quantile regression trained with observation weights, evaluated on test set.

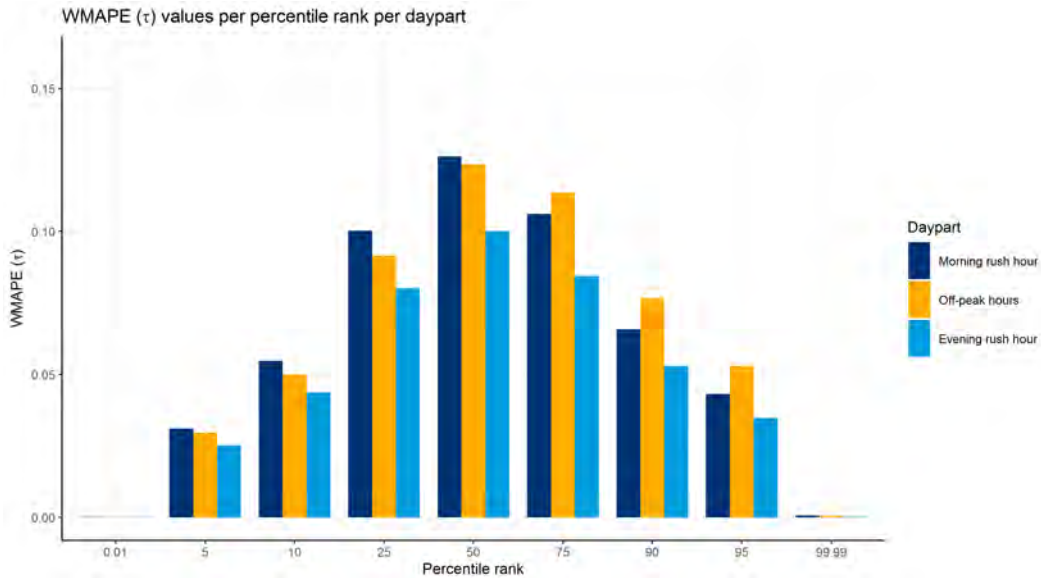
Figure 19 shows the  $WMAPE_{P_{50 \times t}}(\tau)$  values when evaluating the method. The figure shows that again Monday contributes most to the total error, followed by Friday. It also shows that almost all over- and underestimation percentage averages are lower than in Figure 18.



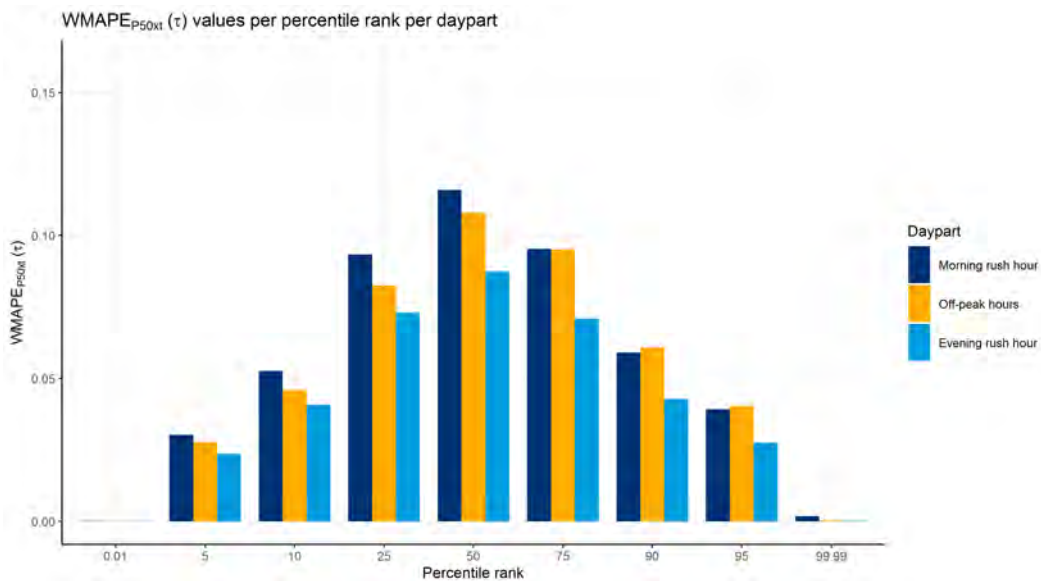
**Figure 19:** WMAPE<sub>P<sub>50 × t</sub></sub>( $\tau$ ) values per percentile rank for the weekdays using quantile regression trained with observation weights, evaluated on test set.

### 4.3.2 Daypart

If we look at the variable Daypart, we see that the ratio of the three dayparts in the data set is: 13.4% morning rush hour, 72.7% off-peak hours and 13.9% evening rush hour. Therefore, the number of observations per level is very skewed, with almost three quarters of the data consisting of off-peak hours. Figure 20 shows the  $WMAPE(\tau)$  values per daypart.



**Figure 20:**  $WMAPE(\tau)$  values per percentile rank for the dayparts using quantile regression trained with observation weights, evaluated on test set.



**Figure 21:**  $WMAPE_{P_{50 \times t}}(\tau)$  values per percentile rank for the dayparts using quantile regression trained with observation weights, evaluated on test set.

