

MASTER

Using travel time predictions based on TomTom's big data in logistical models

van der Hooft, E.J.A.

Award date:
2013

[Link to publication](#)

Disclaimer

This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain

Eindhoven, February 2013

**Using Travel Time Predictions
based on TomTom's Big Data in
Logistical Models**

by
Erik J.A. van der Hooft

BSc Mechanical Engineering (2011)
Student identity number 0566946

in partial fulfilment of the requirements for the degree of

**Master of Science
in Operations Management and Logistics**

Supervisors:

Prof.dr. T. van Woensel, TU/e, OPAC

Prof.dr.ir. W.P.M. Nuijten, TU/e, OPAC

Dhr. George de Boer, TomTom

Using Travel Time Predictions based on TomTom's Big Data in Logistical Models

TUE. School of Industrial Engineering.
Series Master Theses Operations Management and Logistics

Subject headings: Logistics, Big data, Travel Time, Vehicle Routing Problem, Predictions

Abstract

This research report focuses on the need of reliable travel time predictions for logistical models. Although there exist approaches to incorporate stochastic and dynamic travel time information in logistical models, there exists a lack of data to reliably determine these travel time distributions over a whole network. Secondly the effect of stochastic and dynamic travel time estimations on a real life network has not been extensively analyzed. The data used in this report originates from TomTom and comprises a very large number of floating car velocity observations. This dataset enables statistical analysis on virtually every road segment of the road network. Using these observations a method is derived and tested to predict the time dependent travel time distribution. These travel time predictions are then applied in some simple logistical models to verify whether the reliability of the results can be improved. Reliability is hereby defined as the probability that the predicted travel time is not exceeded.

Acknowledgements

I would like to shortly express my gratitude towards my supervisors, who have pointed me in the good direction several times. I would like to thank Tom van Woensel for his guidance throughout this project, and his comments on my preliminary work. Secondly I would like to thank Wim Nuijten for his time to review my work. But I'm also grateful for all the time and critical comments, from a commercial viewpoint, given by George de Boer, and his invitations to several interesting seminars during the last six months.

But last I would like to thank my girlfriend, who took care of me when I busy working on my project.

Contents

ABSTRACT	3
ACKNOWLEDGEMENTS	4
CONTENTS	5
1. INTRODUCTION	6
2. LITERATURE	8
3. PROBLEM DESCRIPTION AND METHODOLOGY	10
3.1 BACKGROUND.....	10
3.2 USING TOMTOM'S BIG DATA.....	11
4. SOLUTION METHOD	13
4.1 DATASET.....	13
4.2 NOTATIONS.....	14
4.3 DETERMINATION OF DATA SET.....	15
4.4 THE RELATION BETWEEN VELOCITY AND TRAVEL TIME.....	15
4.5 COMPARISON BETWEEN ESTIMATED AND REALIZED VELOCITY.....	16
4.6 ROUTE SIMULATION USING SAMPLE OBSERVATIONS.....	17
4.7 PROCESSING THE DATA.....	17
5. ANALYSIS	19
5.1 NUMBER OF OBSERVATIONS.....	19
5.2 DISTRIBUTION OF THE TRAVEL TIME.....	19
5.3 DISTRIBUTION OF VELOCITIES.....	20
5.4 INCLUDING STOPS IN THE TRAVEL SPEED ESTIMATIONS.....	21
5.5 OBSERVED DRIVING SPEED VERSUS YEAR AND TIME OF THE YEAR.....	22
5.6 DRIVING SPEED VERSUS TIME OF THE DAY.....	24
5.7 WEATHER AND DRIVING SPEED.....	25
5.8 WEATHER AND THE TIME OF YEAR.....	25
6. ANALYSIS OF UNRESTRICTED VELOCITIES PER ROAD SEGMENT	27
6.1 PERIOD WITH BEST PREDICTIVE POWER OVER FUTURE TRAVELING SPEED.....	27
6.2 RELIABILITY OF PREDICTED PERCENTILE SCORES.....	28
7. RESULTS FOR TRAVEL TIME PREDICTIONS FOR ROUTES	31
7.1 PREDICTION COMPARED WITH SIMULATED ROUTE.....	31
7.2 REALISTIC ROUTES.....	31
8. IMPLICATIONS FOR THE VEHICLE ROUTING MODEL	34
9. CONCLUSION	36
REFERENCES	37
APPENDICES	39

1. Introduction

The current paradigm of travel time predictions comprises both static and time dependent figures, both are often discrete values, for example based on a postal code matrix or a simple routing algorithm using static velocities per road segment. The time dependent predictions include the dynamic behavior of travel times in their prediction, caused by congestion. The dynamics are often captured in speed profiles, indicating a ratio between the velocity at a certain time interval and the free-flow velocity. However, the travel time for an individual trip is determined by many variables, which cannot all be predicted. Therefore it will be an improvement if travel time predictions are modeled as a stochastic variable, indicating the probability that a trip will last a certain time, or will not last longer than a certain time. Using these dynamic travel time distributions, and their confidence intervals, in logistical models might result in more reliable outcomes. Although the definition of reliability for travel time predictions heavily depends on the application and context, the shape of the distribution all necessary information. A narrow travel time distribution is for example more reliable, as it is more likely to encounter a value close to the expected value. Intuitively it can be assumed that there might be a trade off between the reliability of a route and the average travel time.

When using software to solve planning problems, the software needs travel time approximations between all sets of locations. Depending on the methods, data and software used, the planning software uses estimates from very aggregated, to reliable and realistic. Software can use a straight-line distance or a postal-code matrix to determine the approximation, these methods do not, or only roughly take the actual road network into account. More reliable solutions would arise if the software uses a street map as an input and calculates the best route over the road, using velocity parameters per road. But even better travel time estimations take the time dependency of travel times into account, resulting in 'dynamic travel time estimations'.

However, there exists an important information gap between on-line planning phases, which can rely on live data for their travel time predictions, such as live traffic feeds with information on accidents, reduced speeds or blocked road, and off-line planning phases, which have to rely on historic data alone. Historic data, from various sources, is already widely used to make travel time estimations, but data is often obtained using stationary equipment, such as counting loops or camera's, or from departure-arrival observations for a specific route. The results from these measurements are not generable to the whole road network and therefore have limited applicability. The ideal situation for predicting traveling times, would be if one could measure travel speeds on the whole road, during the whole day, and for every road of the road network. Some examples of this type of data exist, but often the number of observations is limited.

Therefore this paper investigates the predictive power of historic data, which originates from many TomTom Personal Navigation Device (PND) users and includes measures for almost every road. The data covers the metropolitan area of Eindhoven, a city in the south of the Netherlands, over a period of more than 5 years and includes over 3.2×10^9 individual measurements. After presenting a literature background on the problem in Chapter 2, the problem is formally introduced in Chapter 3. The solution method is described in Chapter 4, followed by the analysis of the data in Chapter 5. Then, a method is presented that makes acceptable predictions of future traveling velocities in Chapter 6. The transition from future velocities to travel time estimations, incorporating the stochasticity, is described in Chapter 7. Last it is analyzed whether logistical problems, in particular the vehicle routing model, can actually benefit from the improved travel time predictions in Chapter 8.

The contributions of this report are threefold, first, the data used in this report is quite unique as it is dynamically recorded (*car floating*) data from a very large pool of users, and contains observations for a large part of the road network, including local roads. Secondly, because of the timeframe over which measurements are available, the predictions can actually be tested against control observations. Last, the emphasis is often on improving the heuristics of planning tools or on improving travel time estimations for highways, however, limited research is done on improving travel time estimations that included local roads. Therefore it is investigated whether the reliability of travel time predictions for in city distributions can be improved, based on TomToms' historic data.

2. Literature

Logistics, which is the planning and execution of transporting goods and people, is the backbone of our current society and economy. According to Peeters et al. (2009) logistics provide the Netherlands with 600,000 jobs, divided over 12,000 firms and yearly add about 30 billion Euros to the Dutch economy. Improving efficiency in this sector will have a positive effect on the Dutch economy and is therefore an important field of research. A good introduction on the importance of logistics with an extensive overview on the background in which logistics has developed is given by Rodrigue et al (2006).

This report focuses on travel time prediction for road logistics, as unpredictable travel times are one of the costs associated with logistics. Many companies use software tools to plan their operations, but not all include time dependent congestion in their modeling. An analysis on the costs associated with congestion related delays in a manufacturing and cross docking environment is done by McKinnon (1998). The author states that small and predictable fluctuations in arrival time can be accounted for, but vehicles arriving outside these time windows cause hold ups or inefficient use of vehicles. Another paper by Fosgerau and Karlstrom (2010) describes a relation between the economic value and travel time reliability when both the mean and standard deviation are dependent on the departure time. A real life example, where the Canadian Post adds customer value by decreasing tardy and repeated-deliveries, is given by LeBlanc (2010).

Planning models are an important field within logistic research, as they provide insights and solutions to problems encountered by all companies that have logistic processes. These models are, for example, used to determine the ideal locations to build a warehouse, or couple vehicles, loads and customer locations in cost efficient routes. These models are also widely discussed in literature, an introduction to planning models in freight transportation is given by Crainic and Laporte (1997) and Ghiani et al. (2004). Most planning models use travel time predictions as one of their inputs, the vehicle routing problem (VRP), as introduced by Dantzig and Ramses in 1959, is one of them and this problem often has to be solved before live data is available. The characteristics of the standard VRP are: (1) A single warehouse in which goods are stored and (2) a set of customers that have (3) a set of demands requiring a certain capacity. (4) A set of arcs between the warehouse and customers and between customers, for which (5) a cost for traveling each arc is formulated (based on distance, travel time or fixed travel costs) (Laporte, 2009). An overview of the VRP with many examples is given by Giani et al. (2004). Further additions to the model are also described in literature, such as capacity constraints, which yield the Capacitated Vehicle Routing Problem (CVRP) (Toth & Vigo, 2002) and the application of Time Windows to the VRP (Desrossiers et al., 1995), (Andersen et al., 2009) and (Crainic et al., 1993), resulting in the VRPTW. An overview over algorithms to solve the (C)VRP(TW) are presented by, for example, Gendreau et al. (1994). Because the number of solutions for the VRP increases when applying dynamic travel times, the problem becomes increasingly harder to solve. An application of congestion avoidance by delaying departure times in the VRP is described by Kok et al. (2011).

The calculation or prediction of dynamic travel times, which can be incorporated in the VRP remains an interesting topic of research. The determination of dynamic travel time observations based on a combination of subjective estimates by humans, combined with historic data is presented by Hill and Benton (1992). Estimations based on data from stationary measurement devices, such as counting loops and camera's, are used by, for example, Woensel et al. (2008), Lecluyse et al. (2009) and Chien and Kuchipudo (2003). The estimations in the first paper are

based on a queuing approach, incorporating the stochastic properties of the data, the queuing model is extended in the second paper. The latter paper describes real time travel time predictions for a toll-road, based on a combination of historic and live data.

Since travel times are the result of many factors, a model is introduced by Van Lint et al. (2008) which represents the interplay between demand and supply fluctuations in Figure 2. The changes in the distribution of travel times over the week, for a stretch of freeway, are also described by Van Lint and Van Zuylen (2005) and a graphical overview is given in Figure 1. These distributions of the travel time are based on a trajectory algorithm (van Lint & Van der Zijpp, 2003) using data from inductive detector loops in the road surface. The authors define the reliability of a travel time prediction as the width and skew of the travel time distribution. The wider the distribution, the less reliable a prediction obviously is. However, they found that especially the skew of the distribution has a substantial economic effect. *As in some peak periods the 5% most 'unlucky drivers' incur almost 5 times as much delay as the 50% most fortunate travelers.* In their 2005 paper some metrics are introduced to quantify the reliability as a function of the skew of the distribution.

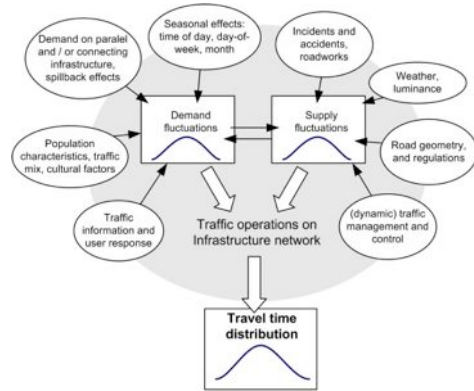


Figure 2: Schematic overview (not exhaustive) of factors influencing the distribution of travel times (van Lint et al., 2008)

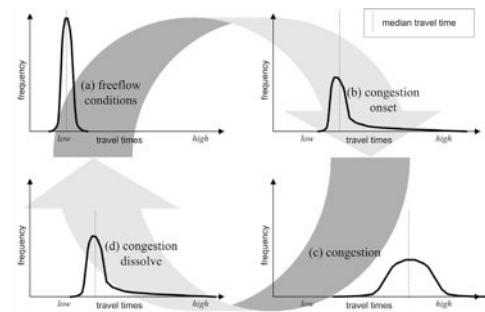


Figure 1: Shape of the day-to-day travel time distribution from free to congested conditions (van Lint & van Zuylen, 2005)

3. Problem description and methodology

3.1 Background

Planning software relies on an Origin Destination matrix, filled with estimated travel times (TTs) between all connected locations. Unreliable or incorrect TT estimations would therefore yield unreliable or incorrect results. However, there is often a lack of data to formulate reliable TT estimations. In the current paradigm, estimations are often based on stationary measurements, observed TTs between fixed locations or a limited number of vehicle based measurements (Lin et al., 2005). Stationary measurement techniques, such as inductive loops and cameras, lack generalizability and accuracy, as it is economically infeasible to measure every road, between every intersection. This problem is most evident for non free-way roads, as these roads form a complex network with many interconnections. Measuring TTs between two fixed locations yields reliable results, but these results cannot be generalized over a larger set of locations. Car based measurements or *car floating data*, in which a car equipped with GPS receiver and data logger, yield the most reliable results, but in most cases the number of probe vehicles is relatively low and the time frame over which observations are available is very limited. Because of the difficulty of obtaining sufficient and accurate data that covers the entire road network, and the lack of generalizability of stationary measurement techniques, the formulation of accurate and reliable OD-matrices, for large sets of locations, is very hard. Therefore there exists a need for a better source of data, which both covers the entire road network and yields a sufficient number of observations, over a sufficient long time period.

Secondly, average TT estimations do not automatically result in more reliable solutions for logistical models. The definition of reliability of TT approximations heavily depends on the context and application, a long haul transporter for which fuel costs should be minimized will have a very different perspective on reliability than an urgent courier. Also, the distributions are potentially heavily skewed (van Lint & van Zuylen, 2005), this implicates theoretically that, using only estimated average travel times, the travel time of an unlucky driver could be located very far from the estimated value. A potential wide and skewed distribution therefore results in very unreliable travel time predictions. An example is given in Figure 4, Figure 4in which the expected arrival time for route 1 (red) is earlier than for route 2 (blue), if the goal is to minimize the yearly average travel time, route 1 would probably be the best choice. But, route 1 has a relative wide and positive skewed distribution, which results in a larger volatility of the travel time per individual trip. Therefore, if a reliable route is defined as the fastest route for 95% of the trips, route 2 would be optimal.

Using the same theoretical background, choosing another departure time based on estimated average travel times, does not include the width and skew. Imagine a narrow distributed TT (free flow) when departing, for

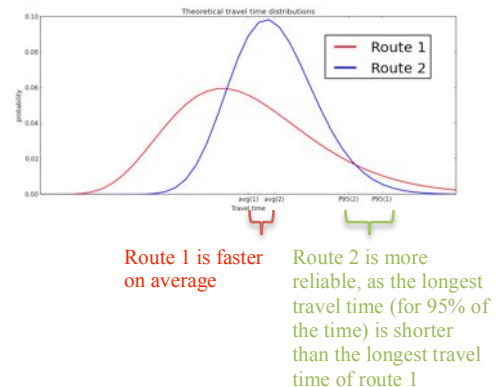


Figure 4: Comparison between the travel time distributions of two routes with differently shaped distributions

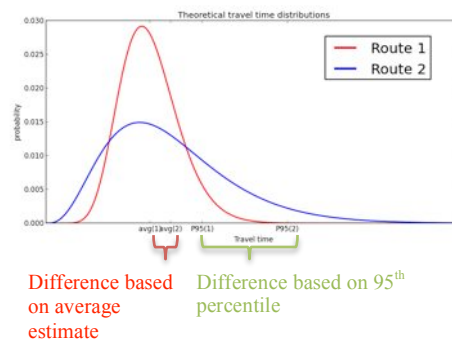


Figure 4: Comparison between the travel time distributions of a route for two different departure times

example, at 7am and a strongly positive skewed distribution at 8am, see also Figure 4. Although the estimated averages might only indicate a relative small difference, there are many individual occasions in which the later departure causes large delays. Therefore comparing the 95th percentile will, for many applications, result in a more reliable prediction.

