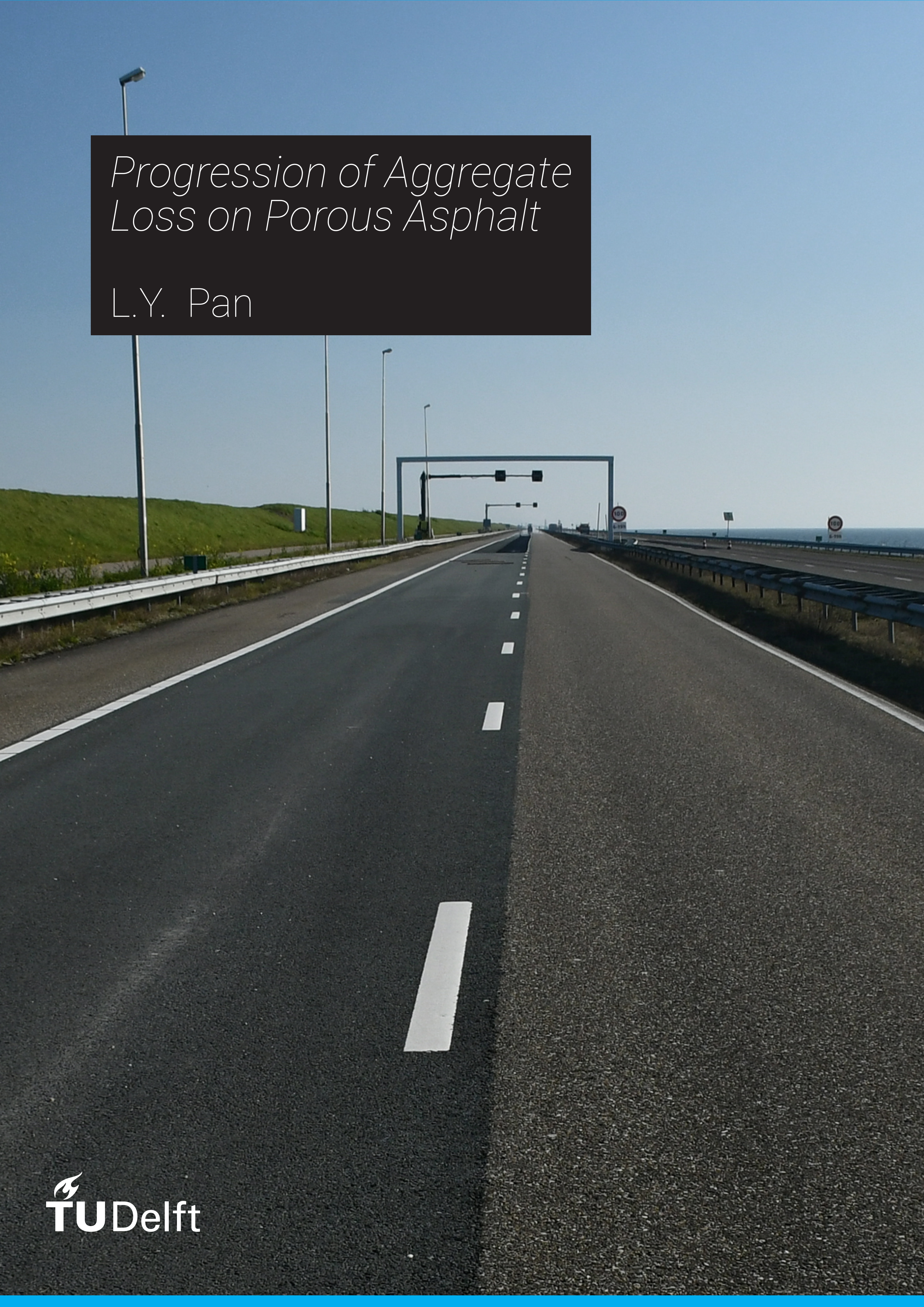

Progression of Aggregate Loss on Porous Asphalt

L.Y. Pan

TUDelft

Rijkswaterstaat
Ministry of Infrastructure and Water Management

Progression of Aggregate Loss on Porous Asphalt

by

Li Yong Pan

to obtain the degree of

Master of Science

in

Applied Mathematics

at the Delft University of Technology,
to be defended publicly on Friday August 28th, 2020 at 02:00 PM.

Student number:	4466411	
Project duration:	November 19, 2019 – August 28, 2020	
Thesis committee:	Prof.dr.ir. G. Jongbloed,	TU Delft, supervisor
	Ir. L. Schouten,	Rijkswaterstaat, supervisor
	Dr. K. Anupam,	TU Delft (Faculty of Civil Engineering & Geosciences)
	Dr. P. Chen,	TU Delft

Delft University of Technology
Faculty of Electrical Engineering, Mathematics & Computer Science
Delft Institute of Applied Mathematics

TU Delft

Preface

This thesis was made as part of and to conclude my master education in Applied Mathematics at the TU Delft in the Netherlands. My research was commissioned by Rijkswaterstaat.

During the course of this thesis, there were several persons who have shown their support greatly. Therefore, I would like to mention and thank them: Léon Schouten not only for his outstanding handling in administrative matters and concise and well-structured feedback on the report, but also for his continuous critical thinking irrespective of mathematical relations; Geurt Jongbloed not only for his excellent supervision and expertise on statistical methods, but also for his infinite great spirit which is motivating throughout; Willem van Aalst not only for providing explanations on the data in person, but also for his willingness to answer emerging questions and most importantly: preparing and delivering the data which opened up the possibility of the research; Thijs Bennis and Stefan Russel not only for their continued interest but also for the feedback on the report and methodology – all of you, thank you!

I would also like to thank my thesis committee members: Geurt Jongbloed, Léon Schouten, Kumar Anupam and Piao Chen for their continued interest and time to evaluate my work. Lastly, I would like to thank my family and friends for their support and providing necessary distraction at times.

L.Y. Pan
Schiedam, August 2020

Abstract

Porous asphalt resides on most top layers of Dutch roads. Scheduling maintenance for these roads is generally dependent on several factors, but ravelling, the loss of aggregates in the top layers, is the main reason for maintenance on Dutch roads. With the recent framework of the DOS-LCMS scheme generating values for aggregate loss in percentages, a prediction for the remaining lifetime of a road section surfaced with porous asphalt with respect to ravelling can be performed. The lifespan for porous asphalt layers is dependent on the most suffered 25% of the section on the respective 100 meter length. The current threshold is set at 10%, implying that road sections of 100 meter need maintenance if more than 25% of the road (75th percentile) measures aggregate loss over 10%. The present work approximates these 75th percentiles throughout the years using parametric and non-parametric approaches, whereafter the estimates of the 75th percentiles are used to construct smooth monotonic increasing convex curves. These curves, which are in fact functions built on P-splines, are then used to perform extrapolation and hence predict the dates on which the threshold is going to be surpassed. The study reveals problems in the raw data which is particularly prominent in the sequence of 75th percentiles, frequently showing a lack of monotonicity and convexity. Putting the monotonicity and convexity constraints on a more granular level were found to be helpful for the predictions and improved the consistency of lifetime predictions over consecutive years.

Contents

List of Abbreviations

AC	Asphalt Concrete
AD	Anderson-Darling
CDF	Cumulative Distribution Function
DLS	Driver Location Sign
DOS	Detectie Oppervlakte Schade (Detection Surface Damage)
KS	Kolmogorov-Smirnov
LCMS	Laser Crack Measurement System
LDA	Linear Discriminant Analysis
LWT	Left Wheel Track
MAD	Median Absolute Deviation
M(I)SE	Mean (Integrated) Squared Error
PA	Porous Asphalt
PDF	Probability Density Function
PRL	Proposed Remaining Lifetime
RWS	Rijkswaterstaat
RWT	Right Wheel Track
SPRS	Strategic Planning Road Surfaces
SW	Shapiro-Wilk
TLPA	Two Layer Porous Asphalt
TSD	Threshold Surpassing Date

1 Introduction

Reliable estimation of road surfacing lifetimes plays a critical role when striving for efficient maintenance planning of roads. The main challenge faced by maintenance planners is deciding when exactly to perform resurfacing. If this moment in time can be defined, the road surfaces can be used to its full potential in terms of longevity — resulting in a greater value per cost ratio.

Dutch highways are primarily ($> 90\%$ [1]) surfaced with a top layer of *porous asphalt* (PA). PA is a permeable road surface type which consists of a mixture of stones, sand, filler material and bitumen, of which the latter is a highly viscous form of petroleum that functions as a binder. The pros of using porous asphalt over *asphalt concrete* (AC) are its reduced noise production, rain drainage, and resistance against corrugation. However, bitumen in PA are more exposed to ageing due to its nature, allowing weather conditions to permeate and cause brittleness. After a certain degree of brittleness, regular traffic flow causes cracks in the bitumen. This ultimately leads to the release or loss of aggregates, which is called *ravelling*. Ravelling is the main damage mechanism in Dutch highways ($> 70\%$ [1]).

Recent advances in automated pavement inspection have opened up new means of analysis. In particular, *Rijkswaterstaat*[1] (RWS) and *TNO*[2] have effectively built on the *Laser Crack Measurement System* (LCMS) provided by the company *Pavemetrics* [2] to apply LCMS in the scheme of ravelling [3]. Ravelling measured in percentage aggregate loss is calculated from the LCMS generated 3D profiles, using *Detectie Oppervlakte Schade* (DOS) (English: Detection Surface Damage) algorithms [4]. A search of the literature revealed few studies which consider DOS-LCMS data for their analyses. Recently Leegwater et al. [5] have used DOS-LCMS data for analysis on asphalt lifetimes. The project led to: a prediction scheme for remaining lifetimes based on machine learning models, a classifying scheme for road segments categorised on exponential curve fitting, and a dashboard to visualise the data for the European project *BE-GOOD*. However, due to the blackbox-nature of the prediction scheme in Leegwater et al. [5], a more fundamentally statistical approach has still not been developed.

[1] Rijkswaterstaat is part of the Dutch Ministry of Infrastructure and Water Management. Specifically, RWS is in charge of the execution of public works and water management.

[2] TNO is a Dutch independent research organisation which focuses on applied sciences.

2

Literature Review

It has previously been observed that the estimation of lifetime distributions can be carried out by applying methods from the branch *survival analysis* [6–8]. Verra et al. [6] uses the concept of the well-known *survival function* given by $t \mapsto S(t)$, which denotes the probability of survival after time t for some subject. In particular, the *Kaplan-Meier estimator* [9] is used to approximate the aforementioned survival function based on independent and identically distributed, right-censored data as follows

$$\widehat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right),$$

in which $d_i > 0$ and n_i respectively denote the number of deaths at time t_i and the number of survivors until t_i. The benefit of this approach is that the Kaplan-Meier estimator is a *non-parametric* statistic: as few assumptions as possible are made about the data [10]; besides the assumption of independent and identically distributed data, there are no heavy assumptions on the underlying data. As a result, non-parametric methods are flexible in terms of applicability; hence for this case as well. By definition of $S(t) = 1 - F(t)$, where $F(t)$ denotes a *cumulative distribution function* (CDF), the estimate of the Kaplan-Meier estimator also generates an estimate of the CDF. Verra et al. [6] then provides estimates by means of common statistics such as the mean, standard deviation and the median for different categories. These categories are defined based on the following: the geographical location of the road (northern, middle, and southern part) in the Netherlands[1], and the position of the lane on the road. The reason for the relevance of the last categorisation can be explained: given that the norm is to drive on the rightmost lane, the endured load over time should inevitably be higher on the rightmost lane which effects longevity; this has indeed been concluded in [6], from which is clear that the lifetimes of the lanes which endures most load are substantially shorter.

Derksen [7] in 2014 and Ebrahimi et al. [8] in 2019 also use survival analysis for their inference. The methodology of Derksen [7] is analogous to the one of Verra et al. [6], but Derksen [7] does not explicitly report the mean, standard deviation and the median per analysis. Ebrahimi et al. [8] takes a different approach: instead of a priori classifying which lane endures most load, they researched the effect of multifarious external variables including the following: traffic volume, posted speed limit, aggregate nominal maximum size, and heavy traffic volume. These were used as categories in a *Cox proportional-hazards model* [11], which in this specific research fundamentally involves constructing several Kaplan-Meier survival curves for different groups based on said categories. Data from Ebrahimi et al. [8] suggests that aggregate nominal size and heavy traffic volume are the significant factors in longevity of the analysed surfaces.

[1] The research of Verra et al. [6] was performed in the Netherlands.

Until now, we have only mentioned work which applies survival analysis methods as their main method for research. Although this approach should not be completely undermined and can give significant insight overall, such expositions are somewhat lacking in applicability. The results from previous studies produce survival probabilities per year for road segments based on their classification. However, it would treat two road sections that belong to the same predefined category[2] but differ significantly in non-predefined influential factors, as having the same survival probability. For example, consider a probable influential factor such as endured load over time. Such a factor is quite troublesome to measure with high accuracy. Two road segments could hence belong to the same category, despite greatly differing in endured load over time, and could not be reclassified since such data about load is unavailable or not representative. In similar fashion, there are many external variables which can be considered that undeniably makes each road segment unique, but can simply not all be captured.

In other words, all inferences from Verra et al. [6] and Derksen [7] and — though in undoubtedly lesser degree — Ebrahimi et al. [8] are fundamentally based on two moments in time: year of construction and year of reconstruction[3]. For the purpose of finding average lifetimes, these certainly suffice and are proper. However, they do not suffice for giving unique lifetime indications per road segment. There is a clear absence of a lifetime measure in all mentioned studies: some statistic which indicates whether a road surface needs maintenance. If such a measure is decided upon, it is possible to monitor that statistic per (fixed) time interval and compute estimations for each road segment. It would also almost completely negate the issues of unjustified classification, as there is no classification to begin with. Combining older and more recent work from Rijkswaterstaat and TNO, it seems feasible to pursue such an approach.

[2] A category in this context would be similar to the proposed categories from the previous studies [6–8].

[3] There is a slight ambiguity in the year of reconstruction, because it can be either be the proposed or the actual year of reconstruction. These two can mismatch for several reasons.

3 Motivation and research questions

This research sets out to investigate the usefulness of data on aggregate loss provided by the DOS-LCMS scheme to assist Rijkswaterstaat in determining asphalt surfacing lifetimes. RWS also aims for efficient construction planning and in addition to their own analyses has already consulted help from external parties. In 2009, RWS started a collaboration with TNO and they have been responsible for implementing and further development of the DOS-LCMS scheme, which has been in use for annual data acquisition since 2012. Furthermore, using DOS-LCMS data combined with the proposed remaining lifetimes from visual inspectors from RWS, TNO has created a *linear discriminant analysis* (LDA) model which provides a means of estimating remaining lifetimes.

Although a model has been established by TNO to predict remaining lifetimes for asphalt surfaces, its current iteration is not based on RWS's predefined aggregate loss threshold for maintenance, or at least in objective manner. To clarify that, we need to specify the actual threshold first. In 2003 and continuing, the threshold was defined to be the following:

Threshold (Verra et al. [6] and DHV et al. [12])

If more than 25% of a road segment (approximately 100 m) measures an aggregate loss percentage of at least 10%, the corresponding road segment needs maintenance.

Clearly, due to the lack of measuring systems before 2012, all proposed remaining lifetimes by the visual inspectors before this date were not established from data. This method — despite the undisputed expertise of the inspectors — is exposed to inconsistencies such as human bias, which induces irreproducibility. The DOS-LCMS implementation, on the other hand, has proven to be within the tolerance bound and can be concluded to be reproducible according to Van Aalst [13]. The LDA model from TNO used judgements of the visual inspectors but, as stated before, these lack reproducibility. Therefore, a helpful question to answer would be: how accurate were the visual inspectors from RWS in their judgements in correspondence with the data?

The threshold mentioned before is certainly not absolute because the planning is ultimately still decided upon by considering the proposed remaining lifetimes from TNO and visual inspection from RWS. However, it does raise the question if and how we can use the DOS-LCMS data to provide reliable estimates of remaining lifetimes. Now that we have data from 2012 — 2019 available, it presents us with the opportunity to perform more in-depth analysis. The following research questions are based on the research aims:

Research Questions

Main Question

Given DOS-LCMS data over multiple years from a road section, how can we predict its corresponding remaining lifetime?

Side Questions

- Under which conditions are the predictions representative based on the official threshold?
- To what extent are possible differences in wheel tracks and lanes portrayed in the data?

The results may be used as a proof of concept and a potential modified approach is not unreasonable. Analysis and producing plots will be done in `R`, which is software used for statistical computing.

4

General Terminology

This Chapter describes the terminology that is common at RWS. Due to the nature of RWS, numerous words and notation that are frequently used are quite straightforward in the Dutch language, but could be ambiguous in the English language for the reader — and myself initially — due to little experience. The Chapter is brief and consists of two parts: we depict the general notation used for a road section and its components, and we provide a short vocabulary list for Dutch readers.

Depictions of a Road Section

Figure 4.1 is a visualisation of a road section and its components.

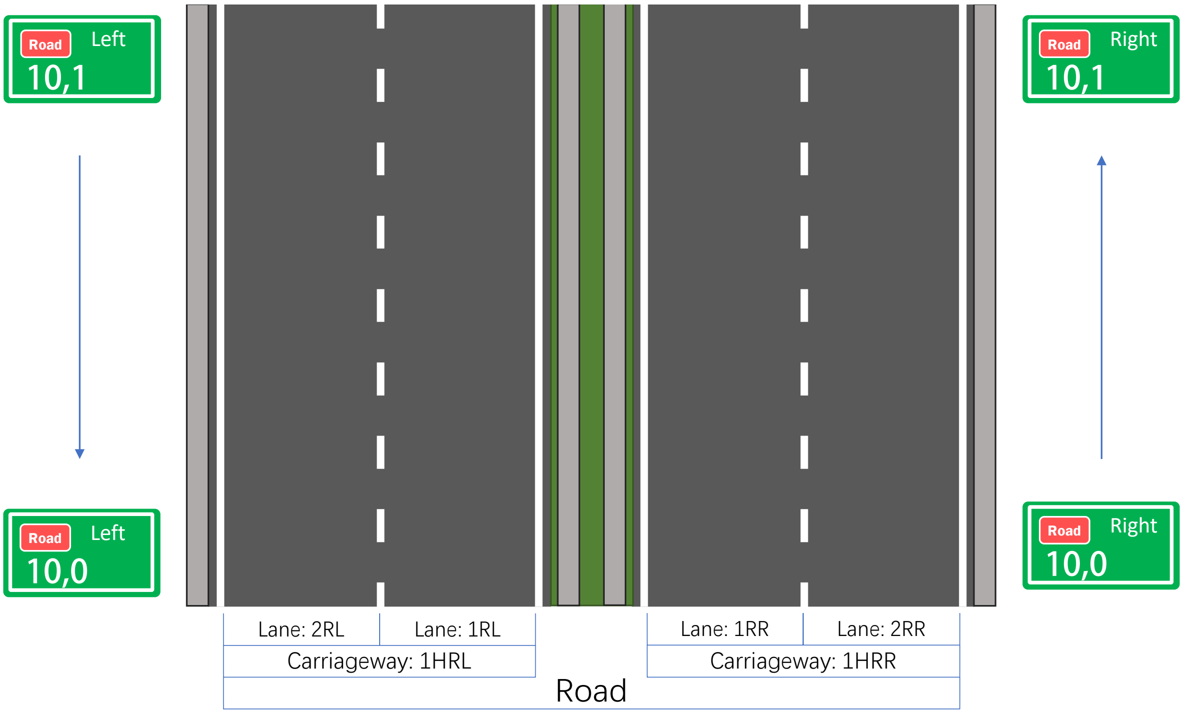

Figure 4.1: A cross-section of a road and corresponding notation, including the driver location signs. The illustration does *not* contain full detailed notation as explained in Driessen et al. [14], but rather annotates the bare minimum demanded for the analyses. The driver location signs with respect to the traffic direction determine if a carriageway is named left or right; an increase indicates a right carriageway, a decrease indicates a left carriageway.

A road or *highway*[1] generally consists of one or more *carriageways*; a carriageway consists of one or more *lanes*. If we would properly join all the road sections as illustrated in Figure 4.1, the union would be the entire road. Keep in mind that there are several configurations for a road (section), but to outline all these would defy the purpose of this thesis. Instead, we refer to Driessen et al. [14] for other possible configurations and even more detailed notation. As for the current notation, note that it is derived from RWS and thus corresponds to Dutch naming. Fortunately, the Dutch words for left (*links*) and right (*rechts*) start with the same letters, which helps with a general rule of thumb in naming, a bare minimum for analysis, stated as follows

Rule of Thumb (BPS)

In BPS notation

- the last letter denotes the carriageway relative to the driver location signs; if travelling forwards leads to an increase (decrease) of the numbers on the hectometre posts it corresponds to a right (left) carriageway,
- the number denotes the corresponding position relative to the centre.

To give an example; if a lane is indicated as ‘3RR’, we can already deduce that it is situated on a (*not* the) right carriageway relative to the driver location signs, and it is the 3rd lane from the middle. However, from this piece of information only it indeed remains unclear which carriageway it is; highways with multiple carriageways associated with the same driving direction do exist. Therefore, in general it is necessary to also declare notation for the carriageway before we can conclude anything about the specific location. This is particularly important when an event or accident has occurred and an explicit description of the position on the road is demanded. An example of BPS notation sufficient for analysis is the following:

Example Notation

A44 - 1HRR - 7.1 - 1RR in BPS notation corresponds to the road section of

- Road: A44
- Carriageway: 1st of the right
- Driver location sign: starting from 7.1 up until 7.2
- Lane: 1st from left-to-right with respect to traffic direction

The final pivotal terms that need to be introduced are the *left* and *right wheel track* (LWT, RWT), shown in Figure 4.2. These are important due to the nature of the provided data, which will be elaborated on in Chapter 5. In general, there is a difference in damage intensity between LWT and RWT of the same lane. Additionally, as indicated before, it shows that the 2RR lane has suffered more from ravelling compared to the 1RR lane — assuming that damage buildup started at roughly the same time. The widths of the LWT and RWT are assumed to be 1 meter. In practice, however, it seems to be smaller than the assumed 1 meter.

[1] Due to personal preference, hereafter we will use ‘road’ rather than ‘highway’.

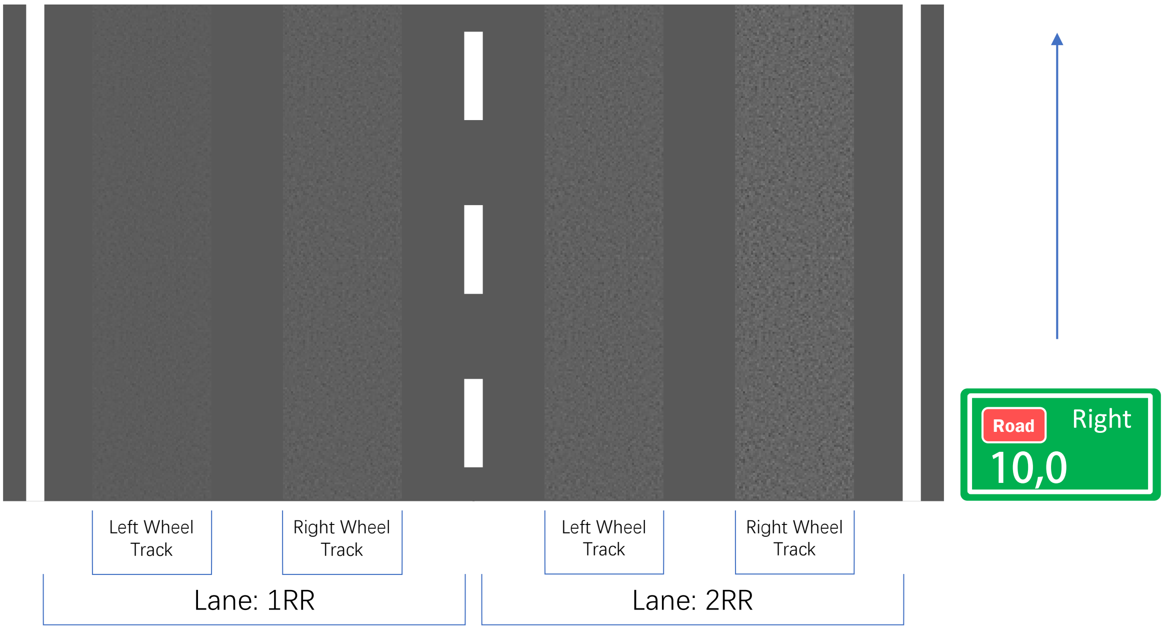

Figure 4.2: A segment of the road section from Figure 4.1, with the assumption that ravelling-buildup started at (roughly) the same time. The illustration shows what it is meant by left/right wheel track, and attempts to portray the discrepancy in severity of damage on these tracks. Although the contrast in severity is well captured for the general case, this situation does not always hold.

Vocabulary (Dutch-English)

Table 4.1: Dutch words and their English translation for terms common in RWS jargon

Dutch	English
weg	road, highway
baan	carriageway
strook	lane
hectometerpaal	driver location sign, hectometre post
ZOAB (zeer open asfalt beton)	PA (porous asphalt)
DAB (dicht asfalt beton)	AC (asphalt concrete)
Groot Onderhoud	pavement rehabilitation
MJPV (Meerjarenplanning Verhadingsonderhoud)	SPRS (Strategic Planning Road Surfaces)
kunstwerk	engineering structure

5 Data

Before conducting analysis and basing conclusions on these, it is mandatory to properly assess the data that we have access to. The data which is available are of the following types: DOS-LCMS, KernGIS, SPRS. This Chapter will discuss the selection, acquisition, format, and their latent complications.

5.1. Road Selection

It was necessary to set specific prerequisites for choosing the roads, such that we could perform sensible analysis. DOS-LCMS data has been gathered for numerous roads and to analyse all these adequately would take a tremendous amount of time. The initial prerequisites for roads were decided upon with Léon Schouten and are enumerated in Table 5.1 with respective reasoning.

Table 5.1: Data prerequisites with corresponding reasoning.

	Prerequisite	Reason
1.	The road surface is of the PA(+) type	There are two main reasons for this specific prerequisite. First of all, over 85% of Dutch roads are of type PA(+), which implies that a proper analysis could translate to the majority of the roads. Secondly, the older iteration of the DOS-LCMS implementation only allows for appropriate aggregate loss detection on PA types, which excludes surface types which have finer texture than PA(+). For completeness sake the most recent DOS-LCMS implementation is capable of handling finer texture, but for analysis we are restricted by the former iteration.
2.	Pavement rehabilitation was done in a specific year.	Restricting to complete pavement rehabilitation in one year allows for more uniform analysis for one road; larger likelihood of finding similar patterns across the road. More crucially, however, is that the initial idea was to investigate differences in phases of severity of aggregate loss: the further back in time the most recent rehabilitation was performed, the higher the degree of severity we can expect.

Prerequisite	Reason
3. Maintenance was executed for a substantial length.	Substantial lengths imply more data and a better means of comparing — thought not independent — on a specific road. It increases the likelihood of a proposed methodology set up for one road section being able to function for other sections of the corresponding road.
4. No changes have been made in road configuration since the pavement rehabilitation.	Road configurations could cause complications in notation if there were any physical changes in road configuration. For example, it is immensely convenient if a road section in the year 2012, say A44 - 1HRR - 7.1 - 1RR, would physically be the same road section as A44 - 1HRR - 7.1 - 1RR in the year 2019. Indeed, in general this does not necessarily hold. If a new lane at the 7.1 driver location sign would be added to the left for which reason whatsoever during 2013 – 2018, the 1RR lane in 2012 would then physically be the same as the 2RR lane in 2019.

Take into consideration that the demands mentioned in Table 5.1 are not an absolute necessity per se, but they definitely are convenient. For the ultimately chosen roads, Léon Schouten found several roads which satisfy the prerequisites and these are summarised in Table 5.2 with additional details. Figure 5.1 shows a map of the topographical locations of the selected roads, and Figures 5.2 to 5.4 display ravelling on these roads; Figures 5.1 to 5.4 are by courtesy of Léon Schouten.

Table 5.2: Final selected roads with details. PRD = Pavement Rehabilitation Date, DLS = Driver Location Sign.

	PRD	DLS	Expected Severity
A44			
1HRR	2002-09-09	2.1 – 7.7	High
A50			
1HRR	2012-09-17	139.9 – 148.4	Low
1HRL	2002-12-31	205.5 – 202.9	High
A6			
1HRR	2005-11-15	280.2 – 288.0	High
1HRR	2005-08-31	288.0 - 295.8	High

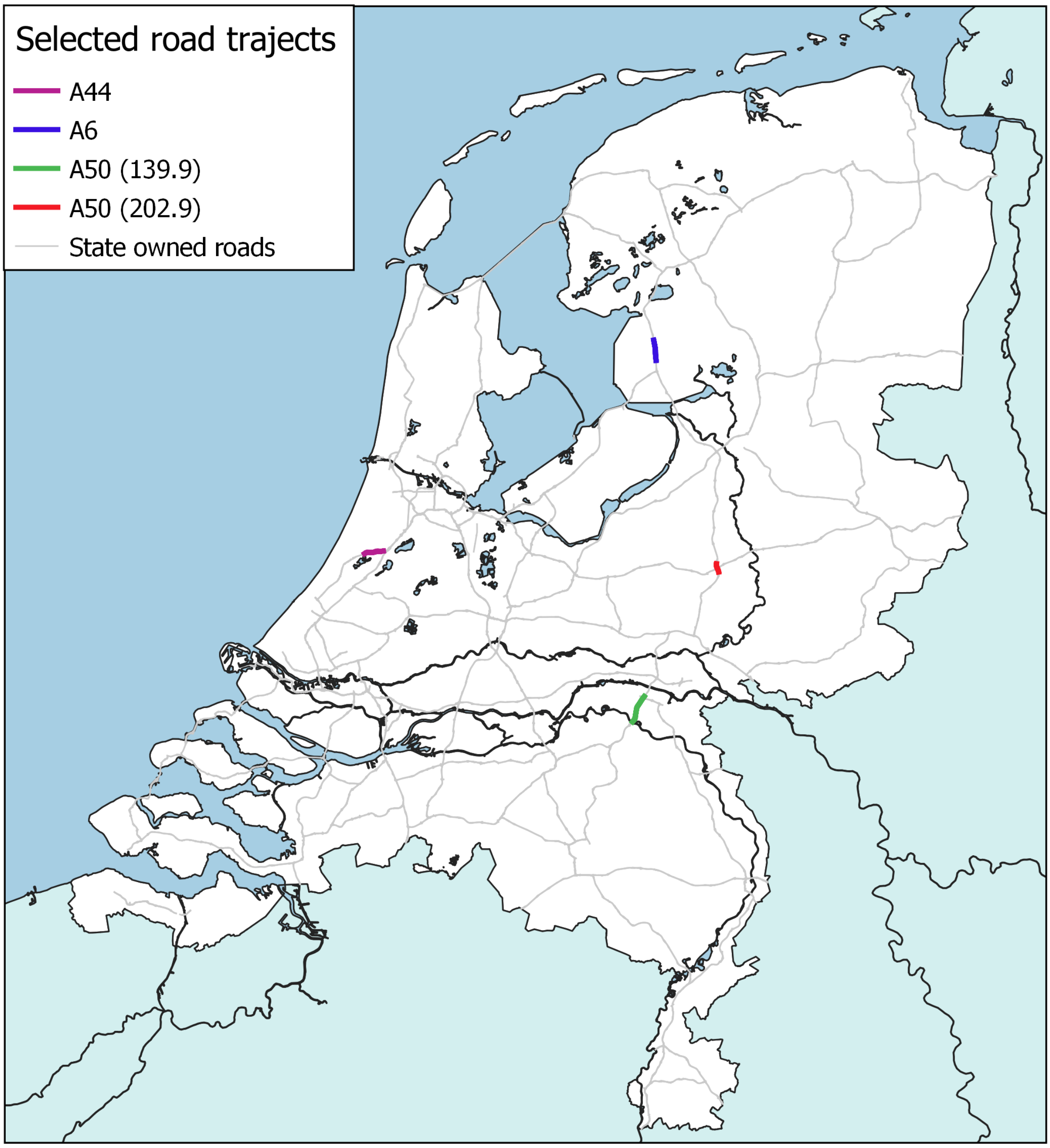

Figure 5.1: A map of the roads owned by the state in which the selected roads from Table 5.2 are highlighted.