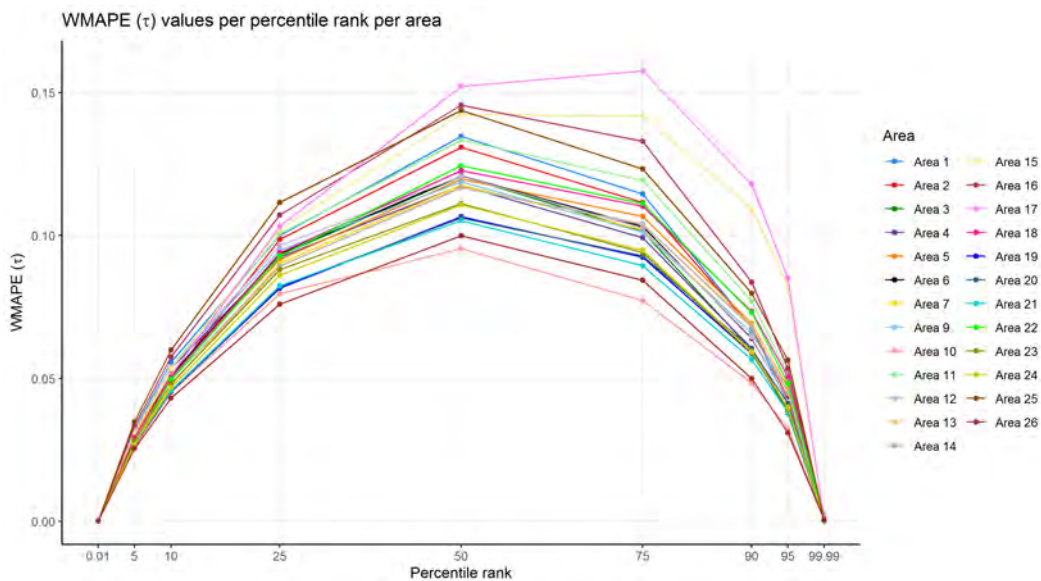
As we can see in Figure 20, for all percentiles the evening rush hour contributes least to the total error. We can also note that for the lower percentiles the morning rush hour contributes most, but for the higher percentiles the off-peak hours contribute most to the total error.

Figure 21 shows the  $WMAPE_{p_{50} \times t}(\tau)$  values per daypart. The figure shows again that the order of contribution between levels is unchanged, but that for the higher percentiles the difference in contribution to the total error between morning rush hour and off-peak hours has become smaller.

### 4.3.3 Area

Since the variable Area distinguishes between 25 areas, the ratio of the areas is shown in Appendix B. The table in the Appendix shows that most observations are in Area 14 (city of Amsterdam/Alkmaar/Purmerend) and Area 20 (city of Utrecht and surroundings) with 16.2% and 9.2% of the total data respectively. The fewest observations are in Area 2 (province of Groningen) and Area 1 (province of Friesland) with respectively 0.7% and 1.1% of the total data. This is probably because train traffic in these provinces is largely provided by train operators other than NS.

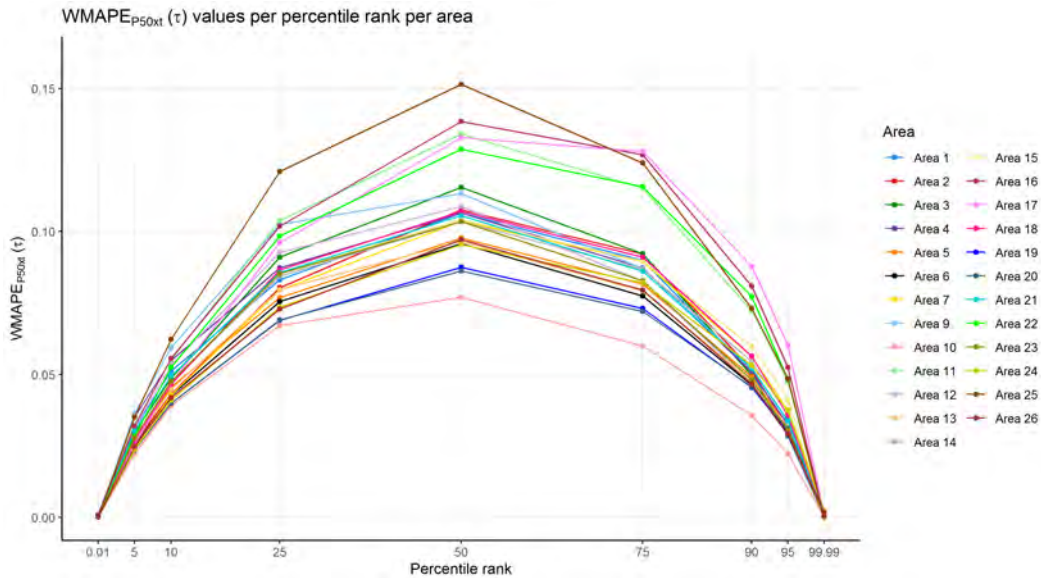
Because a bar plot with 25 bars per percentile rank is too large and too difficult to distinguish between the levels, a line plot is shown in Figure 22 instead. The figure shows an individual line for all areas representing the  $WMAPE(\tau)$  values per percentile rank. It shows that for most areas the lines do not intersect. This means that if we compare two areas with each other, for all percentile ranks the  $WMAPE(\tau)$  values of one area are higher than the values of the other.



**Figure 22:**  $WMAPE(\tau)$  values per percentile rank for the areas using quantile regression trained with observation weights, evaluated on test set.

As can be seen in Figure 22 the smallest contributions to the total error are made by Area 10 (city of Harderwijk/Barneveld/Veenendaal/Ede) and Area 26 (province of Limburg). The largest contributions are made by Area 17 (city of Leiden and surroundings). It can also be seen that for this area and Area 15 (city of Haarlem and surroundings) the  $P_{75}$  is higher than the  $P_{50}$ , which is not the case for all other areas.