From the theoretical description, it follows that neither the average, median nor the variance of the distribution alone, provide sufficient information as the distribution is potentially skewed. Most existing sources of dynamic TTs only incorporate a discrete value per time interval, completely omitting the distribution's shape. Research in which the shape of the TT distribution is incorporated suffers from a lack of data, for example the papers by Van Lint (2003), (2005) and (2008) are only based on a stretch of freeway. Local roads however, might have a totally different profiles as these roads include traffic lights, pedestrian crossings and other disturbances.

Although logistical models are widely discussed in literature, and dynamic travel times are now frequently incorporated in these models, the effects of the travel time distributions have not been intensively investigated. Most research efforts agree that people are risk averse and therefore often prefer routes with a higher mean travel time and small variability over a route with a lower mean travel time but higher variability (Bogers & Van Zuylen, 2004). For example for the VRP, a custom defined level of reliability, might result in different combinations of commodities and demand locations, which results in a longer average TT, but with lower volatility.

3.2 Using TomTom's big data

The solution for the data problem is found by using the *floating car data* gathered by the Dutch company TomTom. TomTom, founded in 1991, has a history of selling navigation products and services since 2001, they introduced their HD Traffic product line in 2007. Since this introduction, users can approve for their travelling data to be shared with TomTom. TomTom has obtained a very large pool of users over the years and this has led to an enormous amount of data which, at the moment, comprises over 4 trillion anonymous, consumer-driven, GPS based and map matched measurements, which are geographically dispersed over the total road network. The data sources are graphically presented in Figure 5.



Figure 5: Different sources contributing to TomTom historic traffic database.

This data opens an opportunity to investigate whether it is possible to solve the problems described before. Therefore the main research question for this report is *whether planning tools, the VRP in particular, can be improved on accuracy and reliability, by using detailed velocity distributions to model departure time depend traveling times.*

The first step was analyzing how the best predictions for future travel times per single road segment, $TT_{pred}^{id}(t)$, could be made using TomTom's historic data, from now on referred to as historic data. Let $TT_{pred}^{id}(t)$ denote the predicted stochastic travel time over edge id at time t . The second step was to summate the predicted distributions per segment into routes to yield the travel time predictions per route $TT_{pred}^R(t)$. The predictions for single segments and routes are tested against observation from a control set, $TT_{control}$, and against estimations from other

sources, $TT_{[source]}$ to determine the reliability and performance of the prediction. Last, it is analyzed whether these predictions can be used to find better solutions for logistical problems. The estimated costs, $E(C_{new})$, using the new predictor variables are compared with the costs using traditional methods $E(C_{[source]})$. This methodology is depicted in Figure 7.

This research follows a phenomenological approach, however, some factors influencing the TT can be controlled for. Therefore the scope of this research, concerning exogenous influences is limited to three factors. This results in a simplified version of the model as introduced by Van Lint et al. (2008), see Figure 2, and is depicted in Figure 6.

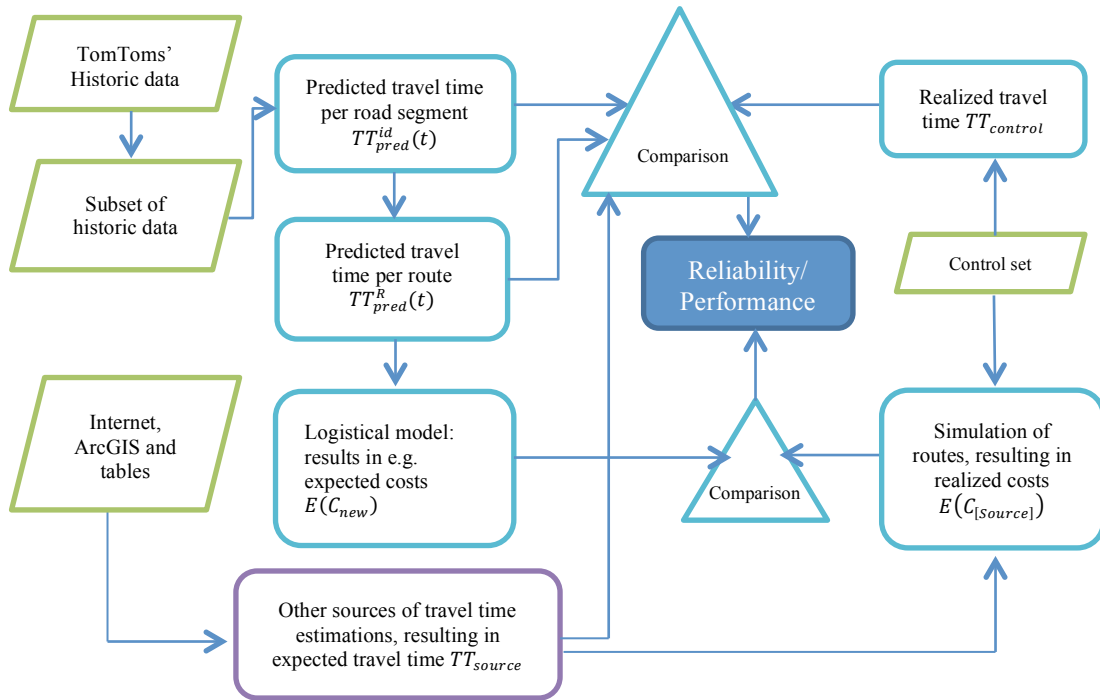


Figure 7: Graphical representation of research methodology

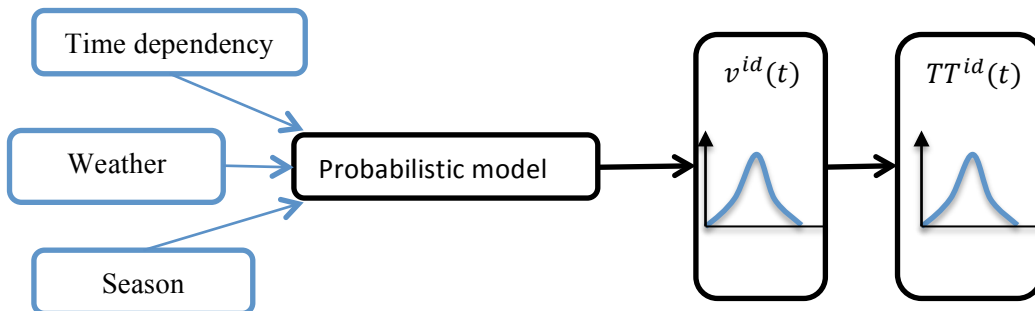


Figure 6: Scope of research concerning factors influencing the distribution of travel times, based on Van Lint et al. (2008)

4. Solution method

4.1 Dataset

The dataset used for this report, which is only a small subset of the total database, contained over 3.2×10^9 individual observations, gathered in a period of just over 5 years for metropolitan area of Eindhoven, a city in the south of the Netherlands. The roads have all been divided into small segments with a unique identifier. The characteristics for one single segment, such as road type, road geometry and speed limit, are constant. Due to the relative small length per segment, it is assumed a speed observation can be generalized over the whole segment. This enables the calculation of the travel times over each segment.

The data source in this report originates from TomTom devices and software, which record *traces* from users that enable information sharing. One trace consists of successive location samples, with a speed attribute to every sample. Precision of GPS devices has its limitations, which cause difficulties matching a trace to a road when roads are located close together. Secondly, although the devices and software are *car centric*, which means they are primarily intended for motorized use, the devices are also used by, for example, cyclist and pedestrians. Therefore TomTom uses a map-matching algorithm to only include measurements that can be reliably matched to a road user on a certain road segment.

The data set contains data from 2006 until October 2012. But for the early years, the number of observations is limited and data is fragmented, therefore no data before 2008 is used in the analysis. The number of observations for the last weeks of 2012 is also limited, as users have to connect their devices to the Internet before any data is shared. Therefore no data later than 2012 week 35 is used in the analysis. In order to verify the results found in this research, the dataset is split in two. The latest data is used as control data. This cut-off date is set to week 36 of 2011, this leaves one year of data for the control set. This is graphically represented in Figure 8.

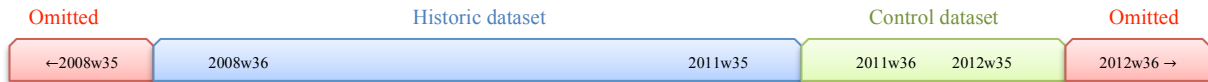


Figure 8: Division of data into Historic and control dataset

Due to the size of the historic dataset and the limited computational power and time available, only a subset of roads has been analyzed. This subset comprises both domestic as an industrial area and includes a stretch of free-way. The resulting dataset contains single velocity observations, for every observation the following information is used in the analysis:

```
[segment_id[-], length[m], frc[-], speed_limit[kph], speed[kph], timestamp[ms]]
```

Each road segment is unidirectional, for a bidirectional road, two segment ids in either direction are used. The functional road class (frc) indicates the type of road, e.g. 0: highway, 2: major road and <2: Local roads, the full list is given in appendix X. The timestamp is the UNIX time, which are the number of seconds that elapsed after January 1st, 1970, recorded in UTC. Last, the length, speed and speed limit are trivial.

In order to perform the analysis, a random sample of observations is drawn from the datasets. To analyze the predictions, the *control sample set* is introduced, which comprises 14,000 randomly selected observations from the Control dataset. Secondly the historic sample set is introduced, which contained over 10,000 observations from the historic dataset.

The length for every road segment is given and there exists an inverse linear relationship between the driving speed and travel time. Because roads of similar road classes have similar legal driving speeds but various lengths, the majority of the analysis is based on velocities, rather than travel times. However analyzing driving speed observations is analogue to analyzing travel times.

4.2 Notations

From the dataset, distributions of the travel speed per road segment can be deduced. Road segments are identified by a unique number, let id denote that unique number and let set I denote the list of segments with $I = \{id_1, \dots, id_q\}$ and $|I| = q$ the total number of segments in the set. The subset used in this report then denoted by I_s with $I_s \subseteq I$. Secondly, let $V_h^{id}(t_1, t_2), \forall id \in I_s$ denote the set of all historic observations taken over period t_1 to t_2 with $t_1 < t_2$, and let $v_h^{id}(t), \forall id \in I_s$ denote a single historic observation with $t_1 \leq t \leq t_2$, then $v_h^{id}(t) \in V_h^{id}(t_1, t_2), \forall id \in I_s$. Next $V_h^{id}(t_1, t_2), \forall id \in I_s$ is divided into smaller subsets containing only the measurements that fall within a certain time of day (TOD) and day of the week (DOW). Let K denote the whole set of intervals, $K = \{k_1, \dots, k_{max}\}$, then k_i describes the unique interval number. Typically, 12 blocks an hour, 24 hours a day and 7 days a week are used thus $1 \leq k \leq 2016 (= k_{max})$. The subsets are then denoted by $V_h^{id^k}(t_1, t_2), \forall k \in K, id \in I$ with,

$$\bigcup_{k=1}^{k_{max}} V_h^{id^k}(t_1, t_2) = V_h^{id}(t_1, t_2), \forall k \in K, \forall id \in I$$

For the remainder, the sets always contain data for an individual segment, therefore the id -label is dropped, resulting in $V_h^k(t_1, t_2)$.

An observation of a vehicle traveling faster than the legal speed limit indicates that *at least* the legal speed limit is possible. Because it would be unethical to predict or promote illegal velocities, all observations are topped to +15% of the legal speed limit. The 15% is chosen to include some margin, as there will also be drivers whose velocity is below the legal speed limit while a free flow is possible. This correction yields the *legal speed corrected dataset*, which only contains observation denoted as $v_{hc}^{id}(t)$. These observations are thus derived from the original observations according to the following equation:

$$v_{hc}^{id}(t) = \begin{cases} \min(v_h^{id}(t), 1.15 \cdot v_{legal\ max}^{id}) & \text{for } rc \in \{0,1,2\}, \quad \forall k \in K \\ \min(v_h^{id}(t), v_{legal\ max}^{id}) & \text{for } frc \notin \{0,1,2\}, \quad \forall k \in K \end{cases}$$

The same notations as for the original sets are followed and thereby the id label is dropped and the sets are split per time interval k , the corrected sets are then denoted by then $V_{hc}^k(t_1, t_2)$. Also the algorithm to adjust the sample size is similar for the corrected sets, resulting in $\tilde{V}_{hc}^k(t_1, t_2), \forall k \in K$.

In order to fit a distribution to the subsets, the number of observations within the subset should at least be 30, therefore a *sample size correction set*,

$$n = 1$$

$$\tilde{V}_{hc}^k(t_1, t_2) = V_{hc}^k(t_1, t_2)$$

$$\text{while } |\tilde{V}_{hc}^k(t_1, t_2)| \leq 30, \forall k \in K:$$

$$k^+ = \begin{cases} k + n & \text{if } k + n \leq \max \\ (k + n) - \max & \text{if } k + n > \max \end{cases}$$

$$k^- = \begin{cases} k - n & \text{if } k - n \geq 0 \\ \max + (k - n) & \text{if } k - n < 0 \end{cases}$$

$$\tilde{V}_{hc}^k(t_1, t_2) += V_{hc}^{k^+}(t_1, t_2) + V_{hc}^{k^-}(t_1, t_2)$$

Algorithm 1: Creation of sample size corrected set

$\tilde{V}_{hc}^k(t_1, t_2)$ is introduced, with $|\tilde{V}_{hc}^k(t_1, t_2)| \geq 30$. If the set $V_{hc}^k(t_1, t_2)$ does not contain the minimal amount of observations, the observations from adjacent subsets are added. Because the Sunday night and Monday morning are also adjacent subsets, there is a correction if $k < 0$ or $k > max$, this results in Algorithm 1. To each set $\tilde{V}_{hc}^k(t_1, t_2)$ a distribution can be fitted. This yields the random variable $v^k(t_1, t_2)$, which is always based on the *sample size corrected* and *legal speed corrected* datasets.

From the analysis later in this paper, it followed that the Gamma distribution gave the best representation of the travel, and as the distribution of $v^k(t_1, t_2)$ depends on the time window, the distribution can be defined shape and scale parameters of the Gamma distribution (α and β):

$$v^k(t_1, t_2) \sim \Gamma(\tilde{\alpha}_h^k(t_1, t_2), \tilde{\beta}_h^k(t_1, t_2)), \quad \forall k \in K$$

For the control set, the same notation is followed where $V_c^{id}(t_3, t_4), \forall id \in I$ denotes this set of control observations taken from a period t_3 to t_4 with $t_1 < t < t_3 < t_4$, for all segments in I . A single observation is denoted by $v_c^{id}(t), t_3 \leq t \leq t_4, \forall id \in I$, thus $v_c^{id}(t) \in V_c^{id}(t_3, t_4), \forall id \in I$. Again the edge identifier is dropped from the notation and the sets are split up in weekly intervals, resulting in observations $v_c^k(t) \in V_c^k(t_3, t_4), \forall k \in K$. Similar to the historic data, the data in the control set is topped to 115% of the legal speed limit. Thus:

$$v_{cc}^{id}(t) = \min(v_c^{id}(t), 1.15 \cdot v_{legal\ max}^{id}), \forall v_{cc}^{id}(t) \in V_{cc}^{id}(t_3, t_4)$$

If only the unrestricted observations are important is that sample observations are **not** included in the set of historic observations, thus $v_c^k(t) \notin V_h^k(t_1, t_2), \forall v_c^k(t) \in V_c^k(t_3, t_4), \forall id \in I$.