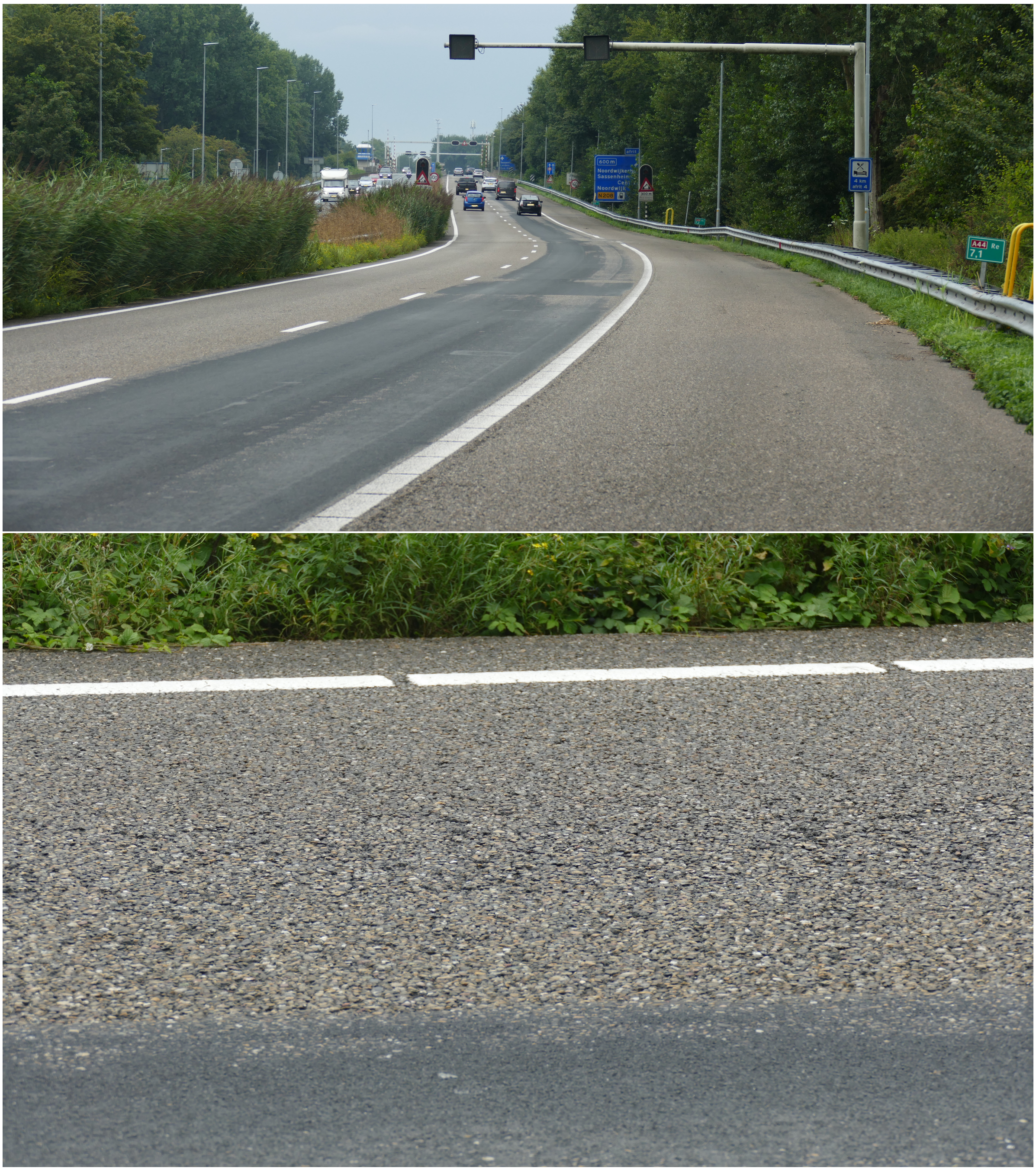

Figure 5.2: Illustration of (ravelling on) the road section starting at hectometer post 7.1 on the carriageway 1HRR of the A44.

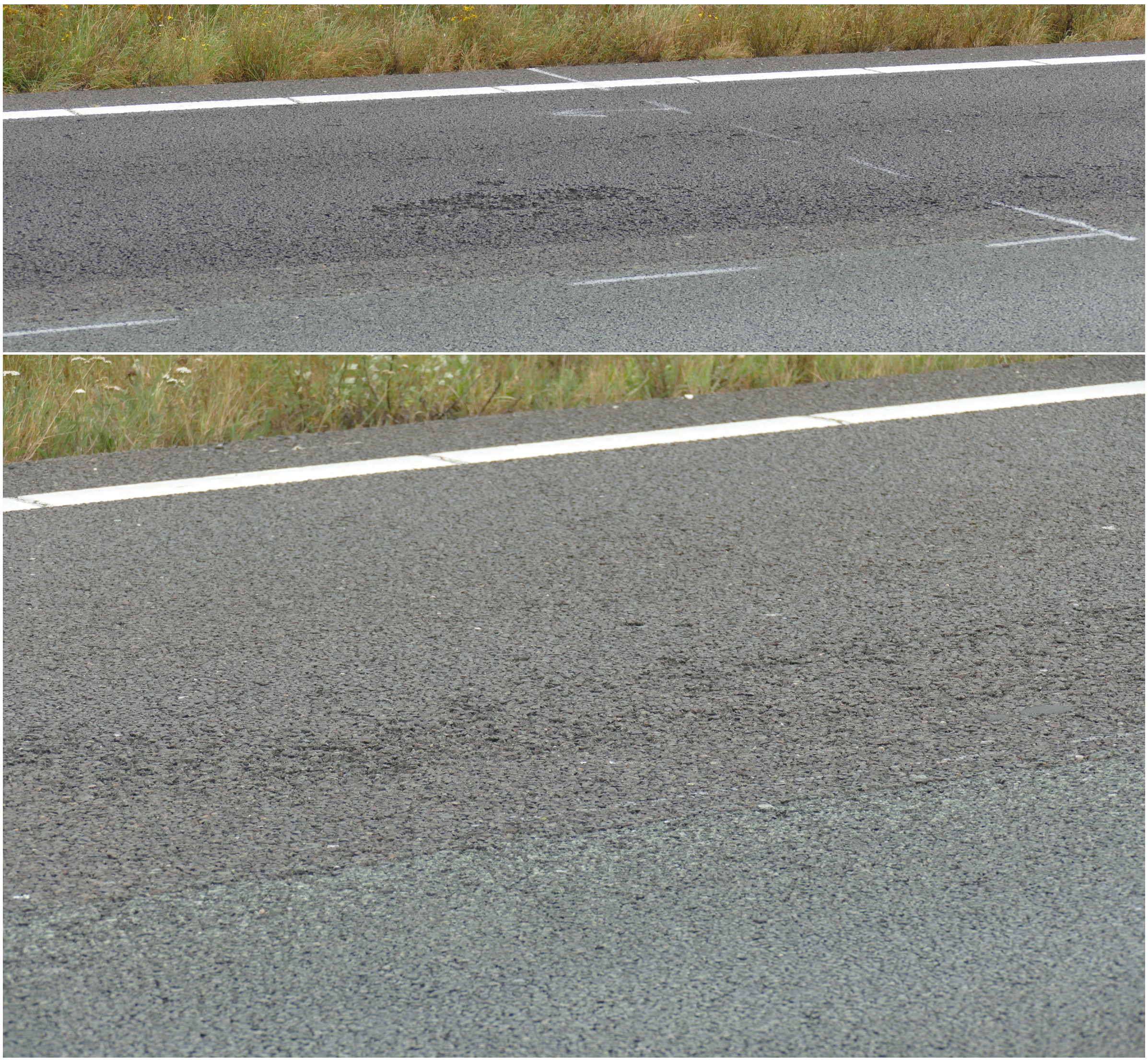

Figure 5.3: Illustration of ravelling on the road sections between driver location signs 203.3-203.5 on the carriageway 1HRR of the A50.

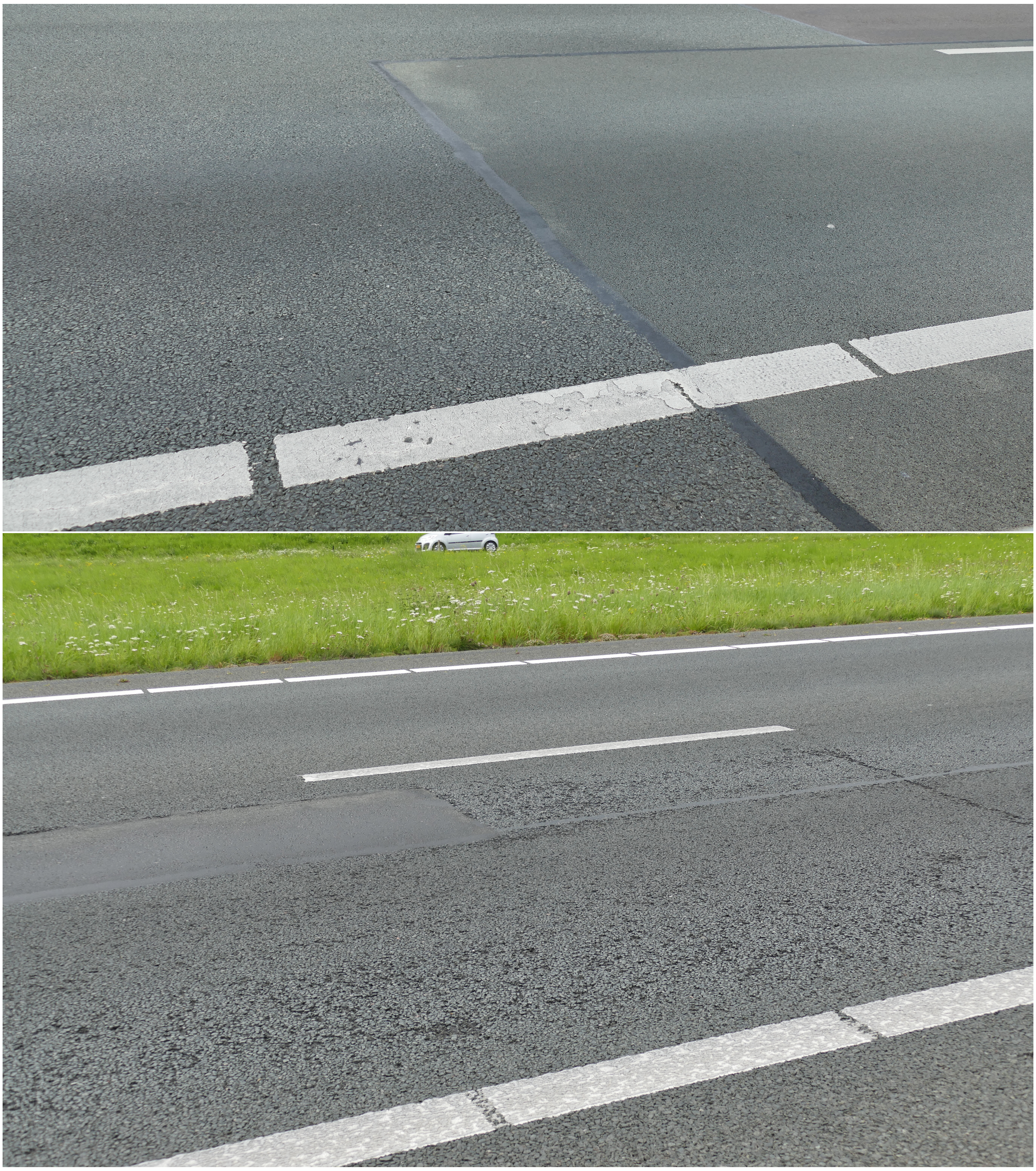

Figure 5.4: Illustration of transition in ravelling between two different surfaces and severe ravelling on the road sections starting respectively at hectometer post 282.7 and 287.3 on the carriageway 1HRR of the A6.

5.2. Detectie Oppervlakte Schade - Laser Crack Measurement System (DOS-LCMS)

According to Van Aalst et al. [3], DOS-LCMS data is based on 3D road surface generation which are measured by high-speed measuring of height profiles using laser triangulation. If generated road surface indicates cavities then the general assumption is that this is due to aggregates being lost.

5.2.1. Acquisition

The acquisition is briefly explained in Van Aalst et al. [3]. Before the 3D surface is generated, some pre-processing steps are necessary. This procedure consists of

- detecting roadmarkings,
- 'flattening' to compensate for road unevenness and vehicle motion,
- removing marks which do no correspond to ravelling damage,
- determining the wheel paths.

The wheel paths (LWT and RWT) are determined by finding the lateral location of maximum damage. When all these steps have been performed, the ravelling is calculated per wheel path per square meter by determining the surface area for which a coin can fit in the 3D surface. The application of a coin finds its use in its size with respect to the size of aggregates in PA. The latter ranges in diameter from 0 to 16 millimeters.

5.2.2. Format

The DOS-LCMS data was given in '.xlsx' files of which the filename had a pattern: [year of measurement]_[coin radius]_[depth coin]. Furthermore, interpreted as a dataframe, its dimensions are $n \times 310$, where n is the number of rows which differs per file. The 310 columns are named and explained in Table 5.3.

Table 5.3: Breakdown of the 310 columns in the DOS-LCMS files. The column names are primarily (based on) Dutch words, but the explanation clarifies what is meant by these.

	Column	Explanation
1.	Weg	Specified road of section
2.	Baan	Specified carriageway of section
3.	Strook	Specified lane of section
4.	HmStart	Starting driver location sign of section
5.	HmStop	Final driver location sign of section
6.	Vehicle	Which vehicle was used to measure the data; 20 = Kiwa-KOAC, 30 = Fugro, 1 = before 2019.
7.	Geldigheid	Remarks on validity of measurements
8.	Errorcode	Code for irregularities; 0 = normal, 10 = too many missing data points usually caused by wet roads
9.	lengte_meting	Length of measurement
10.	Datum_tijd	Date and time of measurement
11 - 110.	sv_i	Aggregate loss in percentages for the i-th meter for $i = 1,\dots,100$ measured across the width of the lane, which means sv_i is **not** the arithmetic mean of svL_i and svR_i
111 - 210.	svL_i	Aggregate loss for LWT in percentages for the i-th meter for $i = 1,\dots,100$

	Column	Explanation
211 - 310.	svR_i	Aggregate loss for RWT in percentages for the i-th meter for $i = 1,\ldots,100$

We would like to address the aggregate loss percentages across the width of the lane. According to Willem van Aalst, the values for lane-wide aggregate loss available to us are not fully pre-processed yet and are barely or not even used at all in general. That is why we will focus on the wheel tracks only. In addition, the threshold found in DHV et al. [12] mentions ravelling in the wheel tracks rather than lane-wide aggregate loss. A subset of a sample of the '2018_6_-2.xlsx' file is given in Table 5.4. It would have been impossible to try and fit complete rows due to the large amount of columns.

5.2.3. Complications

It is inevitable for data that is gathered in practice to be completely free of inconsistencies and such is also the case for the DOS-LCMS data. Data cleaning is required before any useful analysis can be performed and is therefore an important measure [15]. Rather than elaborating on the possible procedures for cleaning the data, for now we will plainly enumerate problems that we have encountered in the DOS-LCMS files. In Chapter 9 we will elaborate on an approach which could help in resolving the non-monotonicity in particular. Below we mention difficulties and why they are classified as such that we need to consider and treat carefully.

1. *Lengths of measurements throughout the years are not exactly the same.*
   Ideally, all the lengths would be fixed close to 100m but that is not the case. This means that assuming the true length of a road section is 100m, a measured length of 96.28m, every aggregate loss percentage corresponds to 0.9628m. However, not only does this cause noise longitudinally — meaning in subsequent road sections — but arguably more regretfully throughout the years for one specific section. In other words: a 1m section defined in 2012 does not necessarily preserve its location until 2019 or even earlier for that matter. Although some margin of error is inevitable, the presence of the inconsistency should still be acknowledged.

2. *No monotonic sequence of aggregate loss per meter per year*
   This obstacle is arguably the most intricate and compelling one. Indeed, it can be deduced from the data that for the i-th meter, there is no guarantee of an increasing percentage of aggregate loss over the years. Assuming that no maintenance nor any rejuvenation mechanism or product has been applied to the section, it defies the logic of the inevitable increasing — or non-decreasing to speak in more general terms — behaviour of aggregate loss and the ageing process. However, the DOS-LCMS framework models the road section in height and the cavities that the aggregates cause could be filled up by some other unknown substances — which in turn can reflect in lower aggregate loss than in reality. If we assume that this is not the main reason for the non-monotonic sequences, possible causes of this obstacle are expected to lie in the questioned precision of the GPS mechanism of the implementation combined with the 1st obstacle. The association of a road surface and the 'real' 100 meter section is not the same in every year: for example, whereas a measurement in 2015 of hectometre post 1.0 indicates the true road section starting at hectometre post 1.0, the measurement in 2016 might very well start 25 meters away from 1.0 (1.025) and then run until 1.125. Another possible explanation can be found in the varying measuring systems and the development of the algorithms to calculate ravelling.

3. *Several observations exists per year for one section.*

Table 5.4: Sample data from ‘2018_6_-2.xlsx’. Notice that columns sv(L/R)_2 – sv(L/R)_99 have been left out, as they would not have fit or the entire table would not have been readable on A4 paper.

Weg	Baan	Strook	HmStart	HmStop	Vehicle	Geldigheid	Errorcode	lengte_meting	Datum_tijd	sv_1	···	sv_100	svL_1	···	svL_100	svR_1	···	svR_100
R007	1HRR	2RR	37.5	37.6	1		0	99.94	27-Feb-2018 12:57:00	1.02		3.91	1.40		5.23	1.26		4.21
R007	1HRR	2RR	37.6	37.7	1		0	100.95	27-Feb-2018 12:57:00	4.28		4.11	5.39		3.55	4.64		4.98
R007	1HRR	2RR	37.7	37.8	1		0	102.98	27-Feb-2018 12:57:00	4.94		5.47	4.49		4.23	6.64		7.22
R007	1HRR	2RR	37.8	37.9	1		0	96.28	27-Feb-2018 12:57:00	4.19		5.03	2.53		4.08	6.14		7.09
R007	1HRR	2RR	37.9	38	1		0	101.56	27-Feb-2018 12:57:00	4.89		4.11	3.79		4.09	6.94		4.51
R007	1HRR	2RR	38	38.1	1		0	100.44	27-Feb-2018 12:57:00	3.81		3.07	3.65		4.34	4.76		2.83
R007	1HRR	2RR	38.1	38.2	1		0	101.20	27-Feb-2018 12:57:00	2.64		3.01	2.90		2.35	3.14		4.98
R007	1HRR	2RR	38.2	38.3	1		0	100.70	27-Feb-2018 12:57:00	2.85		4.36	3.07		5.92	3.70		3.47
R007	1HRR	2RR	38.3	38.4	1		0	99.43	27-Feb-2018 12:57:00	4.86		3.11	6.57		3.52	4.23		3.46
R007	1HRR	2RR	38.4	38.5	1		0	97.40	27-Feb-2018 12:57:00	2.94		5.28	3.60		6.36	2.86		4.99
R007	1HRR	2RR	38.5	38.6	1		0	102.98	27-Feb-2018 12:57:00	5.38		3.54	7.22		4.49	4.44		3.36
R007	1HRR	2RR	38.6	38.7	1		0	108.56	27-Feb-2018 12:57:00	3.61		3.81	4.03		5.27	3.58		3.27
R007	1HRR	2RR	38.7	38.8	1		0	92.33	27-Feb-2018 12:57:00	3.76		3.06	5.80		3.12	2.81		3.46
R007	1HRR	2RR	38.8	38.9	1		0	99.43	27-Feb-2018 12:57:00	3.08		4	3.19		4.20	3.72		4.44
R007	1HRR	2RR	38.9	39	1		0	98.92	27-Feb-2018 12:57:00	3.66		2.87	3.60		2.34	4.22		3.73
R007	1HRR	2RR	39	39.1	1		0	101.46	27-Feb-2018 12:57:00	2.70		2.71	2.94		3.01	3.20		2.99
R007	1HRR	2RR	39.1	39.2	1		0	100.44	27-Feb-2018 12:57:00	3.03		2.36	2.98		2.97	3.69		2.45
R007	1HRR	2RR	39.2	39.3	1		0	98.92	27-Feb-2018 12:57:00	2.34		2.90	2.93		4.32	2.45		2.85
R007	1HRR	2RR	39.3	39.4	1		0	99.94	27-Feb-2018 12:57:00	3.09		2.56	4.33		2.91	2.91		2.96
R007	1HRR	2RR	39.4	39.5	1		0	100.95	27-Feb-2018 12:57:00	2.51		3.33	2.75		3.56	2.99		3.92

It is possible for one road section to have more than one measuring date. For example, for the A7 - 1HRR - 37.5 - 1RR in 2018 there is DOS-LCMS data from February 8th, 13th and 26th. TNO has clarified that the latest measurements are generally the correct ones. Hence this obstacle is not as laborious to deal with, but it should still be dealt with.

5.3. KernGIS

Geographical information system (GIS) is a framework which attempts to capture spatial and geographical data. KernGIS is the database that Rijkswaterstaat uses in which many administrative information is saved and frequently updated. For our purpose it is a useful tool to find the positions of engineering structures. The parts of the road on which these structures lie should be excluded from analysis, as asphalt surfaces on engineering structures should be treated differently. The reason for that is the build-up of wear being different on engineering structures, which could eventually result in less representative figures for ravelling on said road segments. As an example, we will provide an overview of parts from the A44 which have to be excluded for analysis in Table 5.5.

Table 5.5: Parts of the A44 which have to be excluded from analysis.

DLS	Engineering Structure
2.3 - 2.4	viaduct
5.9 - 6.1	bridge
7.5 - 7.7	bridge

5.4. Strategic Planning Road Surfaces

Strategic Planning Road Surfaces (SPRS) [16] contains administrative information on when which parts of the road need maintenance and the financial figures required for execution. The SPRS aims to adhere to quality conditions that serve as safety measures for road users. SPRSs are established partly by expert advisors from Rijkswaterstaat who aim for efficiency across all factors to be considered, such as costs and minimal traffic disruption. The process of establishing an SPRS is described in [16] and is not within our scope of research, apart from the fact that the SPRS recommended maintenance years are dependent on DOS-LCMS measurements as of now. The model from TNO maps the measurements to proposed remaining lifetimes which are considered when the SPRS is formed. Frankly we do not need the proposed remaining lifetimes from the SPRS for our research, but it would be ignorant to not be aware that the current proposed remaining lifetimes are already (partially) based on DOS-LCMS data. It should also be noted that ravelling is not the only damage mechanism which prompts maintenance, so not all lifetimes are based on the state of ravelling. However, in the SPRS the proposed remaining lifetimes are given for each damage mechanism.

6 Methodology

In the remaining Chapters, we will be elaborating on how we establish a method to predict the moment when a road section needs to be resurfaced according to officially documented RWS standards [6]. Recall from Chapter 3 that if more than 25% of the road section measures aggregate loss over 10%, the corresponding road section needs maintenance. In mathematical terms it is equivalent to the *75th percentile* of our data being at least 10%. For the predictions, two fundamental steps need to be considered:

1. Finding the 75th percentile based on the data.

2. Fit the progression of 75th percentiles to a monotonic increasing convex curve.

The first step can be handled in two approaches: the parametric approach and the non-parametric approach. The parametric approach in Chapter 7 is based on the assumption that the data is from some parametric distribution and calculates the 75th percentiles based on the functional properties of the distribution. The non-parametric approach in Chapter 8 does not make any parametric assumptions of distributions. Both approaches view the data at a 100 meter section level and therefore do not explicitly look at the progression on 1 meter level. Additionally, both approaches are *approximations* of the 75th percentile. Naturally there is no guarantee for finding the true 75th percentile.

The second step limits itself to a monotonic curve which abides the at first logical assumption of aggregate loss being monotonically increasing. Furthermore we set a convexity constraint because visual inspectors have seen that the severity of ravelling gradually increases. Using road sections which have been ravelled intensively enough, we can extrapolate to give an indication of when maintenance is required, and compare extrapolations based on the total (most recent) n consecutive percentiles and with $n-i$ percentiles for $i < n$ and i not too large. For example, using the most recent $n = 7$ percentiles could result into a remaining lifetime of 100 days, while using $n-i = 6$ percentiles for $i = 1$ could result into a remaining lifetime of 300 days. That means the prediction needed to be corrected for by 200 days. Let us refer to this approach as the $n-i$ approach hereafter.

Prototyping

Accompanying Chapters 7 and 8 is the data of road section A44 - 1HRR - 7.1 - 1RR which was mentioned as example of BPS notation in Chapter 4. This is not without reason, as this specific road section satisfies the conditions of being a 'proper' prototype to explain the mentioned concepts. The bare minimum conditions would be:

- The intensity of ravelling surpasses or is close to the 10% threshold for the empirical 75th percentile in some year, which allows for the $n-i$ approach for at least $i = 1$.

- Corresponding with the preceding condition, the data admits at least $n = 5$ values of consecutive $q_{0.75}$ estimates for curve fitting.
- The data contains Obstacle 2 of § 5.2.3, such that a workaround can be presented.

A condition such as there being no maintenance executed is not a necessity but would be a convenience by having more quantiles to work with, so it is not stated as a bare minimum condition. However, it was found to be problematic at times to find a monotonically increasing convex curve to less than 4 points. Further research based on road sections which have been resurfaced a relatively long time ago (around 2012-2015) can therefore be troublesome. An important note about the road section is that we have the required measurements from years 2012 – 2019 except for 2013. DOS-LCMS measurements were not gathered in that year for this road section, but this is not an issue for our research and will not be for future research based on the presented work.

With respect to the A44 there were not many road sections in general which we would classify as a proper prototype. In particular the A44 - 1HRR - 7.1 - 1RR is one of the only 7 road sections that have been unaltered in asphalt surface since the pavement rehabilitation in 2002. All other sections have had some sort of maintenance which altered the original surface in 2002, rendering the DOS-LCMS measurements less effective in setting up a lifetime prediction framework. Road sections which have undergone maintenance after the DOS-LCMS measurement date in 2019 could still suffice. However, any road section of which the DOS-LCMS measurements indicate that $q_{0.75} < 10$ could in theory suffice for a lifetime prediction. In that regard even road sections which have recently been resurfaced could be used given that the DOS-LCMS data is available. Whether such a prediction on newly resurfaced parts of the road is sensible to do remains questionable.

Quantiles

The aim is to make an accurate prediction of when the 75th percentile reaches the threshold of 10. Recall that a percentile is related to the more general *quantile*. In particular, for a continuous distribution, a quantile q is a real number that divides the area under the PDF in two parts of set amounts. More generally speaking, let $p \in (0,1)$ and denote q_p as the p-th quantile. For a continuous random variable X with CDF F_X, the p-th quantile is defined to be a value which solves Equation (6.1).

$$F_X(q_p) = p \tag{6.1}$$

For our analysis we will be dealing with continuous random variables and as a consequence, the continuity assumption required for the existence of a solution to Equation (6.1) is satisfied. Equation (6.1) allows us to deduce that finding the 75th percentile is equivalent to solving Equation (6.2)

$$F_X(q_{0.75}) = 0.75 \tag{6.2}$$

In particular for location-scale families, we can denote the p-th quantile q_p as

$$q_p = \mu + \Phi^{-1}(p)\sigma \tag{6.3}$$

where μ and σ respectively are a location and scale parameter[1], and $\Phi^{-1}(p)$ denotes the inverse CDF of the standardised form ($\mu = 0, \sigma = 1$) of the location-scale distribution evaluated at p, or in other words, the $100p$-th percentile of the standardised version of the location-scale distribution. A well-known case is the normal family of distributions. It is common to see Φ denoting the CDF of the standard normal. By concept of Equation (6.3) it is interesting to pursue an approach based on location-scale families amongst others.

[1] μ, σ are *not* necessarily the mean and standard deviation.

Reproduciblity

All analyses have been performed using `R` on version 3.6.1, with packages on their respective versions on 19-11-2019 for the sake of reproducibility. The majority of — if not all — the Figures and Tables in the remaining Chapters can be generated with code provided on this Github page. If the reader would like to reproduce the results, we highly recommend to also install `RStudio`, an integrated development environment for `R`, and to contact Rijkswaterstaat for the terms under which the data can be made available.

7
Modelling: Parametric Approach

Now that we have given an overview of our available data and explained the methodology, we are able to formally introduce mathematical notation and concepts. For this Chapter we will explain concepts with respect to *parametric statistics*. Broadly speaking, the parametric approach for predicting lifetimes is to find parametric distributions which fit our data. Although the idea of fitting distributions might sound elementary, the possible variations and modifications for this process are plentiful.

7.1. Notation and Background

Let W be a road and S_j for $j = 1, 2, \ldots, N_W$ for some $N_W \in \mathbb{N}$ its 100m road sections. For every S_j, we have data from 2012 – 2019[1] for the aggregate loss values per position, per track. That is, for every S_j for every year, we have 3 sets of observations

$$\begin{aligned}\mathbf{x} &= \{x_1, \ldots, x_{100}\}\\ \mathbf{x}^L &= \{x_1^L, \ldots, x_{100}^L\}\\ \mathbf{x}^R &= \{x_1^R, \ldots, x_{100}^R\}\end{aligned}$$

where x_i, x_i^L, x_i^R for $i = 1, \ldots, 100$ respectively denote the aggregate loss percentage lane-wide, on the LWT, and on the RWT for the i-th meter. However, recall that we will focus only on the latter two. For the remaining 2 sets of 100 observations per year that can be considered, we are interested in their respective cumulative distribution functions (CDF) $F_L(x), F_R(x)$ and corresponding probability density functions (PDF)[2] $f_L(x), f_R(x)$. In particular, in this Chapter we assume that any F is the CDF corresponding to P_θ, where P_θ is a distribution and p_θ the corresponding PDF which belongs to a *parametric family*

$$\mathscr{P} = \left\{p_\theta \mid \theta \in \Theta\right\}, \tag{7.1}$$

where $\theta \in \Theta$ is a parameter and $\Theta \subseteq \mathbb{R}^k$ is a *parameter space*.