If we add the observation weights to the evaluation, the differences between the areas become larger. This can be seen in Figure 23. In the figure we see that for the higher percentiles there is a cluster of areas that have a larger contribution to the total error. The areas in question are: Area 17, Area 16 (Schiphol/Hoofddorp), Area 25 (province of Zeeland), Area 22 (Drechtsteden) and Area 11 (province of Flevoland). However, for the lower percentiles and for  $P_{50}$  Area 25 easily stands out as the largest contributor to the total error. The smallest contributions for the percentiles are still made by Area 10.



**Figure 23:**  $WMAPE_{P_{50} \times t}(\tau)$  values per percentile rank for the areas using quantile regression trained with observation weights, evaluated on test set.

# Chapter 5

## Discussion

This chapter first presents the key findings. Afterwards, the interpretation of the results is discussed. Finally, the limitations of the project are explained.

### 5.1 Key findings

This research is aimed at improving the prediction of the percentiles that represent the distribution of a stop-stop prognose in a BDU. For NS, it is most important that the predictions are good for stop-stop prognoses with higher passenger numbers and/or route travel times. In addition, for NS the higher percentiles are the most important percentiles because they are used in the train scheduling. Therefore, the  $WMAPE_{P_{50} \times t}(\tau)$  values are the most important evaluation method, where the observation weights are included in the evaluation. The results show that for this evaluation method, quantile regression trained with observation weights performs best for all higher percentiles. This method therefore seems best suited to predict the percentiles of the number of passengers for stop-stop prognoses in a BDU.

### 5.2 Interpretation of results

The results on the test set show that both the 3.0-prognoseverdeling method and quantile regression achieve smaller errors than the 2.0-prognoseverdeling method. More specifically, this research shows that the greatest reduction in error can be made in the higher percentiles ( $P_{75}$ ,  $P_{90}$  and  $P_{95}$ ). The best result is achieved when observation weights are included. Quantile regression trained with observation weights can achieve about 2.5% to 7.5% better results for the higher percentiles than the 2.0-prognoseverdeling method. This essentially means that this method has a greater chance of reflecting the distribution of a given  $P_{50}$  more accurately, allowing a more appropriate train to be scheduled. The 3.0-prognoseverdeling method seems to perform best for the  $P_{10}$ ,  $P_{25}$  and  $P_{50}$ . This is surprising since this is not the case for the 5-fold cross-validation. On top of that, the 3.0-prognoseverdeling method is not trained with observation weights, but this version of quantile regression is. It would therefore be more logical for quantile regression to perform better

for all percentiles when evaluated with observation weights.

Quantile regression trained without observation weights actually performs worse than the 2.0-prognoseverdeling method for almost all  $P_\tau$  when evaluated with observation weights. It makes sense that this method performs worse than quantile regression trained with observation weights. However, it is not logical that it performs worse than the baseline method, since it does outperform the baseline method when evaluated without observation weights.

If we do consider all observations equally important, then the 3.0-prognoseverdeling method performs best for all percentiles other than  $P_{0.01}$  and  $P_{99.99}$ . This method can improve on the prediction of the distribution of the percentiles by about 2.5% for the lower percentiles and about 5% for the higher percentiles compared to the 2.0-prognoseverdeling method. The fact that the 3.0-prognoseverdeling method performs best is somewhat unexpected, as quantile regression trained without observation weights actually minimises the evaluation function that is used in this study, whereas the 3.0-prognoseverdeling method does not. On top of that, for most percentiles, quantile regression trained with observation weights performs better than trained without observation weights. So we see that adding observation weights makes a big difference in which method performs best. It is therefore very important for NS that the observation weights are chosen such that everything that NS considers important is included in these weights.

We have also split the total errors by factor variable. For the variable Weekday, we see that the largest contribution to the total error comes from Monday and Friday for all percentiles. This could be because the month of April includes Easter Monday and Good Friday. These are special days that look nothing like a normal Monday or Friday. It could be that for the BDU, these days were treated as normal days. Therefore, the predictions and realised numbers of passengers do not match and the errors are larger than usual.

If we look at variable Daypart, we see that the evening rush hour seems to contribute least to the total error for all percentiles. For the lower percentiles, morning rush hour seems to contribute most. The difference between morning and evening rush hour could be due to the fact that people probably take the same train fairly consistently in the morning, but in the evening they might sometimes take a train earlier or later than usual. As a result, there are likely to be higher peaks in terms of passenger numbers in the morning rush hour. For higher passenger numbers, the absolute error is probably larger than for lower passenger numbers. Hence, the error contribution of morning rush hour is likely to be higher than that of evening rush hour. For the higher percentiles the off-peak hours seem to affect the total error most, which might be surprising as the average number of passengers on these stop-stop prognoses is almost twice as low compared to both rush hours as shown in Table 2 in the Methodology. However, it could be that because 72.7% of the data consists of off-peak hours, its combined contribution to the total error is higher than for the evening rush hour.

For the variable Area, we see that the smallest contribution for almost all percentiles is Area 10 (city of Harderwijk/Barneveld/Veenendaal/Ede). The biggest contributors to the total error are Area 17 (city of Leiden and surroundings), Area 25 (province of Zeeland) and Area 16 (Schiphol/city of Hoofddorp). It could be that due to Good Friday and Easter Monday, people go on holiday for a long weekend, causing the error for Schiphol Airport and province of Zeeland to be larger than usual.

### 5.3 Limitations of the research

The limitations of the study will be addressed separately below.

In this study, we only consider data from April 2019. The training and test set are therefore from the same period. It would probably be more valuable to test the methods on a different period than that of the training set, for example training on April 2018 and testing on April 2019. However, this was not possible for this study as there is less than one year of normal data available with the current method as a predictor of percentiles due to the corona pandemic. In the future, this will become possible if there is more than a year's worth of data.