4.3 Determination of data set

If the dataset contains data from adjacent weeks, the dataset can be identified with the values for t_1 and t_2 . But when the dataset contains not-adjacent periods of time, these periods can be identified by vectors \bar{t}_1 and \bar{t}_2 , for which $\bar{t}_1 = \{t_1^I, t_1^{II}, \dots, t_1^i\}^T$ and $\bar{t}_2 = \{t_2^I, t_2^{II}, \dots, t_2^i\}^T$. The i periods are then given by $t_1^I \rightarrow t_2^I$ and further. Possible datasets can contain a number weeks of the most recent historic data, data from the same periods over the last years or data from similar week numbers. The latter is not used in this report as the influence of one single week is very large and errors, due to occasional situations, are not averaged out. In the future this approach might be plausible if more years of data is available.

The prediction for a Monday, week 28 in 2012 might for example be based on the data from: (a) 52 weeks of most recent available data, on (b) data from 3 'summers' or on (c) on data from 3 'weeks 28' only (these weekly datasets are not used in this report). The data examples are depicted in figure X.

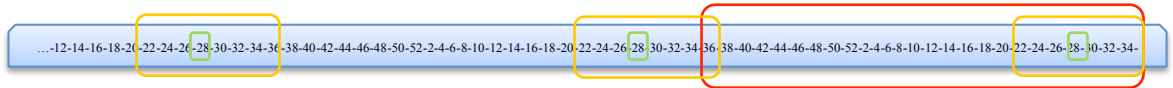


Figure 9: Examples of possible datasets

4.4 The relation between velocity and travel time

As stated before, there exists an inverse linear relation between the velocity and travel time according to $TT^k(t) = \frac{s \cdot 0.06}{v^k(t)} \left[\frac{[m] \cdot 60}{[kph] \cdot 1000} \right]$, with s the length of the road segment. This implicates however that the TT distribution is the inversed distribution of the velocity, which, for the

Gamma distribution, results in the inverse Gamma distribution. Let X denote a stochastic variable for which $X \sim \Gamma(\alpha, \beta)$ then, as the inverse of X is given as $Y = \frac{1}{X}$ for which $Y \sim \Gamma^{-1}\left(\alpha, \frac{1}{\beta}\right)$. A linear product of the inverse Gamma distribution and a constant k is given as $\frac{k}{X} = k \cdot Y$ with $k \cdot Y \sim \Gamma^{-1}\left(\alpha, \frac{k}{\beta}\right)$. Thus a stochastic TT prediction for a road segment, based on historic data, can be denoted as $TT_h^{id^k}(\bar{t}_1, \bar{t}_2)$. If the *id* label is again dropped, this results for every edge in $TT_h^k(\bar{t}_1, \bar{t}_2) = \frac{s \cdot 0.06}{v^k(\bar{t}_1, \bar{t}_2)}$, thus:

$$TT_h^k(\bar{t}_1, \bar{t}_2) \sim \Gamma^{-1}\left(\tilde{\alpha}_h^k(t_1, t_2), \frac{s}{\tilde{\beta}_h^k(t_1, t_2)}\right), \quad \forall k \in K$$

As routes are a concatenation of road segments, the speed distributions for all individual segments are estimated and the length per segment is known, the distribution of the travel time for a route can be calculated as the convolution of its segments' inverse distributions. The sum of inverse Gamma distributed variables however, has no exact solution and there exists, no easy approximation (Witkovsky, 2001). However, using numerical calculations, it is known that the convolution of inverse Gamma distributed variables resembles again an inverse Gamma distribution. Therefore, in this report, a simulation approach was used. The travel time per route was generated 1000 times as the sum of $|R|$ inverse Gamma-random generated variables, with $|R|$ the number of road segments in a route. Over this dataset an inverse Gamma distribution was fitted and the resulting parameters were used to define the distribution for that route. An example is given in Figure 10: Sum of simulated travel times and the Inverse Gamma distribution fit. Figure 10, where the red line indicates the fitted distribution.

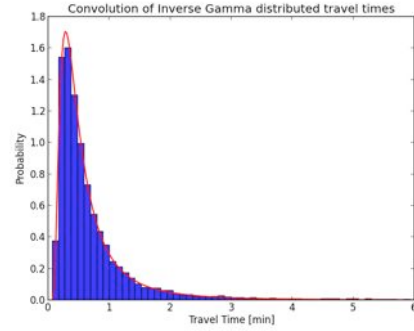


Figure 10: Sum of simulated travel times and the Inverse Gamma distribution fit

4.5 Comparison between estimated and realized velocity

The next step was to determine the distribution of $v^k(\bar{t}_1, \bar{t}_2)$, for different values of \bar{t}_1 and \bar{t}_2 and test these distributions against observations found in the control set. The expected value for the Gamma distribution by $E\left(v^k(\bar{t}_1, \bar{t}_2)\right) = \tilde{\alpha}_h^k(\bar{t}_1, \bar{t}_2) \cdot \tilde{\beta}_h^k(\bar{t}_1, \bar{t}_2)$. The error of the estimation is then defined as the difference between a randomly picked observation from the control set and the predicted mean velocity:

$$e = v_c^k(t) - E\left(v^k(\bar{t}_1, \bar{t}_2)\right)$$

From the predicted distributions of the travel speed, also the percentile scores can be calculated, which predict that $p\%$, $0 \leq p \leq 100$, of the observations will lay within a certain range. If the distributions reliably describe the future observations, $p\%$ of the control observations should actually be located in the $p\%$ confidence interval. The confidence intervals are characterized by an upper and lower bound, $\langle v^-(t, p), v^+(t, p) \rangle$, with

$$P\left(v^-(t, p) \leq v(t) \leq v^+(t, p)\right) = P\left(E(v(t)) \leq v(t) \leq v^+(t, p)\right) = 0.5 \cdot p$$

4.6 Route simulation using sample observations

For a single road segment, the travel time simulation is simply a single observation from the control dataset. But as a route consists of many segments, multiple observations need to be concatenated. Therefore an algorithm was developed that finds a string of observations, which sum represents a fictive vehicle driving that particular route. First the list $R = \{id_1, id_2, \dots, id_{last}\}$ is introduced which contains all road segments in the appropriate order and let n denote the n^{th} segment in list R . Then t_{start}^n denotes the start on which the fictive vehicle starts driving on segment n and $TT^n(t_{start}^n)$ denotes the corresponding travel time. Then the value for the start time at the next edge would be $t_{start}^{n+1} = t_{start}^n + TT^n \pm \Delta t$. A relaxation, $\pm \Delta t$, is introduced because the probability that, for all segments in the route, an observations for id_n is found that exactly matches t_{start}^n , is very small and decreases with the length of the route. Δt is a user defined relaxation parameter, the lower the value of Δt , the more veracious the TT.

The algorithm starts with a user defined route R , value Δt and an initial start time t_{start}^1 . It searches for possible candidate observation that satisfy $t_{start}^1 \pm \Delta t$ and randomly pick one of these observations. Secondly TT^1 is calculated using the segments' length and the next candidate observations are searched for, which have to satisfy $t_{start}^2 = t_{start}^1 + TT^1 \pm \Delta t$. This process continuous until observations for all segments are found, or no observations are found for one of the segments. In the latter case, the algorithm starts again with $(t_{start}^1)_{new} = t_{start}^1 + \Delta t$

4.7 Processing the data

The amount of data required smart processing of the data in order to limit IO-actions (reading writing to a hard disc), to limit total disk space and to keep the amount of RAM within its limits. The process was tweaked to work on a MacBook Pro with a 2.2 GHz Intel Core i7 Processor, 8Gb 1333MHz DDR3 memory, a 750 GB 5400 rpm hard disc and a 120 GB solid state hard disc. PYTHON v2.7 is used as programming language and the GZIP algorithms are used to compress data.

The data is processed in several steps, these steps are depicted in Figure 11. First the data was split into smaller files containing only the observations for one particular week. For each observation only the important data was stored, which were the segment's identifier, velocity and the time of the observation. The observations were not sorted but appended to the end of each file to limit IO-actions. The files were also compressed to save disk space and all other variables were stored in a legend file, including the other characteristics such as the length, the legal speed limit and the functional road class.

In the next step the data was processed per week and sorted per segment, only weeks in the period of 2008 week 36 until 2012 week 35, and segments within the subset were stored. Again, each line of data is appended to the end of each file, limiting the IO-actions. The SSD hard disk was used because of the large number of files. The third step is processing the data per individual segment, such as fitting distributions to the data. The amount of data can now be processed and stored in the computer's RAM, which makes allocation of data, such as sorting, possible. The output is stored in files on the hard disc. The data is processed using dynamic tuples, for which many standard actions are available in the PYTHON language.

For all time or date related actions, the UNIX time stamp is converted into a tuple containing the year, month, day, weekday, hour and minute. Because the time is recorded in GMT, the time should be corrected to the Dutch time zone. A module returning the exact *time delta* is used to ensure accurate transition to the Dutch Time zone. As this time delta is either 1 hour during the

Using Travel Time Predictions based on TomTom's Big Data in Logistical Models

winter, or 2 hours during the summer, and the transition between standard and daylight-saving-time depend on the year, the module uses a database to determine the exact time delta.

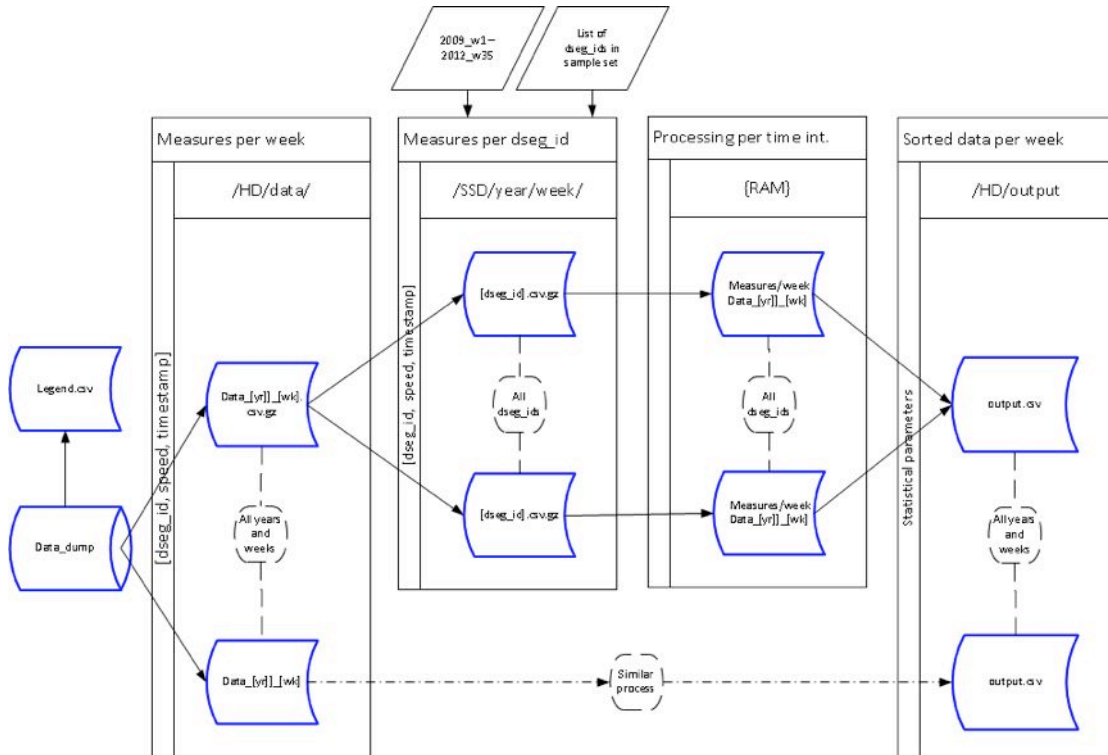


Figure 11: Graphical representation of the data process

5. Analysis

5.1 Number of observations

The dataset contained over 3.2×10^9 individual measurements, divided over roughly 350,000 individual edges. The distribution of the data is depicted in **Error! Reference source not found.**, which depicts (a) the number of observations per TOD ($N=10^6$), (b) per DOW ($N=10^6$) or (c) per week ($N=2.2e9$). There are clearly more observations during daytime, and the majority of observations are taken during the afternoon or evening. The weekend days are also well represented compared to weekdays, although traffic density during the weekend is often less heavy. The weekly number of observations is relative constant, except for ISO weeks 1, 52 or 53. This is due to the fact that these weeks not always contain 7 days.

These properties of the data have two reasons, (1) as a portion of the road users is using a PND, the number of observations will increase with the total number of vehicles on the road. And (2), people tend to make more use of their PNDs when travelling to unfamiliar locations. According to (Wessels, 2012) and (Mulder, 2011), home-work (or reversed) trips are underrepresented in the TomTom data, trips to non-home or non-work locations occur more frequently in the afternoon and weekends. Therefore the assumption is made that the number of observations does not fully correspond with the flow density on the roads.

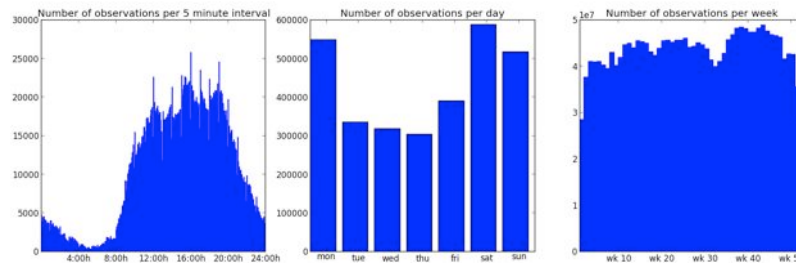


Figure 12: Distribution of observations a) over the day (TOD), b) per day of the week (DOW) and c) per week of the year

5.2 Distribution of the travel time

The travel time distributions for one randomly chosen road segment per frc are plotted in Figure 13. The histograms resemble the distribution found and explained by (van Lint et al., 2008). All distributions have a long positive tail, suggesting that the distributions, on general, are heavily skewed. The red lines are inverse Gamma distributions, which are fitted to the data. The fitted distributions seem to follow the histograms quite good, the theoretical background of using the inverse Gamma is discussed later in this report. The distribution for frc 7 however, which represents the smallest roads, is an outsider that does not seem to follow any known distribution. This is most probably due to the nature of these roads, as these are often very short and are often only used by residents, also, only a

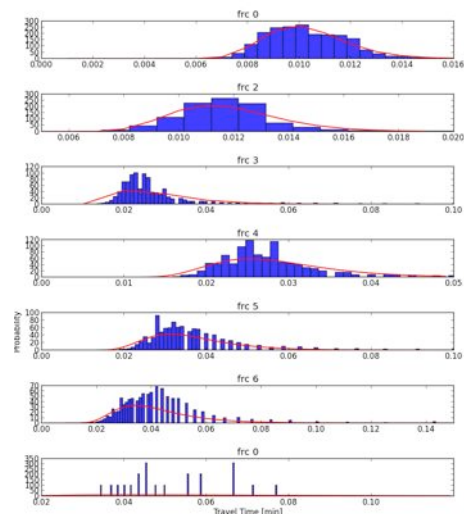


Figure 13: The TT distributions for a randomly chosen road segment per frc and the fitted inverse Gamma distribution

very limited number of observations is available. Contrary to what Figure 13 suggests, not all roads follow a nicely shaped distribution, many road segments seem to be a combination of multiple distributions, as can be seen in Figure X. This behavior can be explained by the characteristics of these road types. On some roads drivers might come across traffic lights, pedestrian crossings or, for example, playing children. However, these obstructions are often of a temporary but recurrent nature. When a driver does not have to stop or slow down, this results in the *unrestricted flow*. If a driver however has to stop or slow down, this results in the *restricted flow*. This is discussed in more detail in Chapter 5.3.

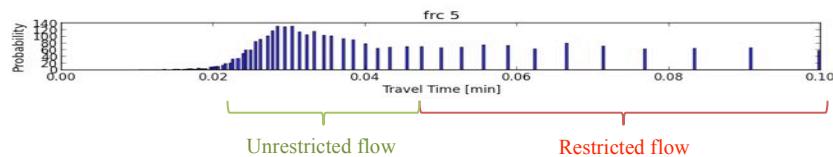


Figure 14: Distribution of a road segment in the effects of both the restricted and unrestricted flows are visible.