Example 7.1 (Parametric Families)**.** *Frequently used parametric families can be divided into several groups based on different characteristics. The amount of parameters that fix the distribution and the distribution being continuous or discrete are the obvious ones. We list cases of parametric families of continuous distributions.*

[1] Available data can slightly vary per road.

[2] We do not consider probability mass functions (PMF) as our data is continuous.

The family of exponential distributions[3] *is parametrised by $\theta = \lambda > 0$ with*

$$\mathscr{P} = \left\{ f(x \mid \lambda) = \lambda e^{-\lambda x} \mid x \geq 0,\, \lambda > 0 \right\} \tag{7.2}$$

The family of normal distributions is parametrised by $\theta = (\mu, \sigma)$ where $\mu \in \mathbb{R}$ and $\sigma > 0$ are the respective location and scale parameters with

$$\mathscr{P} = \left\{ f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) \mid x, \mu \in \mathbb{R},\, \sigma > 0 \right\}. \tag{7.3}$$

In particular, the normal family is also a location-scale family.

The family of logistic distributions is parametrised by $\theta = (\mu, s)$ where $\mu \in \mathbb{R}$ and $s > 0$ are the respective location and scale parameter with

$$\mathscr{P} = \left\{ f(x \mid \mu, s) = \frac{\exp(-(x-\mu)/s)}{s(1+\exp(-(x-\mu)/s))^2} \mid x, \mu \in \mathbb{R},\, s > 0 \right\}. \tag{7.4}$$

Similar to the normal family, the logistic family is a location-scale family.

The generalised gamma distribution is parametrised by $\theta = (a, d, p)$ where $a, d, p > 0$ are its parameters and

$$\mathscr{P} = \left\{ f(x \mid a, d, p) = \frac{p/a^d}{\Gamma(d/p)} x^{d-1} e^{-(x/a)^p} \mid x, a, d, p > 0 \right\} \tag{7.5}$$

where Γ denotes the gamma function.

There are several reasons to consider a so-called parametric model, but for our case the convenience and interpretation are what stand out the most. Although convenience is a subjective term, the accessibility of established methods such as goodness-of-fit provide proper means of justifying steps in analysis. For example, the *Anderson-Darling* test [17] and *Shapiro-Wilk* test [18] are known to be among the most powerful tests for normality [19] and could therefore be of use in justifying the assumption of normality.

Correlation and Dependence

An important note about our DOS-LCMS data is the possibility of presence of sample *autocorrelation.* Whereas regular correlation is measured between two different variables, autocorrelation is a measure of dependency between one variable and a τ-lagged version of itself. For an ordered set of observations X_k with $k \geq 1$, the τ-lagged version of X_k is given by $X_{k-\tau}$ for $k > \tau$. The potential problem of a high degree of serial correlation clashes with the assumption of independent observations, which is frequently used in statistical theory. That is in our context, if we have information of aggregate loss in the second meter x_2, a strongly autocorrelated sample would imply that aggregate loss in the third meter x_3 (as an example) is highly dependent on x_2 and vice versa. Notice that τ in our context indicates the jump in 1m sections. Sample autocorrelation function (ACF) plots could help us in detecting if there is any dependency for lag τ, where τ denotes the order of the considered difference. For any set of observations, the autocorrelation at lag $\tau = 0$ is 1, since logically it compares two equal sets. For an i.i.d. (independent and identically distributed) set of observations and $\tau > 0$, however, all autocorrelations should not differ significantly from zero for the i.i.d assumption to hold.

For the remainder, we assume *stationarity* of the data, such that sample autocorrelations are only influenced by the lag. Let us illustrate ACF plots for an i.i.d. sample and the $W = A44, S = 1\text{HRR}^{1\text{RR}}_{7.1}$, which can both be assumed stationary. For the i.i.d. sample we consider a random sample of size $n = 100$ from a standard uniform distribution $U(0,1)$. Its sample autocorrelation function is given in Figure 7.1. It indeed shows that at lag $\tau = 0$ the sample is correlated with itself and for larger

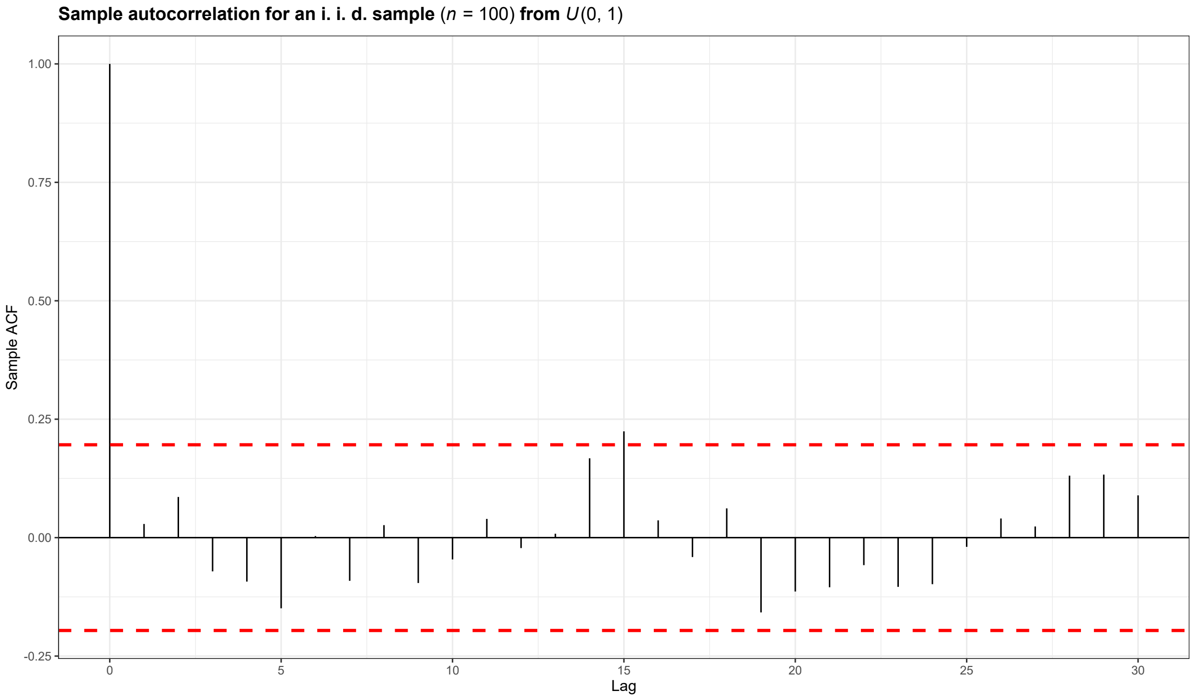

Figure 7.1: Sample autocorrelation function plot for a random sample ($n = 100$) from $U(0,1)$ where the red lines indicate the 95% confidence interval for ρ.

values of τ, it is generally well within the 95% confidence interval of ρ. If we take our example road section, we end up with Figure 7.2. Some key points from Figure 7.2 are the following:

- there indeed seems to be some form of serial dependence,
- the LWT seems to be most consistent with a lag of around $\tau = 3$,

Let us elaborate on these findings. The first point can be argued by observing that the general pattern we see in Figure 7.2 shows that the autocorrelation at lag $\tau \in \{1,2,3\}$ are not within the 95% confidence interval (red dashed lines) for ρ. The second point is established by noticing that for the LWT, $\tau = 3$ is within or close to the confidence interval of being statistically significantly near 0; for the other cases this is not as apparent and consistent. What this means for the continuation of the parametric approach are several things. First of all, due to serial dependency we should be very careful with relying on the drawn conclusions and results. However, we have only considered one road section. If for other road sections — which we unfortunately cannot all elaborate on in great detail — the autocorrelation is less present, then the results from the parametric approach can be more relied on. For the remainder of the Chapter we will continue using the i.i.d. assumption, while being mindful of the serial dependence.

7.2. Finding a suitable distribution

The essence of finding a suitable distribution is the following: it provides a method of approximating the true 75th percentile. In turn, as a consequence of using parametric models, it allows for an exact and closed-form expression of the 75th percentile which is our statistic of interest, but the exact and closed-form expression do not necessarily have to be utilised.

[3] The family of exponential distributions are part of the more general class of exponential families, which in terminology is not to be confused with parametric families in general.

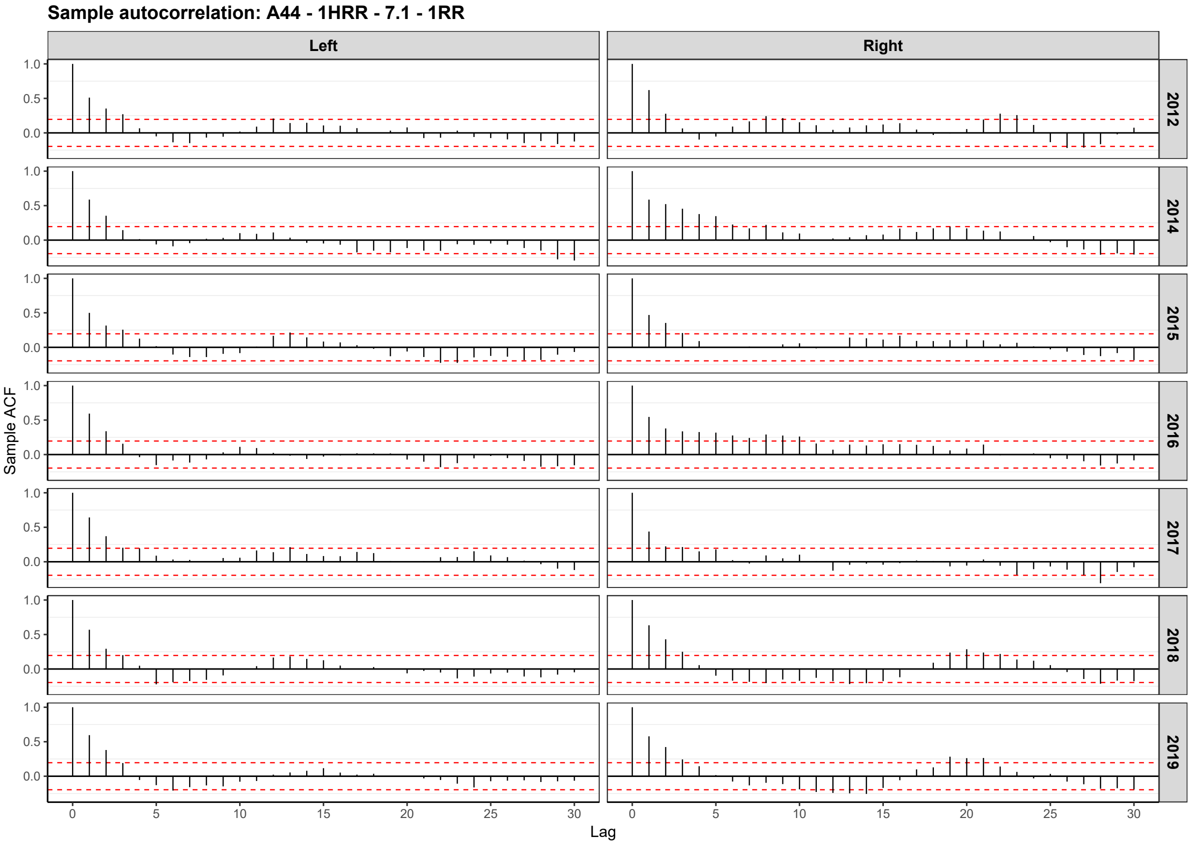

Figure 7.2: Plots of autocorrelation functions for $W = A44, S = 1\text{HRR}^{1\text{RR}}_{7.1}$ from 2012 – 2019 except for 2013 faceted on track and year. $\tau > 30$ has been checked but showed no significant result, hence those are not included and cause no clutter.

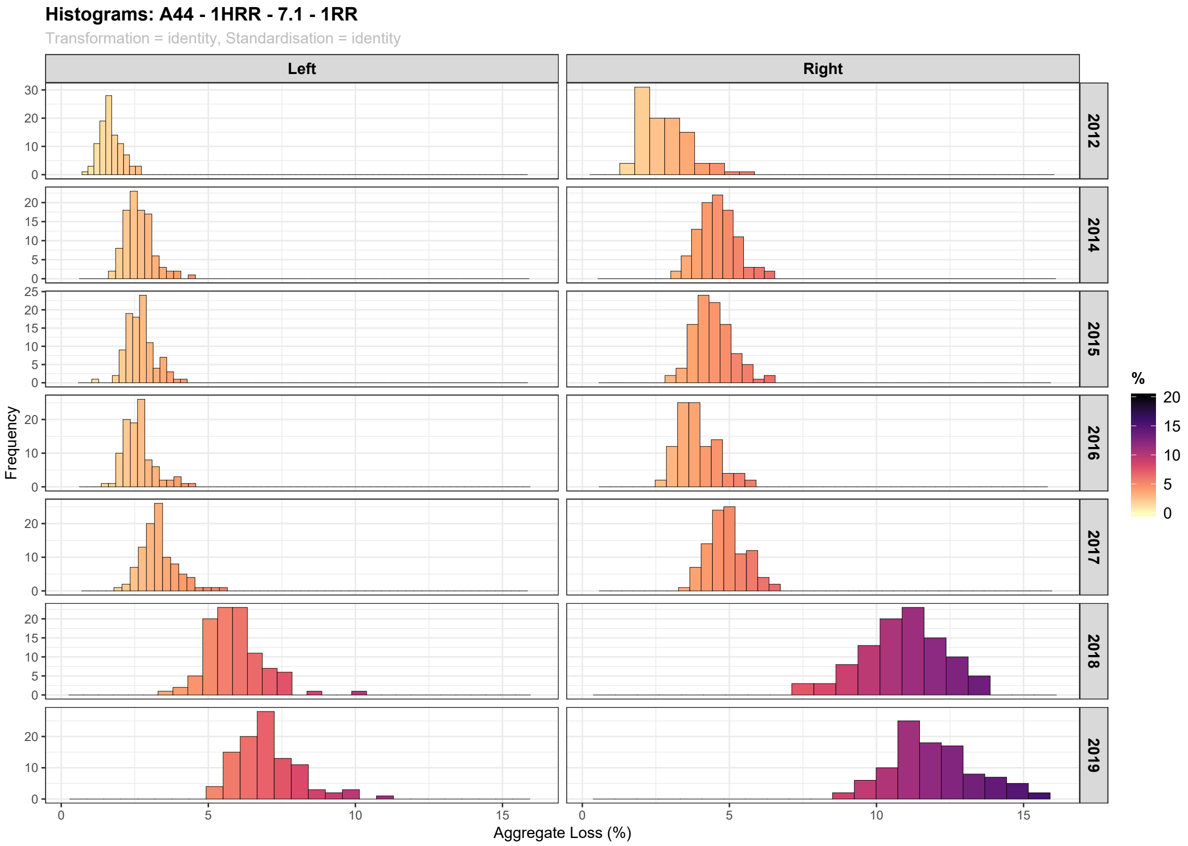

Figure 7.3: Histograms for the progression of aggregate loss on $W = A44, S = 1\mathrm{HRR}_{7.1}^{1\mathrm{RR}}$ from 2012 – 2019 except for 2013.

7.2.1. Histograms

To find suitable distributions, a first elementary attempt would be to observe the respective histograms. Figure 7.3 shows the plotted histograms for the $W = A44$ and $S = 1\mathrm{HRR}_{7.1}^{1\mathrm{RR}}$, ordered by year and position.

Illustrations such as Figure 7.3 allow for a brief overview of the road section and the progression of aggregate loss throughout the years per position. The key points for this specific section are the following:

- there is a clear difference in magnitude between the LWT and RWT,
- in general the histograms *seem* to have a certain degree of symmetry but also show some right-skewness,
- there is no clear monotonic progression of aggregate loss as presumed in § 5.2.3,
- although the non-decreasing behaviour is not apparent, over the course of 7 years (2012 – 2019) the aggregate loss has increased.

The third key point requires some further explanation; if the aggregate loss values per meter would be monotonic throughout the years, the histograms as depicted in Figure 7.3 would need to shift towards the right, or at the very least stay in the exact same position in subsequent years. In 2015 – 2016 the shift towards the left is arguably the most clear for both wheel tracks, which means it invalidates the implied shift towards the right if the sequences were monotonic. Another way to look at it is to only observe the minimum and/or maximum bin values. By definition, the minimum and/or maximum bin value needs to increase in subsequent years for monotonicity per meter.

Although it is quite effortless to interpret a plot such as Figure 7.3, there are pivotal remarks about the current representation and the approach of histograms to find suitable distributions. First of all, the perspective can change considerably when changing bin widths. To accommodate, the bin widths used in Figure 7.3 are based on the *Freedman-Diaconis rule* [20]. It is based on the following formula for the bin width:

$$\text{bin width} = 2n^{-1/3} \cdot \mathrm{IQR}(\mathbf{x}) \tag{7.6}$$

where n is the sample size and $\mathrm{IQR}(\mathbf{x})$ the interquartile range of sample $\mathbf{x} = \{x_1, \ldots, x_n\}$. Equation (7.6) implies that the bin width[4] increases if the sample size n decreases and/or the $\mathrm{IQR}(\mathbf{x})$ increases. As $n = 100$ for all cases, it means a sample $\mathbf{x}_H$ with higher IQR relates to a larger bin width compared to a sample $\mathbf{x}_L$ with lower IQR. For example, the RWT has clear distinct bin widths for the years 2017 and 2018. Nevertheless, it must be said that there is not necessarily a rule for bin widths which would or could completely dispose of every skewed view of the data, as is also the case here.

Secondly, the collection of histograms as of now do not properly show off each individual histogram. This is not necessarily a problem and was to be expected, as the idea of the format is to provide a quick glance at the data for one road section. The main cause of this problem is due to the histograms being plot for a fixed horizontal domain, namely on $[0, c]$ for some real $c > 15$, linked with the varying bin widths from the first remark. The link can be explained: suppose that we would like to know if the data is normally distributed purely by observing the histograms; depending on the relative size of the histogram, some histograms might ‘tend more towards’ a normal distribution. To clarify: the histogram of 2018 from the RWT is relatively great in plot size and that helps to classify it as a reasonable histogram of a sample from a normal distribution at first glance. The histogram of 2012 from the LWT, however, is relatively much smaller in size, but could very well be a sample from some normal distribution as well. Figure 7.3 does not portray which sample of these two would be a better fit to some respective normal distribution.

7.2.2. Data Transformation

Although histograms directly plot from the non-transformed data are clearly lacking in ultimately deciding which parametric family suits best, it does allow us to find simple graphically interpretable characteristics such as the presence or lack of symmetry. As mentioned before, the varying bin widths and the canvas on which all of the histograms were plot are a problem in deciding a suitable distribution based on the entirety of the histograms. One way to solve this issue is to transform the data; doing so will result in more comparable histograms. There are numerous methods to transform data with, so it would be impossible to mention them all. Regardless, we will try to give an idea of available techniques. All of these are based on a (continuous) function transform $t(\cdot)$ on a sample $\mathbf{x} = \{x_1, \ldots, x_n\}$ which can be written with slight abuse of notation as

$$t(\mathbf{x}) = \{t(x_1), ..., t(x_n)\}.$$

Standardisation

Let us start with a well-known method which is likely familiar to anyone who is acquainted with some introductory course about Probability and Statistics: standardisation. It is a method which combines a location shift with a scaling adjustment. For a sample $\mathbf{x} = \{x_1, \ldots, x_n\}$, the standardised value z_i of an observation x_i is given by

$$z_i = t(x_i) := \frac{x_i - \bar{x}}{\mathrm{sd}(\mathbf{x})}, \tag{7.7}$$

where $\bar{x}$ and $\mathrm{sd}(\mathbf{x})$ are respectively the sample mean and sample standard deviation. We will demonstrate a possible application of standardisation.

[4] This should not be confused with the *amount of bins*, which is dependent on the spread of the data.

Example 7.2 (Exam Scores)**.** *Consider a French French teacher working in a Dutch secondary school which attempts to implement a French assessment style. The conventional Dutch scoring norm ranges from 1 to 10, but the newly styled exam ranges from 1 to 20 as is usual in France. Two of the teacher's students, Jeremy and Viola, are in different classes: Jeremy (class A) takes a regular-assessed exam, while Viola (class B) takes the French-styled exam. After taking the exams, the teacher would like to know which of Jeremy (score 9.0) and Viola (score 18.0) performed better relative to their classmates. The teacher acquired the sufficient data to answer the question and denotes the following*

$$\begin{aligned} \bar{x}_A = 6.1, \quad \mathrm{sd}(\mathbf{x}_A) = 1.2 &\implies z_J = \frac{9.0 - 6.1}{1.2} \approx 2.42 \\ \bar{x}_B = 14.6, \quad \mathrm{sd}(\mathbf{x}_B) = 1.4 &\implies z_V = \frac{18.0 - 14.6}{1.4} \approx 2.27 \end{aligned}$$

The teacher concludes that relative to their respective classmates, Viola performed better than Jeremy.

Example 7.2 does not tell us much more without further information on the data. We could apply it to our data to find out if the i-th meter in year t_1 has suffered relatively more aggregate loss than in year t_2 compared to the other meters, but this is not exactly what we are looking for now — although it could be interesting for future research. Instead, we would like to find out more about the distributions when we standardise the data. However, rather than directly applying the introduced standardisation, we will make use of its robust counterpart:

$$z_i = t(x_i) := \frac{x_i - \mathrm{median}(\mathbf{x})}{\mathrm{MAD}(\mathbf{x})} \tag{7.8}$$

where $\mathrm{MAD}(\mathbf{x})$ is the *median absolute deviation* of sample $\mathbf{x}$ defined by

$$\mathrm{MAD}(\mathbf{x}) = \mathrm{median}_i(|x_i - \mathrm{median}(\mathbf{x})|). \tag{7.9}$$

Notice how this transformation is similar to the classical standardisation. The advantage of such robust estimators is that they work better than the classical ones for data which contain outliers. If we apply Equation (7.8) to the $W = A44, S = 1\mathrm{HRR}_{7.1}^{1\mathrm{RR}}$, we obtain Figure 7.4.

Figure 7.4 makes it much more workable to detect a potentially suitable family of distributions which fit our data. One major difference between Figures 7.3 and 7.4 apart from the scale, is the bin widths. In Equation (7.6), we know that sample size n is constant. Yet now as a result of standardisation, the IQR of every sample will generally be relatively close to each other. Any difference between these would likely be unnoticeable at first glance due to its magnitude.

Recall that we wish to find families of distributions for each of the LWT and RWT purely based on the available data. Whether this is justifiable on physical basis can be argued, but based on the data, the family of distributions between these positions do not have to coincide. For the sake of explaining the concept, let us start with viewing the data for the LWT. From Figure 7.4 it appears to be that most of the histograms have a degree of positive skewness, and this is actually not even exclusive to the LWT. Now, there are multiple ways to proceed in our pursuit of which we will name two:

1. Perform a transformation prior to standardisation

2. Fit asymmetric distributions to the standardised data

For the remainder we will only work with the first approach. Furthermore, hereafter we refer to 'standardisation' as standardisation using the median and the median absolute deviation.

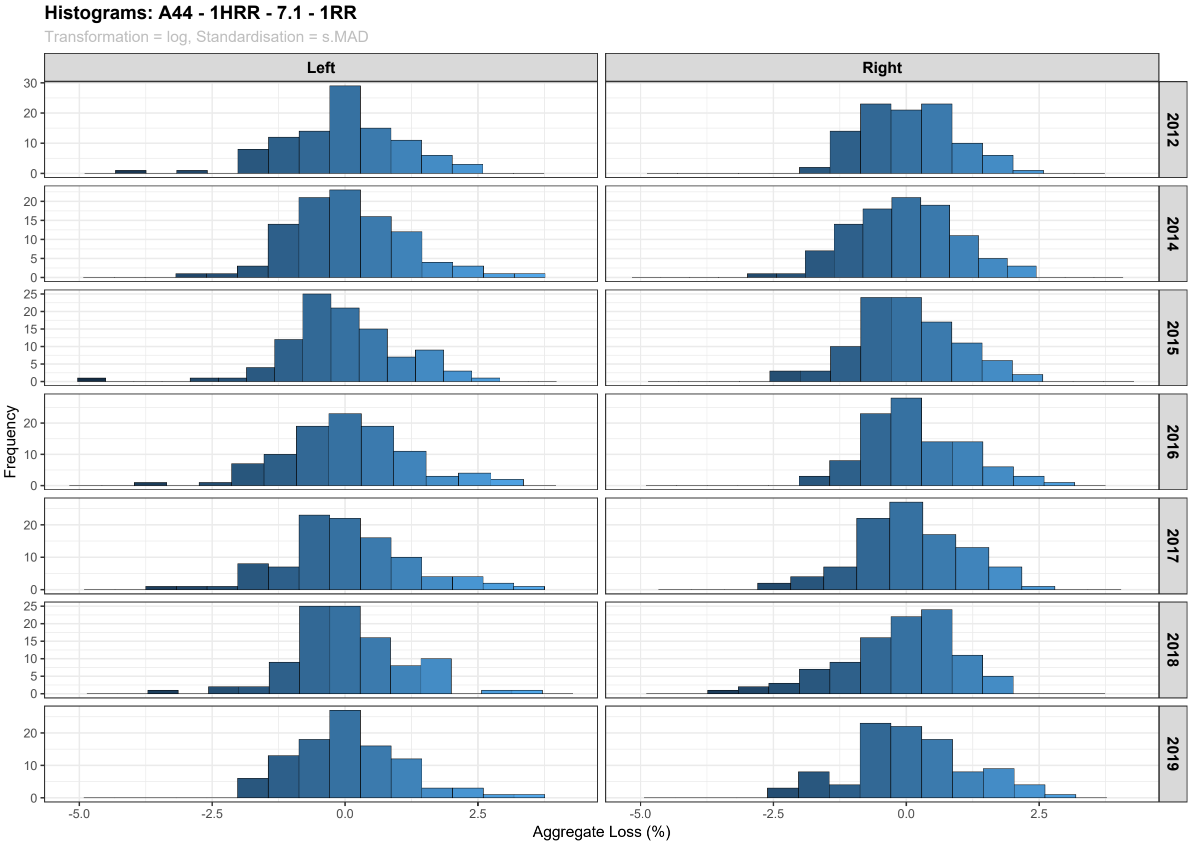

Figure 7.4: Histograms for the standardised values of the progression of aggregate loss on $W = A44, S = 1\text{HRR}_{7.1}^{1\text{RR}}$ from 2012 – 2019 except for 2013.

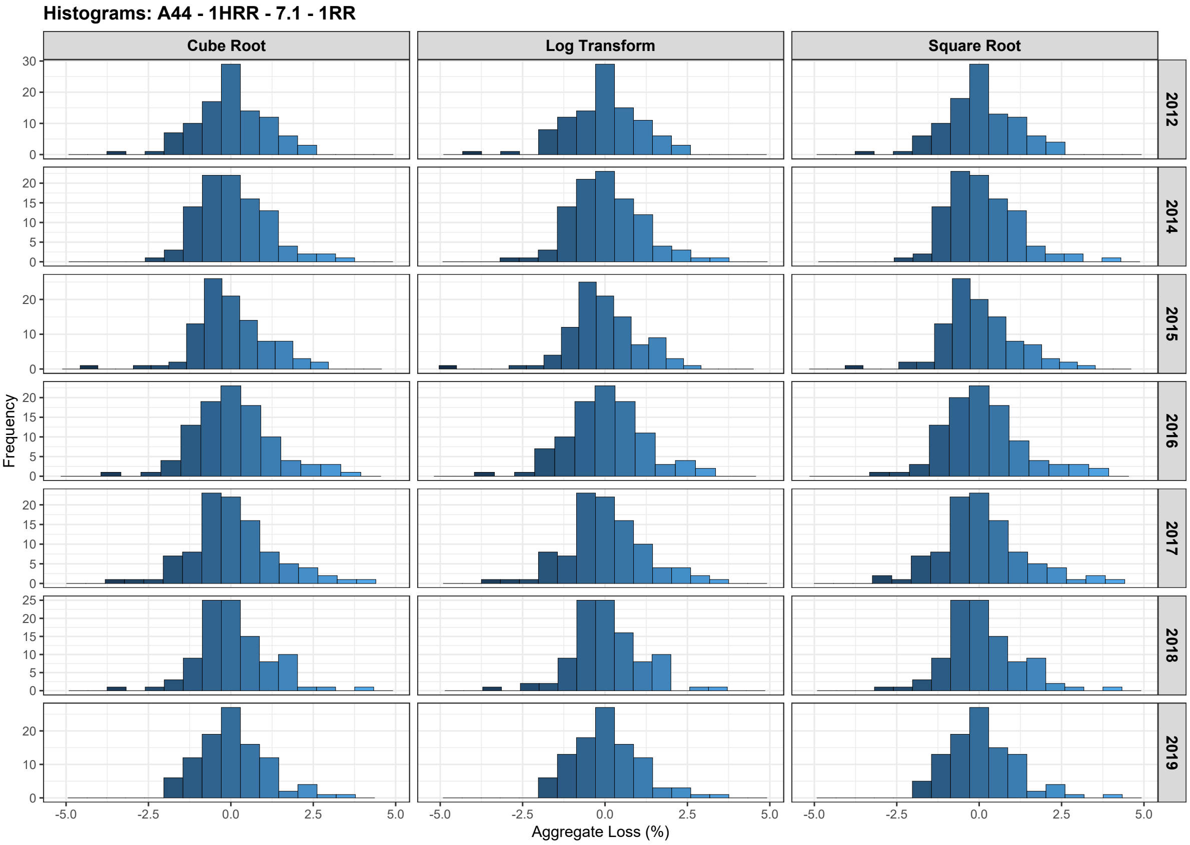

Figure 7.5: Histograms of transforms prior to standardisation of of the progression of aggregate loss on the LWT of $W = A44, S = 1\text{HRR}_{7.1}^{1\text{RR}}$ from 2012 – 2019 except for 2013.

Transformation Prior to Standardisation

One of the mentioned key points of Figure 7.3 were that the histograms seemed to have a certain degree of symmetry. Conversely, this was merely based on the first impression. With the new representation in Figure 7.4, multiple histograms tend to look more right-skewed than initially observed. This is not a problem per se, but if the assumption of normality is justifiable, the vast knowledge of normal distributions could be of great help. In situations where we try to transform right-skewed non-negative data to symmetrical data, useful mappings are $x \mapsto \sqrt{x}$, $x \mapsto \sqrt[3]{x}$, and $x \mapsto \log(x)$.

Two of the mentioned function mappings do have requirements which the data should meet, which can be deduced from the respective domains. Specifically, the square root-transform requires data to be positive; the log-transform requires strictly positive values, while the cube root-transform can handle any values from the real line. Mapping-wise $x \mapsto \sqrt[3]{x}$ is valid for negative-values, but it is often not used for negative data. Figure 7.5 shows the new histograms of data which is transformed (cube root, log-transform, and square root) followed by a standardisation with the median.