Specific occasions that apply only to the month of April may have affected the methods. For example, in the Netherlands, Good Friday, Easter and King's Day are in April, which look very different from normal days in terms of train traffic. This might have influenced the variable selection and the evaluation since in this study these special days were included as normal days.

This study does not include data on weekends, because NS has indicated that it is more important to focus on weekdays rather than weekend days. In addition, the weekend may cause problems in quantile regression because it creates a dependency between the variables Weekday and Daypart. This is because in the weekend no distinction is made between morning rush hour, off-peak hours and evening rush hour. On the other hand, weekdays cannot have the weekend daypart.

International trains are not considered in this study. NS has stated that the estimates of the realised number of passengers for these trains in 2019 is not reliable enough. Should these be reliable enough in the future, they could be included in a follow-up study.



# Chapter 6

## Conclusions

This chapter first presents the main conclusions, where we provide an answer to the research question. After this, the recommendations for future research are presented.

### 6.1 Main conclusions

In this study, we examine various methods for determining the distribution of passenger forecasts. The research question of this study is: “How can the estimation of certain percentile values in the probability distribution of a predicted number of passengers for a combination of train number, day of the week and route in a BDU be improved?”.

Based on analysis of past passenger forecast data, we can conclude that both quantile regression and the 3.0-prognoseverdeling method perform better as predictors of the percentiles than the current method and the 2.0-prognoseverdeling method. The results show that the prediction method used in 2019 on average overestimated the lower percentiles and underestimated the higher percentiles. This can be seen in Figure 8 where the line for the current method is not a straight line but a curve instead. If we look at the lines in this plot for the other methods, we see that the percentages of these methods are very close to the percentile ranks. This implies that the methods predict the percentiles well on average. Because a correct prediction for the stop-stop prognoses with higher  $P_{50}$  forecast and/or route travel time is considered more important than for other stop-stop prognoses, the most important evaluation metric are the  $WMAPE_{P_{50} \times t}(\tau)$  values, where the observation weights are included. On top of that, the higher percentiles ( $P_{75}$ ,  $P_{90}$  and  $P_{95}$ ) are more important to NS than the lower percentiles as the higher percentiles are the ones mainly considered when making or adjusting a train schedule. For the higher percentiles, quantile regression achieves the best result with an improvement of about 2.5% to 7.5% compared to the 2.0-prognoseverdeling method. In conclusion, the use of quantile regression trained with observation weights is a way to improve the probability distribution of a predicted number of passengers for a combination of train number, day of the week and route in a BDU.

## 6.2 Recommendations for future research

The biggest recommendation is to investigate whether it is possible to switch to quantile regression for determining the percentiles as it shows promising results, especially when evaluated with observation weights. The fact that quantile regression can take observation weights into account when determining the best coefficients for the model is a big advantage over the other methods. If the method is to be implemented, then further research will certainly be required as this project did not investigate weekends and, of course, there are passenger forecasts for weekend days as well. In addition, in this project we have looked at only one month of data. It would probably be interesting to look at data of a full BDu. Lastly, we have not considered international trains, so if predictions of percentiles for these are desired, this should also be investigated.

The variable selection for quantile regression now focuses only on the  $P_{90}$ . However, it could be the case that for different  $P_{\tau}$ s a different combination of variables would yield the best results. This would require independent variable selection for all desired percentiles, searching for the best combination of variables for each percentile individually.

Should it be that implementing quantile regression is not possible or that an improvement is needed on a shorter term, the recommendation would be to implement the groups of the 3.0-prognoseverdeling method to replace those of the 2.0-prognoseverdeling method.

The  $P_{50}$  groups of the 3.0-prognoseverdeling method are now defined manually. A different approach for this is to use deciles. In that case, all stop-stop prognoses are distributed in ten  $P_{50}$  groups so that all groups are of approximately equal size in terms of stop-stop prognoses (Hayes, 2021). Another approach could be to use the seat capacities of trains as break-off points for the  $P_{50}$  groups.

There is still room for improvement in the 3.0-prognoseverdeling method in terms of missing data. If it is now the case that there are no stop-stop prognoses within a certain combination of Weekday, Daypart, Variable size  $P_{50}$  group, Area and Rolling stock, but there are stop-stop prognoses of this combination in a new data set, then the prediction of the 3.0-prognoseverdeling for this stop-stop prognose is the average of all stop-stop prognoses in the training set. Of course, other techniques can be used for this, such as using the average of the best comparable combination of variable levels as a prediction. This of course requires defining when combinations of variable levels are comparable.

Taking into account how close a prediction is to the seat capacity of a train could also be useful to add to the observation weights. E.g., if a rolling stock unit has a capacity of 150 seats, then accurate predictions for a  $P_{90}$  of 151 passengers or 299 passengers are much more important than accurate predictions for a  $P_{90}$  of 225 passengers.

Another method that has been examined very briefly is quantile random forest. This method also seems to be a suitable predictor of the percentiles. Given the limited

time span of this study, we decided not to investigate this method thoroughly, but instead to improve the variable selection of the 2.0-prognoseverdeling method (which is the 3.0-prognoseverdeling method). In this project, quantile random forest was performed once, achieving results that are slightly less good than the quantile regression method. With some proper adjustments, comparable results can probably be achieved. This could be done as a future research project.