5.3 Distribution of velocities

The velocity, rather than the TT, is analyzed because the TTs are dependent on the various lengths of the segments. Therefore the velocities of road segments from the same frc are easier to compare and velocity measures are also easier to interpret than an arbitrary TT. The first step is to verify whether the velocities from the historic observations follow a known distribution. This should be a continuous distribution, which is able to describe data with different expected values and standard deviations. Looking at the histograms in **Error! Reference source not found.**, in which a selection of one randomly selected segment per frc¹ is presented, using the data from 2010, the data appears to be bell-shaped with a long tail and, for some of the local roads. For frc 3, the histogram shows a spike at low velocities and a long tail towards higher speeds.

Distributions that could fit this type of data are for example the Normal, Logistic or (family of) Gamma distribution. As the velocity is defined as a random variable and the distribution of some road segments appears to follow the well-known bell-shape, the Normal distribution is included in the analysis. The positively skewed distributions for the smaller roads, with lower speed, might indicate that a skewed distribution better represents these observations, therefore the Gamma distribution is also analyzed. Other distributions might be usable as well, but it is assumed that either the Gamma or Normal distribution are well suited to describe the distribution of the driving speed observations. For example, the logistic distribution was found to be narrower but cannot represent a skewed distributed variable. Also the other distributions from the Gamma-family, such as the Exponential ($\alpha = 1$), Erlang ($\alpha \in \mathbb{N}$) or Chi-squared can be expressed as a special case of the Gamma distribution itself and are therefore omitted.

Theoretically, the differences between the Normal and Gamma distributions should, according to the Central Limit Theorem, diminish for higher expected values. For the local roads, with lower average velocities, the differences between the Gamma en Normal distribution are more apparent. Moreover, the Gamma distribution satisfies the non-negative property, which holds that for X with $X \sim \Gamma(\alpha, \beta)$ that $P(X < 0) = 0$. In the case of driving velocities this is a very welcome

¹ There are no roads with frc 1 in the data set for Eindhoven and surroundings.

property, as the velocity is measured as a positive variable only. Consequently, both the Normal and Gamma distributions are used for further analysis.

For local roads, frc 3, 6 and 7, no known distribution describes the shape of the histogram. This is obvious caused by a large number of observations below 10 [kph]. These slow observations, below 10 [kph] were visible for more local roads and can be explained the by the restricted flow, as introduced in Chapter 5.2. In Figure 15 the histograms for the higher frc's are plotted with (blue) and without (red) observations below 10 [kph]. Removing the low velocities resulted in a much better fit for the Gamma distribution on the unrestricted data.

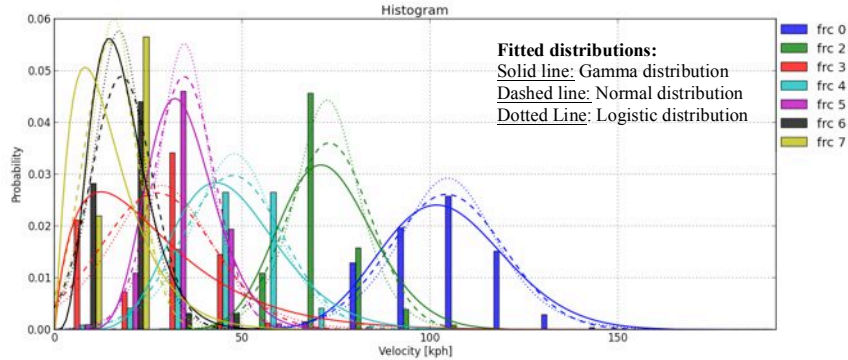


Figure 16: Distributions for a selection of road segments, one per frc, and the fit of the Gamma, Normal and Logistic regression.

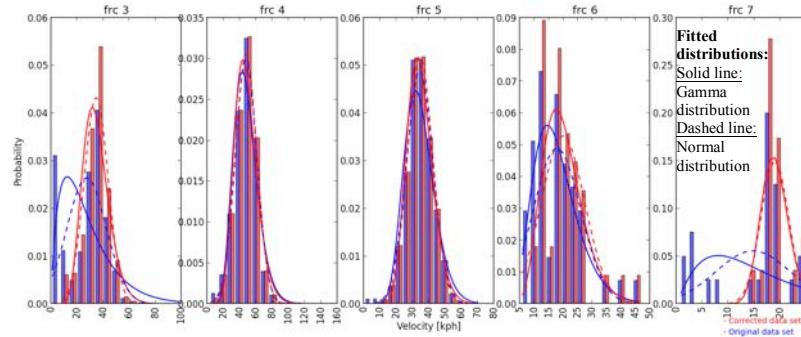
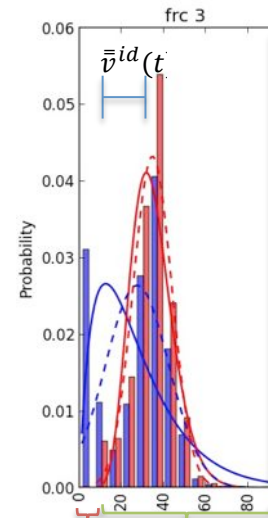


Figure 15: The distributions for the higher frc's including (blue) and excluding (red) the observations below 10 [kph]

5.4 Including stops in the travel speed estimations

In the analysis it was found that for some roads the velocity either follows an unrestricted distribution, or there exists a restriction and the velocity is very low. Let $p_{restricted}^{id}$ denote the probability a vehicle has to stop or significantly slow down on road segment id . For one single road segment, this results a velocity that either follows the restricted distribution, or the unrestricted distribution. The overall average driving $\bar{v}^{id}(t)$ will be a velocity between the mean restricted and unrestricted speed, but the probability this exact driving speed is encountered is very small. Let $v_r^k(t_1, t_2)$ denote the restricted velocity and $v_{u,r}^k(t_1, t_2)$ the unrestricted velocity, then $v^k(\bar{t}_1, \bar{t}_2)$ can be characterized by the probability tree depicted in Figure 18.



Restricted flow Unrestricted flow
Figure 17: Restricted and unrestricted velocity observations

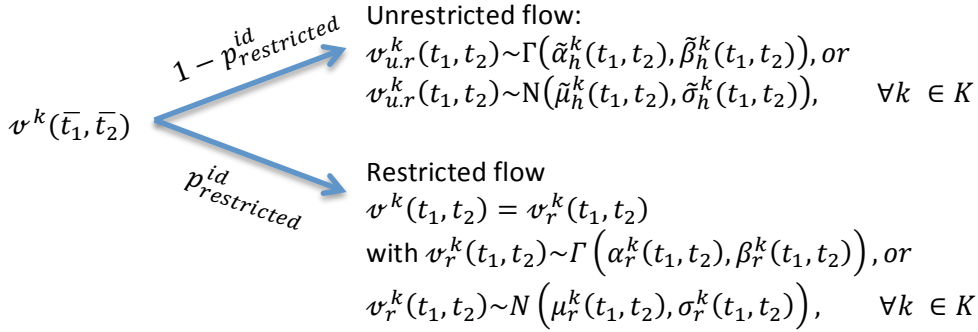


Figure 18: Probability tree of the restricted and unrestricted velocity

The shape of the distribution, for a road segment that contains both restricted and unrestricted velocities, might seem problematic. For a route however, there is a probability that one has to stop on any of the edges, but the exact stopping location is not that important. Thus, as long as the effect of each potential stop is taken in to account the TT-prediction for a route is still reliable. Therefore the distribution for each road segment can be defined as the weighted sum of the restricted and unrestricted velocity. The weighted convolution, $Y = \sum_{i=1}^k w_i \cdot X_i$ with w_i the corresponding weights, for Normally distributed variables is given by a mixture distribution, for the Gamma distribution no (exact) solution for the weighted convolution is known. An exact solution in the situation $\beta_i = \beta, \forall i$ is described in (Di Salvo, 2006) and this appears to be a Gamma-like distribution.

Because of the complex nature of the convolutions, it was first tested which distribution best described the unrestricted velocities, this part of the analysis is presented in Chapter 6. The best distribution appeared to be the Gamma distribution and also the restricted velocities can be expressed as a Gamma distribution, with a small value for the scale parameter. Therefore the convolution will be expressed as the weighted sum of Gamma distributed variables, which resembles again a Gamma distribution. As no exact solution to this convolution exists, a simulation approach could be used to estimate the parameters. However, this would yield the same result as if a Gamma distribution was directly fitted to the dataset containing both the restricted and unrestricted velocities. Altogether it is therefore assumed that, within a route, a Gamma distribution fitted over the entire dataset yields the desired distribution, with the requirement that the unrestricted velocities are well represented by a Gamma distribution.

In the remainder of this report, the datasets will be identified as *full dataset* if all data, both from restricted and unrestricted observations are used, and *unrestricted dataset* if all observations below 10 [kph] are omitted.

5.5 Observed driving speed versus year and time of the year

A graphical representation of the data without corrections, for the same selected road segment as in Figure 16, is presented in Figure 19. Per graph, the boxplots for one segment is displayed, representing data from 2009, data from 2010 or data from 2011. Some differences in medians and box sizes are found, thus the observed velocities are not constant over the years. The boxplots in Figure 20 represent the same road segments, but this time the data is divided in four different times of the year (TOY). The year is divided into (1) summer, (2) spring, (3) fall and (4) winter and these periods roughly correspond with the seasons, the see **Error! Reference source not found.**

for exact definitions. In Figure 20, also some differences are seen between the boxplots, although less than the intra-year differences.

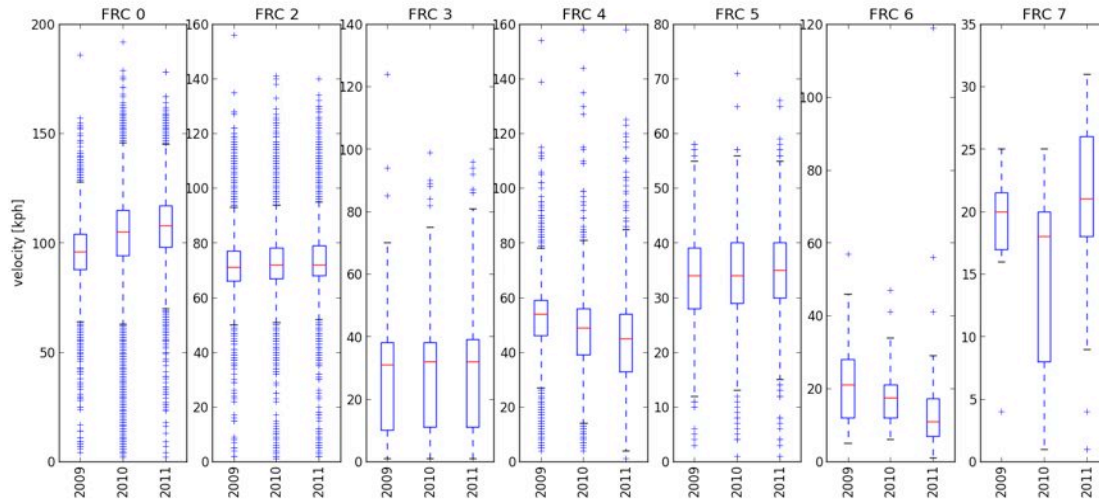


Figure 19: Boxplot of the data for a selection of road segments, divided per year

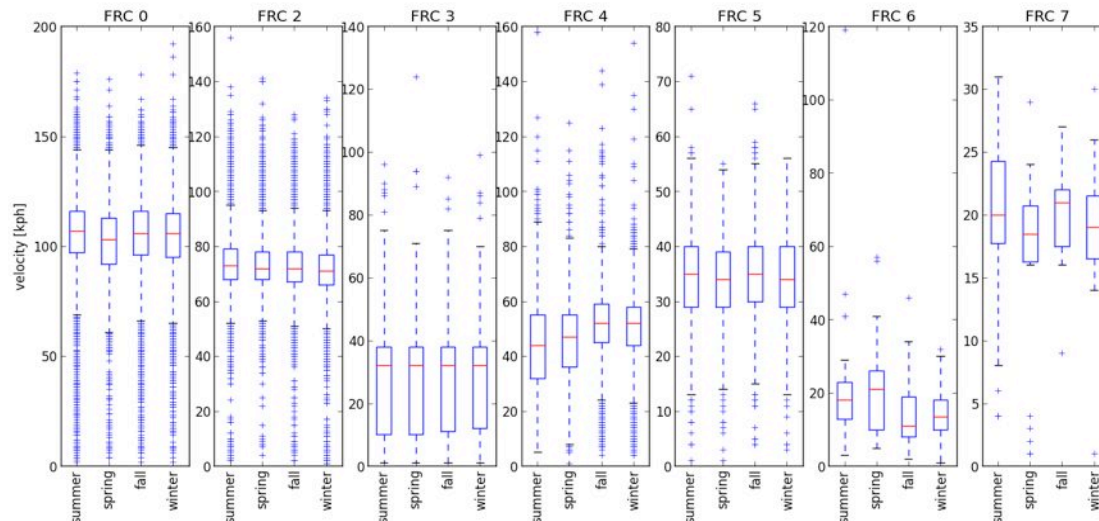


Figure 20: Boxplot of the data for a selection of road segments, divided per season

These findings are also tested for a larger set of observations. Therefore the historic sample set was analyzed for correlations between the velocity and time related variables. The time related variables are the timestamp, year and TOY. In contrast to what the boxplots showed, the resulting correlations indicate that the TOY has a significant predictive value over the driving speed for frc 0, 2, 3, 4 and 5 ($.001 \leq \alpha \leq .041 < .05$). However, the coefficients indicate that effects are limited, as only about 4% of the variance in the driving speed is explained by the TOY ($r \leq 0.2$). It was also found that freeways and major roads were negatively influenced by the TOY, thus the driving speeds show a minor, but significant, decrease when the TOY-number increases. Striking though is that the local roads have a minor, but again significant, *positive* correlation with the time of year, indicating that speeds increase with the TOY-number. This opposite effect might be explained by the fact that the population of road users on these roads are less homogenous, which means that the population comprises both motorized and non-motorized road users, such as

cyclists and pedestrians. During the summer or spring the number of non-motorized road users will probably increase, which in turn attenuates the velocity. Last, the speeds on the smallest local roads, $frc > 5$, appear not to be influenced by the time of year.

In the examples for segments in the boxplots, differences between years were found. For the whole dataset, these differences are only significant for the highways, $frc 0$ ($\alpha = .000 < .05$) and the larger local roads, $frc 4$ ($\alpha = .007 < .05$). The positive correlation coefficient ($r \approx 0.2$) indicates that the speed slowly increased over the years on the highways, which might be caused by improvements made to the road network. For the local roads the speeds slowly decreased ($r \approx -0.1$) over the years. This might be caused by traffic calming and media attention to drive carefully on local roads. The exact causes are however outside the scope of this report.

Future velocities might thus by seasonal and in- or decrease over the years. Therefore travel time predictions based on different datasets, defined by \bar{t}_1 and \bar{t}_2 , are used for further analysis. It should be analyzed whether it is beneficial split the data per TOY to incorporate seasonality. Secondly the effect of the time period, from which the data is sourced, on the reliability of the velocity predictions should be analyzed.

5.6 Driving speed versus time of the day

The differences and distributions of the data over the day is of great importance for this paper, as it underpins the necessity of dynamic travel times. Figure 21 and Figure 22 depict the data over 2010, for an edge of $frc 0$, for different times of the day. From Figure 21, it follows that for a particular road segment, the velocities are not constant over the day, also, the number of outliers increases with the number of observations. From Figure 22 it appears that the subsets can still be approximated by a Gamma or Normal distribution. The correlation between the hour and speed is also tested for the historic sample set, to investigate whether the velocity is time dependent in general. Because the speed will not have a linear correlation to the variable 'hour', a dummy variable is introduced, see **Error! Reference source not found.** The dummy variable is expected to roughly indicate hours of decreasing speed. For $frc 0, 2$ and 4 , a significant correlation is found ($.000 \leq \alpha \leq .006 < .05$) between the dummy variable and velocity. The relation is a negative one, which, in line with the expectations, indicates that the driving speeds decrease for a higher value of the dummy variable. However the correlation coefficients are very small and for the other road classes, there is no significant difference between the velocities for the different times of the day.