From Figure 7.5 all three transformations do not seem to differ too much. There is one interesting thing to note which is exclusive to the log-transform[5] for the data of this section. Although data transformation seemed to be almost indistinguishable from Figure 7.5, a key difference between the represented transforms are its relation to distributions. Indeed, if we can assume normality after applying the log-transform, then we found out that our initial data can in fact considered to be from a *log-normal* distribution. If a random variable $X \sim \text{Log-normal}(\mu, \sigma^2)$, then $\ln(X) \sim \mathcal{N}(\mu, \sigma^2)$. That is to say, if the logarithm of a random variable X is normally distributed with parameters μ, σ^2, then

[5] Geurt Jongbloed rightfully points out that the square of a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ exists in the form of the rescaled non-central χ^2 distribution, but it is less easy to work with.

X is log-normally distributed with parameters μ and σ^2. This notation and definition imply that the parameters of the normal distribution are decisive in denoting the distribution.

7.2.3. Goodness of Fit

This section is dedicated to finding a suitable distribution, yet we have not defined what we mean by *suitable*. In broad terms, we could ask ourselves if our model assumptions pass some criteria and if they do, we could call the model suitable. It is quite common and good practice for statisticians to formulate suitability in this context very carefully. Rather than saying: 'The data follows $X \sim P_\theta$.', the formulation should and commonly is more similar to 'There is not enough evidence to say that $X \sim P_\theta$ does not hold.' — which is much more conservative in tone. Indeed, even if there is not enough evidence to say $X \sim P_\theta$ is not the case, it does not necessarily mean that $X \sim P_\theta$ is in fact true. We adapt to the conservative mindset and define a distribution to be suitable if we can not find enough evidence to think otherwise.

Now it remains to see how we can find such evidence. Continuing with the previous findings, we would like to know if our data of the LWT is log-normally distributed for some $\mu \in \mathbb{R}, \sigma^2 > 0$. Model diagnostics help in determining if the statement could be true. In actuality we have already illustrated one method — observing the histograms — to see what kind of distribution could be suitable. Exactly this approach has led us to the log-normal distribution as an initial guess. A final unexplored option for the histograms would be to plot several fitted distributions along the bins, but this could still provide misleading visuals. Instead we will now focus on the more representative diagnostics, such as plotting the empirical cumulative distribution function (ECDF) and probability plots (Q-Q plot).

ECDF

The ECDF will actually prove to be of even greater importance in the discussion of the non-parametric approach in Chapter 8, but for diagnostics it is also relevant. For a random sample $X_1, \ldots, X_n$ from some unknown distribution F, the ECDF $\hat{F}_n$ is a piecewise function defined as

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\}, \tag{7.10}$$

where $\mathbf{1}$ denotes the indicator function defined as

$$\mathbf{1}\{X_i \leq x\} = \begin{cases} 1 & X_i \leq x, \\ 0 & \text{else.} \end{cases} \tag{7.11}$$

In plain English, the ECDF estimates the probability of X being less than or equal to x by counting the fraction of observations X_i being less than or equal to x. Recall that a CDF evaluated at x is equivalent to the probability of the associated random variable X being less than or equal to x. For a small sample size n the ECDF will have few yet relatively large *jump discontinuities*, but as $n \to \infty$ the number of jumps increase but their individual heights become smaller, ultimately resulting in a seemingly smoother curve compared to the ECDF corresponding to small n. The last remark is effectively a property of the ECDF as estimator of the true CDF. Indeed, the *Glivenko-Cantelli* theorem [21] tells us that $\hat{F}_n$ converges to F uniformly, that is

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0. \tag{7.12}$$

This means that with probability 1 (almost surely), the maximum of the differences of the ECDF and true CDF evaluated at all values $x \in \mathbb{R}$ goes to 0 as n goes to infinity. In terms of estimators, we call the ECDF $\hat{F}_n$ a *consistent* estimator for F.

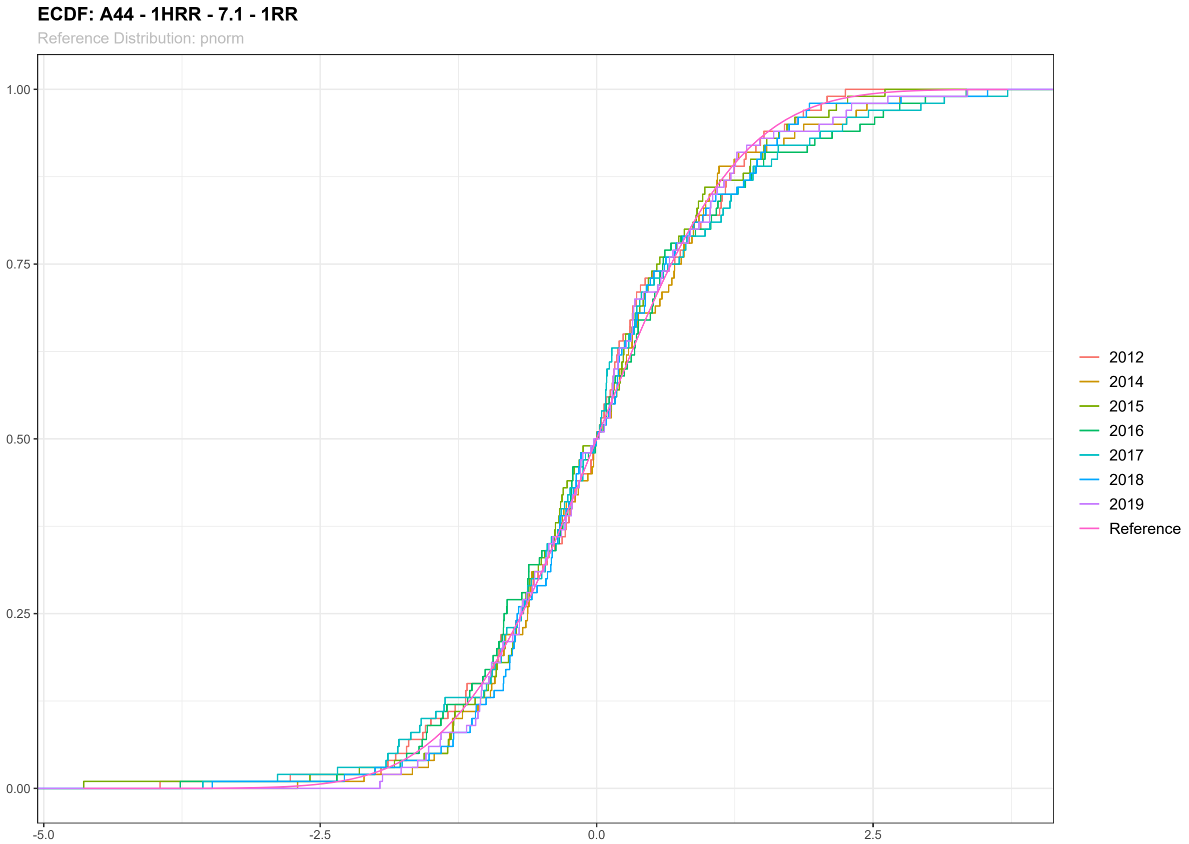

Figure 7.6: CDF of $\mathcal{N}(0,1)$ as a reference curve and ECDFs of standardised log-transformed data of aggregate loss on the LWT of $W = A44, S = 1\text{HRR}_{7.1}^{1\text{RR}}$ from 2012 – 2019 except for 2013.

As to the relevance of the ECDF in our diagnostics: if we plot the ECDF against the CDF which we would like to test for and they generally do not overlap well, there is evidence to believe that the sample is not from the chosen reference CDF. Let us view the ECDF plots for $W = A44, S = 1\text{HRR}_{7.1}^{1\text{RR}}$ in Figure 7.6.

Figure 7.6 shows ECDFs per year and a reference CDF corresponding to the standard normal. Understandably the plot might seem a bit clustered, but that really is an indication of the reference distribution — the standard normal in this case — being a decent representation at first sight.

Now, Figure 7.6 is *not* persuasive enough to say that the log-transformed data is absolutely not normal. However, a critical look at the plots do reveal that in the right tail there might be some mismatch as well as in the left tail; the right tail is more important for us though. Figure 7.6 treats each year separately, hence if we truly want to categorise the entirety of the LWT as coming from one distribution — standardised that is — we should also examine what the effect of aggregating all values from all years results into. That is exactly what Figure 7.7 presents. From Figure 7.7 it does seem more apparent that the initial finding was true — the right tail seems to somewhat deviate from a standard normal and even the left tail shows signs of deviation. We can use other diagnostics to see if that conclusion might be correct.

Probability Plots

The idea of Q-Q plots is comparing the quantiles of two different distributions or samples. If these sets of quantiles approximate a linear relation as in $y = ax + b$ for some $a, b \in \mathbb{R}$ and where y and x represent the quantiles, then we could say that the respective samples could very well be from the location-scale family generated by the distribution whose quantiles are used on the horizontal axis. In particular for $a = 1$ and $b = 0$ it would indicate that it is very likely that the samples stem from the

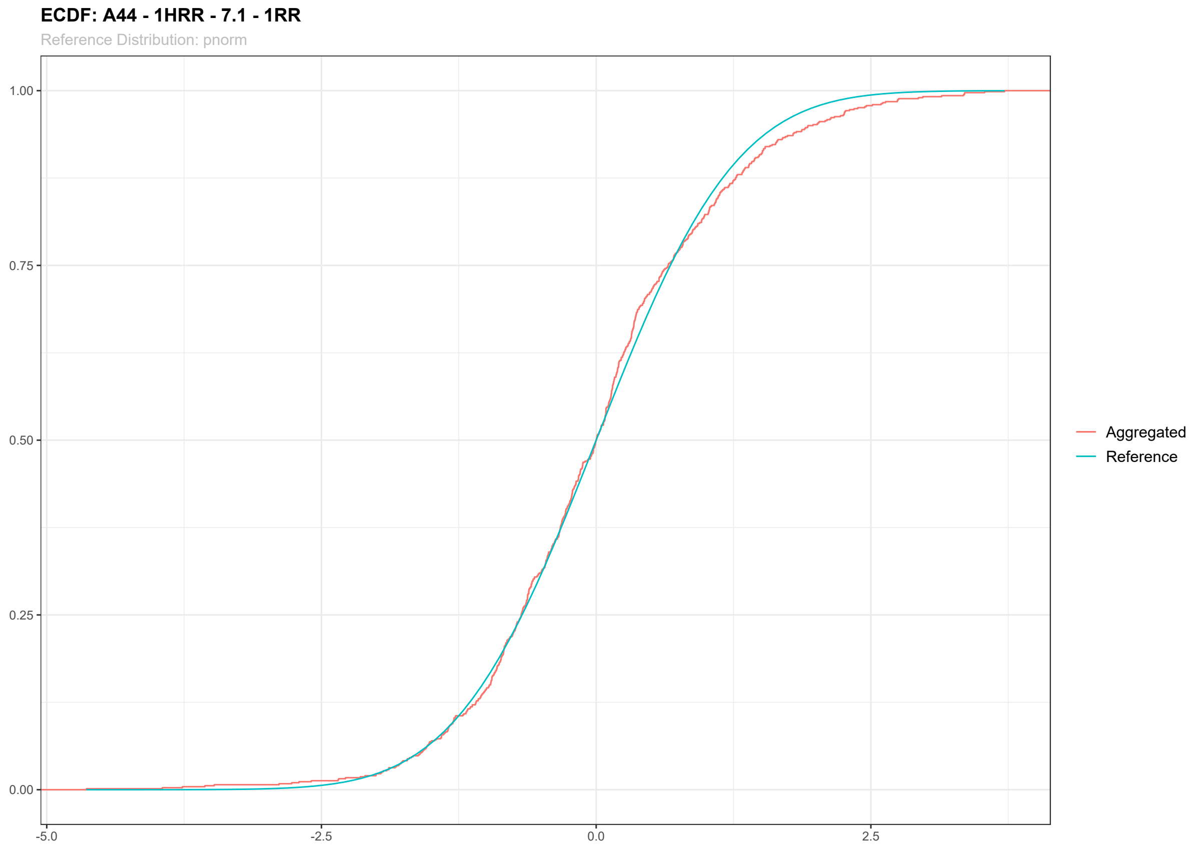

Figure 7.7: CDF of $\mathcal{N}(0,1)$ and an aggregated ECDF of the standardised log-transformed data of aggregate loss on the LWT of $W = A44, S = 1\text{HRR}^{1\text{RR}}_{7.1}$ from 2012 – 2019 except for 2013.

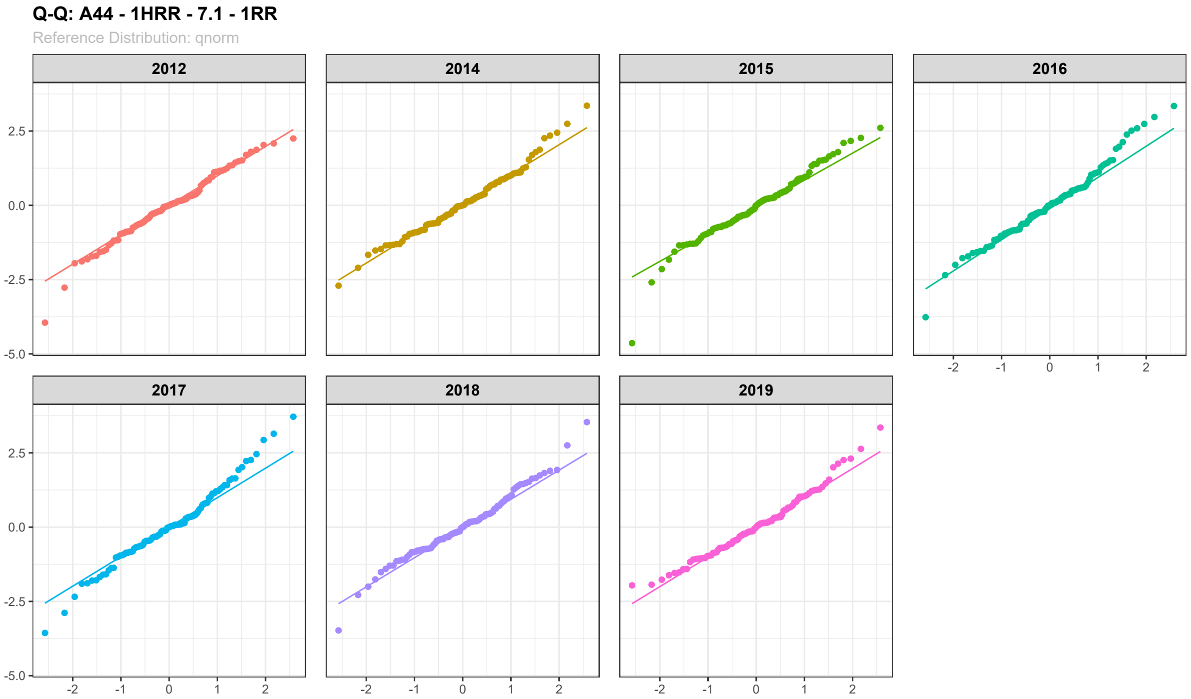

Figure 7.8: Q-Q plot of a $\mathcal{N}(0,1)$ distribution against the standardised log-transformed data of aggregate loss on the LWT of $W = A44, S = 1\mathrm{HRR}_{7.1}^{1\mathrm{RR}}$ from 2012 – 2019 except for 2013.

same distribution with equal fixed parameters. If the relation from y and x is far from linear, then it is feasible to assume that the samples do not come from the same family of distributions. Figure 7.8 shows the Q-Q plots per year.

The Q-Q plots do confirm that our data deviates to some extent in the right tail and to a lesser degree in the left tail from the proposed normal distribution: for every plot the plotted points in the tails of the domain are not aligned with the line as properly as the other points. In particular, Figure 7.8 generally shows that the right and left tails are heavier than the normal distribution. A distribution which is known to have this property and is comparable to the normal is the *logistic* distribution [22 and 23]. In order to compare the standard normal with a logistic distribution we should be cautious. The standard normal has parameters mean $\mu_N = 0$ and variance $\sigma^2 = 1$, but the standard logistic has parameters $\mu_L = 0$ and scale $s = 1$ [22 and 23]. The location parameters coincide, but while σ^2 is the variance for the standard normal, the variance for a logistic distribution is given by $s^2\pi^2/3$. Hence for appropriate comparison we need to compensate for the inflated variance of the logistic. As a result we should compare the standard normal to a logistic distribution with $\mu_L = 0$, $s = \sqrt{3/\pi^2}$: see Figure 7.9 for the ECDF and Q-Q plots for the specified logistic distribution.

Figure 7.9 is plot on a smaller scale so the visuals can be slightly misleading, but it indeed looks to fit the right tail marginally better for the individual years compared to the standard normal. For the sake of explaining the concepts of goodness-of-fit hereafter we will continue to work with the normal, but for the results via the parametric approach in Chapter 10 we will consider the logistic distribution too.

Hypothesis Testing

We will briefly clarify what kind of hypothesis tests we will perform. In our situation, we would like to know whether our data comes from some parametric distribution. The usual one-sample test

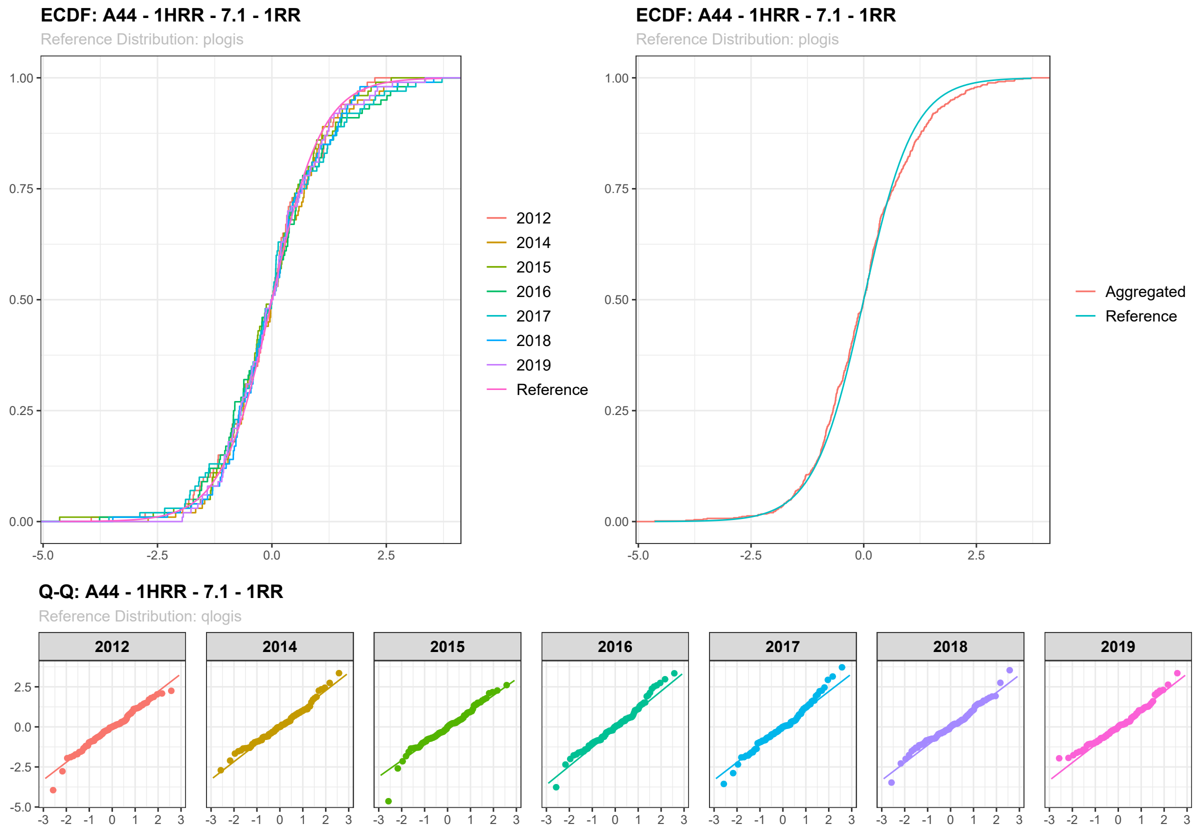

Figure 7.9: ECDF and Q-Q plot of Logistic$(0, \sqrt{3/\pi^2})$ with standardised log-transformed data of aggregate loss on the LWT of $W = A44$, $S = 1\mathrm{HRR}_{7.1}^{1\mathrm{RR}}$ from 2012 – 2019 except for 2013.

that we will see and use can be formulated with only the null hypothesis H_0

$$H_0: \text{ The data follows } X \sim F_\theta,$$

where F_θ is some parametric distribution. The null hypothesis is rejected if the p-value that corresponds to the test statistic is lower than some significance level $\alpha \in (0,1)$. In most scientific studies, $\alpha = 0.05$ is used to perform the testing and we will adapt to this convention. The test statistic on the other hand is inherently dependent on which test is performed. A well-known non-parametric test [6] is the *Kolmogorov-Smirnov* [24] test (KS-test). It makes use of the previously defined empirical cumulative distribution function $\hat{F}_n$ in the statistic

$$D_n = \sup_x |\hat{F}_n(x) - F(x)|. \tag{7.13}$$

The idea of test statistic D_n is quite simple: it considers the maximum of differences of the true distribution and the empirical distribution evaluated at every point; if this is small, then one could say that the true CDF and the ECDF coincide. However, Razali and Wah [19] have shown that the KS-test is quite weak in power, that is, it requires a relatively large sample size to reject the null hypothesis properly. This means that the associated p-values from KS-tests would generally be very high for a small sample size, meaning that we do not reject H_0 even though the sample that was used was clearly drawn from a distribution different from the test distribution. It is therefore to no surprise that statistically more powerful tests such as the Shapiro-Wilk (SW-test) and Anderson-Darling test (AD-test) should be chosen over the KS-test.

Indeed, if testing for normality, Razali and Wah [19] have shown that the SW-test and AD-test perform excellently and 'much better' than the KS-test. Specifically, Razali and Wah [19] used non-normal distributions which for relative small sizes are relatively difficult to distinguish from some normal distribution. For example, in case of testing if samples from Beta$(2,2)$[7] are normal at a significance level of $\alpha = 0.05$ and sample size $n = 100$, the SW-test and AD-test are able to correctly reject H_0 for approximately 45% and 35% of the cases respectively, where the KS-test only rejects H_0 in 10%. If we increase the sample size to $n = 200$, the values are respectively around 90%, 65%, and 20%. In particular for normality testing, the KS-tests are simply too weak for justifying the assumption of normality if there is no information on the true population.

If we perform the SW, AD, and KS-test for normality, we can combine the results into Table 7.1.

[6] The use of a non-parametric test would not make our approach any 'less' parametric.

[7] The Beta distribution with parameters $\theta = (a,b)$ for $a = b \geq 2$ shows similarity in shape with $\mathcal{N}(0.5, 0.25)$.

Table 7.1: p-values given by SW, AD, and KS-tests for normality for the standardised log-transform of LWT of $W = A44, S = 1\text{HRR}_{7.1}^{1\text{RR}}$. A notion on the critical values used: the assumption of independence is made, so these resulting p-values are not completely representative as there is some dependence.

Sample	SW	AD	KS
2012	0.156	0.424	0.727
2014	0.271	0.276	0.998
2015	0.089	0.160	0.964
2016	0.174	0.120	0.843
2017	0.205	0.058	0.623
2018	0.144	0.130	0.873
2019	0.101	0.224	0.792
Combined	0.000	0.000	0.220

Table 7.1 admits logical yet intriguing results. There are a few things to note:

1. The KS-test admits higher p-values than the SW and AD-tests.
2. At a significance level $\alpha = 0.05$, assumption of normality can justifiable be assumed to a certain degree for single-year samples.
3. The null hypothesis is rejected for the union of all single-year samples, except for the KS-test.

Let us elaborate on each item. Firstly, it was mentioned before that the KS-test generally results into higher p-values due to its weak statistical power. The SW and AD-tests on the other hand produce much lower p-values whilst $p > 0.05$ which indicate that accepting the null hypothesis should still be considered with care. Secondly, for low sample sizes such as $n = 100$, it is even challenging for powerful tests to properly reject the null hypothesis. Recall that we opted for a logistic distribution after observing Figure 7.8. Undoubtedly we have no idea if our reconsideration of the data not being normal after the log-transform is completely true after all, but these results would indicate that normality is not a far-fetched assumption. Thirdly, the rejected null hypothesis for the union of samples could somewhat be expected. After observing Figure 7.7 we saw that the right tail of the union deviates slightly and given the new sample size $n = 7 \cdot 100$, the SW and AD-tests are powerful enough to properly reject the null — coinciding with our remark. The exception is for the KS-test, but we have stated its weakness numerous times already and this point confirms it to some degree. However, Geurt Jongbloed rightfully points out that a significant deviation for larger sample sizes is not necessarily a relevant deviation.

7.3. Methods to Compute the Quantiles

For the remainder of the parametric approach with respect to $W = A44, S = 1\text{HRR}_{7.1}^{1\text{RR}}$, we will assume that the single-year samples come from the log-normal family.

7.3.1. Fitting the Distribution

There are several parametric methods to fit a parametric distribution to data. The two arguably most well-known ones are *Maximum Likelihood Estimation* (MLE) and *Method of Moments*.

Maximum Likelihood Estimation

When observing the joint density $f(\mathbf{x}\,|\,\theta)$ for a sample $\mathbf{x}$, the *likelihood function* is defined as

$$L(\theta) = L(\theta\,|\,\mathbf{x}) = f(\mathbf{x}\,|\,\theta). \tag{7.14}$$

Equation (7.14) should really be regarded the way it is suggested by its notation: it is a function of θ (rather than the observed $X_i = x_i$ for $i = 1,\dots,n$) given $\mathbf{x}$. For $L(\theta\,|\,\mathbf{x})$ the value $\mathbf{x}$ is viewed as fixed whereas for $f(\mathbf{x}\,|\,\theta)$ the parameter vector θ is viewed as fixed. The idea of the MLE method is to maximise this function and pick θ such that it maximises $L(\theta)$, that is,

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, L(\theta), \tag{7.15}$$

and $\hat{\theta}$ is called the Maximum Likelihood Estimator of θ. The assumption of $\mathbf{x}$ being i.i.d. implies that the joint density is simply a product of univariate densities. In those cases, the *loglikelihood* function is often considered and defined by

$$l(\theta\,|\,\mathbf{x}) = \ln L(\theta\,|\,\mathbf{x}). \tag{7.16}$$

By continuity and ln being an increasing function, the parameter θ which maximises the likelihood-function also maximises the loglikelihood function. The idea of applying the log-transform for the likelihood function is that the partial derivatives with respect to θ_i are generally effortless to derive for the loglikelihood.

Example 7.3 (MLE of Log-normal Parameters)**.** *Let $X_1,\dots,X_n$ be i.i.d. with density*

$$f(x_i\,|\,\theta) = x_i^{-1}\phi(\ln(x_i)\,|\,\mu,\sigma) \tag{7.17}$$

where ϕ is the density function for the normal distribution. The loglikelihood function is given by

$$\begin{aligned} l(\theta\,|\,\mathbf{x}) &= \ln L(\theta\,|\,\mathbf{x}) \\ &= \ln\left(\prod_{i=1}^{n} \frac{1}{x_i\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln(x_i)-\mu)^2}{2\sigma^2}\right)\right) \\ &= -\sum_{i=1}^{n}\ln(x_i) - \frac{1}{2}n\ln(2\pi\sigma^2) - \sum_{i=1}^{n}\frac{(\ln(x_i)-\mu)^2}{2\sigma^2}. \end{aligned} \tag{7.18}$$

Partial derivatives are found by

$$\frac{\partial l}{\partial \mu}(\mu,\sigma\,|\,\mathbf{x}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\ln(x_i)-\mu) \tag{7.19}$$

and

$$\frac{\partial l}{\partial \sigma^2}(\mu,\sigma\,|\,\mathbf{x}) = -\frac{n}{2\sigma^2} + \sum_{i=1}^{n}\frac{(\ln(x_i)-\mu)^2}{2\sigma^4}. \tag{7.20}$$

Now setting these equations equal to 0*, we find an MLE for μ*

$$\sum_{i=1}^{n}\ln(x_i) - n\mu = 0 \iff \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\ln(x_i) \tag{7.21}$$

and an MLE for σ^2

$$\sum_{i=1}^{n}(\ln(x_i)-\mu)^2 - n\sigma^2 = 0 \iff \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(\ln(x_i)-\mu)^2. \tag{7.22}$$

Thus the MLE for θ is given by $\hat{\theta} = (\hat{\mu},\hat{\sigma}^2)$. Alternatively if we consider $\tilde{X}_1 = \ln(X_1),\dots,\tilde{X}_n = \ln(X_n)$ with density $\phi(\tilde{x}_i\,|\,\mu,\sigma)$, the MLE method could also be applied to the log-transformed data to find estimates for μ and σ^2 using the density of a normal distribution.

Method of Moments

The j-th moment of a random variable X is denoted as $\mathbb{E}(X^j)$. The idea of the Method of Moments is that the parameter vector θ of the underlying distribution can be expressed by its moments, that is, if $\theta_1, \dots, \theta_k$ are the parameters then

$$\mathbb{E}(X^j) = m_j(\theta_1, \dots, \theta_k) \tag{7.23}$$

for $j = 1, \dots, k$. Equation (7.23) admits a system of k equations of k unknowns, which therefore can be solved to find all values for θ_j. It requires the knowledge of the structures of the m_j. This is related to characteristic functions of random variables, as they (as their name suggests) completely define a distribution. Moments can also be approximated by sample data: for a set of observed values $x_1, \dots, x_n$ of X, the k-th sample moment is defined as

$$\mathbb{E}(x^k) = \frac{1}{n} \sum_{i=1}^{n} x_i^k. \tag{7.24}$$

Example 7.4 (Method of Moments for Log-normal Parameters)**.** *Let $X_1, \dots, X_n$ be a sample and let the moments be defined by*

$$\mathbb{E}(X^k) = \exp(k\mu + k^2\sigma^2/2). \tag{7.25}$$

The system of equations that it admits is

$$\omega_1 \equiv \mathbb{E}(X) = \exp(\mu + \sigma^2/2), \tag{7.26}$$

$$\omega_2 \equiv \mathbb{E}(X^2) = \exp(2\mu + 2\sigma^2). \tag{7.27}$$

Applying a log-transform on Equations (7.26) and (7.27) yields linear equations for μ and σ^2. Multiplying Equation (7.26) by 4, subtracting Equation (7.27) from Equation (7.26) and dividing by 2 gives the Moment estimator for μ

$$\hat{\mu} = \ln\left(\frac{\omega_1^2}{\sqrt{\omega_2}}\right), \tag{7.28}$$

while multiplying Equation (7.26) by 2 and subtracting Equation (7.26) from Equation (7.27) gives the Moment estimator for σ^2

$$\hat{\sigma}^2 = \ln\left(\frac{\omega_2}{\omega_1^2}\right). \tag{7.29}$$

Thus the Method of Moments estimator for θ is given by $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$. Similar to Example 7.3, this approach can be applied to data which is log-transformed already. In that case the result of Method of Moments estimator is similar to the MLE in Equations (7.21) and (7.22).