The  $P_{\min}$  and especially the  $P_{\max}$  are tricky percentiles to include in the desired percentiles and require a different approach. As these percentiles are not defined for quantile regression, the  $P_{0.01}$  and  $P_{99.99}$  are chosen instead. In the 2.0-prognoseverdeling method and 3.0-prognoseverdeling method, the  $P_{\min}$  of a given group is defined as the minimum of the number of realised passengers. This will not cause any problems since the number of passengers for the  $P_{\min}$  is limited by the fact that it cannot be negative. However, the  $P_{\max}$  is highly affected by outliers and must therefore be handled with care. In the 2.0- and 3.0-prognoseverdeling methods, we thus choose a linear extrapolation from the  $P_{99.5}$  by adding the difference between the  $P_{99.5}$  and  $P_{99}$  to this to get a prediction for the  $P_{\max}$ .

It could be that the  $P_{50}$  of a stop-stop prognose predicted by quantile regression or the 3.0-prognoseverdeling method is higher or lower than the  $P_{50}$  predicted by NS. This may cause the distribution around the NS  $P_{50}$  to become very odd. A possible solution is to raise or lower all percentile predictions by the difference between the NS  $P_{50}$  and the quantile regression or 3.0-prognoseverdeling method  $P_{50}$ .

# Appendix A

## Numbers of missing data

In this appendix, the exact differences in numbers of stop-stop prognoses between the SOFA dataset and the passenger forecast dataset are presented. On top of that, possible explanations for this are provided.

The SOFA dataset contains 619,687 stop-stop prognoses about the realised number of passengers. There is no missing data in this dataset itself. The passenger forecast dataset contains 642,703 stop-stop prognoses about the passenger predictions in terms of  $P_\tau$  with  $\tau \in \{\text{min}, 5, 10, 25, 50, 75, 90, 95, \text{max}\}$ . In this dataset, there are 1,552 stop-stop prognoses with missing percentile values. Most of these missing data (1,513 stop-stop prognoses) are on the train route Gouda - Gouda Goverwelle and vice versa. NS has confirmed that something went wrong when the data for this route was added to the dataset.

The datasets are merged into a dataset containing 600,637 stop-stop prognoses. Most of the stop-stop prognoses are present in both sets, but this is not the case for the following amount of stop-stop prognoses: 18,550 stop-stop prognoses are missing in the SOFA dataset and 40,014 stop-stop prognoses are missing in the passenger forecast dataset.

There are a number of reasons why a stop-stop prognose can be present in the SOFA dataset, but not in the passenger forecast dataset. For example, it is possible that a train has to stop at a certain station due to a disruption, or if for some reason an extra stop has to be added. For these stop-stop prognoses there are no passenger forecasts, but there is a realised number of travelled passengers. Trains may also be cancelled for any reason. In this case, the number of passengers for the stop-stop prognoses of this train are predicted, but there is no realised number of passengers. In that case, these stop-stop prognoses are present in the SOFA dataset, but not in the passenger forecast dataset.

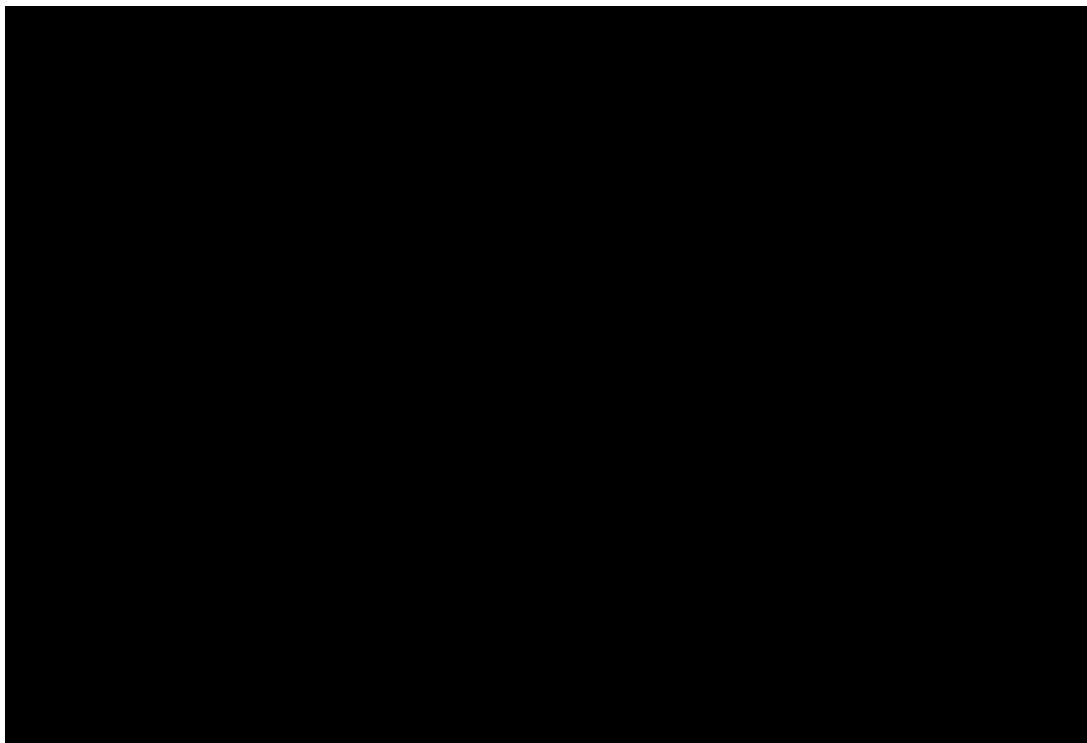
There are also a few special days in the month of April, namely: Good Friday and Easter Monday. It could be that a forecast for these special days was made as if it were a normal Friday or Monday. The numbers of missing stop-stop prognoses for these days are: 592 and 2,765 respectively for the SOFA dataset and as much as 699 and 13,128 respectively for the passenger forecast dataset.