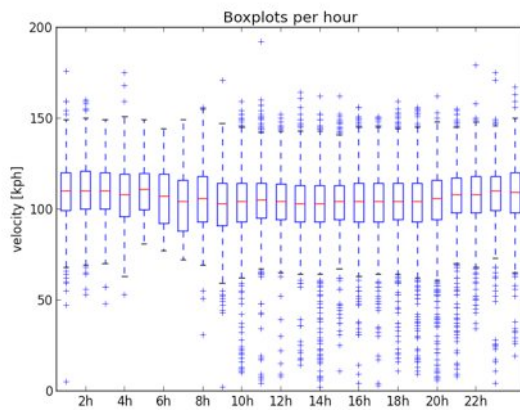


Figure 21: Boxplots of the data, divided per hour, for an edge of $frc 0$

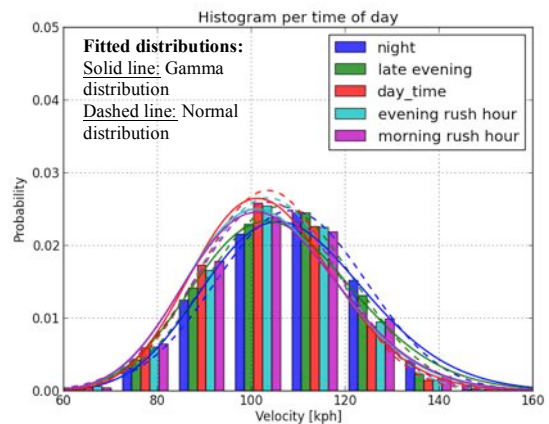



Figure 22: Histograms for the distributions for different times of the day

In spite of the fact that the results from this analysis do not yield conclusive evidence that the velocities are dynamic, the assumptions is that travel time should be modeled as a dynamic variable. The reason for these results might be that congestion occurs at different times for different days. There are probably also differences between roads. Therefore the time dependent behavior of the velocity is further analyzed in this report.

Dummy Variable for Time of Day				
Time window	Start [hour]	End [hour]	Period	Expected driving velocity
1	22	5:59 (next day)	Night	
2	19	21:59	Late evening	
3	10	15:59	Day-time	
4	16	18:59	Evening rush hour	
5	6	9:59	Morning rush hour	

'Start' and 'End' hours are found using trail-and-error to optimize correlation
Table 1: Definitions of Time of day (TOD)

5.7 Weather and driving speed

One of the factors expected to have its influence on the driving speed, and for which historic data is available, is the weather. From the Dutch weather institute KNMI, a dataset was downloaded, containing the following measures from weather station Eindhoven:

- Wind speed (daily average [0.1 m/s])
- Temperature (daily average [0.1 dgr Celcius])
- Min ground temperature (height 10 cm)
- Sunshine (daily sum, [0.1 hours])
- Rain (daily sum [0.1 mm])
- Minimal visability {0,...,89} (from 100 meters until >70km)
- Clouds {1,...,9} (from fully visible to fully invisible sky)
- Relative Humidity (daily average [%])

The correlation between the velocity and the daily weather variables was tested for the historic sample set and it was found that for frc 0-4, the average and minimal temperature had a significant correlation ($.001 \leq \alpha \leq .038 < .05$) with the driving speed, although the effects were very limited ($r < 0.1$). Also the humidity seems to have its influence, despite the fact that the humidity did not correlate significantly with the visibility. On the whole, these results indicate that it might be beneficial to correct the data for weather influences.

Of course it can be expected that some combination of weather variables might have an even greater influence on the velocity, for example the combination of freezing temperatures and rain, which might result in snow or icy roads. Therefore these two variables are combined into a binary variable and the correlation between this variable and the velocity is tested. The results were only significant for the larger roads, frc 0 and 2 ($.000 \leq \alpha \leq .002 < 0.05$), but the effects were limited, as only 1% of the variance was explained by this variable.

$$Slippery_roads = \begin{cases} 1 & \text{if } avg_temp < 3 \text{ and } rain > 20\% \\ 0 & \text{else} \end{cases}$$

5.8 Weather and the time of year

Because weather predictions are another field of research, including these forecast in driving speed estimation is not feasible for long-term predictions. But the weather roughly follows the local climate, which pattern is a yearly cycle. Therefore the TOY might be a predictor for the average weather and could be used to correct for some of the variance caused by the weather. As the average and minimal temperatures per week had the strongest correlation with the velocity,

the TOY is based upon chronologically successive weeks, which average and minimal temperatures fall within a certain range.

Both the week and the time of year might therefore have a strong correlation with the weather. If true, the week number or the time of year could be used to incorporate the influence of the weather in the model. Because the ISO week numbers are based on the calendar, not on the weather, this numbering is not useful. And thus the weeks are renumbered based on the (1) average temperature, (2) the average minimal temperature and (3) the absolute minimum temperature. The correlations between these renumbered weeks. The correlations between the TOY and weather variables were tested. Significant result were found between the TOY variables and the all weather-variables ($\alpha < .05$), except between for the rain.

Time of the year			
Time window	Start t_1	End t_2	Period
1	week 21	week 37	Summer
2	week 11	week 20	Spring
3	week 38	week 48	Fall
4	week 1	week 10	Winter
	week 49	week 53	

Table 2: Definitions of Time of Year (TOY)

The combination of the findings that the weather has its influence on the velocity, and the weather is correlated with the TOY, might partially explain the seasonality found in Chapter 5.5. Altogether this strengthens the argument to analyze the predictive power of data split per TOY.

6. Analysis of unrestricted velocities per road segment

6.1 Period with best predictive power over future traveling speed

Different sets of data, represented by \bar{t}_1 and \bar{t}_2 , resulted in different estimations for the velocity. In order to compare the predictive power of these different datasets, the errors of the estimated value, $e^k(t) = v_c^k(t) - E(v^k(\bar{t}_1, \bar{t}_2))$ and Mean Square Error (MSE) are calculated for the all observations in the control sample set. The errors for both the Normal and Gamma distributions were very similar, this can be explained by the fact that the average velocity for many roads is larger than 30 [kph] and it was already established that the Gamma distribution resembles the Normal distributions for mean values above 30.

First, the overall characteristics of the error distribution was investigated, a negative or positive mean of the error distribution indicates a biased estimator and, in practical sense, means that the driving speed is respectively over- or underestimated. The standard deviation indicates the width of the error distribution, a narrower width indicates that the predictions are more reliable, in the sense that it is more likely that a single observation of the driving speed equals the expected value. The errors are normally distributed, except for the errors of frc 7. An example of these errors, using a data set containing 2 years of recent data, is given in Figure 23.

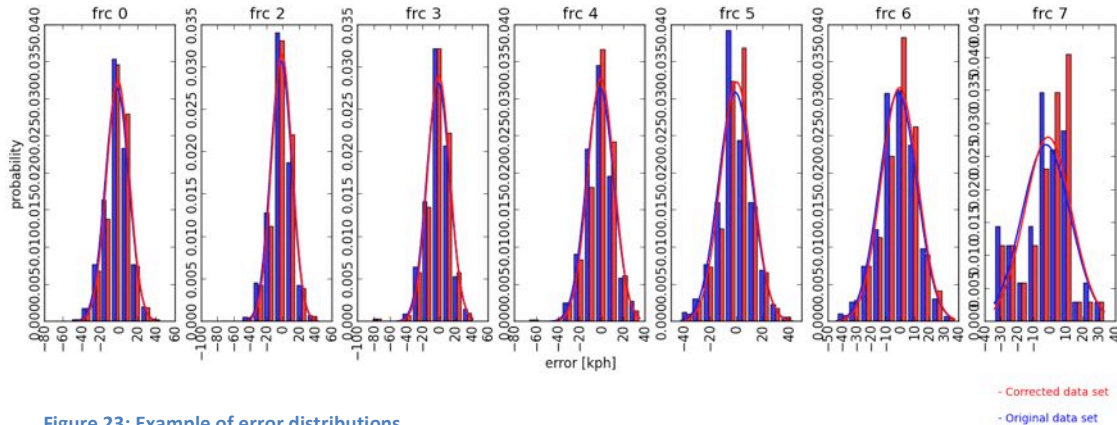


Figure 23: Example of error distributions

The next step was comparing the results between datasets. The resulting error distributions for the Gamma distribution are given in Appendix 2, the results for the Normal distribution were similar for the decimal places presented in the table. It is clear that the use of split data yields a much better result than using recent data, as the average error, MSE and standard deviation are lower. This smaller average error is most probably caused by the build-in seasonal and weather-corrections. For the datasets containing recent data, the set containing two years of data obviously performs best. The explanation for the lesser performance when using 13 or 26 weeks of data is that these datasets only contain observations from the summer period, as the recent period ends at 2011 week 35. Therefore the predictions based on these datasets overestimate the velocities for the major roads in the wintertime. As a result it is chosen no longer to use the datasets containing only 13 or 26 weeks of recent data in this report.

The average errors for the split are relative close together, although the speed corrected, unrestricted dataset seemed to perform best. This is of course a logical result from the fact that the

unrestricted sample dataset itself only contains unrestricted observations. Surprisingly, the dataset containing both the restricted and unrestricted observations, resulted in the lowest overall MSE and standard deviation. Altogether, the data set containing 2 years' worth of restricted and unrestricted split data has the best performance based on the error analysis.

Last, it was also checked whether the errors were time related, as that would mean the prediction for a certain TOY, time of the week or TOD should be adjusted. The results, using 2 years of split data, are depicted in Figure 24. The error does not seem to be time related, although the line for frc 7 is whimsical, but this is due to the low sample size for this class. The same effect is noticeable in the early morning, as only limited observations are made during these hours. On average, the absolute error equals about 10 [kph]. Similar graphs have been made based on other datasets, which resulted in similar outcomes.

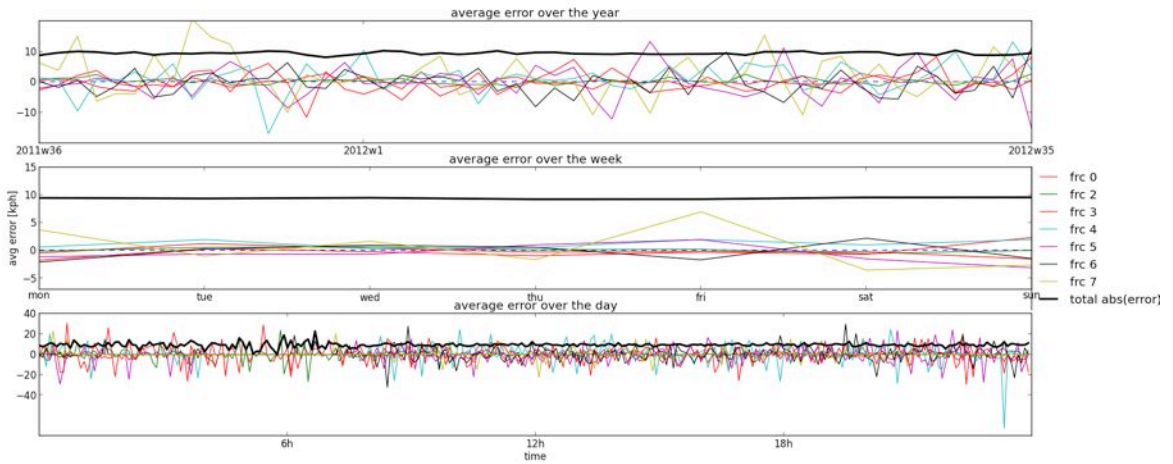


Figure 24: Errors over time, a) error over the year, b) error over the week and c) error over the day

6.2 Reliability of predicted percentile scores

From the predicted distributions of the velocity, confidence intervals can be calculated, which predict that $p\%$ of the observations will lay within a certain range. If the distributions reliably describe the future observations, the predicted and actual percentiles should match. As the Normal probability density function is symmetric, the observations should be symmetrically distributed above and below the mean. The Gamma distribution however, can be positively skewed, which indicates that the number of observations above the mean can transcend the number of observations below the mean. This results in a range, characterized by $(v^-(t, p), v^+(t, p))$, with

$$P(v^-(t, p) \leq v(t) \leq E(v(t))) = P(E(v(t)) \leq v(t) \leq v^+(t, p)) = 0.5 \cdot p$$

A graphical representation for a road segment of frc 0 and 4, are given in Figure 25. In these figures the velocity is plotted against the TOD, for a Tuesday and Saturday. The predictions are either based on the Normal, or the Gamma distribution, as is indicated on the left. The dark lines represent the expected velocities based on 52 weeks of most recent corrected data, the light green area should contain 50% of the observations, the light blue area 95%. A random selection of observations from the control set are colored green or orange when they are located within the 50% or 95% area, or red when located outside both areas. The differences between the Gamma and Normal distributions are very small for the example of frc 0, as was expected as the average

speed vastly exceeds the 30 [kph]. For the example of frc 4 however, the larger tail of the Gamma distribution clearly captures more observations within the 95% range.

The accuracy of the confidence intervals is also analyzed for the entire control sample set. For each observation in this set, it is determined in which confidence interval the observation is located. The distribution that best predicts the confidence intervals is preferred, as that indicates that the distribution reliably predicts the shape of the future velocity. The results are given in Appendix 3 and are clearly in favor of the 1 and 2 years of recent data, containing both restricted and unrestricted observations. The confidence intervals from the other distributions seem to be too narrow.

The results from the analysis point out two potential datasets that perform best. The uncorrected split data, containing both restricted and unrestricted observations and the most recent data containing 2 years of data. The former dataset results in estimations with a small average error and the smallest error distribution, and thus the predicted values are the most accurate. The latter dataset however, best describes the shape of the future velocity distribution. As the remainder of this report focuses on the effect of the distribution's width and skew of the TT distributions, it is chosen to use the full dataset containing 2 years of most recent data. The full dataset is used because possible stops should be incorporated in the TT predictions for routes. The average overestimation of the velocity found in Chapter 6.1 is a disadvantage of this dataset that is taken for granted. Last, the choice for the Gamma distribution is strongly supported by the fact that the Normal distribution was unable to describe the shape of future observations.

Using Travel Time Predictions based on TomTom's Big Data in Logistical Models

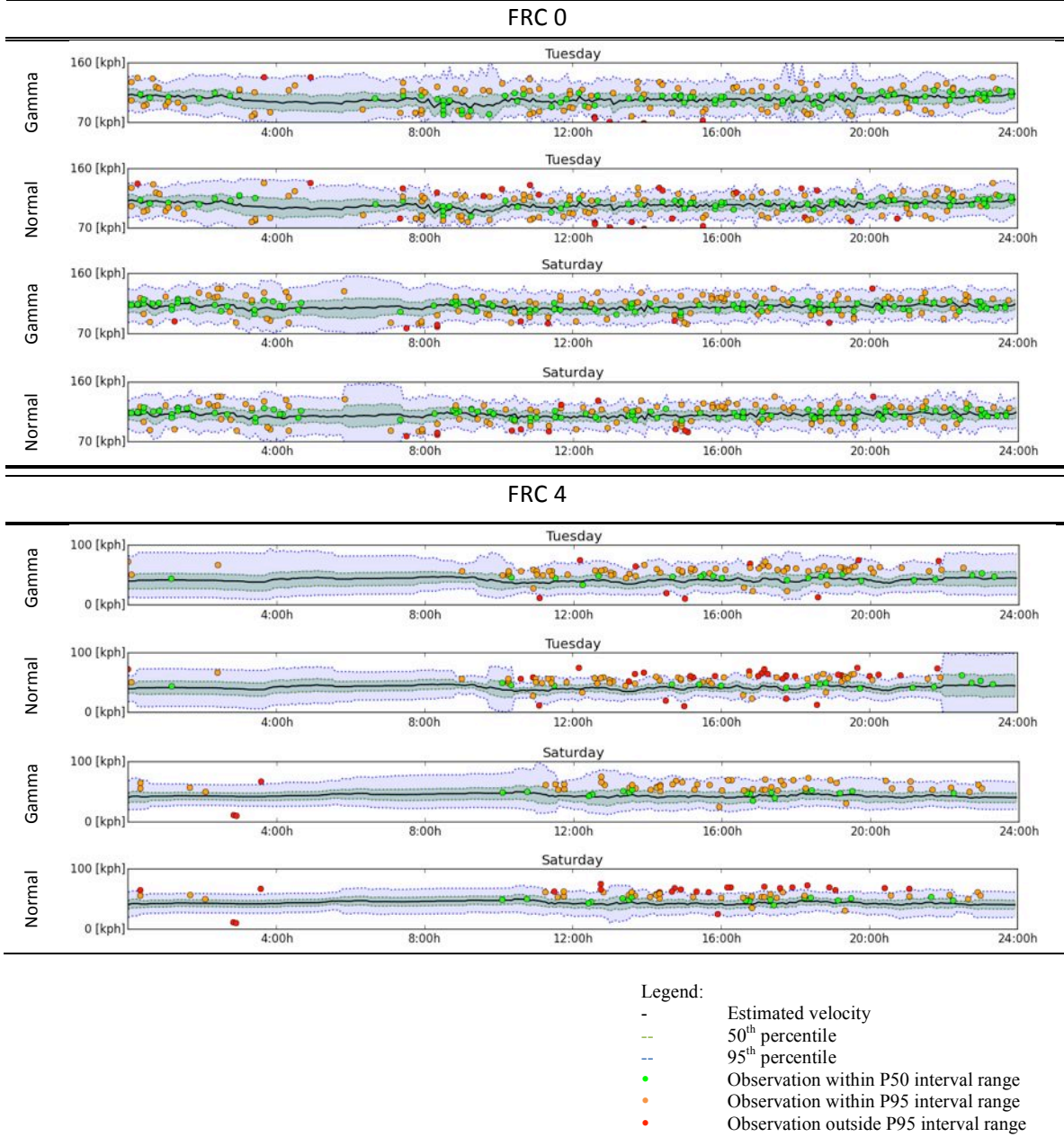


Figure 25: Graphical representation of velocity distributions for two road segments

7. Results for travel time predictions for routes

7.1 Prediction compared with simulated route

Using the results from the analysis per road segment, it is possible to calculate the dynamic travel time distributions for a route. As it was found that the driving speeds are very well represented by a Gamma distribution, it is implicated that the travel times follow an inverse-Gamma distribution. This is again in line with the findings by Van Lint et al. (2008), as the inverse Gamma distribution is positively skewed. This dynamic TT distribution can then be compared to observed travel times for these routes. The results of a dynamic travel time distribution for a randomly chosen, non existing, route, for two randomly picked days (15 and 16 May 2012) are presented in figure X. The travel time is given as a function of the departure time. The observed travel times are generated using the simulation algorithm described in Chapter 4.6 and are colored to indicate whether the observation falls within the 50th percentile (green), 95th percentile (orange) or outside the 95th percentile (red). In order to find a sufficient number of simulated routes, an accuracy of $\pm 30min$ is used, which means that an observed traveling time for a segment should be within a 30-minute range from the actual time. The simulation algorithm found routes between 6 am and 10 pm for the specified dates.