7.3.2. Extracting Quantiles

Assume that we have found the parameter vector θ which corresponds to the best fit for a certain parametric distribution in some family. It implies that we have a fixed CDF $F_\theta(x)$ and PDF $f(x \mid \theta)$. To find q_p, we can solve either one of the following equations:

$$F_\theta(q_p) = p \tag{7.30}$$

$$\int_{-\infty}^{q_p} f(x \mid \theta)\, \mathrm{d}x = p \tag{7.31}$$

Observe that as we are working with parametric distributions, we can attempt to solve this analytically, but it also allows us to use numerical methods for finding q_p. The latter can be done by subtracting p from both sides of Equations (7.30) and (7.31) and use root-finding algorithms, such as Newton's method [25].

If the parametric distribution originates from a location-scale family, then we can use Equation (6.3). The standardised version of the distribution, that is with some location and scale parameter not necessarily limited to the mean μ and standard deviation σ, admits its own p-th quantile. By multiplying with the scale parameter used to standardise and adding the location parameter used to standardise, we have another method to calculate the p-th quantile. Clearly we are interested in the case that $p = 0.75$.

8

Modelling: Non-parametric Approach

The idea of this Chapter is similar to Chapter 7 but with a significant change in assumption. *Non-parametric* statistics does not use models or methods based on parametric families. Akin to the parametric case we will present quantile estimations. We shall look at two different approaches:

- Empirical estimations,
- Density estimations

Non-parametric approaches are not limited to only these two; an example is the *Average Group Quantile* method introduced by Heidelberger and Lewis [26] which is particularly useful for estimating more extreme quantiles ($p > 0.9$) and for relatively large datasets. Due to time constraints we will focus on the mentioned ones. We will begin with some general remarks on notation.

8.1. Notation and Background

8.1.1. Order Statistics

The ECDF $\hat{F}_n$ has been defined before and can be used again and another distribution-free concept to consider are *order statistics*. For random variables $X_1, \ldots, X_n$ its order statistics are denoted $X_{(i)}$ such that

$$X_{(1)} \leq X_{(2)} < \ldots \leq X_{(n-1)} \leq X_{(n)}. \tag{8.1}$$

In case there are no ties in the data, that is $X_i \neq X_j$ for all $i \neq j$, all the inequalities in Equation (8.1) change to strict inequalities. Note that $X_{(1)}$ and $X_{(n)}$ are by definition the minimum and maximum; $X_{(1)} = \min_i X_i$, $X_{(n)} = \max_i X_i$.

As for parametric models, they also come with a set of disadvantages which should be acknowledged. These directly translate to the advantages of the non-parametric models. In particular the following points are relevant:

- parametric models are risky in the sense of misspecification,
- it can be quite cumbersome to find a suitable parametric model,
- non-parametric methods are generally less sensitive to outliers.

8.1.2. Location-Scale Approach

Again we recall Equation (6.3) and rethink how we can use it with some location-scale family. If we assume some starting density in the first year of ravelling progression and that aggregate loss should progress monotonically over time, it is not irrational to believe that this starting density shifts over time without losing *too* much of its original shape. This starting density then admits its own quantiles and these can take the role of $\Phi^{-1}(p)$ in Equation (6.3). Hence finding a representative starting density is key in this approach of approximation.

For taking into account the various shapes of the densities throughout the years, one idea is to aggregate the standardised data from all years. The standardisation is performed with the median and MAD for each year and the aggregation should culminate in a density which is centered near 0 with MAD close to 1. This derived density then serves as the starting density from which we deduce $\Phi^{-1}(p)$, which combined with the median, MAD and Equation (6.3) admit a p-th quantile.

8.2. Empirical Estimation

Let $X_1, \dots, X_n$ be some sample from a continuous distribution $F_X(x)$. Then we can use order statistics for a point estimator of any p-th quantile. Recall Equation (6.1) and consider the case that F in Equation (6.1) is not continuous or strictly increasing. Then the generalised inverse

$$F^{-1}(p) = \inf\{q : F(q) \geq p\} \tag{8.2}$$

is used and Equation (8.2) implies that $F^{-1}(p)$ is the p-th quantile q_p. In particular this can be applied with the ECDF $\hat{F}_n$ and we find by continuity of F_X that there are no ties in the data, hence the quantile q_p can be point estimated by

$$\hat{q}_n(p) = \hat{F}_n^{-1}(p). \tag{8.3}$$

Now since F_X is continuous, we can estimate the p-th quantile of F_X by its order statistic by

$$\hat{F}_n^{-1}(p) = X_{(i)} \iff p \in \left(\tfrac{i-1}{n}, \tfrac{i}{n}\right] \tag{8.4}$$

In particular for $n = 100$ and $p = 0.75$ we find that

$$\hat{q}_{100}(0.75) = X_{(75)}. \tag{8.5}$$

The observant reader will recognise that in the trivial case that $n = 100$, the non-parametric point estimator for $q_{0.75}$ is simply the 75th smallest observation. The approximation of quantiles using order statistics can be directly applied to the aggregated standardised data, that is, we can calculate the standardised sample p-th quantile and perform extrapolation on these quantiles using Equation (6.3).

8.3. Density Estimation

In density estimation we attempt to estimate a density of interest f from independent observations $X_1, \dots, X_n$ which are from density f. It is crucial to realise that f is not observable, just as it was the case in Chapter 7. In fact, in Chapter 7 we attempted to estimate the density by approximating it with some data-dependent member of a parametric family of densities. In the non-parametric case, however, we need to use another strategy for a density estimator which we denote in similar fashion as the ECDF: $\widehat{f}_n$. The goal is to find an $\widehat{f}_n$ *smooth* enough to be representative for the data.

A Remark on Smoothness of Functions

To introduce smoothness as blatantly as we did neglects the mathematical interpretation. Smoothness can be broadly defined and for comparison sake, it would be practical to be able to characterise if some functions are 'smoother' than others. This can be achieved by notion of the *differentiability classes* C^k for $k \in \mathbb{N} \cup \{0\}$. The class C^k serves as a classification for functions and the properties of their derivatives: its derivatives up until order k must exist and be continuous. A function of class C^0 is simply continuous everywhere, but a function of class C^1 is a function which is C^0 (continuous everywhere) and of which its 1^{st} derivative exists and is continuous (continuously differentiable). Some properties are that C^{k+1} is contained in C^k ($C^{k+1} \subsetneq C^k$), and the strict requirement for the k-th derivative to be continuous to be classified as C^k — existence of the k-th derivative is *not* sufficient. The two extremes of smoothness could be defined by C^0 and C^∞, of which the former is the class of continuous functions and of which the latter is the classification for functions which have derivatives of all order. An example of a non-trivial[1] case which is C^∞ is the function $f(x) = e^x$ with domain $\mathbb{R}$. A trivial case is the class of polynomials on domain $\mathbb{R}$: all polynomials have derivatives of any order.

Bias-Variance Trade-Off

First we need to give proper definitions of *bias* and *variance* of a function estimator $\widehat{f}(x)$, because these properties are essential for finding an optimal estimator. The bias of an estimator is the difference between the expected value of an estimator and the true value of the parameter (or function) being estimated. At first thought one would like this to be small, but this point of view changes when one realises that low bias generally comes with worse estimates outside the original domain compared to estimators which have slightly higher bias. The variance of an estimator is a measure for sensitivity to small deviations from the original domain. For example, if an estimate $\widehat{f}(0.5)$ differs significantly from $\widehat{f}(0.6)$, the function estimator $\widehat{f}$ suffers from high variance. The bias and variance of $\widehat{f}(x)$ with respect to the true function $f(x)$ are respectively given by

$$\text{bias}(x) = \mathbb{E}\left(\widehat{f}(x)\right) - f(x) \tag{8.6}$$

$$\text{Var}(x) = \text{Var}\left(\widehat{f}(x)\right). \tag{8.7}$$

In terms of how well an estimator performs, we can fall back to some kind of *loss function* L. The L_p loss function is defined by

$$L_p\left(\widehat{f}(x), f(x)\right)(x) = \left|\widehat{f}(x) - f(x)\right|^p. \tag{8.8}$$

It is clear that a desirable feature would be for L_p to be small. Let us fix $p = 2$ and introduce the *Mean Squared Error* (MSE) which is a term which in fact combines the bias and variance:

$$\begin{aligned} \text{MSE}(x) &= \mathbb{E}\left[L_2\left(\widehat{f}(x), f(x)\right)(x)\right] \\ &= \underbrace{\left(\mathbb{E}\left[\widehat{f}(x)\right] - f(x)\right)^2}_{\text{bias}(x)^2} + \underbrace{\text{Var}\left(\widehat{f}(x)\right)}_{\text{Var}(x)}. \end{aligned} \tag{8.9}$$

Equation (8.9) leads to the *bias-variance trade-off*: we would like the MSE to be small, but the contribution of the bias and variance somehow needs to be managed. One reason to use the L_2 loss function is its relation to the bias-variance trade-off for the MSE. Notice that both bias and variance are denoted as functions, but that is a simple consequence of our objective of estimation. More importantly is the fact that the MSE alone is not a direct measure of $\widehat{f}$ being a good estimator because of its dependence on x. As a result one should opt for the *Mean Integrated Squared Error* (MISE) given by

$$\text{MISE} = \int_{\mathbb{R}} \text{MSE}(x)\,\mathrm{d}x. \tag{8.10}$$

Equation (8.10) is actually proportional to the average function value of MSE(x). For two distinct estimators $\widehat{f}_1$, and $\widehat{f}_2$ which admit $\text{MSE}_1(x)$ and $\text{MSE}_2(x)$, one can justify picking the estimator with lowest MISE.

[1] One could argue that the exponential function is frankly quite trivial as an example, but in this sense we mean that it is not a polynomial.

8.3.1. Histograms

The histograms introduced in Chapter 7 are in fact density estimators. The bin width $h = 1/m$ for some $m \in \mathbb{N}$ of histograms is called the *bandwidth* in the context of smoothing. The view of the histogram changes drastically as h varies and it can therefore be called a smoothing parameter. The histogram estimator for f is given by

$$\widehat{f}_n(x) = \sum_{j=1}^{\#\text{bins}} \frac{\sum_{i=1}^{n} \mathbf{1}\{x_i \in B_j\}}{nh} \mathbf{1}\{x \in B_j\} = \sum_{j=1}^{\#\text{bins}} \frac{\widehat{p}_n}{h} \mathbf{1}\{x \in B_j\}, \tag{8.11}$$

where $\widehat{p}_j = Y_j/n$ and Y_j denotes the number of observations in bin B_j. Let us explain the idea of Equation (8.11). First we should recognise a property which it shares with the ECDF: it is a sum of modified indicator functions, of which each individual modfied indicator function is uniquely defined by its corresponding bin B_j with length h. That is, $\widehat{f}_n$ approximates the density f per bin B_j. The value that it assigns to bin B_j then depends on $\widehat{p}_n/h$, the fraction of the total observations in B_j. h serves as the parameter which controls the size of the bins by definition, but also as the normalising parameter such that $\widehat{f}_n$ integrates to 1, as a property of probability density functions. Wasserman [10] motivates why the histogram estimator is not unrealistic as a choice for a density estimator; for small h and $x \in B_j$ the histogram estimator is unbiased

$$\mathbb{E}(\widehat{f}_n(x)) = \frac{\mathbb{E}(\widehat{p}_j)}{h} = \frac{\int_{B_j} f(t)\,\mathrm{d}t}{h} \approx \frac{f(x)h}{h} = f(x), \tag{8.12}$$

where

- the first equality is a property of expectations,
- the second equality is due to $\widehat{p}_j$ being binomially distributed with parameters $n, \mathbb{P}(X \in B_j)$,
- the third equality is by the Mean Value Theorem (for integrals) and f being continuous.

We will not be using the histogram estimator despite the fact that it is approximately unbiased from Equation (8.12), because there is a better alternative.

8.3.2. Kernels

We use the notions as in Wasserman [10]. A *kernel* is a smooth function K which is non-negative on its domain and satisfies two Equations:

$$\int_{\mathbb{R}} K(x)\,\mathrm{d}x = 1, \tag{8.13}$$

$$\int_{\mathbb{R}} xK(x)\,\mathrm{d}x = 0. \tag{8.14}$$

The requirements imply that any PDF with mean 0 can be classified as a kernel. Examples of kernels used in practice are visualised in Figure 8.1 and defined by

$$\text{(Rectangular)} \quad K(x) = \frac{1}{2} \cdot \mathbf{1}\{|x| \le 1\} \tag{8.15}$$

$$\text{(Triangular)} \quad K(x) = (1 - |x|) \cdot \mathbf{1}\{|x| \le 1\} \tag{8.16}$$

$$\text{(Epanechnikov)} \quad K(x) = \frac{3}{4}\left(1 - x^2\right) \cdot \mathbf{1}\{|x| \le 1\} \tag{8.17}$$

$$\text{(Biweight)} \quad K(x) = \frac{15}{16}\left(1 - x^2\right)^2 \cdot \mathbf{1}\{|x| \le 1\} \tag{8.18}$$

$$\text{(Gaussian)} \quad K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{8.19}$$

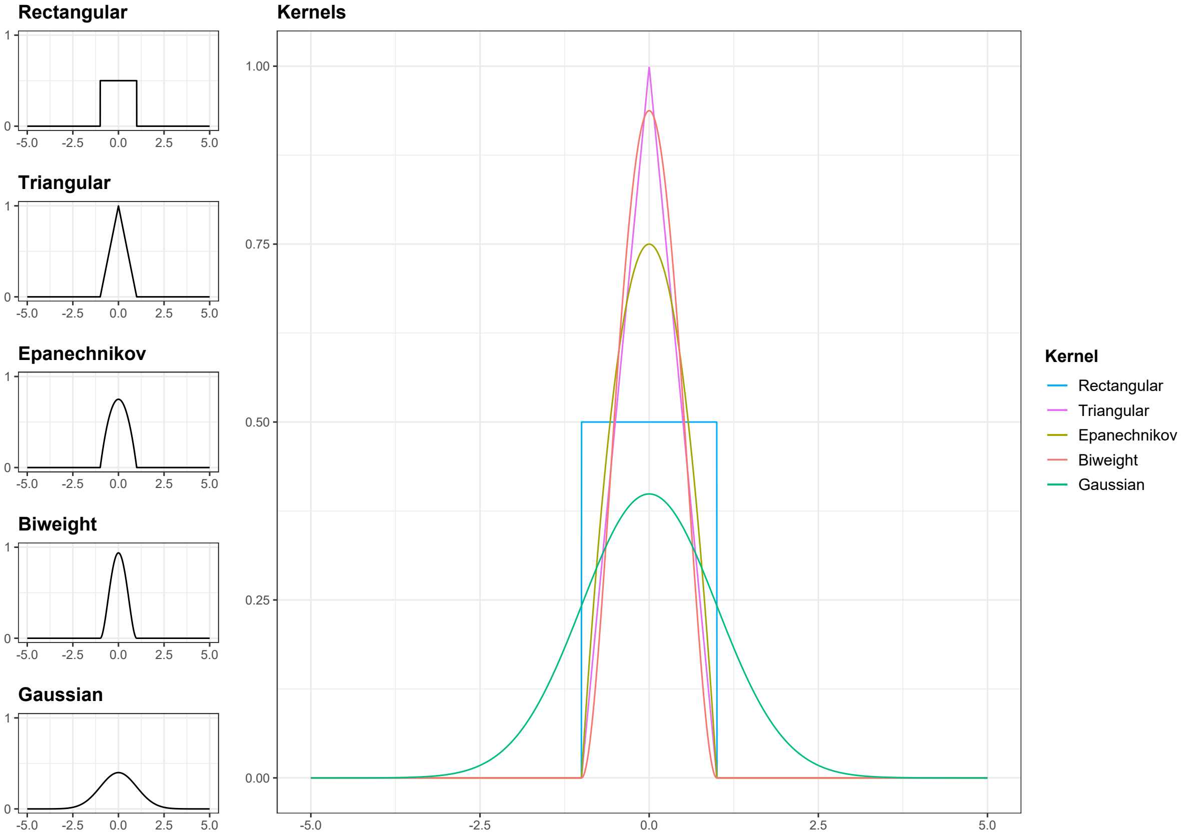

Figure 8.1: Plots of the rectangular, triangular, Epanechnikov, biweight, and Gaussian kernel.

The kernel density estimator is defined by

$$\widehat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\frac{1}{h}K\left(\frac{x-X_i}{h}\right)}_{=\tilde{K}_{i,h}}. \tag{8.20}$$

Let us try to understand what Equation (8.20) enforces by carefully considering values for x. From the conditions of kernels we know that any kernel is centered. For $x = X_i$ the term $\tilde{K}_{i,h}$ does *not* vanish; instead it reaches its peak value. So if we imagine the real positive half-line and position our data X_i accordingly, we are placing 'modified' kernels $\tilde{K}_{i,h}$ at each X_i position. By taking $1/n$ we are taking the local (weighted) average at x for terms $\tilde{K}_{i,h}$ for $i = 1,\ldots,n$. Even though we mention it last, the bandwidth h is extremely important in this process! First consider $h \to 0$: one should visualise that the modified kernel is actually being 'squeezed' from the sides towards X_i, which causes the peak value to increase since $1/h \to \infty$, but as a consequence has decreased the length of the interval of support[2]. The squeezing causes the local (weighted) average to depend on less modified kernels and more on the data point which is closest to x. On the other hand for $h \to \infty$, we are spreading the distribution by pushing it down form the peak value until it eventually becomes a straight line $y \equiv 0$, since the term $1/h \to 0$: see Figure 8.2. In terms of smoothing, for $h \to 0$ we are undersmoothing while for $h \to \infty$ we are oversmoothing. According to Wasserman [10], the choice of K is not crucial but the choice of the bandwidth is. One thing that we would like to add to that is that Wasserman probably meant to say: from the smoother kernels, the choice of the kernel is not that fundamental. If one opts for the rectangular kernel, the density estimator will genuinely not be as smooth as desired; this will be shown at the end of this Section.

[2] The support is the closure of the set of values for which a function is *not* 0.

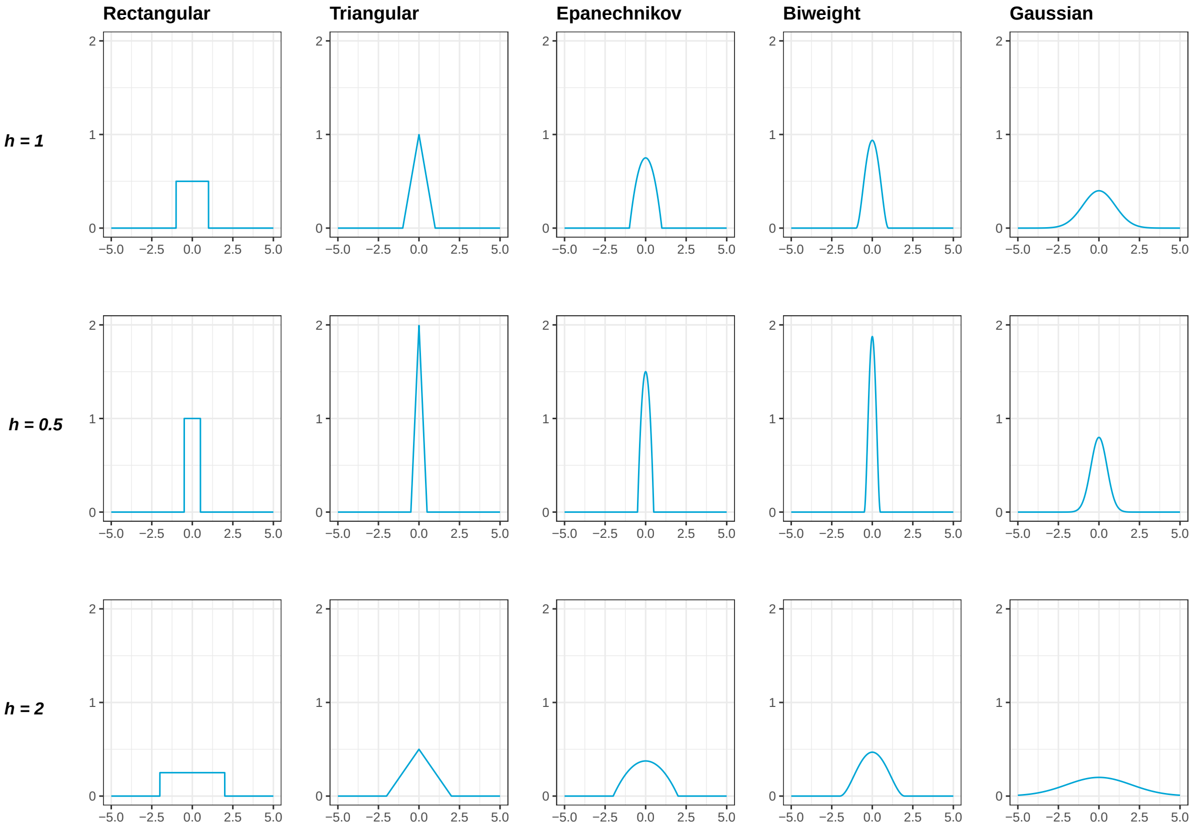

Figure 8.2: Plots of the rectangular, triangular, Epanechnikov, biweight, and Gaussian kernel with varying bandwidths h.

It is now time to see how the kernels for the same bandwidth h estimate the density: see Figure 8.3. The $\widehat{f}_n$ were fit using all the mentioned kernels and it is clear that the rectangular kernel density estimator is the odd-one-out. The density curve is much more rigid than any of the other estimators, but this was to be expected since the rectangular kernel is not as smooth. The Gaussian kernel is convenient as the antiderivative of it can be used for approximating quantiles, which we will show at the end of this Section. Another more relevant topic is how the bandwidth h was chosen exactly. The essence of the bandwidth h was not emphasised too long ago, so it needs some clarification for depicting in Figure 8.3.

The optimal value h^* for the smoothing parameter h is sometimes called an *oracle* [10]. Several approaches exist for finding h^* of which we will mention one: see [10] for other methods. For finding an oracle we can consider the L_2 loss function introduced before. Now that we know that the density estimator $\widehat{f}_n$ is in fact dependent on the bandwidth h, we can use the integrated squared error loss function to write

$$\begin{aligned}\mathrm{ISL}(h) &= \int_{\mathbb{R}} (\widehat{f}_n(x \mid h) - f(x))^2 \,\mathrm{d}x \\ &= \int_{\mathbb{R}} \widehat{f}_n(x \mid h)^2 \,\mathrm{d}x - 2\int_{\mathbb{R}} \widehat{f}_n(x \mid h) f(x) \,\mathrm{d}x + \int_{\mathbb{R}} f(x)^2 \,\mathrm{d}x \qquad (8.21)\end{aligned}$$

For finding the oracle h^* the last term in Equation (8.21) can be neglected, as this does not depend on h. A *cross-validation* estimator of ISL(h) [10] is given by

$$\widehat{\mathrm{ISL}}(h) = \int_{\mathbb{R}} \widehat{f}_n(x \mid h)^2 \,\mathrm{d}x - \frac{2}{n} \sum_{i=1}^{n} \widehat{f}_{(-i)}(X_i \mid h) \qquad (8.22)$$

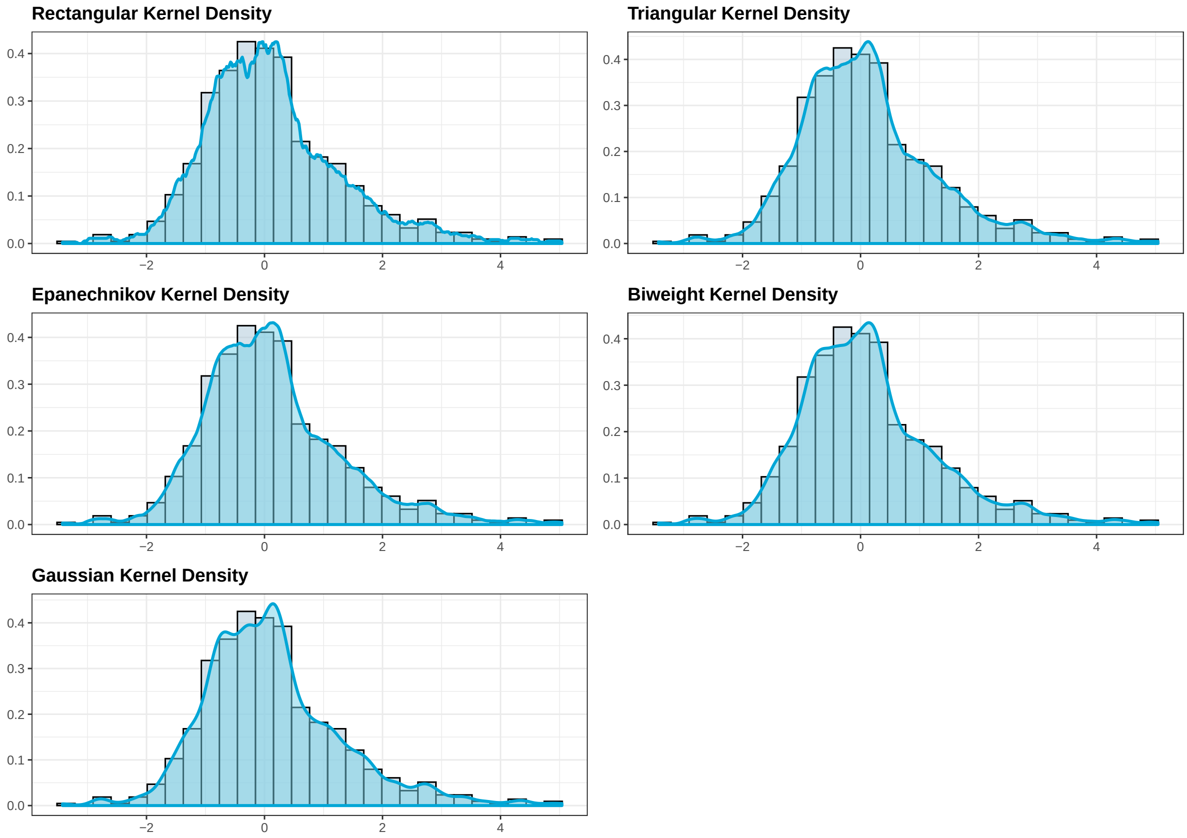

Figure 8.3: Kernel density estimates with similar bandwidth h of the standardised data of the LWT from $W = A44, S = 1\text{HRR}^{1\text{RR}}_{7.1}$ of all years, using rectangular, triangular, Epanechnikov, biweight, and Gaussian kernels.

where $\widehat{f}_{(-i)}$ is the density estimator calculated without the i-th observation. Our oracle h^* is the value which minimises Equation (8.22) and hence (attempts to) minimise ISL(h).

Now that we clarified the choice of the bandwidth and finally have a density estimate, it is possible to approximate the quantiles. It is based on finding a solution to

$$\widehat{F}_n(x) - p = 0 \Longleftrightarrow \frac{1}{n}\sum_{i=1}^{n} \kappa\left(\frac{x - X_i}{h}\right) - p = 0 \tag{8.23}$$

where $\widehat{F}_n(x)$ is the antiderivative of $\widehat{F}_n$ and $\kappa(x) = \int_{-\infty}^{x} K(t)\,\mathrm{d}t$. Here the Gaussian kernel is convenient since its antiderivative is the CDF of the standard normal which is implemented in computing software. Although in terms of convenience, solving Equation (8.23) really is equivalent to solving

$$\int_{-\infty}^{q_p} \widehat{f}_n(x)\,\mathrm{d}x = p, \tag{8.24}$$

for which we can approximate the solution numerically as well. For Equation (8.24) we do not need an intermediate step in the form of (analytically) finding the antiderivative.

9

Extrapolation

The theory from Chapters 7 and 8 leads to approximations of the 75th percentiles for a specific road section along the years. Now it remains to find ways to predict when a threshold has been reached with knowledge of the current progression of aggregate loss. In this Chapter, we will discuss methods to provide predictions for reaching this level by the use of *shape constrained additive models* (SCAMs). The reason for this is that by using shape constraints, we can circumvent the problem of non-monotonicity as hinted in Chapter 5.

9.1. Generalised Additive Model (GAM)

The concept of SCAMs are based on *generalised additive models* (GAMs). GAMs in turn are linked to the less broad term of *generalised linear models* (GLMs). For the remainder of Chapter 9 we will consider the univariate case. Recall that for ordinary linear regression, the response variable has a normal (error) distribution. That is, for the response variable Y_i and predictor variable X_i, for simple linear regression it holds for $i = 1, \dots, n$ that

$$\mathbb{E}(Y_i) = \mu_i = \beta_0 + X_i\beta_1, \quad Y_i \sim \mathcal{N}(\mu_i, \sigma^2) \tag{9.1}$$

where $\beta_0, \beta_1 \in \mathbb{R}$ and σ^2 constant. Simple linear regression is a specific case of GLMs which implies that the assumptions in GLMs are more flexible. Instead of being restricted to the normal distribution, GLMs come with *link functions* which expresses the relation of the expected $\mathbb{E}(Y_i)$ with the linear predictor $\beta_0 + X_i\beta_1$. This leads us to a relaxed version of Equation (9.1) in

$$\mathbb{E}(Y_i) = \mu_i, \quad g(\mu_i) = \beta_0 + X_i\beta_1, \quad Y_i \sim P(\mu_i, \sigma^2(\mu_i)), \tag{9.2}$$

The comprehending reader will notice that in the case of a Gaussian distribution, the link function g is the identity function $z \mapsto z$ and P is the Gaussian family — de facto the relaxation of Equation (9.1). Equation (9.2) implies that

$$\mathbb{E}(Y_i) = g^{-1}(\beta_0 + X_i\beta_1) \tag{9.3}$$

$$\text{Var}(Y_i) = \text{Var}(g^{-1}(\beta_0 + X_i\beta_1)), \tag{9.4}$$

meaning that for GLMs the variance can be seen as a function of the mean. Relaxations on GLMs result in what we are ultimately going to use for obtaining results: GAMs. The framework for GAMs can be denoted by

$$\mathbb{E}(Y_i) = \mu_i, \quad g(\mu_i) = \beta_0 + s(X_i), \quad Y_i \sim EF, \tag{9.5}$$

where s is some *smooth* function and EF belongs to some *exponential family* of distributions. The significant difference between Equation (9.2) and Equation (9.5) is the freedom in the smooth function s applied to X_i. This could very well be simply a scalar transform as $X_i\beta_1$ admits, but in the use-case for GAMs this is uncommon.