# Appendix B

## Details of variable Area

A table showing the full name of the area, in which RIP region it is located (Randstad, invloedsgebied or periferie), the percentage of the dataset and the five-number summary per area. Area 8 does not have a percentage and five-number summary as the train traffic in this area is provided by a train operator other than NS. Area 27 are the border crossings, which are not included in this study.

**Table 9:** The full name, RIP, percentage of the dataset and five-number summary of the realised number of passengers per area.



# Appendix C

## Step-up steps for quantile regression without observation weights

This appendix provides the steps for the step-up method used for the quantile regression model trained without observation weights. Table 10 shows the  $\hat{\beta}(0.9)$  values and whether all variables are significantly different from zero for a quantile regression model with each explanatory variable separately.

**Table 10:** The  $\hat{\beta}(0.9)$  values (in number of passengers) and significance of variables for a quantile regression model with each explanatory variable separately.

Explanatory variable	$\hat{\beta}(0.9)$	All variables significant?
$\log(P_{50})$	33,206.71	Yes
$P_{50}$ group (100)	43,176.39	Yes
$P_{50}$ group (variable)	34,979.62	Yes
Weekday	77,703.46	Yes
Daypart	71,782.92	Yes
Time (sin & cos)	76,969.30	Yes
RIP	76,178.70	Yes
Area	74,703.64	Yes
Rolling stock	72,127.63	Yes

Table 10 shows that the  $P_{50}$  forecast variables are the most influential variables as their  $\hat{\beta}(0.9)$  values are about twice as small as those of the other variables. Of the  $P_{50}$  forecast variables, the logarithm seems to perform best. This variable is therefore chosen to form the basis of the model. Table 11 shows the results for a quantile regression model with  $\log(P_{50})$  and each remaining explanatory variable separately.

Table 11 shows that variable Area achieves the lowest error value. To test whether the model including Area performs significantly better than the model without Area, the  $F$ -test is performed, resulting in an  $F$ -statistic of 746.71, with corresponding  $p$ -value of  $< 10^{-15}$ . As the  $p$ -value is smaller than the significance level of  $\alpha = 0.05$ ,

**Table 11:** The  $\hat{\beta}(0.9)$  values (in number of passengers) and significance of variables for a quantile regression model with  $\log(P_{50})$  and each remaining explanatory variable separately.

Explanatory variables	$\hat{\beta}(0.9)$	All variables significant?
$\log(P_{50})$ , Weekday	33,204.94	Yes
$\log(P_{50})$ , Daypart	33,126.95	Yes
$\log(P_{50})$ , Time (sin & cos)	33,127.19	Yes
$\log(P_{50})$ , RIP	32,402.32	Yes
$\log(P_{50})$ , Area	31,798.18	Yes
$\log(P_{50})$ , Rolling stock	33,121.89	Yes

we can conclude that the model including Area performs significantly better than the model without Area. Therefore, Area should be included in the model. Table 12 shows the results for a quantile regression model with  $\log(P_{50})$ , Area and each remaining explanatory variable separately.

**Table 12:** The  $\hat{\beta}(0.9)$  values (in number of passengers) and significance of variables for a quantile regression model with  $\log(P_{50})$ , Area and each remaining explanatory variable separately.

Explanatory variables	$\hat{\beta}(0.9)$	All variables significant?
$\log(P_{50})$ , Area, Weekday	31,794.68	Yes
$\log(P_{50})$ , Area, Daypart	31,724.31	Yes
$\log(P_{50})$ , Area, Time (sin & cos)	31,748.11	Yes
$\log(P_{50})$ , Area, RIP	31,784.46	Yes
$\log(P_{50})$ , Area, Rolling stock	31,758.13	Yes

Table 12 shows that variable Daypart achieves the lowest error value. Again, the  $F$ -test is performed, resulting in an  $F$ -statistic of 473.45, with corresponding  $p$ -value of  $< 10^{-15}$ . As the  $p$ -value is smaller than the significance level of  $\alpha = 0.05$ , we can conclude that the model including Daypart performs significantly better than the model without Daypart. Therefore, Daypart should be included in the model. Since variables Daypart and Time both cover the time of departure and Daypart performs better, Time is not included in the model. Table 13 shows the results for a quantile regression model with  $\log(P_{50})$ , Area, Daypart and each remaining explanatory variable separately.

Table 13 shows that variable Weekday achieves the lowest error value. Again, the  $F$ -test is performed, resulting in an  $F$ -statistic of 123.18, with corresponding  $p$ -value of  $< 10^{-15}$ . As the  $p$ -value is smaller than the significance level of  $\alpha = 0.05$ , we can conclude that the model including Weekday performs significantly better than the model without Weekday. Therefore, Weekday should be included in the model. Table

**Table 13:** The  $\hat{\beta}(0.9)$  values (in number of passengers) and significance of variables for a quantile regression model with  $\log(P_{50})$ , Area, Daypart and each remaining explanatory variable separately.

Explanatory variables	$\hat{\beta}(0.9)$	All variables significant?
$\log(P_{50})$ , Area, Daypart, Weekday	31,624.88	Yes
$\log(P_{50})$ , Area, Daypart, RIP	31,695.95	Yes
$\log(P_{50})$ , Area, Daypart, Rolling stock	31,683.04	Yes

14 shows the results for a quantile regression model with  $\log(P_{50})$ , Area, Daypart, Weekday and each remaining explanatory variable separately.

**Table 14:** The  $\hat{\beta}(0.9)$  values (in number of passengers) and significance of variables for a quantile regression model with  $\log(P_{50})$ , Area, Daypart, Weekday and each remaining explanatory variable separately.