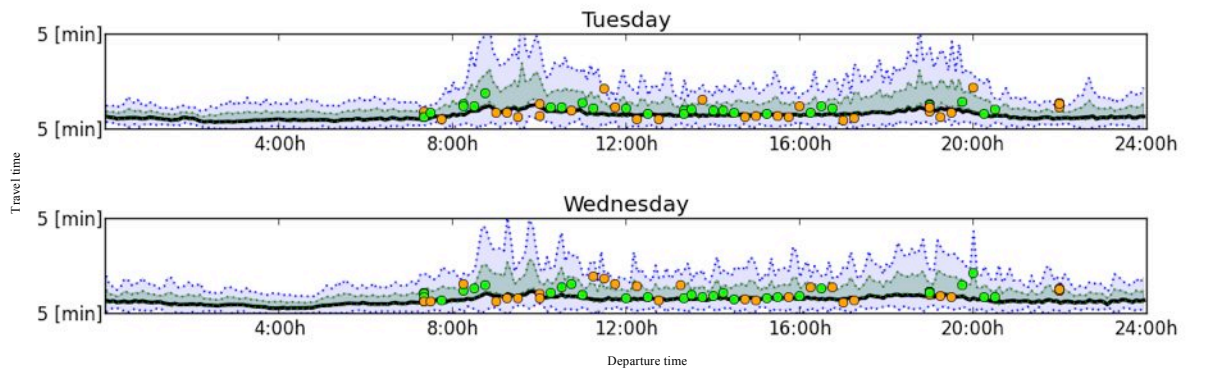


Figure 26: Graphical representation of the travel

Legend:

- Estimated Travel Time
- 50th percentile
- - - 95th percentile
- Observation within P50 interval range
- Observation within P95 interval range
- Observation outside P95 interval range

The expected travel time is higher during daytime and the longest travel times are, as expected, found during the morning and evening rush hours, these times are also far more volatile, as shaded area is wider. The positive skew is represented by the larger area above the expected value, in comparison to the area below the expected value. And, in line with what was expected from literature, the skew is much more severe during congested hours than during free flow hours.

7.2 Realistic routes

The simulated routes in chapter 7.1 only contained a limited number of edges, for which sufficient data per day was available to simulate the route for a specific day. However, realistic routes are longer and often include roads of different road classes, therefore two randomly chosen routes between two sets of addresses are created, the former route uses a large main road and the

latter stays within the city borders. Figure 27 depicts a small segment of the map of Eindhoven with Route 1 (red, 1 → 2) and Route 2 (blue, 3 → 4). The directions, distances and some static travel time estimations are given in Table 3.



Figure 27: Subsection of Eindhoven's roadmap

Distances:	Route 1	Route 2
Straight line	3,016 [m]	1,251 [m]
Over road	4830.7 [m]	2041.2 [m]
Postalcode start	5652XJ	5616KD
Postalcode finish	5655JW	5652NX

Static Travel time:	Route 1	Route 2
Google	8 [min]	6 [min]
TomTom (internet)*	7 [min]	6 [min]
TomTom (static)	5.512 [min]	2.517 [min]
Postalcode table	7 [min]	4 [min]

*middle of the day, no congestion

Route 1 Welschapsedijk – Warmelo (red) (65 edges)	
Directions:	Distance [m]
Start at 1	
Go northeast on Welschapsedijk	64.7
Turn right on Noord Brabantlaan	234
Turn right onto ramp to N2	559.7
Bear right on N2	423.8
Bear right onto ramp to Meerveldhovenseweg	1877.3
Turn left on Meerveldhovenseweg	379.9
Turn right on Ulenpas	516
Turn right on Twickel	229.3
Turn left on Warmelo	477
Turn left to stay on Warmelo	67.8
Finish at 2	1.3

Route 2 Schoenerstraat - Vigliuslaan (blue) (43 edges)	
Directions:	Distance [m]
Start at 1	
Go west on Schoenerstraat toward Botterstraat	103.2 m
Turn right on Tjalkstraat	150.1 m
Turn left at Klipperstraat to stay on Tjalkstraat	85.1 m
Turn left on Strijpsestraat	140.6 m
Turn right on Beukenlaan	29.1 m
At fork keep left on Beukenlaan	112.4 m
At roundabout, take 1st exit to proceed north on Beukenlaan	448 m
Make sharp right	7.1 m
At fork keep left	112.8 m
Turn right on Noord Brabantlaan	622.8 m
Turn right on van der Muydenstraat	8.7 m
Turn left on Noord Brabantlaan	72.9 m
Turn right on Vigliuslaan	148.4 m
Finish at 2	

Table 3: a) Static travel time estimations, b) and c) the driving directions for Route 1 and 2

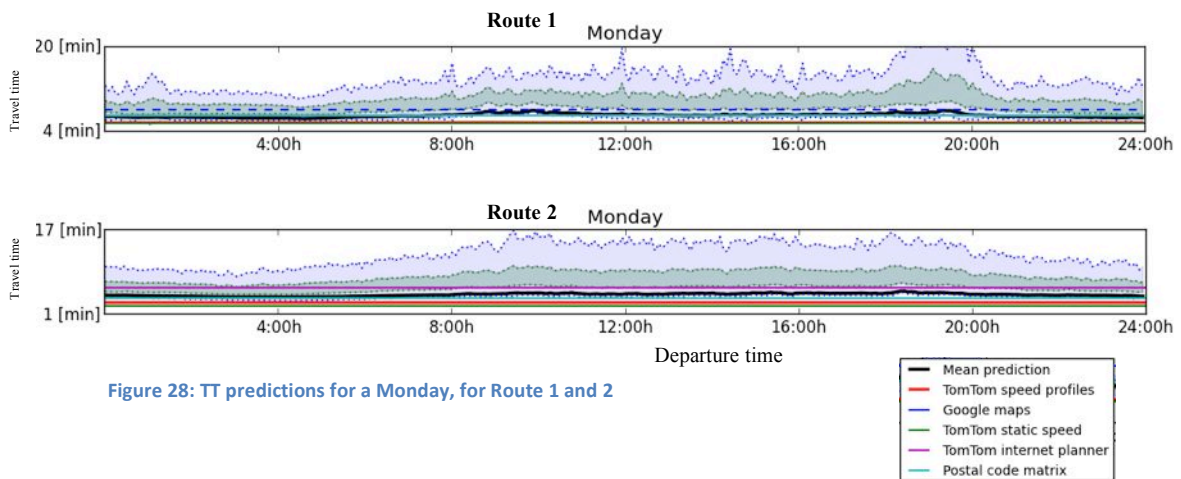


Figure 28: TT predictions for a Monday, for Route 1 and 2

Figure 28 depicts the travel time predictions for routes 1 and 2. The straight lines represent static travel time predictions based on two internet-planners: Google maps (blue) and TomToms' online route planner (Purple). The third prediction is based on TomTom's routable maps (Green), the travel times are calculated using the maximal legal speed and should theoretically identify the lower bound of all predictions. The last static prediction is based on a Postal Code Matrix² (Cyan). A dynamic travel time prediction, based on TomToms' speed profiles (Red) resulted in a virtual straight line, as speed profiles were only provided for a very small number of road segments.

For Route 1, the internet-planners' predictions either coincide with the expected peaks (google maps) or expected lower value (TomTom internet planner) of the expected prediction. As the TomTom internet-planner is set to 1 pm and congestion was not taken into account, the results are in line with the findings in this report. TomToms' static speed calculation does underestimate the travel time, as it does not incorporate stops and assumes that a vehicle always travels at the legal velocity.

For Route 2, which is a more inter-city route, the expected travel time shows a flatter profile. The 95th percentile shows that the distribution is much wider during daytime then during nighttime. However, the distribution seems to be relative flat over the day. The internet-planners from both Google and TomTom coincide and have a longer estimate travel time than the static and dynamic speeds obtained from TomToms' routable map. Both internet-planners seem to overestimate the travel time. And both estimations based on TomToms' routable map seem to underestimate the travel time.

On a whole, the estimated travel time, based on 2 years of recent data, is in line with the estimations from other sources, but the differences in the width and shape of the distributions are fully omitted by the other sources' predictions.

² <http://www.eenmanierom.nl/afstand-te-berekenen-in-kilometers-tussen-postcodes-voor-je-reiskostendeclaratie>

8. Implications for the Vehicle Routing Model

In this section of the report, it is investigated, for a small subsection of Eindhoven, whether information obtained from travel time distributions result in more reliable solutions for the VRP. First the vehicle routing problem (VRP), comprising 1 distribution center, 4 vehicles and 11 demand locations is introduced in Figure 29. Using ESRI ArcGIS, the VRP was solved such that all demand locations are assigned to exactly one vehicle and should be visited in a particular order. The resulting routes were optimized solely on time.

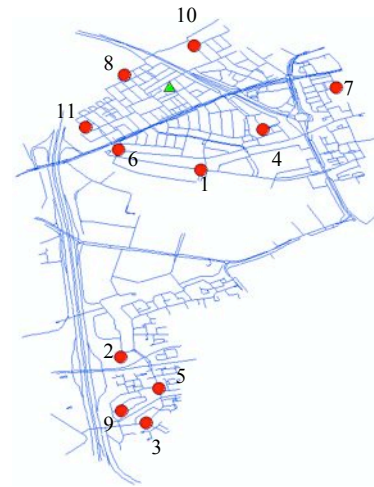


Figure 29: VRP for subsection of Eindhoven

As a reference, the VRP was solved using the legal speed limit as velocity per road segment. The resulting routes are depicted in Figure 30, including the order in which the locations should be visited. Then the VRP was solved using the expected travel times for a Monday at 7am, 8am, 9am, 10 am and 3pm. This resulted in small changes for the route, as well as small changes in the expected travel time per route. The only major change was a reversed sequence for demand locations 5, 3 and 10. It can therefore be concluded that time dependent travel times do have an effect on the solutions found by the VRP. An example of a change in route is given in Figure 31. Let the routes, which are the result of the VRP when the estimate velocities are used, be denoted as *average estimate centric routes for time x*.

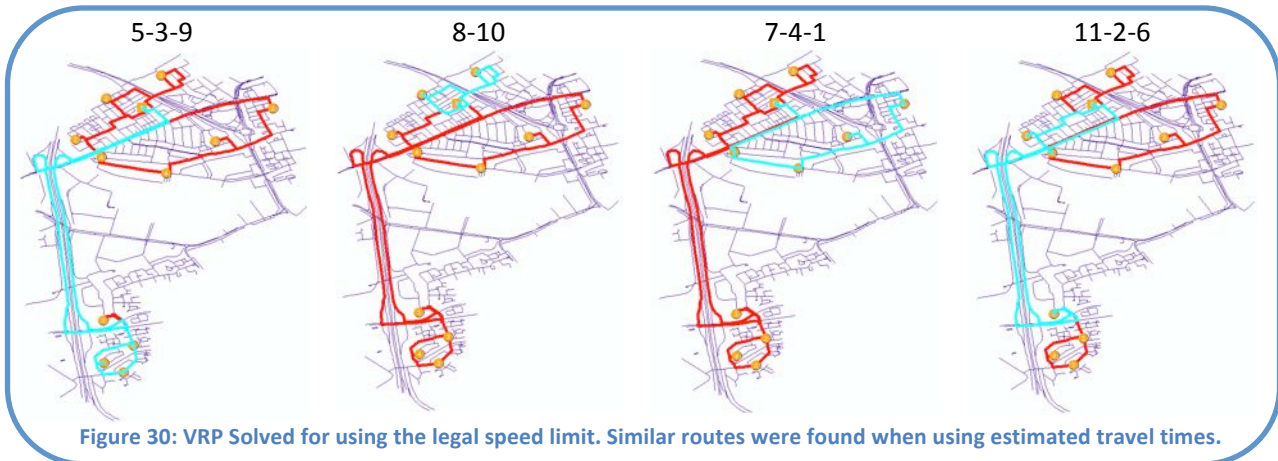


Figure 30: VRP Solved for using the legal speed limit. Similar routes were found when using estimated travel times.

It got more interesting when the TT estimations were based on the 95th percentile velocity. The allocation of the demand locations per route changed and a transition from the small local roads and freeway towards the main local roads was visible, see Figure 33. This indicates a that the smaller local roads have a relative higher penalty. Let these routes be denoted as the *p95 estimate centric routes for time x*.