9.2. Shape Constrained Additive Model (SCAM)

Now that we have defined GAMs, we can introduce the shape constrained additive models as they were worked on in detail by Pya and Wood [27]. Pya and Wood provide an extension of Equation (9.5) to account for constraints — such as the combination of monotonicity and convexity/concavity — as

$$g(\mu_i) = \beta_0 + s(X_{1,i}) + m(X_{2,i}), \quad Y_i \sim EF(\mu_i, \psi), \tag{9.6}$$

where g is a known smooth monotonic link function, m is an (unknown) shape constrained smooth function, and ψ a ‘scale’ parameter, which in bare minimum should express the relation between the mean and variance [27]. This extension assumes that there are several predictor variables, namely $X_{1,i}$ and $X_{2,i}$ (and perhaps more). For our case, however, we only need to consider one predictor variable, which means that the function s in context of Equation (9.6) is abundant. Our predictor variable is based on the date of measurement; if future research on the topic of ravelling would like to consider the addition of another predictor variable, then this can be achieved by adding another smooth function s_j or m_j depending on the assumptions. The shape constraints in m are the defining features for SCAMs.

9.2.1. Constructing Functions

The SCAMs that we will ultimately be using to find a function which represents the progression of aggregate loss is based on the idea of constructing a function by combining *basis functions* [28]. Say that we want to approximate a function $R(x)$ by a set of basis functions $\phi(x)$. Then the *basis function expansion* of R is given by

$$\hat{R}(x) = \sum_{k=1}^{K} c_k \phi_k(x), \tag{9.7}$$

where $c_k \in \mathbb{R}$ for $k = 1, \dots, K$ are coefficients. A basis function expansion is hence fundamentally a linear combination of the basis functions. Depending on the characteristics of $R(x)$ — in particular its (non-)periodicity — the ϕ_k are often chosen from a particular group of basis functions: the Fourier basis or the *spline* basis. The DOS-LCMS data that has been gathered until now imply a non-periodic function that we would like to approximate. In the far future with measurements of a road section going through multiple maintenance, it would be possible to opt for a Fourier basis. However as of now the wiser choice would be to stick with the spline basis system.

Ramsay et al. [28] rightfully point out how basis systems are not new and are more common than one might imagine. In particular Ramsay et al. recall how the class of polynomials are a concrete example of the use of a basis system. Indeed, any polynomial $P(x)$ is a linear combination of monomials $\phi_k(x) = x^k$ for $k \in \mathbb{N}$, that is,

$$P(x) = \sum_{k=0}^{K} c_k x^k. \tag{9.8}$$

Splines

Splines are piecewise-defined functions for which every piece is a polynomial: a piecewise polynomial function. A spline $\xi(x)$ of order p on an interval $[a, b]$ with non-decreasing *knots* $\{x_i\}_{i=1}^{n-1}$ and $x_0 = a, x_n = b$ can be defined as

$$\xi(x) = \begin{cases} \xi_0(x) & x \in [x_0, x_1], \\ \vdots & \\ \xi_{n-1}(x) & x \in [x_{n-1}, x_n], \end{cases} \tag{9.9}$$

where ξ_j on $[x_j, x_{j+1}]$ are polynomials of order p — which implies a degree of $p-1$, and a smooth connection between the ξ_j and its derivatives is required in the knots. A non-trivial case is realised for $p = 2$, and the spline would then consist of linear polynomials while its derivative is a step function, which in terms of smoothness leaves a lot to be desired. Another non-trivial but more frequently used case is attained for $p = 4$. In that case we are dealing with *cubic* splines which are useful due to their smoothness in both the regular function fit as well as its derivative which is fit piecewise by quadratic polynomials.

B-splines

Within the use of splines there are actually several basis systems for constructing spline functions [28]. The most commonly used are the *B-splines* which gained attention in 1978 by work of De Boor and was revised quite recently in 2001 by De Boor [29]. B-splines are basis functions for spline functions. The naming intuitively indicates that the basis functions ϕ_k are splines as well which is indeed the case. Multiples of splines remain spline functions as well as sums and differences of splines remain splines, which means that any linear combination of spline (basis) functions are splines as well [30]. For the scope of our research it is not necessary to get into all the details of the underlying basis system, but we will try to give the required minimum knowledge of how B-splines are constructed for our purpose.

Adhering to De Boor [29], it helps to first define the B-splines of order 1 $B_{i,1}(x)$ for a knot sequence $\{x_i\}_{i=1}^n$ which are in fact indicator functions

$$B_{i,1}(x) := \begin{cases} 1 & \text{if } x_i \leq x < x_{i+1}, \\ 0 & \text{else.} \end{cases} \tag{9.10}$$

Some other properties also need to be satisfied, of which we will focus on one; the others can be found in De Boor [29] for the interested reader. A defining property for B-splines is that for every value of x in the respective domain, the values of the basis functions should add up to 1; mathematically this can be denoted as

$$\sum_{i=1}^{n} B_{i,1}(x) = 1, \text{ for all } x. \tag{9.11}$$

Recall that these are the basis functions, hence this constraint does not imply that functions with values greater than 1 cannot be fit. De Boor [29] introduced the recurrence formula for higher-order B-splines:

$$B_{i,p} := \omega_{i,p} B_{i,p-1} + (1 - \omega_{i+1,p}) B_{i+1,p-1} \tag{9.12}$$

for which the coefficients $\omega_{i,p}$ are defined as

$$\omega_{i,p}(x) := \begin{cases} \dfrac{x - x_i}{x_{i+p-1} - x_i} & \text{if } x_i \neq x_{i+p-1}, \\ 0 & \text{else.} \end{cases} \tag{9.13}$$

Now the polynomial structure might not be as apparent from the definitions at first sight. It certainly is absent for the B-splines of order $p = 1$ — although you could say that the number 1 is a polynomial of degree 0 — but for $p > 1$ it is more clear. The B-splines of order $p = 2$ can be deduced from Equations (9.12) and (9.13) and are given by

$$B_{i,2}(x) := \frac{x - x_i}{x_{i+1} - x_i} B_{i,1}(x) + \left(1 - \frac{x - x_{i+1}}{x_{i+2} - x_{i+1}}\right) B_{i+1,1}(x). \tag{9.14}$$

Equation (9.14) shows a polynomial of degree 1: the coefficients of $B_{i,1}(x)$ and $B_{i+1,1}(x)$ are both of degree 1 — the term x is the highest order monomial — and $B_{i,1}(x)$, $B_{i+1,1}(x)$ are indicator functions. Now with Equation (9.12) it should be more clear that for order $p = 2$, a term containing x is being multiplied by another term containing x, resulting in a polynomial of degree 2. With recurrence the polynomial structure becomes more clear and it is evident that a B-spline of order p corresponds to a piecewise polynomial function with degree $p - 1$— similar to regular splines. For more details of properties of B-splines we refer to De Boor [29].

9.2.2. Constructing a Smooth Monotone Convex Function

We started § 9.2 with the approach from Pya and Wood [27]. When we continue with their approach, we can construct the monotonically increasing smooth function m from Equation (9.6) using B-splines as basis. It allows us to write

$$m(x) = \sum_{j=1}^{d} \gamma_j B_j(x), \tag{9.15}$$

in which d is the dimension of the basis and B_j are B-spline basis functions of order $p \geq 2$ to account for a bare minimum of smooth functions. If we want m to be smooth monotone increasing, then a bare minimum condition is for the first derivative of m to be non-negative, $m'(x) \geq 0$ for all x in some domain. According to formulas in [29], this is satisfied when the first order differences for γ_j are non-negative: $\gamma_j \leq \gamma_{j+1}$ for all $j = 1, \ldots, d$. Hence if we find an increasing[1] sequence of spline coefficients γ_j, the resulting smooth function m is monotone. Pya and Wood [27] opt for a clever re-parameterisation by defining

$$\boldsymbol{\beta} := (\beta_1, \beta_2, \ldots, \beta_d), \tag{9.16}$$

$$\tilde{\boldsymbol{\beta}} := (\beta_1, e^{\beta_2} \ldots, e^{\beta_d}), \tag{9.17}$$

$$\boldsymbol{\Sigma} := \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 1 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 1 & 1 & \ldots & 1 \end{pmatrix}, \tag{9.18}$$

$$\mathbf{X}_i := (B_1(x_i), B_2(x_i), \ldots, B_d(x_i)), \tag{9.19}$$

where $\boldsymbol{\Sigma}$ is $(d \times d)$ and imposing

$$\boldsymbol{\gamma} := \boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}, \tag{9.20}$$

which results in

$$\gamma_j = \beta_1 + \sum_{l=2}^{j} e^{\beta_l} \tag{9.21}$$

$$g(\mu_i) = m(x_i) = \mathbf{X}_i \boldsymbol{\Sigma} \tilde{\boldsymbol{\beta}} \tag{9.22}$$

for which the β_j are unknown unconstrained parameters. As $e^x > 0$ for every $x \in \mathbb{R}$, it is clear that for all j, $\gamma_j \leq \gamma_{j+1}$.

Penalisation

In the context of smoothing a paradigm which is often considered is *penalisation* [10 and 30 and 31]. It is important for finding a regression function which is representative for our data and hence the population which we want to model, even though we acknowledge that it is nearly impossible to find the 'true' behaviour. In case the regression function is an underfit or overfit of the data, then it respectively fails to capture a general pattern or ignores the general pattern and fails to account for variability. This is closely related to the bias-variance trade-off mentioned earlier, that demonstrates how finding optimal model parameters such that the bias and variance of the model lead to a minimal model error is of crucial importance. An extremely biased fit does not use the data (constant function, hence lacks variance), but a model with high variance and low bias is often not able to filter (random) noise from the data. Penalisation attempts to achieve a balance between underfitting and overfitting.

[1] Non-decreasing would also be fine, but then one has to take into account that the resulting function is not strictly monotonic.

P-splines

In the scope of B-splines, Eilers and Marx [31] introduced P-splines (penalised B-splines) to account for penalties in B-spline fits. Eilers and Marx penalise the differences in basis coefficients of a B-spline basis. Pya and Wood [27] adapt that idea for their shape constrained models. Pya and Wood state how the β_j for $j \geq 2$ are really just ln differences in γ_j, and we can see this as follows:

$$\begin{aligned} e^{\beta_j} &= \gamma_j - \left(\beta_1 + \sum_{l=2}^{j-1} e^{\beta_l}\right) \\ e^{\beta_j} &= \gamma_j - \gamma_{j-1} \\ \beta_j &= \ln\left(\gamma_j - \gamma_{j-1}\right), \end{aligned} \tag{9.23}$$

where the first and second equality are due to Equation (9.21). Since β_j represent (ln) differences of the spline coefficients γ_j, it maintains the notion of Eilers and Marx [31] to impose a penalty on these. A general way to incorporate this is to introduce a matrix $\mathbf{D}$ and penalise for $||\mathbf{D}\boldsymbol{\beta}||^2$, where $||\cdot||$ is the Euclidean norm: for a vector $\mathbf{x} = (x_1, \ldots, x_n)$ the Euclidean norm is defined by

$$||\mathbf{x}|| := \sqrt{\sum_{i=1}^{n} x_i^2}, \tag{9.24}$$

and $\mathbf{D}$ is the matrix on which conditions can be specified to account for shape constraints [27], and of which its dimensions are also dependent on the constraints. We want to be able to measure differences in adjacent β_j's and by penalising on $||\mathbf{D}\boldsymbol{\beta}||^2$, as this controls the *wiggliness* of the curve fit, of which we will soon show an explicit illustration.

Minimisation Objective

So far we have defined all necessary terms to state the objective to minimise. To compute the smooth increasing convex function m, we know from Equation (9.15) that after fixing the order of the B-splines, it only requires appropriate values for γ_j — which implies that we need to find an appropriate vector of values $\boldsymbol{\beta}$. If data points are denoted (x_i, y_i) for $i = 1, \ldots, n$, $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{X}$ is the matrix such that $X_{ij} = B_j(x_i)$, then we need to find

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}||^2 + \lambda ||\mathbf{D}\boldsymbol{\beta}||^2, \tag{9.25}$$

where λ is a smoothness parameter. It controls the bias-variance trade-off [31 and 32]: if $\lambda \to 0$ the roughness is neglected, while for $\lambda \to \infty$ the roughness plays an essential role in the minimisation objective. Now it only remains to find the constraint for $\boldsymbol{\Sigma}$, $\mathbf{D}$ such that the resulting function m is increasing and convex. The $\boldsymbol{\Sigma}$ defined before was an example of a simple monotone increasing m. For convexity we also need $m''(x) \geq 0$ besides the $m'(x) \geq 0$ constraint. By [29] the former and latter are satisfied if

$$\gamma_j - \gamma_{j-1} \geq 0, \quad \text{for } j = 3, \ldots, d \tag{9.26}$$

$$\gamma_j - 2\gamma_{j-1} + \gamma_{j-2} \geq 0, \quad \text{for } j = 3, \ldots, d. \tag{9.27}$$

Equation (9.26) actually holds for $j = 1, \ldots, d$ but since both Equations (9.26) and (9.27) need to be satisfied simultaneously, $j = 3, \ldots, d$ is the proper condition. These are satisfied if the $(d \times d)$-matrix

$\mathbf{\Sigma}$ is defined as

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 2 & 1 & 0 & \cdots & 0 \\ 1 & 3 & 2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & d-1 & d-2 & d-3 & \cdots & 1 \end{pmatrix} \tag{9.28}$$

and the $((d-3)\times d)$ smoothness matrix $\mathbf{D}$ as

$$\mathbf{D} = \begin{pmatrix} 0 & 0 & 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}. \tag{9.29}$$

The penalty $\lambda||\mathbf{D}\boldsymbol{\beta}||^2$ is now given by

$$\lambda||\mathbf{D}\boldsymbol{\beta}||^2 = \lambda\left((\beta_3-\beta_4)^2+\cdots+(\beta_{d-1}-\beta_d)^2\right). \tag{9.30}$$

From Equation (9.30) it is clear that the singular bands of 1 and −1 in Equation (9.29) could be swapped around: the squares of the differences do not change in Equation (9.30). If $\lambda \to 0$ then the curve is wiggly, while for $\lambda \to \infty$ the fit is not wiggly: see Figure 9.1. For $\lambda = 10$ the fit looks to be linear, for $\lambda = 10^{-6}$ the fit is wiggly, and $\lambda = 1$ produces a decent estimate between overfitting and underfitting.

Akin to the bandwidth h in Chapter 8, there exists several optimisation schemes for the smoothness parameter λ. Again we can consider some cross-validation optimiser with respect to λ, or opt for the *Akaike Information Criterion* (AIC) which is another prediction error measure. As for any criterion, we want to minimise the prediction error and the value for λ which fulfils this will be considered as the 'optimal' penalty. We will not provide the details here, but refer to [27] for the interested reader.

9.3. Extrapolation Methods

Assuming that we have fit a smooth monotonic curve to our percentile data points, we can attempt extrapolations based on the curve. However, extrapolations with splines is cumbersome due to the behaviour of splines outside of their support. For both methods we would like to address that we should be cautious with extrapolating far into the future.

9.3.1. Linear Extrapolation

In general the predictive behaviour outside of the given knot range is not representative: the constructed smooth function is an interpolation method, meaning that behaviour outside of the support is generally hard if not impossible to predict. Nonetheless, Pya [32] has implemented a method to predict outside of the provided support in the `R` package called `scam`. Its implementation as of now is using *linear* extrapolation. The slope of the line is determined by the value of the first derivative at the end point.

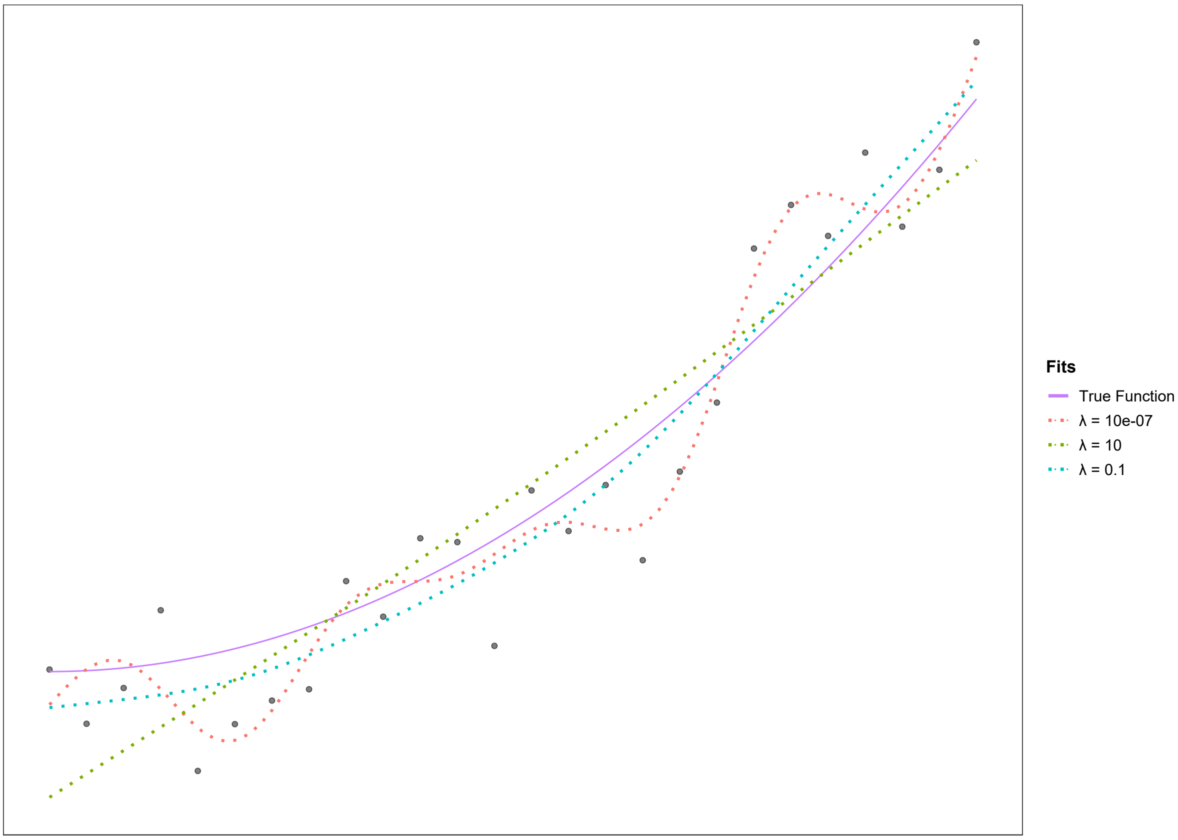

Figure 9.1: Simulation of points from $y(x) = x^2/3 + \varepsilon$ with noise term $\varepsilon \sim U(-2,2)$ with P-spline fits using various smoothness penalties λ.

As a consequence for our results this will mean that for a smooth and rather convex curve fit, a linear extrapolation will be conservative in its predictions: in other words, when looking for the value for which the curve surpasses a threshold, the linear extrapolation will predict this value generally later than a continuation of the curve would have predicted. In particular for curves which have not been fit to be increasing convex[2], which is possible for a small amount of data points where monotonicity leaves a lot to be desired, it could cause unnatural predictions for lifetimes (25+ years). Yet, we can argue that when the curve fit is simply a straight line, it was difficult to find a representative increasing convex curve to begin with. This would also indicate that according to the state of the road up till now, maintenance does not seem necessary. In such cases we can opt to say that for a remaining lifetime of great value, we can categorise it as 'no maintenance needed in the coming 5 years'. This categorisation is similar to what Rijkswaterstaat has been using now.

9.3.2. Polynomial Extrapolation

The linear implementation by Pya in the `scam` package is understandable yet somewhat unsatisfying for our research in particular. The conservative approach as explained before can be circumvented by implementing a *polynomial* extrapolation. Do note that it only applies in cases such as these, where we assume to have knowledge on the progression of some quantity (aggregate loss). The idea is to use the smooth curve and its interpolated results. By definition we know there must exist a decent polynomial approximation to the smooth spline curve of order p (degree $p-1$). Then by fitting the interpolated results, we have a new model which of which the natural extrapolation is based on the curve fit. The interpolated results in this context are the approximated values of the 75th percentile on every single day between the original domain, which ranges from the first measurement date until the last measurement date. In general a polynomial fit does not have to satisfy the constraints imposed before, but by providing many data points — one data point per day — it is possible to achieve such a fit. The model defined by this has two advantages over the linear extrapolation. Firstly, the method is less conservative and represents the assumption of smooth convex monotonicity by continuation of the curve fit. Secondly, the polynomial extrapolation within a short range from the original provided domain is almost equal to the linear extrapolation: at best we do not lose any information by choosing the polynomial over the linear version. The problems mentioned for the linear extrapolation, however, are still present in the polynomial extrapolation. As long as the estimated 75th percentiles do not show a pattern in line with our assumption, the prediction will remain skewed independent of the type of extrapolation used. The only disadvantage and risk using the polynomial extrapolation is that the extrapolation outside the domain does not always continue the curve as we expect it to go. Specifically, after following the trend of increasing convexity outside of the provided domain, we could for some reason see that the polynomial extrapolation goes downwards. This is not contrary to the constraint that we set, as these were put on the provided domain. One way to deal with such an outcome is to simply choose the linear extrapolation if this occurs.

[2] Here we mean that the curve fit is not simply a straight line, which lacks curvature.

10
Results and Discussion

10.1. Results for Prototype

Let us start with a reminder that the build-up for all the results for our prototype ($W = A44, S = 1\text{HRR}_{7.1}^{1\text{RR}}$) can be applied to the any other road section which satisfies similar conditions as mentioned in Chapter 6. The only difference in applying the results for other road sections is in the parametric approach, where the data was manually observed and suitable parametric distributions were suggested.

10.1.1. Quantile Estimates

From the parametric approach, there was reason to believe that either

- the raw unmodified data is log-normal, or
- the log-transformed data is logistic.

For the log-normal approximation we can simply perform MLE or Moment estimators for the parameters of the distribution, which we then use to calculate the quantiles. Though for the logistic approach, we need to log-transform the data and perform MLE/Moment estimation for the parameters and then take the inverse log function (exp) for the quantile on the proper scale.

From the non-parametric approach, we can aggregate the standardised data (done per individual year) and

- use the empirical quantile,
- estimate the quantile from a kernel density estimator by performing root-finding algorithms.

In the case of the kernel estimator we chose the Gaussian kernel. Furthermore the non-parametric approaches are not limited to only this prototype which is a significant advantage over the parametric estimates. The estimates for all approaches are noted in Table 10.1 and plot in Figure 10.1. From Figure 10.1 the estimates from all four approaches seem to be extremely close to each other. Especially all estimates besides the log-normal are not distinguishable from the linear interpolation in Figure 10.1. If we take a look at Table 10.1 then we can confirm that the log-normal estimates are clearly dominating for all seven estimates, while the other estimate interchange the position of 2$^{\text{nd}}$ highest estimate. Although we do say 'clearly dominating', the differences are of an extremely small order between the highest and lower estimates. The interpretation of this observation is that

- investing time in finding a suitable parametric distribution results in similar numbers as a distribution-free approach does, which in turn could imply;

Table 10.1: Estimates for quantiles and parameters of the LWT of $W = A44, S = 1\text{HRR}^{1\text{RR}}_{7.1}$, where MLE was performed for the parameters of the parametric distributions.

	Parametric						**Non-parametric**			
	Log-normal			*Logistic*					*Empirical*	*Kernel*
Date	$\ln(\mu)$	$\ln(\sigma)$	$q_{0.75}$	μ	s	$q_{0.75}$	med.	MAD	$q_{0.75}$	$q_{0.75}$
2012-07-27	0.49	0.23	1.90	0.49	0.126	1.88	1.65	0.34	1.88	1.89
2013 data was not measured for the prototype.										
2014-03-20	0.96	0.17	2.93	0.95	0.096	2.88	2.58	0.44	2.89	2.89
2015-03-19	0.98	0.18	3.00	0.98	0.097	2.96	2.66	0.42	2.96	2.96
2016-02-27	0.95	0.19	2.95	0.95	0.106	2.89	2.57	0.42	2.86	2.86
2017-02-14	1.18	0.18	3.69	1.18	0.098	3.62	3.26	0.47	3.58	3.59
2018-03-27	1.76	0.16	6.50	1.76	0.089	6.40	5.79	0.88	6.40	6.40
2019-03-30	1.94	0.15	7.68	1.93	0.083	7.56	6.90	0.98	7.58	7.59

- investing time in the parametric approach is too time consuming for analysis on a greater scale.

It raises the question what estimates for $q_{0.75}$ a *non*-suitable distribution would admit. From Figure 7.3 we believe it is fair to classify exponential distributions and uniform distributions as bad fits for the data. If resulting quantile estimates using non-suitable distributions end up similar to quantile estimates from the suitable distributions, there is more reason to neglect the parametric approach. Parameter estimation for these non-suitable parametric distributions to the unmodified data are given in Table 10.2. A comparison between Tables 10.1 and 10.2 instantly shows that the $q_{0.75}$ estimates from Table 10.2 are much higher compared to Table 10.1. Even more it shows how investing time in finding an appropriate parametric family leads to estimates closer to the non-parametric ones. Despite all this, we should acknowledge that we simply do not and cannot know what the ‘true’ value of $q_{0.75}$ is or should be, although the similar estimates from Table 10.1 do seem persuasive as reasonable approximations.

10.1.2. Predictions of Remaining Lifetime

Aside from quickly visualising the estimated quantiles, Figure 10.1 is not useful for predictions: a linear extrapolation based on linear interpolation of $q_{0.75}$ throughout the years will clash with what we expect to be a smooth monotonic progression of aggregate loss. Specifically this means that a prediction of $q_{0.75}$ of a date which is approximately one year later than the date of the last measurement (2019-03-30) is based on the slope of the line segment from 2018-03-27 till 2019-03-30. Although the final prediction *could* be based on linear extrapolation, the slope of the last line segment is not based on the general pattern of progression. Using the `scam` package and its underlying functions, we can fit the $q_{0.75}$ estimates as the response variable with our date of measurements as the predictor variable. From Figure 10.1 we should expect that the fitted smooth monotone curves should be very similar in form because the point estimates are close. That is exactly what Figure 10.2 illustrates and more. Using a P-spline basis with dimension $d = 4$ and order $p = 4$ (cubic splines) we

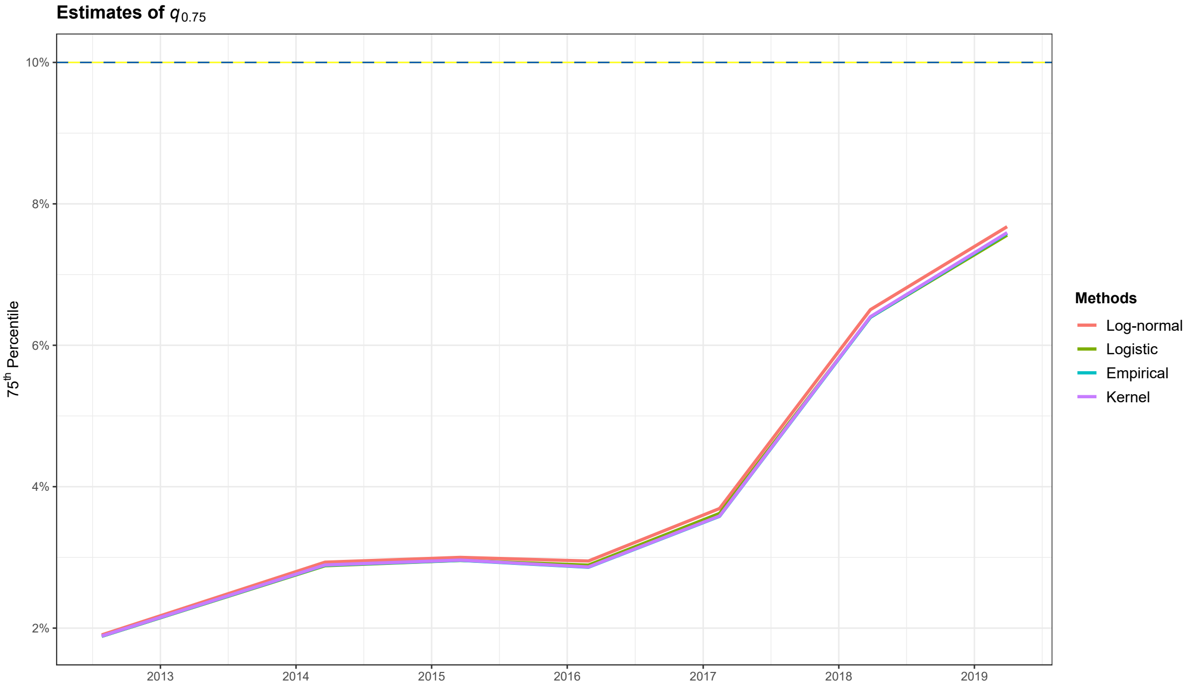

Figure 10.1: Linear interpolation of $q_{0.75}$ estimates from 2012 – 2019 plotted for the LWT of $W = A44, S = 1\text{HRR}^{1\text{RR}}_{7.1}$. The dashed horizontal line is the RWS threshold.

Table 10.2: Estimates for quantiles of the LWT of $W = A44, S = 1\text{HRR}^{1\text{RR}}_{7.1}$, where MLE was performed for parameters of non-suitable parametric distributions.