Explanatory variables	$\hat{\beta}(0.9)$	All variables significant?
$\log(P_{50})$ , Area, Daypart, Weekday, RIP	31,577.44	No
$\log(P_{50})$ , Area, Daypart, Weekday, Rolling stock	31,622.35	No

Table 14 shows that for both RIP and Rolling stock the error is less than in Table 13, but not all variables are significantly different from zero anymore. This indicates that a model with one of these variables does not outperform a model without these variables. Hence, the best found quantile regression model trained without observation weights consists of the variables  $\log(P_{50})$ , Area, Daypart and Weekday.



# Appendix D

## Step-up steps for quantile regression with observation weights

This appendix provides the steps for the step-up method used for the quantile regression model trained with observation weights. Table 15 shows the  $\hat{\beta}_{P_{50} \times t}(0.9)$  values and whether all variables are significantly different from zero for a quantile regression model with each explanatory variable separately.

**Table 15:** The  $\hat{\beta}_{P_{50} \times t}(0.9)$  values (in number of passengers<sup>2</sup> minutes) and significance of variables for a quantile regression model with each explanatory variable separately.

Explanatory variable	$\hat{\beta}_{P_{50} \times t}(0.9)$	All variables significant?
log(P <sub>50</sub> )	33,566.84	Yes
P <sub>50</sub> group (100)	43,678.33	Yes
P <sub>50</sub> group (variable)	35,063.00	Yes
Weekday	95,131.80	Yes
Daypart	82,977.94	Yes
Time (sin & cos)	97,751.11	Yes
RIP	93,300.24	Yes
Area	91,491.07	Yes
Rolling stock	82,296.87	Yes

Table 15 shows that the P<sub>50</sub> forecast variables are the most influential variables as their  $\hat{\beta}_{P_{50} \times t}(0.9)$  values are about twice as small as those of the other variables. Of the P<sub>50</sub> forecast variables, the logarithm seems to perform best. This variable is therefore chosen to form the basis of the model. Table 16 shows the results for a quantile regression model with log(P<sub>50</sub>) and each remaining explanatory variable separately.

Table 16 shows that variable Area achieves the lowest error value. To test whether the model including Area performs significantly better than the model without Area, the  $F$ -test is performed, resulting in an  $F$ -statistic of 172.40, with corresponding

**Table 16:** The  $\hat{\beta}_{P_{50} \times t}(0.9)$  values (in number of passengers<sup>2</sup> minutes) and significance of variables for a quantile regression model with  $\log(P_{50})$  and each remaining explanatory variable separately.

Explanatory variables	$\hat{\beta}_{P_{50} \times t}(0.9)$	All variables significant?
$\log(P_{50})$ , Weekday	33,591.48	Yes
$\log(P_{50})$ , Daypart	33,485.37	Yes
$\log(P_{50})$ , Time (sin & cos)	33,498.92	Yes
$\log(P_{50})$ , RIP	32,722.83	Yes
$\log(P_{50})$ , Area	32,352.88	Yes
$\log(P_{50})$ , Rolling stock	33,423.66	Yes

$p$ -value of  $< 10^{-15}$ . As the  $p$ -value is smaller than the significance level of  $\alpha = 0.05$ , we can conclude that the model including Area performs significantly better than the model without Area. Therefore, Area should be included in the model. Table 17 shows the results for a quantile regression model with  $\log(P_{50})$ , Area and each remaining explanatory variable separately.

**Table 17:** The  $\hat{\beta}_{P_{50} \times t}(0.9)$  values (in number of passengers<sup>2</sup> minutes) and significance of variables for a quantile regression model with  $\log(P_{50})$ , Area and each remaining explanatory variable separately.

Explanatory variables	$\hat{\beta}_{P_{50} \times t}(0.9)$	All variables significant?
$\log(P_{50})$ , Area, Weekday	32,373.15	Yes
$\log(P_{50})$ , Area, Daypart	32,203.35	Yes
$\log(P_{50})$ , Area, Time (sin & cos)	32,310.44	Yes
$\log(P_{50})$ , Area, RIP	32,350.02	Yes
$\log(P_{50})$ , Area, Rolling stock	32,365.05	Yes

Table 17 shows that variable Daypart achieves the lowest error value. Again, the  $F$ -test is performed, resulting in an  $F$ -statistic of 307.03, with corresponding  $p$ -value of  $< 10^{-15}$ . As the  $p$ -value is smaller than the significance level of  $\alpha = 0.05$ , we can conclude that the model including Daypart performs significantly better than the model without Daypart. Therefore, Daypart should be included in the model. Since variables Daypart and Time both cover the time of departure and Daypart performs better, Time is not included in the model. Table 18 shows the results for a quantile regression model with  $\log(P_{50})$ , Area, Daypart and each remaining explanatory variable separately.

Table 18 shows that variable Weekday achieves the lowest error value. Again, the  $F$ -test is performed, resulting in an  $F$ -statistic of 63.98, with corresponding  $p$ -value of  $< 10^{-15}$ . As the  $p$ -value is smaller than the significance level of  $\alpha = 0.05$ , we can conclude that the model including Weekday performs significantly better than the

**Table 18:** The  $\hat{\beta}_{P_{50} \times t}(0.9)$  values (in number of passengers<sup>2</sup> minutes) and significance of variables for a quantile regression model with  $\log(P_{50})$ , Area, Daypart and each remaining explanatory variable separately.

Explanatory variables	$\hat{\beta}_{P_{50} \times t}(0.9)$	All variables significant?
$\log(P_{50})$ , Area, Daypart, Weekday	32,115.50	Yes
$\log(P_{50})$ , Area, Daypart, RIP	32,167.54	Yes
$\log(P_{50})$ , Area, Daypart, Rolling stock	32,225.84	Yes

model without Weekday. Therefore, Weekday should be included in the model. Table 19 shows the results for a quantile regression model with  $\log(P_{50})$ , Area, Daypart, Weekday and each remaining explanatory variable separately.