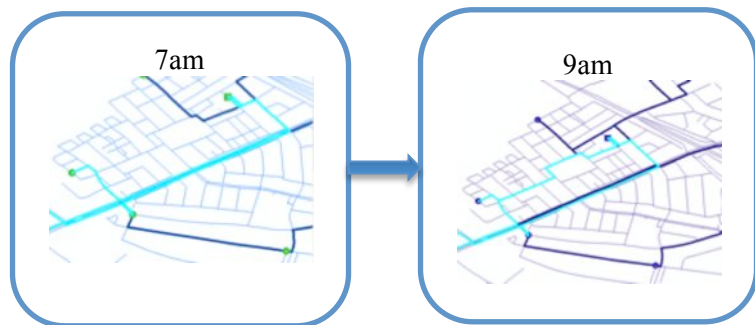
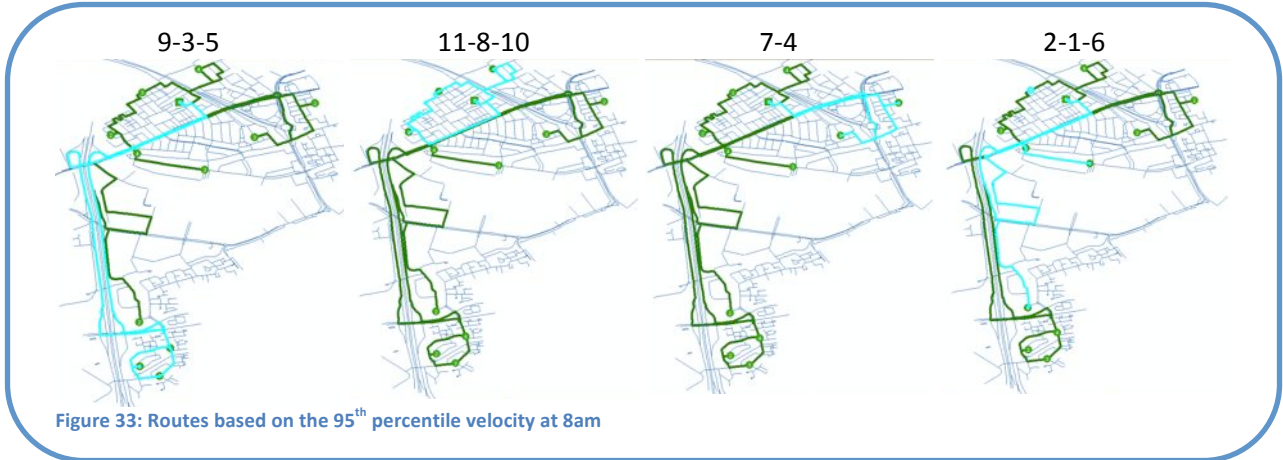
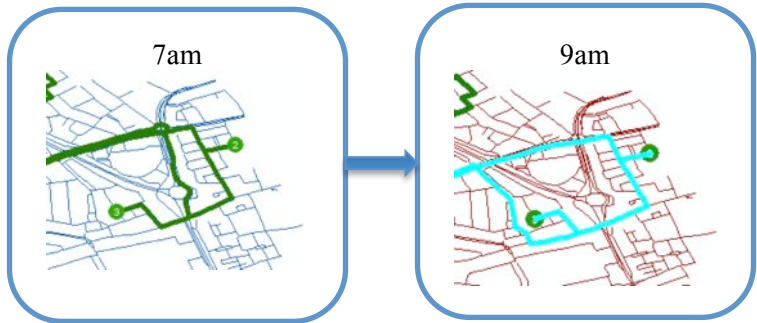


Figure 31: Example of a change in route between 8am and 9am, while using the estimated travel times as input.



The main question remains whether the *average estimate centric routes for time x* are indeed less reliable than the *p95 estimate centric routes for time x*. Therefore, two parts of the suggested routes are identified that have the same start and end point, see Figure 33. For these two routes the distributions is calculated for a Monday at 8am.



The resulting distributions are presented in Figure 35, and the characteristics are exactly as expected. The expected travel time is minimal for the *average estimate centric route* and the 95th percentile is minimized for the *p95 estimate centric route*. The distance between the expected value and 95 percentile score is an indication of the width of the distribution. Therefore it can be concluded that, at least for this example, the *p95 estimate centric route* resulted in a more reliable route. The differences are very small however, due to the short length of the routes.

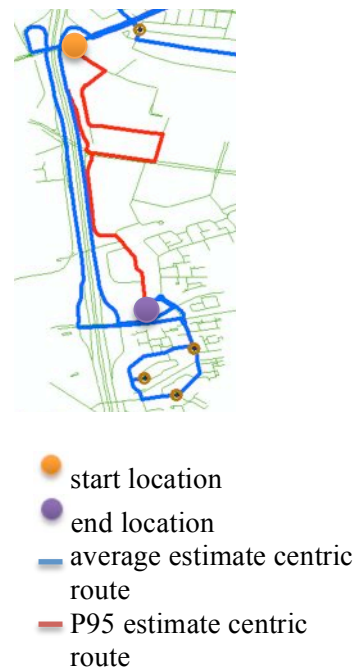
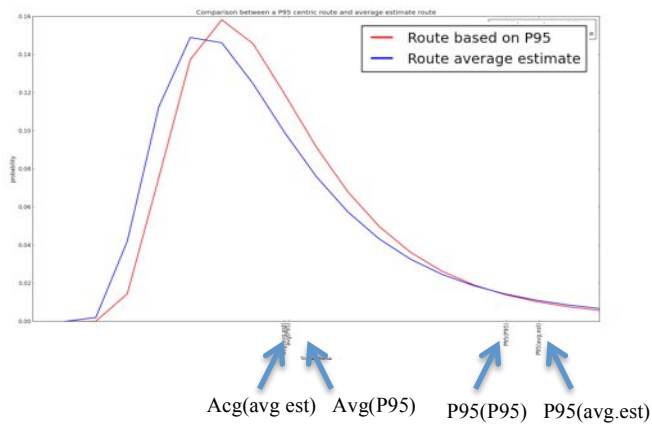


Figure 34: Two parts of a route with similar start and end point

Figure 35: Resulting distributions for the average estimate centric route and the p95 estimate centric route on a Monday at 8am

9. Conclusions and recommendations

As travel times are stochastic variables and the shape of their distributions changes over the day, the traditional discrete travel time prediction, either static or dynamic, does not provide sufficient information. Finding reliable data, which can be used to predict the future travel time distribution is however a bottleneck. In this report a dataset from TomTom is used and it is shown that this data has a good predictive power over the future distribution of the travel times. The predicted distributions have been tested against observations from a control set and proved to be accurate up to frc 6. This is an enormous improvement compared to other available data sources.

For the logistical models, the VRP in particular, it can be concluded that the use of, for example, the 95th percentile of the velocity distribution resulted in routes with a lower volatility. On the other hand, the average travel time increased. The tradeoff between reliability and the, on average, shortest route should be considered by the user. Therefore, the overall conclusion is that planning tools, the VRP in particular can benefit from velocity distributions.

However, a lot more research should be conducted to investigate the influence of the travel time distributions on the VRP. In this report only a few simple examples are presented. Secondly the data comprised only a very small subset of the total dataset, and although the general line of this report might be generalizable, the data and suggested applications should be investigated on a much larger scale. In this research, only the 95th percentile score is used, however, there might be another percentile score which has a much better balance between risk evasiveness and average travel time. Also, the observations marked as *future observations* were all made within a one year period since the end of the dataset used as historic data. It would be interesting to know if predictions over a longer time would still be reliable.

An important shortcoming during this report was a solution method that uses stochastic travel times directly as its input. The methods available were only able to solve the VRP with discrete variables. The devious method used in this report, using ESRI ArcGIS, was time consuming, took a lot of manual labor and was far from ideal. It is therefore recommended to use a specialized solution method in order to compare the effects of travel time distributions on a larger scale.

The last hurdle that had to be taken for this report was the amount of data. Now that an initial proof is provided, it is interesting to further analyze the potential of TomTom's Big Data. There will be better, more sophisticated analysis techniques, which can be used to further improve its predictive power or include more exogenous variables.

References

- Ahuja, R.K., Mehlhorn, K., Orlin, J.B. & Tarjan, R.E., 1990. Faster Algorithms for the Shortest Path Problem. *Journal of Association of Computing Machinery*, 37, pp.231-23.
- Andersen, T., Crainic, T.G. & Christiansen, M., 2009. Service network design with asset management: Formulations and comparative analysis. *Transportation Research Part C*, pp.197-207.
- Belfiore, P. & Yoshizaki, H.T.Y., 2009. Scatter search for a real-life heterogeneous fleet vehicle routing problem with time windows and split deliveries in Brazil. *European Journal of Operational Research*, 199, pp.750-58.
- Bemmelen, M.M.K.J.v., 2012. *Modeling and Solving a Real-life Load Building and Routing Problem in the Retail Industry*. Master thesis. Eindhoven: TU/e Technical University of Eindhoven.
- Bogers, E.A.I. & Van Zuylen, H.J., 2004. The importance of reliability: the Importance of reliability in route choice in freight transport for various actors on various levels. In *Proceedings of the European Transport Conference...*, 2004.
- Chien, S.I. & Kuchipudo, C.M., 2003. Dynamic Travel Time Prediction with Real-Time and Historic Data. *Journal of Transportation engineering*, 129(6), pp.608-16.
- Crainic, T.G., Gendreau, M. & Dejax, P., 1993. Dynamic and stochastic models for the allocation of empty containers. *Operations research*, 41(1), pp.102-26.
- Crainic, T.G. & Laporte, G., 1997. Planning models for freight transportation. *European Journal of Operational Research*, 97, pp.409-38.
- Desrosiers, J., Dumas, Y., Solomon, M.M. & Soumis, F., 1995. Time constrained routing and scheduling. In *Handbooks of OR & MS*. Elsevier Science B.V. pp.35-139.
- Di Salvo, F., 2006. The exact distribution of the Weighted Convolution of two. *Atti della XLIII Riunione Scientifica SIS*, pp.511-14.
- Dijkstra, E.W., 1959. A Note on Two Problems in Connexion with Graphs. *Numerische Mathematlk*, 1, pp.269 - 271.
- European-Union, 2006. *Regulation 561/2006*.
- Feillet, D., Dejax, P. & Gendreau, M., 2005. Traveling Salesman problems with profits. *Transportation Science*, 39(2), pp.188-205.
- Fosgerau, M. & Karlstrom, A., 2010. The value of reliability. *Transportation Research Part B*, pp.38-49.
- Gendreau, G., Hertz, A. & Laporte, G., 1994. A Tabu Search Heuristic for the Vehicle Routing Problem. *Management Science*, 40(10), pp.1276-90.
- Ghiani, G., Laporte, G. & Musmanno, R., 2004. *Introduction to Logistics Systems Planning and Control*. West Sussex: John Wiley & Sons Ltd.
- Golden, B., 1976. Shortest_path Algorithms: A Comparison. *Operations Research*, 24(6), pp.1164-68.
- Golden, B., Bodin, L., Doyle, T. & Steward, W.j., 1980. Approximate Traveling Salesman Algorithms. *Operations Research*, 28(3), pp.694-711.
- Hill, A. & Benton, W.C., 1992. Modelling Intra-City Time-Dependent Travel Speeds. *The Journal of the Operational Research Society*, 43(4), pp.343-51.
- Ichoua, S., Gendreau, M. & Potvin, J.Y., 2003. Vehicle dispatching with time-dependent travel times. *European Journal of Operations Research*, 144(2), pp.379-96.
- Jansen, B., Swinkels, P.C.J.T.G.J.A., Antwerpen de Fluiter, B.v. & Fleuren, H.A., 2004. Operational planning of a large-scale multi-modal transportation system. *European Journal of Operations Research*, 156, pp.41-53.

- Kok, A.L., Hans, E.W. & Schutten, J.M.J., 2011. Optimizing departure times in vehicle routes. *European Journal of Operational Research*, 210, pp.579-87.
- Kuo, Y., Wang, C. & Chuang, P., 2009. Optimizing goods assignment and the vehicle routing problem with time-dependent travel speeds. *Computer & Industrial Engineering*, pp.1385-92.
- Laporte, G., 2009. Fifty years of vehicle routing. *Transportation Science*, 43(3), pp.408-16.
- LeBlanc, J., 2010. Customer experience Improvements build customer value. *Customer Strategist*, 2(1), pp.12-15.
- Lecluyse, C., Woensel, T.v. & Peremans, H., 2009. Vehicle Routing with Stochastic Time-Dependent Travel Times. *4OR: A QUARTERLY JOURNAL OF OPERATIONS RESEARCH*, 7(4), pp.363-77.
- Lin, H., Taylor, M.A.P. & Zito, R., 2005. A review of Travel_time prediction in transport and logistics. *Proceedings of the Asia Society for Transportation Studies*, pp.14433-1448.
- Malandraki, C. & Daskin, M.S., 1992. Time dependent vehicle routing problems: Formulations properties and heuristic algorithms. *Transportation Science*, 26(3), pp.185-200.
- McKinnon, P.A.C., 1998. *The Impact of Traffic Congestion on Logistical Efficiency*. Heriot-Watt University and the Institute of Logistics.
- Mulder, W., 2011. *The bias of FCD*. Bachelor Thesis. NHTV Internationale Hogeschool Breda.
- Peeters, p.d.C., Bouwman, i.T. & Hendrickx, F., 2009. Wegvervoer en logistiek. *Wegvervoerenlogistiek*, 3, pp.11-182.
- Rodrigue, J., Comtois, C. & Slack, B., 2006. *The Geography of Transport*. London and New York: Routledge, Taylor & Francis group.
- Steward, T., Strijbosch, L., Moors, H. & van Batenburg, P., 2007. *A simple approximation to the convolution of gamma distributions*. Tilburg: Tilburg University.
- Steward, T., Strijbosch, L., Moors, H. & van Batenburg, P., sept 2007. *A simple approximation to the convolution of gamma distributions*. Tilburg: Tilburg University.
- TomTom, n.d. *At your service, market analysis on delivery times*. TomTom.
- Toth, P. & Vigo, D., 2002. Models, relaxations and exact approaches for the capacitated vehicle routing problem. *Discrete applied mathematics*, 123, pp.487-512.
- van Bemmelen, M.M.K.J., 2012. *Modeling and Solving a Real-life Load Building and Routing Problem in the Retail Industry*. Master thesis. Eindhoven: TU/e Technical University of Eindhoven.
- van Lint, J.W. & Van der Zijpp, N.J., 2003. Improving a travel-time estimation algorithm by using dual loop detectors. *Transportation Research Record: Journal of the Transportation Research Board*, 1855(1), pp.41-48.
- van Lint, J.W.C. & van Zuylen, H.J., 2005. Monitoring and predicting freeway travel time reliability: Using width and skew of day-to-day travel time distribution *Transportation Research Record: Journal of the Transportation Research Board* 1917.-1 (2005): 54-62. *Transportation Research Record: Journal of the Transportation Research Board* , 1917(1), pp.54-62.
- van Lint, J.W.C., van Zuylen, H.J. & Tu, H., 2008. Travel time unreliability on freeways: Why measures based on variance tell only half the story. pp.258-77.
- Wessels, R., 2012. *Data Mining in TomTom's Big Data: Trip Origin and Destination Classification using Machine Learning*. Masters' Thesis. University of Twente.
- Witkovsky, V., 2001. Computing the Distribution of a linear combination of inverted Gamma Variables. *Kybernetika*, pp.79-90.
- Woensel, T.v., Kerbache, L.P.H. & Vandeale, N., 2008. Vehicle routing with dynamic travel times: a queueing approach. *European Journal of Operational Research*, 186(3), p.990-1007.
- Zhan, F.B., 1997. Three Fastest Shortest Path Algorithms on Real Road. *Journal of Geographic Information and Decision Analysis*, 1(1), pp.70-82.