	Exponential	*Uniform*
Date	$q_{0.75}$	$q_{0.75}$
2012-07-27	2.32	2.17
	No 2013 data	
2014-03-20	3.68	3.77
2015-03-19	3.74	3.39
2016-02-27	3.66	3.65
2017-02-14	4.61	4.69
2018-03-27	8.19	8.33
2019-03-30	9.74	9.72

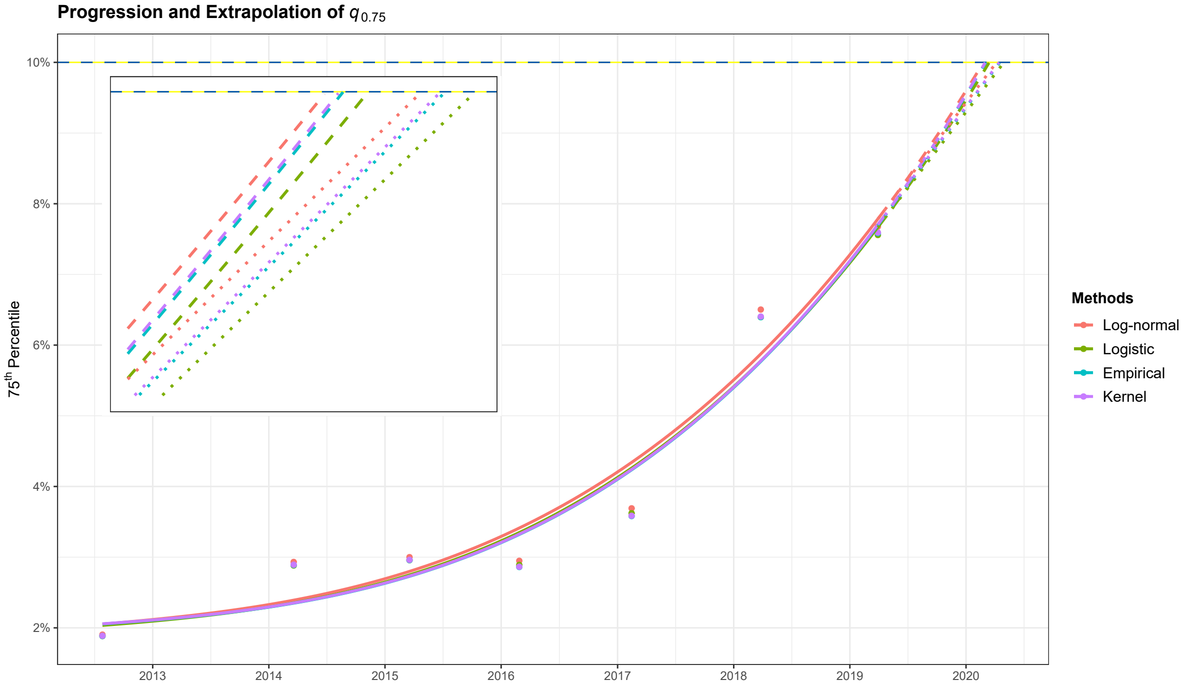

Figure 10.2: Linear and polynomial extrapolations of $q_{0.75}$ estimates from 2012 – 2019 plotted for the LWT of $W = A44$, $S = 1\text{HRR}^{1\text{RR}}_{7.1}$ after fitting smooth monotone curves with P-splines as basis with dimension $d = 4$ and order $p = 4$ (cubic spline). The curved dashed lines represent the polynomial extrapolation, while the dotted lines represent the linear extrapolation. The dashed horizontal line is the RWS threshold. The top-left corner contains a zoomed in version of the clutter on the top-right.

find smooth monotone and convex curves as explained in Chapter 9. Figure 10.2 shows the smooth convex monotone curves and both the linear and polynomial extrapolation. We warn the reader for the clutter that is shown in Figure 10.2, but have provided a zoomed-in section of the clutter and also would like to use it to prove the point made in Chapter 9: the closer the polynomial extrapolation is to the provided dates, the more it resembles the linear extrapolation. In addition it proves another point made about conservatism: the linear predictions of the proposed remaining lifetimes (PRL) are slightly farther in the future. For the remainder and the reader's eyesight, however, we will refrain from plotting as much as we did in Figure 10.2 and provide plots corresponding to one method. The estimated PRL based on *only* the LWT per method are given in Table 10.3. From Table 10.3 it is evident that the differences in estimated PRL are minimal: the largest difference for the polynomial extrapolation is 18 days, whereas 23 is the largest for the linear; both per the log-normal and the logistic method.

A Remark on the Proposed Remaining Lifetimes

For Rijkswaterstaat and the contractors who perform maintenance on roads, one specific day might be a harsh cut-off and we acknowledge that it is not very desirable. However, the current implementation does not allow for a mathematically justified confidence interval for the PRL. Future research could very well explore the possibilities of such intervals.

10.1.3. Predictions with the $n - i$ approach

With respect to the $n - i$ approach, it raises the question how the extrapolations would have been if we removed the last few quantiles from the sequence of quantiles. The corresponding curves and

Table 10.3: Estimates for the proposed remaining lifetime (PRL) of the $W = A44, S = 1\text{HRR}^{1\text{RR}}_{7.1}$ based on the LWT.

	Threshold Surpassing Date[1]		**PRL**	
Method	*Polynomial*	*Linear*	*Polynomial*	*Linear*
Log-normal	2020-02-24	2020-04-06	331 days	373 days
Logistic	2020-03-13	2020-04-29	349 days	396 days
Empirical	2020-03-03	2020-04-16	339 days	383 days
Kernel	2020-03-01	2020-04-14	337 days	381 days

Table 10.4: Differences in the estimated proposed remaining lifetime (PRL) of the $W = A44, S = 1\text{HRR}^{1\text{RR}}_{7.1}$ based on the LWT and excluding the most recent quantile estimate.

	Threshold Surpassing Date		**ΔPRL**	
Method	*Polynomial*	*Linear*	*Polynomial*	*Linear*
Log-normal	2019-09-16	2020-02-14	161 days	52 days
Logistic	2019-10-06	2020-03-16	159 days	44 days
Empirical	2019-10-08	2020-03-21	147 days	26 days
Kernel	2019-10-07	2020-03-19	146 days	26 days

numbers for $i = 1$ are given in Figure 10.3 and Table 10.4 for only the log-normal[2] plots in Figure 10.3. The last column of Table 10.4 denotes how much later the threshold is reached compared to the fit with all quantiles: a ΔPRL of 52 means that data up until 2018 implied a surpassed threshold 52 days **before** the estimate including 2019 data. For this particular prototype, PRLs calculated using 2019 data too were farther in the future than using data up until 2018. Furthermore the ΔPRL are surprisingly small: if the estimate from 2019 can be used as a 'correct' reference, then in 2018 we would only have been off by 52 days at max for the prototype.

It should be acknowledged that there is no way to classify if the PRL estimate from 2019 is more accurate than the one from 2018. The only remark we can add is that small differences between these estimates from consecutive years is beneficial as it admits consistency throughout extrapolations over the years.

For $i = 1$ the predictions seem to be reasonable, but for $i = 2$ this changes drastically. Notice that by removing the quantile estimates in 2018 and 2019, the remaining quantile estimates do not seem to be a representation of a proper monotone increasing convex curve. If the estimates were more in line with the usual behaviour, the predictions are not as extreme. To showcase this, we manually modify the values for the log-normal approach: the problem mainly lies in the three intermediate values in 2014 – 2016. Figure 10.4 shows how forcing a curve through the manipulated points and the other points leads to a hugely improved fit over the latter linear and polynomial fits — which in this case coincide. Figure 10.4 should help with emphasising the value in accuracy of

[2] The choice for the log-normal is arbitrary: due to the small differences in estimates any other approach would also have brought the point across, yet it makes Figure 10.3 less cluttered.

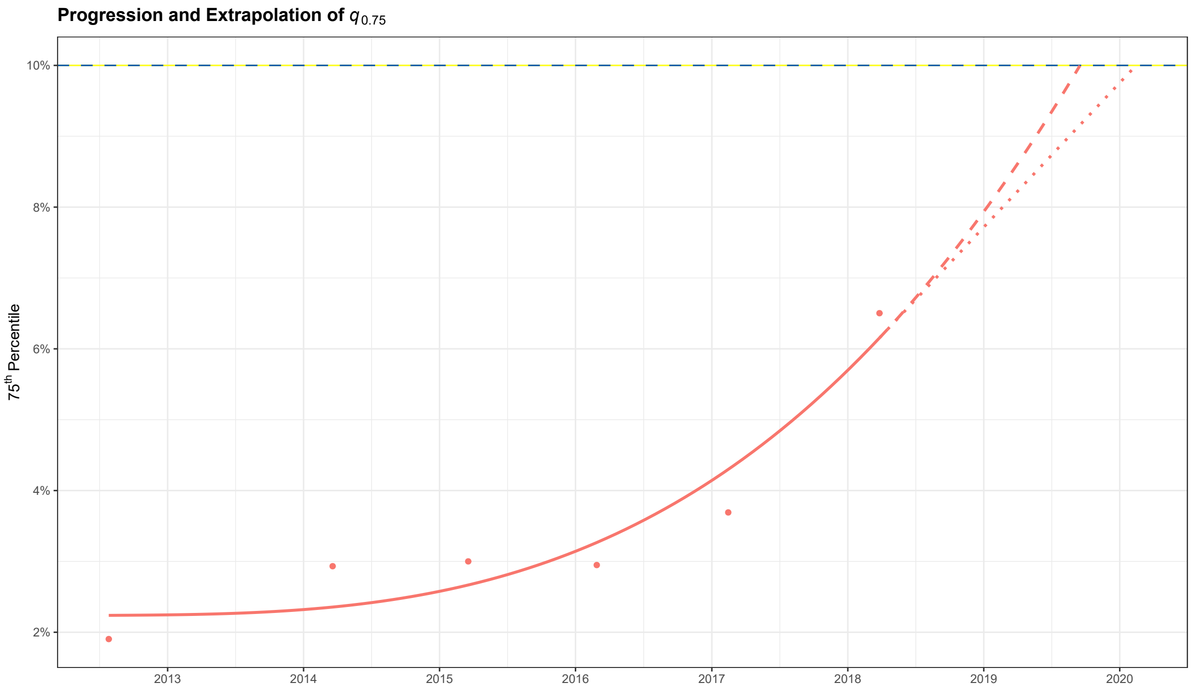

Figure 10.3: Linear and polynomial extrapolations of log-normal $q_{0.75}$ estimates from 2012 – 2018 plotted for the LWT of $W = A44, S = 1\text{HRR}^{1\text{RR}}_{7.1}$ after fitting smooth monotone curves with P-splines as basis with dimension $d = 4$ and order $p = 4$ (cubic spline). The curved dashed line represents the polynomial extrapolation, while the dotted line represents the linear extrapolation.

the GPS mechanism: the curve generally tends more towards the logical assumption (montone increasing convex) as the accuracy increases, hence towards a more accurate PRL approximation as the accuracy increases. Despite that, Figure 10.4 also illustrates that the further away the last used estimated quantile is from the threshold, the more the predicted PRL is off from the reference PRL (2019). Furthermore we see how the conservative linear extrapolation differs more than the polynomial extrapolation with respect to the reference PRL: the former indicates a PRL around the third quarter of 2023, while the latter indicates a PRL during the second quarter of 2020 — much more in line with the reference PRL.

Whereas the PRL estimates in Table 10.3 between the polynomial and linear predictions shows small differences, these differences are much more apparent in Table 10.4. The reason for that is that the threshold of 10 was closer under the conditions of the latter; it agrees with how the polynomial and linear predictions are close when the predictor variable is near the support of the smooth curve.

10.1.4. Predictions using the RWT

All results mentioned up till now were for the LWT of the prototype road section. For maintenance, the official documents which state the threshold [6 and 12] consider both wheel tracks. This can be easily translated to the data we have at hand, while the lane-wide aggregate loss is much more ambiguous in terms of definitions. The reason for the initial focus on the LWT was arbitrary: in Chapter 7 we opted to show concepts for the LWT as the general ideas are similar for the RWT.

The naive approach – not bad per se — of incorporating RWT data is to find the date for which the data averaged over the LWT and RWT surpass the threshold. It is the most obvious manner for the resulting PRL to be based on both wheel tracks and in fact is mentioned in DHV et al. [12]. However, we do not agree with this methodology as taking the average in such a case will generally result in lower quantiles for $p > 0.5$. That is, averaging will diminish the observed severity in the individual wheel tracks. This is contrary to the threshold, which implicitly looks at the *worst* performing 25%

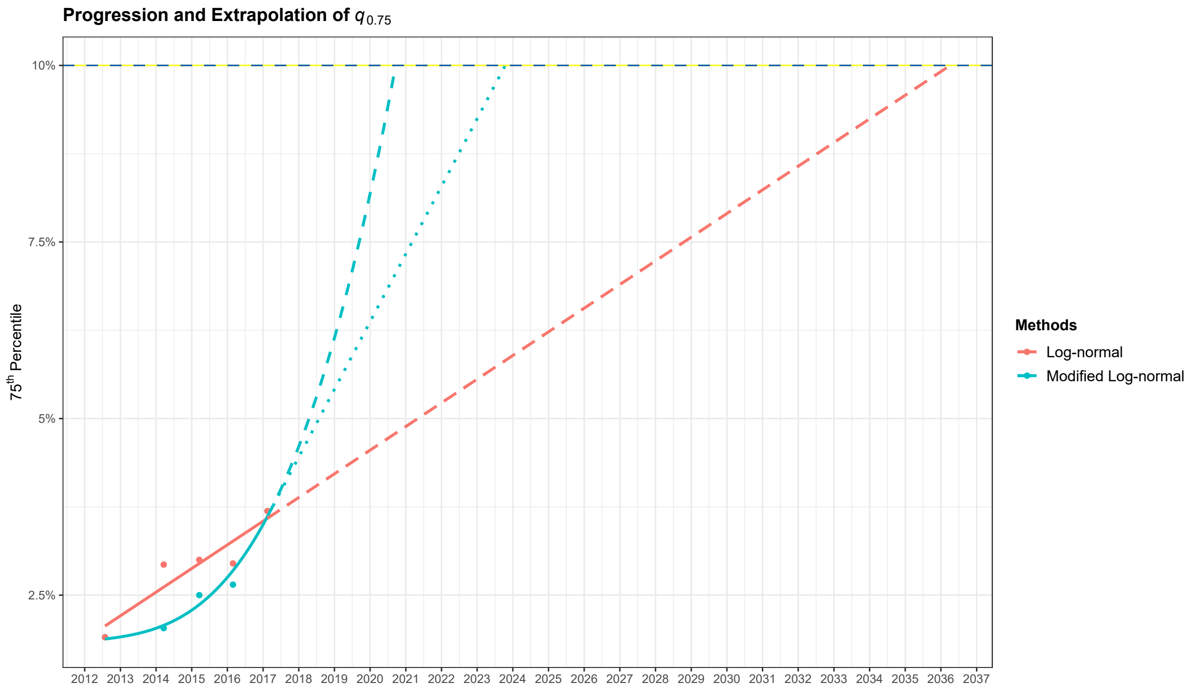

Figure 10.4: Linear and polynomial extrapolations of log-normal and modified log-normal $q_{0.75}$ estimates from 2012 – 2017 plotted for the LWT of $W = A44, S = 1\text{HRR}^{1\text{RR}}_{7.1}$ after fitting smooth monotone curves with P-splines as basis with dimension $d = 4$ and order $p = 4$ (cubic spline). The curved dashed lines represent the polynomial extrapolation, while the dotted lines represent the linear extrapolation. For the regular log-normal estimates these extrapolations coincide. The dashed horizontal line is the RWS threshold.

of the road section.

Rather, the approach we would commend is to group the data for the LWT and RWT and perform density estimation on the pooled data. A PRL based on the joint data is naturally based on both wheel tracks, and we do not downplay the severity of ravelling by averaging: this happens specifically if the severities in both wheel tracks are not of the same order of magnitude. Nevertheless we do respect the respective interpretation of users of DOS-LCMS data and their decisions in how the data should be interpreted, but we choose a different practice. In this case, by observation of a location difference in the LWT and RWT from Figure 7.3, the density will likely be a *mixed*-density or at the very least be bimodal. Figure 10.5 indeed shows the bimodality by aggregating the LWT and RWT data.

For a potential parametric approach for the aggregated data, in general it is possible to find a mixed parametric density. For the sake of presenting results for multiple road sections, however, such a step is an obstacle. The logical circumvention is then to perform non-parametric methods in order to present the progression of aggregate loss. In particular when observing that our suitable parametric densities result in very similar quantile estimates as the non-parametric approach with respect to the prototype, the latter approach is more favourable.

A familiar problem emerges if we were to use the mixed data from the prototype. Mixing the data results in estimates which showed no monotonic **convex** increasing pattern, shown in Table 10.5 and Figure 10.6. In particular, the threshold of 10 would have been reached in 2018 already, while the preceding $q_{0.75}$ estimates would have indicated a maintenance required over 14(!) years later seen in Figure 10.6. This should yet again stress the significance of a pattern of convexity for a proper PRL prediction. For these estimates specifically, it is unsettling to see how the values in years 2014 – 2016 steadily decrease. This problem will be prominent in a later Section as well, but an attempt to tackle it along with other emerging problems will be given in § 10.3.

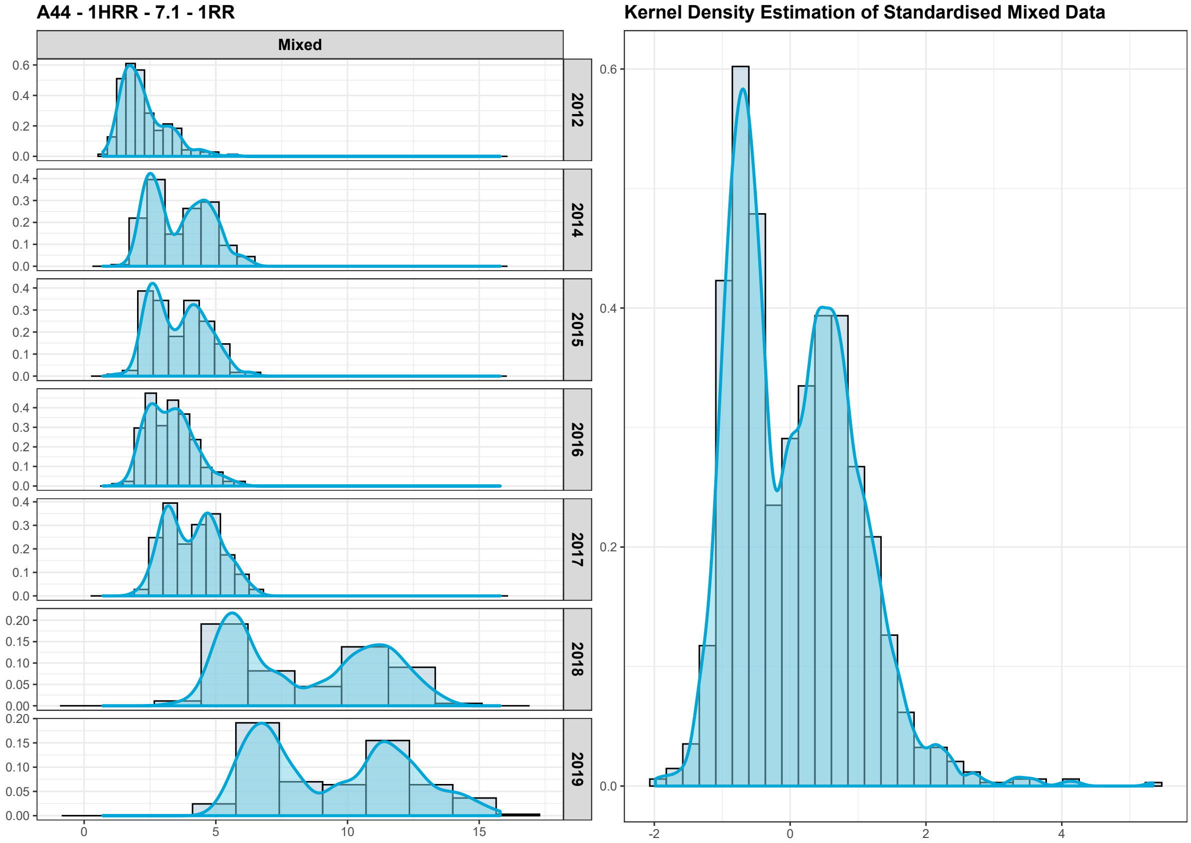

Figure 10.5: Kernel density approximations of mixing LWT and RWT from 2012 – 2019 plotted for $W = A44, S = 1\text{HRR}_{7.1}^{1\text{RR}}$. The left plot shows density approximations per year, while the right one shows the density estimation of the aggregated standardised data over all respective years.

Table 10.5: Kernel $q_{0.75}$ estimates of $W = A44, S = 1\text{HRR}_{7.1}^{1\text{RR}}$

Date	LWT	RWT	Mixed
2012-07-27	1.89	3.31	2.49
No 2013 data			
2014-03-20	2.89	5.08	4.67
2015-03-19	2.96	4.87	4.41
2016-02-27	2.86	4.21	3.86
2017-02-14	3.59	5.31	4.99
2018-03-27	6.40	12.00	10.25
2019-03-30	7.59	12.70	12.18

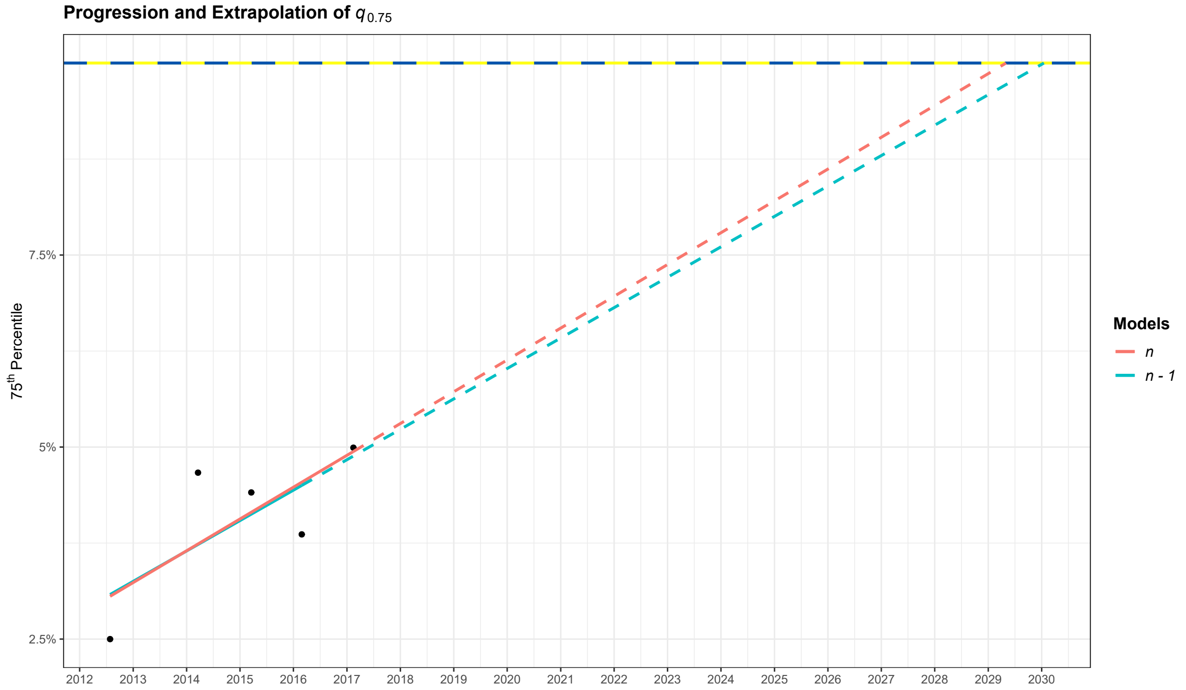

Figure 10.6: Polynomial extrapolations of kernel $q_{0.75}$ estimates for mixed LWT and RWT data from 2012 – 2017 plotted for $W = A44, S = 1\text{HRR}^{1\text{RR}}_{7.1}$ after fitting smooth monotone curves with P-splines of dimension $d = 4$ and order $p = 4$ (cubic spline). The curved dashed lines represent the polynomial extrapolations. The dashed horizontal line is the RWS threshold.

10.1.5. Comparing Wheel Tracks and Lanes

For the prototype we have seen that the LWT and RWT showed a difference significant enough to see that the mixed density is multimodal. It brings up the issue of whether this can be observed systematically or not. Additionally we can analyse the differences per lane given that there are at least two lanes for a road section. To illustrate the difference we will explore the sample year $q_{0.75}$ values of the data for both wheel tracks for every road section and every year that is available for $W = A44$. For a non-skewed representation of the results, we exclude the road sections associated with engineering structures and also exclude measurements from all road sections after maintenance has been performed[3].

The results for $W = A44$ are summarised in Figure 10.7. It is clear that for $W = A44$, the aggregate loss of the LWT *generally* is lower than on the RWT for the 1RR. It is slightly harder to see for the 2RR lane. As by logical reasoning the 2RR lane seems to suffer more from aggregate loss than the 1RR lane due to the nature of the traffic[4]. From 2016 onward there seems to be no sign of 2RR data, but this corresponds to the fact that (partial) maintenance had been executed in 2016.

Figure 10.7 also indicates that our previously determined maintenance dates from the administration might be incomplete if not wrong, or that the DOS-LCMS data we used was invalid. Notice how in 2016 and 2017 for the 1RR the sample year empirical $q_{0.75}$ seem to be close to 0 for both LWT and RWT, indicating that maintenance has been executed. However, we have already excluded these measurements as a precondition based on our prior information, which makes the inclusion of such values peculiar.

Unlike for the contrasts per lane, the discrepancy between the wheel tracks cannot be explained by traffic flow only. Even though the phenomenon cannot be directly clarified from the data we have

[3] Usually the exact date of maintenance is complicated to retrieve. That is why we subtract one year from the actual date and fix this date whereafter measurements are dropped.

[4] The rightmost lane(s) have to endure more load from trucks and by traffic rules (in the Netherlands) in general.

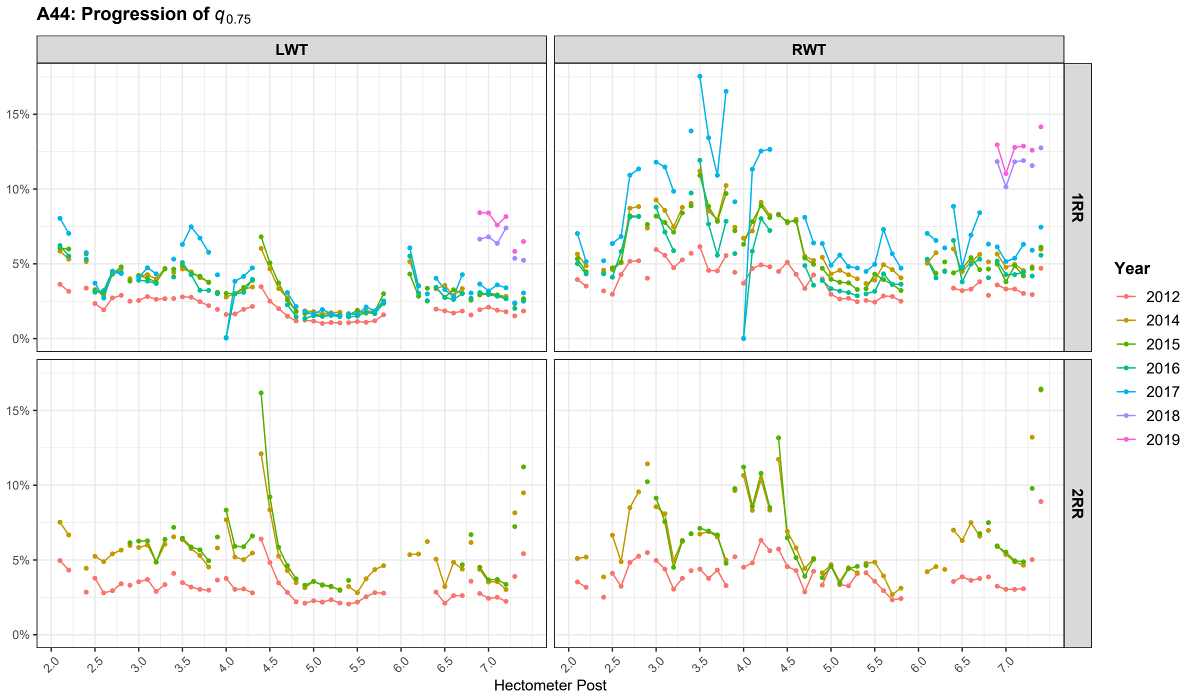

Figure 10.7: Sample empirical $q_{0.75}$ values per year per wheel track and per lane plotted for the road sections of $W = A44$.

at our hands and therefore is also out of scope, Léon Schouten has proposed that the cause might lie in the *cross slope* of a road. The cross slopes are introduced such that (rain)water or liquids in general can drain from the road. If this proposition turns out to be true, we would be able to deduce the direction of which the cross slope drains water towards by viewing the most ravelled wheel track.

10.2. Results for Roads

Now that the prototype has been discussed, we will attempt to extend these results to the corresponding road and even other roads. In particular we would like to see what the PRL estimates give and if there is consistency with the $n-i$ approach. The road sections which are considered have to satisfy the properties mentioned in Chapter 6 and which we will reiterate:

- The intensity of ravelling surpasses or is close to 10% for the empirical quantile in some year, allowing comparisons with the $n-i$ approach for at least $i = 1$.
- There are at least $n = 5$ values of consecutive $q_{0.75}$ estimates in order to fit the curve for the n and $n-1$ percentiles.
- The data contains Obstacle 2 of § 5.2.3, such that a workaround can be presented.

The first condition is quite easily satisfied if we already found one prototype of some road, since the state of the road across its sections is quite comparable conditioned on the sections which have not undergone maintenance.

We will adhere to the interpretation of the LWT and RWT being the most important indicators and decide to look at the combined values of the LWT and RWT per year. As location differences of the LWT and RWT values cannot be assumed to be near 0, our current framework with the parametric approach will not be able to provide reasonable estimates. Therefore we will approximate values of $q_{0.75}$ per year by the non-parametric approach using the kernel method[5]. Smooth convex

[5] The empirical estimates were also checked, but showed minimal differences to the kernel estimates.

monotonic curves will be fit to these estimates using P-splines of order 4. The predictions will be based on polynomial extrapolation, as it performs as linear extrapolation at worst.

10.2.1. A44

The first road to consider will be $W = A44$, associated to our prototype. It had its most recent pavement rehabilitation on 2002-09-09 — or sometime close to this date for that matter. The DOS-LCMS data that is available to us for the $W = A44$ is from one stretch of road. The intermediate maintenance it has undergone is given in Table A.2 in Appendix A.

Table 10.6 provides an overview of estimates for $W = A44$ and its road sections $S_j = 1\text{HHR}_j^{1\text{RR}}$. It should be viewed in combination with Figure 10.9, of which a PDF file illustrating each plot in greater detail can be found on the Github page. Recall that Table 10.6 and Figure 10.9 are based on the joint data of LWT and RWT per year. That might clarify discrepancies between visualisations in Figure 10.9 and the ones shown earlier in Figures 10.2 and 10.3.

From both Table 10.6 and Figure 10.9 we generally see how the computed $q_{0.75}$ estimates are not sufficiently convex in pattern. Additionally the $n-1$ prediction suffers from inconsistency with respect to the n prediction due to this lack of convexity in the estimates. The lack of consistency can be deduced by observing the values of $|\Delta\text{PRL}|$: the lower this measure, the more consistent the prediction is in consecutive years. Possible solutions to tackle this problem will be discussed in § 10.3 along with other emerging problems.

Table 10.6: Lifetime predictions for $W = A44$ and its road sections using kernel approximations and polynomial extrapolations. n denotes the number of consecutive $q_{0.75}$ values used for the calculation, PRL = Proposed Remaining Lifetime (days relative to the date of the last percentile used), ΔPRL = difference in PRL between prediction n and $n-1$ (negative implies that prediction $n-i$ was $|\Delta\text{PRL}|$ further in future time).