**Table 19:** The  $\hat{\beta}_{P_{50} \times t}(0.9)$  values (in number of passengers<sup>2</sup> minutes) and significance of variables for a quantile regression model with  $\log(P_{50})$ , Area, Daypart, Weekday and each remaining explanatory variable separately.

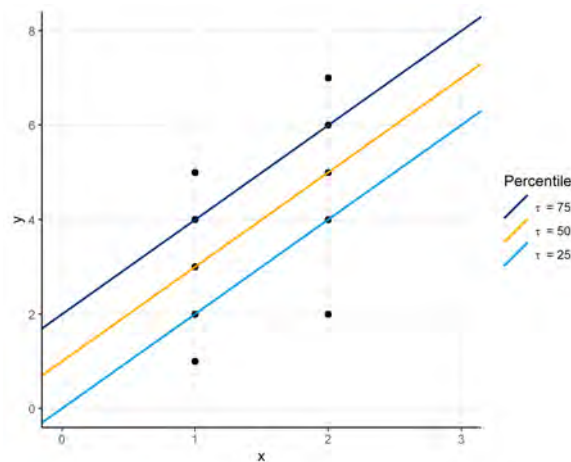
Explanatory variables	$\hat{\beta}_{P_{50} \times t}(0.9)$	All variables significant?
$\log(P_{50})$ , Area, Daypart, Weekday, RIP	32,071.66	No
$\log(P_{50})$ , Area, Daypart, Weekday, Rolling stock	32,131.90	Yes

Table 19 shows that for RIP the error is less than in Table 18, but not all variables are significantly different from zero anymore. For variable Rolling stock, the  $\hat{\beta}_{P_{50} \times t}(0.9)$  value is higher than that of the model without Rolling stock. This indicates that a model with any of these variables does not outperform a model without these variables. Hence, the best found quantile regression model trained with observation weights consists of the variables  $\log(P_{50})$ , Area, Daypart and Weekday.

# Appendix E

## Example of differences in minimums

In this example we show why the different percentiles cannot be compared with each other. We do this by showing that for different percentiles, there are different minimum errors that can be achieved. Figure 24 shows a small data set with ten data points and three prediction lines running through the data at exactly the 25th, 50th and 75th percentiles.



**Figure 24:** Example dataset of ten data points with prediction lines at exactly  $P_{75}$ ,  $P_{50}$  and  $P_{25}$ .

If we then calculate the  $WMAPE(\tau)$  values according to Equation 3.2 in the Methodology for the three different percentiles, we get the following:

$$WMAPE(25) = 0.25 \times (1 + 2 + 3 + 1 + 2 + 3) + 0.75 \times (1 + 2) = 5.25,$$

$$WMAPE(50) = 0.5 \times (1 + 2 + 1 + 2) + 0.5 \times (1 + 2 + 1 + 3) = 6.5,$$

$$WMAPE(75) = 0.75 \times (1 + 1) + 0.25 \times (1 + 2 + 3 + 1 + 2 + 4) = 4.75.$$

So we see that, although the prediction lines in Figure 24 perfectly pass through the desired percentiles, the  $WMAPE(\tau)$  values still differ from each other between the percentiles. We can therefore conclude that the  $WMAPE(\tau)$  values are not comparable between the percentiles.

# Bibliography

- Berrar, D. (2019). Cross-validation. *Encyclopedia of bioinformatics and computational biology*, 1:542–545. pages 11, 27
- Chockalingam, M. (2007). Forecast accuracy and safety stock strategies. *Demand Planning*, 6(11):2003. pages 13
- Foi, A. (2009). Optimization of variance-stabilizing transformations. *Preprint, 2009b*, 94:1809–1814. pages 9
- Glen, S. (2021). Percentiles, percentile rank & percentile range: Definition & examples. <https://www.statisticshowto.com/probability-and-statistics/percentiles-rank-range/>. Accessed: 15-08-2021. pages 3
- Govers, B. (2011). Regie in de (randstad)knoop. pages 8
- Hayes, A. (2021). Decile. <https://www.investopedia.com/terms/d/decile.asp>. Accessed: 23-08-2021. pages 42
- Heckert, N. A., Filliben, J. J., Croarkin, C. M., Hembree, B., Guthrie, W. F., Tobias, P., Prinz, J., et al. (2002). *Handbook 151: NIST/SEMATECH e-Handbook of Statistical Methods*. NIST Interagency/Internal Report (NISTIR). pages 16
- Huang, M. L. and Rat, R. (2017). A new weighted quantile regression. *Cogent Mathematics & Statistics*, 4(1):1357237. pages 23
- Hyndman, R. J. and Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365. pages 16
- Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50. pages 13
- Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of economic perspectives*, 15(4):143–156. pages 3, 19, 20
- Koenker, R. and Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448):1296–1310. pages 14, 28
- Kolassa, S., Schütz, W., et al. (2007). Advantages of the mad/mean ratio over the mape. *Foresight: The International Journal of Applied Forecasting*, pages 40–43. pages 13, 27

London, I. (2016). Encoding cyclical continuous features - 24-hour time. pages 10

Pardoe, I., Simon, L., and Young, D. (2021). The Hypothesis Tests for the Slopes. <https://online.stat.psu.edu/stat501/lesson/6/6.4>. Accessed: 17-07-2021. pages 20, 21

PTV Group (2021). Public transport planning with ptv visum. <https://www.ptvgroup.com/en/solutions/products/ptv-visum/>. Accessed: 14-08-2021. pages 2