Appendices

Description of Functional Road Classes (FRC)

FRC	Description	In data
0	Motorway, Freeway, or Other Major Road	Yes
1	a Major Road Less Important than a Motorway	No
2	Other Major Road	Yes
3	Secondary Road	Yes
4	Local Connecting Road	Yes
5	Local Road of High Importance	Yes
6	Local Road	Yes
7	Local Road of Minor Importance	Yes
8	Other Road	No

Appendix 1: Description of Functional road classes

Error distributions, using data from most recent data or split data

$\Delta(t_1, t_2)$			Frc 0 (N=3730)	Frc 2 (N=7891)	Frc 3 (N=495)	Frc 4 (N=399)	Frc 5 (N=461)	Frc 6 (N=574)	Frc 7 (N=51)
13 weeks of data $t_1 = '11w23$ $t_2 = '11w35$	Uncorrected data	Mean	-2.8794	-2.8005	-2.4875	-2.4125	-1.4350	-2.0272	-2.4931
		Std dev	12.5089	12.8460	13.9325	12.5540	12.3696	12.7996	14.8365
		MSE	164.6815	171.3509	197.3087	167.6663	160.5013	168.8313	256.0750
	Corrected & unrestricted data	Mean	-2.1719	-2.1679	-2.0003	-1.6412	-0.6458	-1.2758	-1.7124
		Std dev	12.2839	12.6837	13.7130	12.2130	12.1956	12.7227	14.5232
		MSE	161.9022	168.8797	197.0709	171.1957	163.6015	165.4934	239.9674
26 weeks of data $t_1 = '11w10$ $t_2 = '11w35$	Uncorrected data	Mean	-2.7185	-2.6376	-2.5678	-2.2188	-1.1826	-2.1008	-2.4755
		Std dev	12.4327	12.7912	13.8775	12.7714	12.5112	12.6409	14.4409
		MSE	166.9229	172.6206	201.6329	172.5601	167.8979	168.2659	255.1680
	Corrected & unrestricted data	Mean	-2.0068	-1.9955	-2.0422	-1.4983	-0.4034	-1.2717	-1.4852
		Std dev	12.1737	12.6121	13.6031	12.4083	12.2237	12.4983	14.1577
		MSE	167.8893	173.6693	203.0382	169.0304	173.0522	170.7265	252.8965
52 weeks of data $t_1 = '10w36$ $t_2 = '11w35$	Uncorrected data	Mean	-3.0570	-2.9877	-2.8178	-2.6015	-1.6227	-2.3419	-3.0116
		Std dev	12.5555	12.8528	14.0172	12.7599	12.6654	12.7305	14.7426
		MSE	155.5280	164.1141	189.1312	156.9877	154.8169	164.8141	242.4342
	Corrected & unrestricted data	Mean	-2.2719	-2.2518	-2.2630	-1.7818	-0.7297	-1.4319	-2.0519
		Std dev	12.2274	12.6341	13.6638	12.3524	12.2801	12.5924	14.5286
		MSE	152.1575	161.4117	186.8490	160.4912	155.2045	159.8996	226.4461
104 weeks of data $t_1 = '09w36$ $t_2 = '11w35$	Uncorrected data	Mean	-2.6233	-2.5323	-2.1317	-1.8639	-1.2896	-1.9511	-2.6733
		Std dev	12.6917	12.9891	14.1930	12.6544	12.8986	12.9253	14.8582
		MSE	154.5965	163.2631	189.0103	159.8810	157.3961	162.5346	241.5669
	Corrected & unrestricted data	Mean	-1.6873	-1.6734	-1.4276	-0.9931	-0.2047	-0.9223	-1.3830
		Std dev	12.2802	12.6504	13.7424	12.1663	12.3471	12.6434	14.2549
		MSE	153.5710	161.4296	187.8093	155.1461	158.3451	161.3213	227.1403
2 years of SPLIT data	Uncorrected data	Mean	0.7320	0.8007	0.7431	1.3048	0.9511	0.7241	0.8303
		Std dev	11.3847	11.4999	11.6662	11.8693	11.5533	11.4545	9.9732
		MSE	130.1259	132.8832	136.5145	142.3856	134.2350	131.5930	98.9463
	Corrected & unrestricted data	Mean	0.0389	-0.0401	0.2377	0.5342	1.3969	0.7139	-0.1906
		Std dev	12.2168	12.6072	13.6150	12.4188	12.5295	12.8102	13.8249
		MSE	149.1744	157.4786	182.1423	159.0131	160.7210	164.3904	200.6802
3 years of SPLIT data	Uncorrected data	Mean	0.7783	0.8424	0.7642	1.3935	0.9849	0.7811	0.8758
		Std dev	11.4698	11.6023	11.7761	11.9201	11.6739	11.4997	9.9440
		MSE	132.1526	135.3177	139.1944	143.9378	137.1803	132.7883	99.0833
	Corrected & unrestricted data	Mean	-0.0924	-0.1270	0.1491	0.4330	1.2074	0.6211	-0.2207
		Std dev	12.4735	12.8730	13.9808	12.4718	12.8376	13.1163	13.4648
		MSE	155.5207	164.3991	192.9133	160.0212	167.9626	170.9591	197.6967

Legend: Lowest mean, lowest standard deviation and lowest Mean Square Error (MSE)

Appendix 2: Error between estimation and observations from the unrestricted control sample set.

Using Travel Time Predictions based on TomTom's Big Data in Logistical Models

Percentile scores

FRC	GAMMA									NORMAL									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	10%	20%	30%	40%	50%	60%	70%	80%	90%	
1yr uncorr	0	10.14%	20.96%	29.90%	40.84%	51.01%	62.47%	71.57%	82.09%	91.08%	8.06%	16.47%	24.19%	31.98%	40.82%	50.05%	59.48%	69.27%	81.10%
	2	10.67%	21.18%	32.09%	42.82%	53.61%	64.33%	74.11%	82.85%	90.51%	8.92%	17.92%	26.65%	35.70%	44.82%	53.45%	62.16%	70.37%	78.79%
	3	6.95%	15.33%	23.89%	33.87%	45.45%	56.33%	69.34%	80.57%	88.77%	4.28%	8.91%	13.01%	17.47%	22.82%	28.34%	35.29%	45.99%	56.15%
	4	9.45%	17.66%	24.13%	32.84%	43.53%	58.71%	71.39%	81.84%	91.79%	7.46%	13.68%	21.64%	27.86%	35.57%	41.54%	50.25%	59.20%	71.64%
	5	11.29%	21.34%	30.34%	43.03%	56.08%	65.08%	72.84%	83.60%	90.48%	7.94%	14.99%	22.05%	28.92%	38.98%	49.56%	56.97%	67.55%	76.90%
	6	12.77%	23.08%	33.22%	43.54%	53.19%	62.19%	72.83%	83.80%	88.87%	10.97%	18.17%	25.86%	33.99%	40.26%	49.10%	57.12%	66.61%	78.23%
	7	11.94%	20.90%	34.33%	40.30%	50.75%	62.69%	65.67%	71.64%	80.60%	5.97%	16.42%	23.88%	34.33%	40.30%	47.76%	52.24%	59.70%	71.64%
2yrs uncorr	0	10.12%	20.15%	29.63%	40.52%	50.88%	61.83%	72.29%	82.92%	92.82%	7.02%	14.28%	21.09%	28.16%	35.45%	43.30%	52.27%	61.69%	73.38%
	2	11.23%	23.10%	34.13%	45.71%	56.06%	67.00%	76.36%	84.68%	92.08%	7.70%	16.24%	24.25%	32.70%	40.91%	49.16%	58.10%	66.42%	76.20%
	3	6.60%	14.97%	24.06%	33.51%	45.99%	57.22%	69.70%	80.93%	89.66%	3.57%	7.84%	12.83%	17.47%	22.64%	25.67%	32.62%	42.07%	54.72%
	4	8.21%	16.67%	25.62%	37.06%	49.50%	61.69%	75.73%	83.33%	90.30%	5.97%	15.17%	22.89%	31.09%	38.81%	45.52%	51.00%	61.19%	72.64%
	5	11.64%	24.69%	33.86%	43.92%	56.26%	64.73%	75.84%	84.66%	91.18%	7.76%	16.93%	23.46%	29.81%	39.33%	47.97%	56.26%	66.14%	77.60%
	6	10.95%	22.55%	35.62%	45.10%	56.70%	65.20%	73.86%	83.66%	90.52%	8.66%	17.81%	26.63%	33.99%	41.67%	48.20%	58.01%	67.81%	79.58%
	7	13.24%	26.47%	35.29%	44.12%	57.35%	64.71%	66.18%	73.53%	83.82%	8.82%	17.65%	26.47%	32.35%	38.24%	47.06%	63.24%	64.71%	67.65%
1yr corr	0	12.23%	22.43%	31.05%	41.07%	52.51%	63.85%	73.00%	81.79%	90.77%	9.23%	17.85%	24.63%	31.31%	40.19%	49.43%	59.63%	69.31%	80.65%
	2	8.01%	16.28%	26.08%	35.54%	45.25%	56.19%	65.84%	75.11%	83.31%	4.76%	9.06%	14.98%	21.01%	26.18%	32.13%	38.70%	46.26%	55.04%
	3	6.78%	13.92%	22.16%	30.77%	40.48%	49.08%	59.34%	69.05%	77.11%	4.21%	6.96%	11.54%	16.30%	22.34%	27.11%	32.05%	37.55%	48.17%
	4	7.02%	10.37%	18.06%	26.76%	31.10%	42.47%	51.84%	59.53%	70.23%	4.00%	6.00%	11.67%	16.00%	21.33%	25.00%	30.33%	38.67%	46.33%
	5	8.19%	14.94%	20.72%	29.16%	37.59%	46.27%	56.14%	67.23%	76.14%	5.54%	9.40%	15.90%	19.76%	26.27%	30.84%	37.35%	43.61%	54.22%
	6	7.85%	14.78%	21.17%	28.65%	33.76%	41.24%	49.64%	61.13%	72.26%	6.93%	11.50%	16.42%	19.34%	25.36%	31.39%	35.40%	41.79%	50.91%
	7	2.63%	5.26%	10.53%	13.16%	18.42%	23.68%	31.58%	36.84%	47.37%	2.50%	2.50%	5.00%	5.00%	10.00%	10.00%	15.00%	17.50%	22.50%
2yr corr	0	11.45%	21.88%	32.30%	43.87%	53.04%	63.12%	73.42%	83.28%	92.55%	8.82%	14.89%	21.19%	28.29%	34.71%	41.70%	50.63%	61.05%	72.74%
	2	8.03%	16.92%	25.81%	35.59%	45.24%	56.58%	66.62%	75.25%	84.37%	4.66%	10.53%	15.36%	20.45%	25.38%	31.41%	37.87%	44.72%	54.21%
	3	6.04%	14.47%	23.26%	32.23%	40.29%	50.00%	59.52%	70.33%	77.84%	3.11%	7.33%	11.17%	15.75%	20.70%	27.11%	32.78%	38.83%	46.15%
	4	7.02%	12.37%	17.06%	24.08%	33.11%	40.80%	51.84%	58.86%	69.23%	2.33%	7.00%	11.33%	13.67%	18.67%	22.00%	28.33%	37.67%	46.33%
	5	6.67%	15.80%	23.46%	28.89%	38.77%	47.41%	56.79%	69.14%	78.02%	5.41%	9.83%	15.72%	20.15%	24.82%	30.47%	36.36%	42.51%	54.30%
	6	8.03%	14.53%	22.56%	30.21%	34.80%	42.26%	51.82%	62.14%	72.47%	6.30%	10.69%	13.74%	19.27%	25.57%	30.73%	37.21%	41.41%	49.62%
	7	2.44%	9.76%	12.20%	14.63%	19.51%	26.83%	29.27%	34.15%	43.90%	2.38%	4.76%	9.52%	11.90%	11.90%	14.29%	19.05%	19.05%	26.19%
t.o.y. 2yrs corr	0	8.38%	16.62%	26.20%	36.97%	46.78%	55.94%	67.08%	78.41%	89.41%	7.44%	15.07%	22.69%	32.61%	40.97%	49.47%	59.72%	70.00%	82.04%
	2	7.43%	15.27%	23.23%	30.92%	39.81%	48.73%	57.40%	66.32%	74.52%	5.17%	10.25%	15.66%	20.73%	26.40%	32.09%	38.53%	46.22%	54.51%
	3	6.53%	15.14%	23.24%	31.07%	38.90%	49.87%	57.44%	67.36%	75.46%	4.70%	9.92%	13.84%	17.75%	24.28%	30.55%	35.25%	42.82%	53.00%
	4	1.95%	7.42%	14.45%	21.88%	28.91%	39.06%	49.61%	61.33%	73.44%	2.71%	5.04%	9.69%	13.18%	20.54%	24.42%	31.78%	41.09%	50.78%
	5	7.73%	13.60%	21.33%	29.87%	37.87%	48.00%	57.60%	66.67%	77.33%	4.79%	9.57%	14.36%	19.95%	26.60%	32.71%	39.63%	48.94%	57.71%
	6	6.48%	14.46%	21.95%	28.43%	36.16%	43.14%	51.62%	59.60%	71.07%	5.97%	10.70%	15.42%	18.41%	25.37%	29.85%	35.57%	41.79%	51.99%
	7	9.09%	9.09%	9.09%	13.64%	18.18%	18.18%	18.18%	22.73%	27.27%	8.70%	8.70%	8.70%	8.70%	8.70%	17.39%	17.39%	17.39%	17.39%
t.o.y. 3yrs corr	0	9.24%	18.12%	27.29%	37.39%	47.35%	57.44%	67.86%	79.26%	89.32%	8.22%	15.65%	23.87%	32.25%	40.80%	49.87%	59.66%	70.40%	81.96%
	2	7.38%	15.20%	23.92%	33.41%	42.10%	50.94%	60.15%	68.64%	76.84%	4.30%	8.82%	13.80%	19.09%	24.29%	29.84%	35.81%	41.92%	50.44%
	3	7.23%	14.29%	24.05%	31.83%	40.69%	50.09%	58.41%	68.54%	77.58%	4.52%	8.68%	15.55%	19.89%	24.95%	28.57%	35.26%	42.13%	49.37%
	4	5.79%	12.12%	18.18%	24.24%	33.06%	42.42%	50.69%	59.50%	71.90%	2.47%	6.58%	10.68%	17.26%	22.47%	27.40%	33.70%	41.92%	51.23%
	5	6.40%	13.76%	23.06%	31.40%	38.76%	48.64%	59.50%	67.25%	76.94%	2.90%	7.74%	13.15%	19.73%	26.11%	33.08%	40.43%	50.29%	58.61%
	6	6.79%	13.76%	21.43%	27.35%	35.37%	42.86%	49.65%	60.63%	71.78%	4.52%	9.04%	13.74%	19.13%	24.70%	28.35%	34.78%	40.52%	49.74%
	7	5.00%	7.50%	15.00%	20.00%	27.50%	27.50%	32.50%	37.50%	8.06%	16.47%	24.19%	31.98%	40.82%	50.05%	59.48%	69.27%	81.10%	
t.o.y. 2yrs uncorr	0	9.43%	18.60%	28.27%	37.70%	48.26%	57.15%	68.30%	79.37%	9.43%	8.14%	16.05%	24.06%	32.34%	40.35%	48.97%	58.53%	68.81%	81.26%
	2	7.98%	15.78%	23.32%	31.43%	40.07%	48.19%	55.96%	64.17%	7.98%	5.05%	9.87%	14.92%	19.79%	24.75%	30.09%	35.50%	42.55%	50.67%
	3	6.42%	15.69%	23.71%	31.73%	43.49%	54.55%	68.98%	79.14%	6.42%	4.63%	8.02%	11.05%	13.90%	20.86%	26.02%	33.33%	39.75%	48.48%
	4	5.75%	11.75%	16.50%	22.75%	32.00%	43.00%	54.00%	65.25%	7.75%	4.23%	8.46%	11.69%	14.18%	18.16%	26.12%	30.85%	39.30%	50.00%
	5	8.47%	17.99%	26.28%	35.27%	47.44%	59.79%	69.31%	80.25%	8.47%	5.29%	10.05%	15.34%	22.05%	29.45%	34.39%	43.21%	52.20%	62.26%
	6	8.20%	15.57%	27.21%	33.77%	42.95%	51.80%	62.13%	72.30%	8.20%	3.60%	10.31%	15.22%	19.64%	25.20%	31.59%	35.84%	41.90%	51.55%
	7	10.71%	14.29%	21.43%	23.21%	32.14%	41.07%	41.07%	46.43%	10.71%	1.75%	7.02%	8.77%	10.53%	12.28%	15.79%	21.05%	26.32%	31.58%
t.o.y. 3yrs uncorr	0	9.38%	18.35%	28.10%	37.53%	47.90%	57.25%	68.52%	79.53%	9.38%	8.06%	15.73%	23.93%	32.07%	40.11%	48.87%	58.18%	68.79%	8.06%
	2	7.73%	15.46%	23.89%	32.29%	41.31%	49.78%	58.06%	65.94%	7.73%	4.98%	9.46%	14.46%	19.06%	24.26%	29.45%	34.73%	41.47%	4.98%
	3	6.42%	13.90%	22.82%	31.19%	43.85%	54.90%	68.63%	79.14%	6.42%	3.92%	7.66%	11.23%	15.69%	21.03%	26.56%	33.16%	39.75%	3.92%
	4	7.23%	12.22%	17.71%	25.44%	32.92%	41.40%	54.11%	64.84%	7.23%	4.23%	7.96%	12.19%	15.92%	21.39%	25.87%	31.34%	40.05%	4.23%
	5	7.41%	18.52%	26.63%	36.33%	48.85%	58.55%	69.31%	80.42%	7.41%	4.41%	10.58%	15.17%	22.05%	27.69%	34.57%	43.21%	52.38%	4.41%
	6	8.20%	15.25%	26.39%	34.59%	43.11%	52.13%	61.80%	71.48%	8.20%	4.75%	9.00%	14.57%	19.64%	24.22%	30.77%	35.84%	41.41%	4.75%
	7	9.26%	14.81%	22.22%	25.93%	37.04%	46.30%	48.15%	51.85%	9.26%	0.00%	3.64%	7.27%	7.27%	16.36%	20.00%	21.82%	29.09%	0.00%

Appendix 3: Percentile scores per dataset