C.way	Hm	Lane	n	$q_{0.75}$	Age	Date	TSD	PRL	ΔPRL
1HRR	2.1	1RR	5	7.49	14.44	2017-02-14	2021-03-11	1486	-568
1HRR	2.2	1RR	5	6.01	14.44	2017-02-14	2025-01-11	2888	-12
1HRR	2.4	1RR	5	5.68	14.44	2017-02-14	2026-06-04	3397	157
1HRR	2.5	1RR	5	5.57	14.44	2017-02-14	2022-07-16	1978	-4734
1HRR	2.6	1RR	5	5.34	14.44	2017-02-14	2027-09-04	3854	-246
1HRR	2.7	1RR	5	9.92	14.44	2017-02-14	2017-10-02	230	-857
1HRR	3	1RR	5	9.09	14.44	2017-02-14	2019-03-15	759	-595
1HRR	3.2	1RR	5	9.08	14.44	2017-02-14	2018-07-31	532	-2053
1HRR	3.7	1RR	5	8.86	14.44	2017-02-14	2019-09-12	940	-4730
1HRR	3.9	1RR	5	7.98	14.44	2017-02-14	2020-04-25	1166	-2668
1HRR	4.1	1RR	5	9.12	14.44	2017-02-14	2018-12-14	668	-1179
1HRR	4.2	1RR	5	9.91	14.44	2017-02-14	2017-06-01	107	-108
1HRR	4.7	1RR	5	6.45	14.44	2017-02-14	2021-05-19	1555	-11055
1HRR	4.8	1RR	5	5.42	14.44	2017-02-14	2025-08-01	3090	-24218

C.way	Hm	Lane	n	$q_{0.75}$	Age	Date	TSD	PRL	ΔPRL
1HRR	4.9	1RR	6	9.47	15.56	2018-03-27	2018-07-12	107	-1336
1HRR	5	1RR	5	4.52	14.44	2017-02-14	2031-08-23	5303	-14421
1HRR	5.1	1RR	5	5	14.44	2017-02-14	2023-05-20	2286	-17354
1HRR	5.2	1RR	6	9.06	15.56	2018-03-27	2018-09-08	165	-1968
1HRR	5.3	1RR	6	7.79	15.56	2018-03-27	2019-01-29	308	-1945
1HRR	5.4	1RR	6	7.80	15.56	2018-03-27	2019-02-09	319	-3979
1HRR	5.5	1RR	6	9.36	15.56	2018-03-27	2018-09-20	177	-9355
1HRR	5.6	1RR	5	5.70	14.44	2017-02-14	2024-05-11	2643	-2890
1HRR	5.7	1RR	6	9.42	15.56	2018-03-27	2018-07-13	108	-1417
1HRR	5.8	1RR	6	8.80	15.56	2018-03-27	2018-10-13	200	-4841
1HRR	6.4	1RR	5	7.29	14.44	2017-02-14	2019-05-25	830	-2183
1HRR	6.5	1RR	6	9.69	15.56	2018-03-27	2018-08-14	140	-8019
1HRR	6.6	1RR	5	5.74	14.44	2017-02-14	2025-09-22	3142	-692
1HRR	6.8	1RR	6	4.28	15.56	2018-03-27	2034-11-04	6066	2020
1HRR	6.9	1RR	5	5.59	14.44	2017-02-14	2026-09-11	3496	-502
1HRR	7.1	1RR	5	4.96	14.44	2017-02-14	2029-06-13	4502	-264
1HRR	7.2	1RR	5	6.04	14.44	2017-02-14	2020-10-06	1330	-3791
1HRR	7.3	1RR	5	5.20	14.44	2017-02-14	2023-11-03	2453	-3874
1HRR	7.4	1RR	5	6.33	14.44	2017-02-14	2024-12-04	2850	-771

10.2.2. A50

A second road we wish to analyse is $W = A50$, of which we in fact have two stretches of the road. The first stretch ranges from hectometer 139.9 – 148.8 and has low expected severity, and the second stretch ranges from hectometer 202.9 - 205.5 of which the expected severity is higher. We will treat these two separately.

139.9 – 148.4

According to KernGIS, the pavement rehabilitation date was around 2012-09-17. As the available DOS-LCMS data is from 2012 – 2019, the initial expectation is that we will not see a high intensity of ravelling. This indeed turns out to be the case seen in Figure 10.8. It visualises how for each road section defined by its starting hectometer value, the empirical quantile of the sample computed from each year barely increases. For example, the peak for $S = 1\text{HRR}^{1\text{RR}}_{144.6}$ indicates that for the 1RR lane from hectometre post 144.6 till 144.7, the sample (empirical) $q_{0.75}$ of that year for the LWT was around 4%. The hectometres for which no lines are drawn at all correspond to the engineering structures which, as we recall, have different build-up of wear. Figure 10.8 includes road sections with measurements which were taken at least one year later than its most recent mainte-

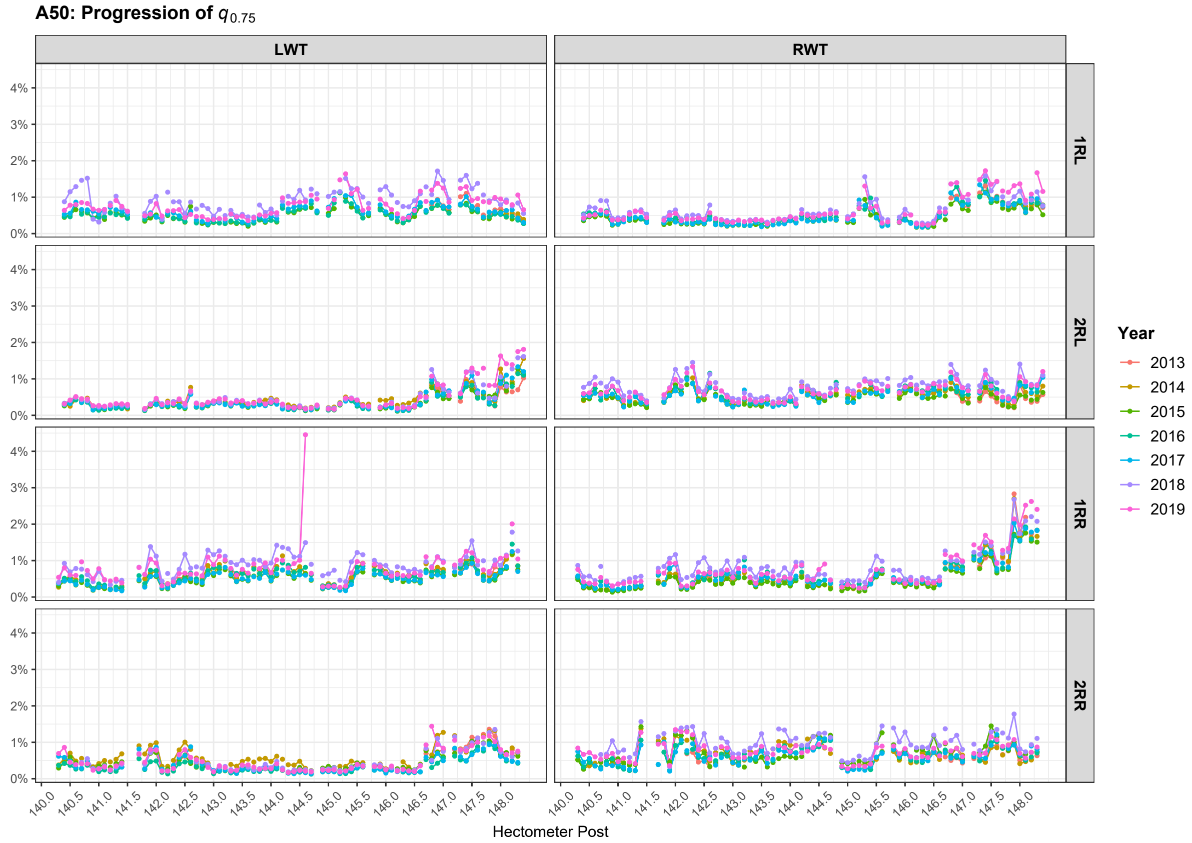

Figure 10.8: $q_{0.75}$ values per year per wheel track and per lane plotted for $W = A50$ and road sections $S_{ij} = 1\text{HRR}^{\text{i}}_{j}$ for $i \in \{1\text{RL}, 2\text{RL}, 1\text{RR}, 2\text{RR}\}$ and $139.9 \leq j \leq 148.4$.

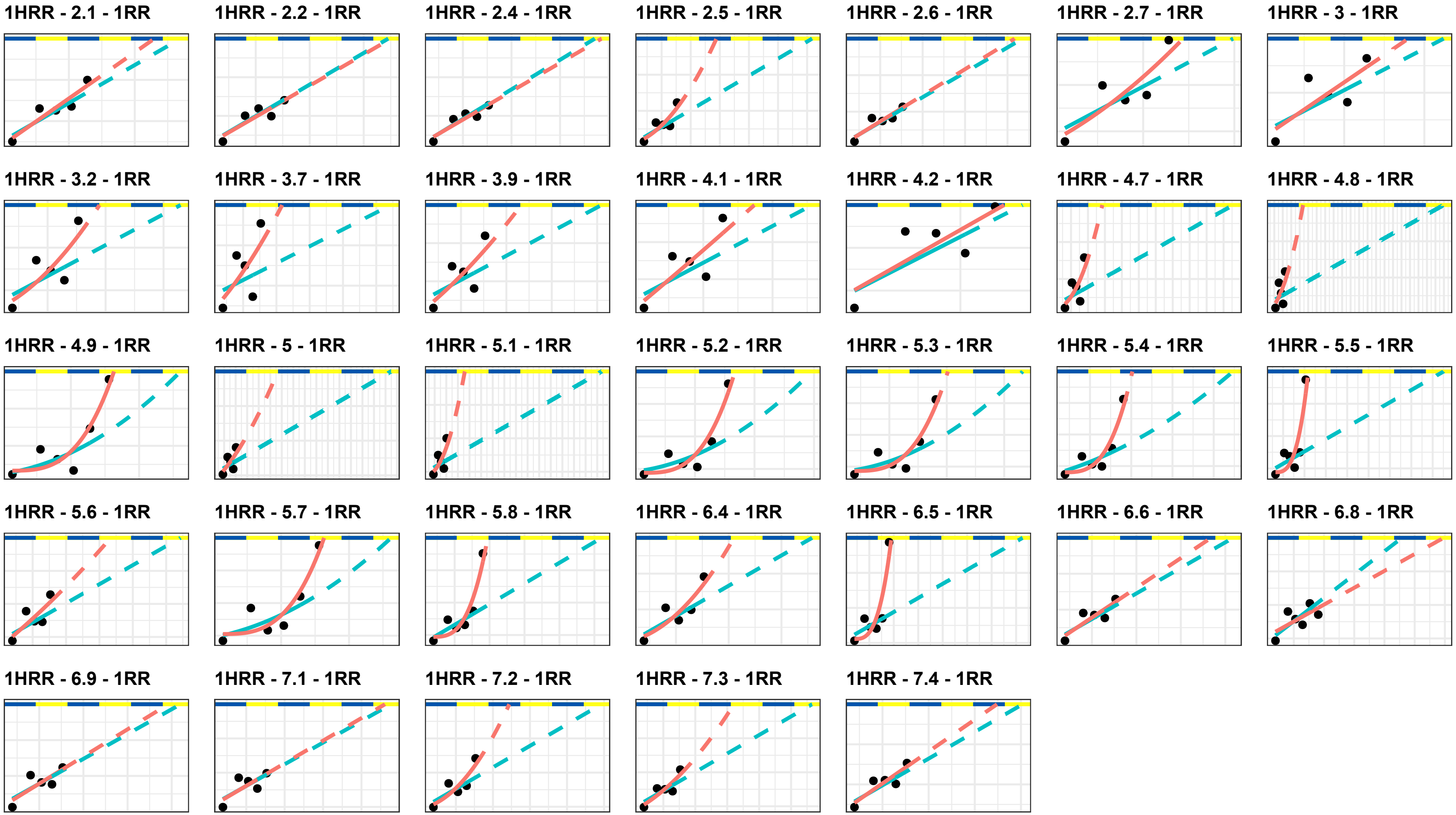

Figure 10.9: Polynomial extrapolations of $q_{0.75}$ estimates plotted for $W = A44$ and its road sections based on the joint data of LWT and RWT. The blue curves represent the fits with $n-1$ $q_{0.75}$ values, while the red curves use n $q_{0.75}$ values; all curves are fit using P-splines of order 4. The dashed curves are the extrapolations based on the solid curves. The dashed horizontal lines represent the RWS threshold.

nance. The reason for this, as Léon Schouten suggests, is that the DOS-LCMS measurements can be unreliable for new surfaces. Aside from the odd value of $S = 1\text{HRR}^{1\text{RR}}_{144.6}$ in 2019 the road sections from this stretch of the road seem not to show an increase significant enough to be able to apply our prediction scheme.

To confirm or disprove this observation, let us consider $S = 1\text{HRR}^{1\text{RR}}_{145.5}$. This road section has undergone its last maintenance on 2013-03-19 according to Table B.2 in Appendix B. In Figure 10.10 we have fit a smooth monotone convex curve using P-splines of order 4 and we see something surprising. Notice that for a prediction using all available $n = 6$ estimates of $q_{0.75}$, the prediction seems less realistic than using $n = 5$ estimates of $q_{0.75}$. This can be explained by the value of the last $q_{0.75}$ estimate, which is relatively much lower than initially forecast in the preceding year. It goes to show that a single additional point estimate can change the prediction drastically for a small set of data points, which is yet another problem which will be discussed in § 10.3. We will not provide the predictions of lifetimes in Tables of Figures for this stretch of the road as it makes little sense to do it for these low values as observed in Figure 10.10.

202.9 – 205.5

If once again we rely on KernGIS, the date of the pavement rehabilitation was around 2002-07-12. It indicates that we can expect more intensely ravelled road sections. Figure 10.11 indeed does show that this is the case for lane 1RL, whereas 2RL shows significantly less aggregate loss. It implies that we should expect to see more sensible predictions for the 1RL than the 2RL lane, if the latter admits predictions at all. $S = 1\text{HRL}^{1\text{RL}}_{204.8}$ is an interesting outlier because in Figure 10.11, the jump in aggregate loss is the most extreme in both the LWT and RWT. We show the progression on a meter level in Figure 10.12. The plot could either signal the margin of error in GPS accuracy of the measurements, or it could mean that the build-up of aggregate loss gradually shifts along the lane. More importantly though is the incredibly high values for aggregate loss, reaching over 40% on the RWT. A quick look-up in the provided DOS-LCMS data does not show multiple measurements for this year, so assuming that the measurements are valid, it certainly does explain our observations from Figure 10.11. An overview of lifetime predictions for this stretch of $W = A50$ given in Figure 10.15, and Table 10.7 in Appendix B.

Table 10.7: Lifetime predictions for $W = A50$ and its road sections using kernel approximations and polynomial extrapolations. n denotes the number of consecutive $q_{0.75}$ values used for the calculation, PRL = Proposed Remaining Lifetime (days relative to the date of the last percentile used), ΔPRL = difference in PRL between prediction n and $n-1$ (negative implies that prediction $n-i$ was |ΔPRL| further in future time).

C.way	Hm	Lane	n	$q_{0.75}$	Age	Date	TSD	PRL	ΔPRL
1HRL	203.2	1RL	5	5.59	14.52	2017-07-04	2019-09-09	797	722
1HRL	203.3	1RL	5	4.13	14.52	2017-07-04	2022-12-10	1985	-1979
1HRL	203.4	1RL	5	4.96	14.52	2017-07-04	2020-11-07	1222	-2564
1HRL	203.5	1RL	5	7.44	14.52	2017-07-04	2018-08-22	414	275
1HRL	203.6	1RL	5	5.02	14.52	2017-07-04	2021-09-24	1543	801
1HRL	203.7	1RL	6	9.02	15.51	2018-07-02	2018-11-15	136	-1546
1HRL	203.8	1RL	6	8.48	15.51	2018-07-02	2018-12-22	173	-1517
1HRL	203.9	1RL	6	5.53	15.51	2018-07-02	2020-03-02	609	-1417

C.way	Hm	Lane	n	$q_{0.75}$	Age	Date	TSD	PRL	ΔPRL
1HRL	204	1RL	6	5.03	15.51	2018-07-02	2020-05-23	691	-8326
1HRL	204.1	1RL	6	6.05	15.51	2018-07-02	2020-03-08	615	-3621
1HRL	204.2	1RL	5	4.80	14.52	2017-07-04	2019-10-28	846	-562
1HRL	204.3	1RL	6	9.22	15.51	2018-07-02	2018-09-25	85	-788
1HRL	204.4	1RL	6	8.88	15.51	2018-07-02	2018-12-06	157	-502
1HRL	204.5	1RL	5	4.09	14.52	2017-07-04	2021-11-26	1606	-74
1HRL	204.6	1RL	5	5.74	14.52	2017-07-04	2021-06-23	1450	987
1HRL	204.7	1RL	5	7.58	14.46	2017-06-12	2018-07-03	386	161
1HRL	204.8	1RL	5	9.95	14.46	2017-06-12	2017-06-29	17	37
1HRL	205.5	2RL	5	3.18	14.99	2017-07-04	2021-01-15	1291	-278

10.2.3. A6

The last road which we wish elaborate on is $W = A6$. Figure 10.13 allows us to gain an idea of its situation. The most staggering part in the visualisation is $S = 1\text{HRR}^{1\text{RR}}_{285.8}$. A meter level plot is given in Figure 10.14. Aside from the non-optimal alignment of the ‘bobble’ centered at the 72nd meter, the maximum aggregate loss on meter level from 2017 to 2018 and 2019 seems to have increased by a factor of 2(!). From the perspective of $q_{0.75}$, the sample quantile (empirical) on the RWT increased from approximately $q_{0.75} = 5$ to approximately $q_{0.75} = 12.8$ in 2018, and $q_{0.75} = 20.5$ in 2019. The difference between the LWT and RWT for this road section is stark, and especially in 2019: whereas the LWT seems to only suffer extreme aggregate loss at meters 40 to 50, the RWT shares the same extremeness for meters 15 to 50. This could partially be explained by Figure 10.14: the difference in ravelling build-up is quite evident. While the LWT showed the ‘bobble’ much clearer and in earlier stages, its other meters were quite undamaged. The RWT on the other hand did not show this ‘bobble’ as clearly, but already depicted that a longer stretch of its track underwent ravelling build-up.

For $W = A6$, we found a total of 270(!) possible extrapolations. In line with the first stretch of $W = A50$, we will only show the results for the most ravelled sections. Specifically we condition on road sections which suffered more than 6% aggregate loss in either wheel track and in either lane for some year. The interested reader can view the PDF file on the Github page which contains all of the extrapolations. Having said that, Figure 10.16 shows the extrapolations for the remaining sections and Table 10.8 contains the exact corresponding values.

Table 10.8: Lifetime predictions for $W = A6$ and its road sections using kernel approximations and polynomial extrapolations. n denotes the number of consecutive $q_{0.75}$ values used for the calculation, PRL = Proposed Remaining Lifetime (days relative to the date of the last percentile used), ΔPRL = difference in PRL between prediction n and $n-1$ (negative implies that prediction $n-i$ was |ΔPRL| further in future time).

C.way	Hm	Lane	n	$q_{0.75}$	Age	Date	TSD	PRL	ΔPRL
1HRR	280.3	1RR	8	6.73	13.41	2019-04-12	2022-01-23	1017	-398

C.way	Hm	Lane	n	$q_{0.75}$	Age	Date	TSD	PRL	ΔPRL
1HRR	280.8	1RR	8	5.94	13.41	2019-04-12	2021-01-07	636	-136
1HRR	283.8	1RR	8	5.79	13.41	2019-04-12	2020-12-19	617	-124
1HRR	283.8	2RR	5	5.60	10.34	2016-03-16	2018-12-24	1013	-3423
1HRR	285.8	1RR	7	7.76	12.48	2018-05-07	2019-02-08	277	-291
1HRR	285.8	2RR	5	5.56	10.34	2016-03-16	2019-03-31	1110	-13961
1HRR	288.6	1RR	8	3.54	13.62	2019-04-12	2023-01-30	1389	119
1HRR	288.7	1RR	8	3.49	13.62	2019-04-12	2023-05-25	1504	128
1HRR	288.7	2RR	8	7.55	13.63	2019-04-14	2019-11-11	211	141
1HRR	288.8	1RR	8	3.48	13.62	2019-04-12	2023-07-01	1541	-10
1HRR	288.8	2RR	8	9.14	13.63	2019-04-14	2019-05-25	41	92
1HRR	288.9	1RR	8	3.51	13.62	2019-04-12	2023-08-26	1597	111
1HRR	288.9	2RR	8	6.61	13.63	2019-04-14	2020-03-09	330	176
1HRR	290.4	1RR	8	2.58	13.62	2019-04-12	2024-12-01	2060	255
1HRR	290.4	2RR	8	4.83	13.63	2019-04-14	2021-02-23	681	177
1HRR	291.1	1RR	8	2.42	13.62	2019-04-12	2025-01-27	2117	433
1HRR	291.1	2RR	8	5.01	13.63	2019-04-14	2021-02-27	685	150
1HRR	293.4	1RR	8	3.26	13.62	2019-04-12	2024-11-07	2036	-505
1HRR	293.4	2RR	8	4.97	13.63	2019-04-14	2021-02-16	674	139
1HRR	293.5	1RR	8	3.63	13.62	2019-04-12	2023-08-26	1597	-146
1HRR	293.5	2RR	8	5.57	13.63	2019-04-14	2020-09-06	511	151
1HRR	293.7	1RR	8	2.66	13.62	2019-04-12	2025-06-15	2256	99
1HRR	293.7	2RR	8	4.08	13.63	2019-04-14	2022-02-28	1051	21
1HRR	293.8	1RR	8	2.48	13.62	2019-04-12	2026-08-22	2689	-9378
1HRR	293.8	2RR	8	4.71	13.63	2019-04-14	2021-03-12	698	262
1HRR	293.9	1RR	8	2.55	13.62	2019-04-12	2025-08-29	2331	266
1HRR	293.9	2RR	8	5.30	13.63	2019-04-14	2020-09-12	517	258
1HRR	294.0	1RR	8	3.18	13.62	2019-04-12	2024-07-02	1908	-118
1HRR	294.0	2RR	8	4.80	13.63	2019-04-14	2021-02-27	685	223

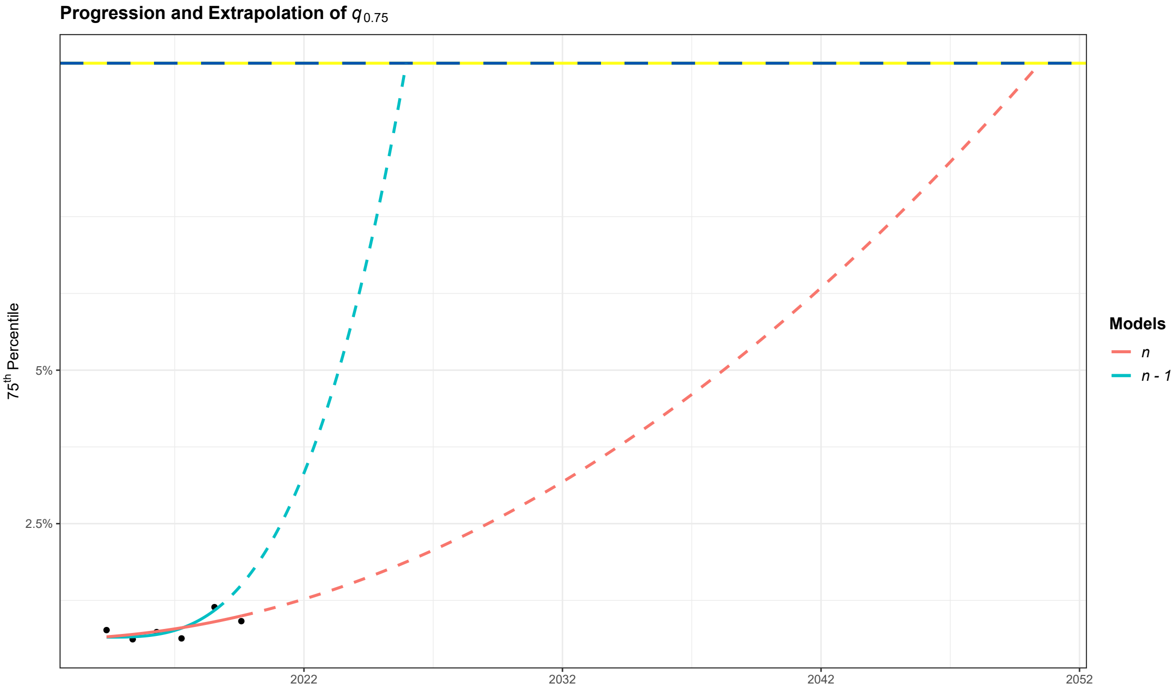

Figure 10.10: Polynomial extrapolations of kernel $q_{0.75}$ estimates from 2013 – 2019 plotted for the the mixed data (LWT and RWT) of $W = A50, S = 1\text{HRR}^{1\text{RR}}_{145.5}$ after fitting smooth monotone curves with P-splines as basis with dimension $d = 4$ and order $p = 4$ (cubic spline). The dashed curved lines represent the extrapolations, while the dashed horizontal line is the threshold from RWS.

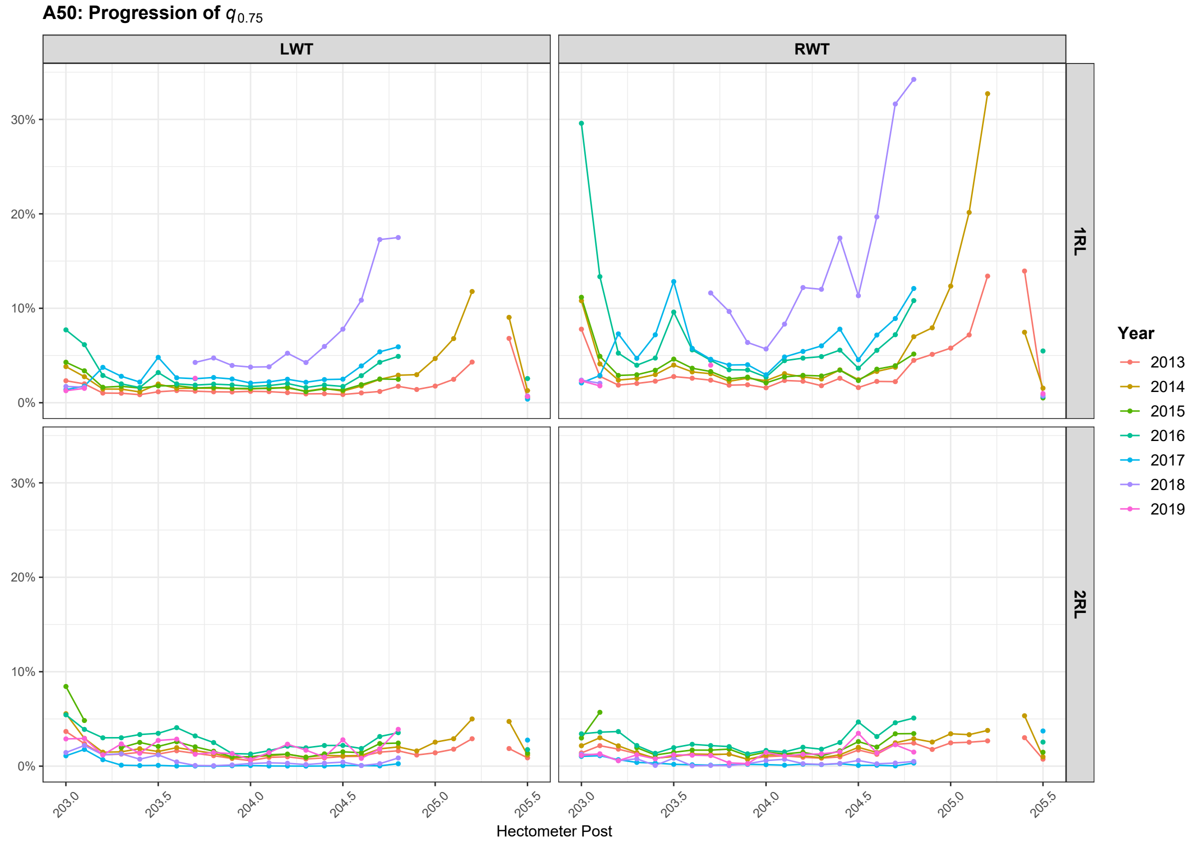

Figure 10.11: Sample empirical $q_{0.75}$ values per year per wheel track and per lane plotted for the road sections of $W = A50$.

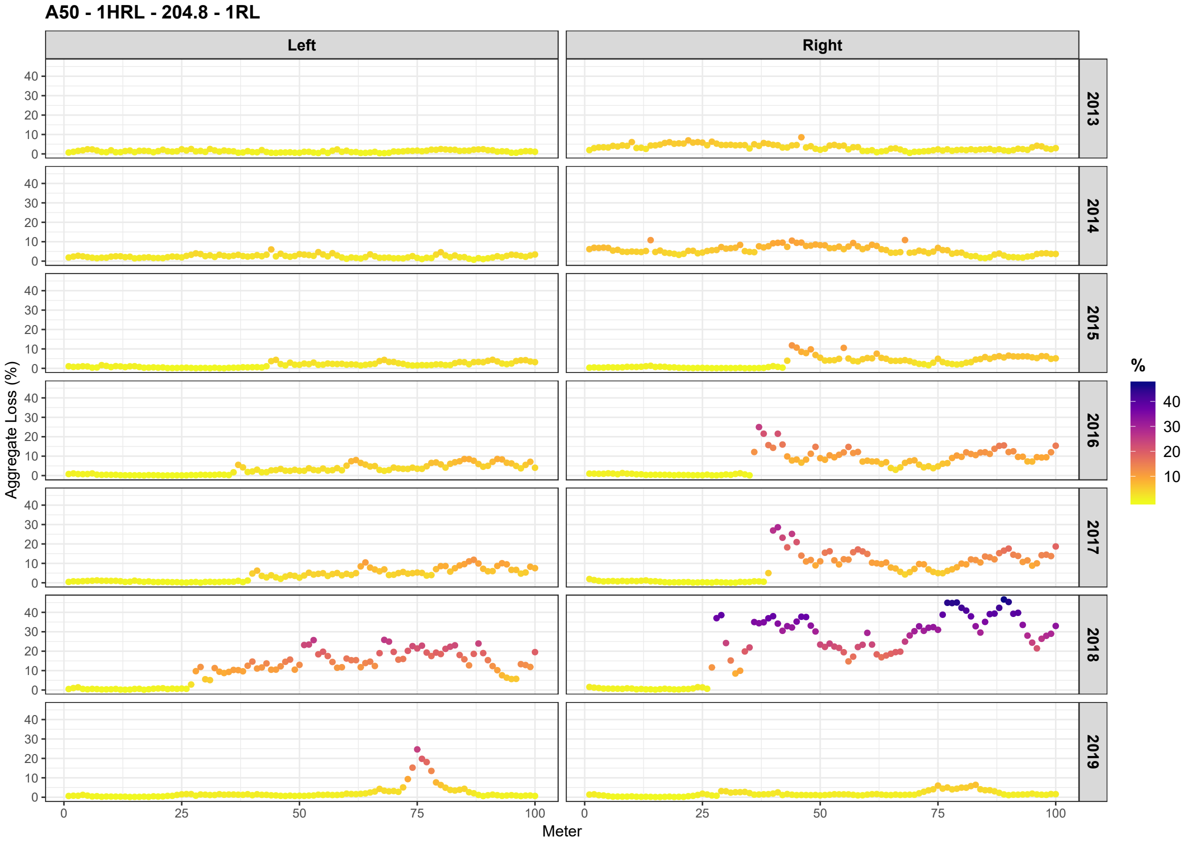

Figure 10.12: Aggregate loss values per year per wheel track plotted for $W = A50, S = 1\mathrm{HRL}^{1\mathrm{RL}}_{204.8}$.

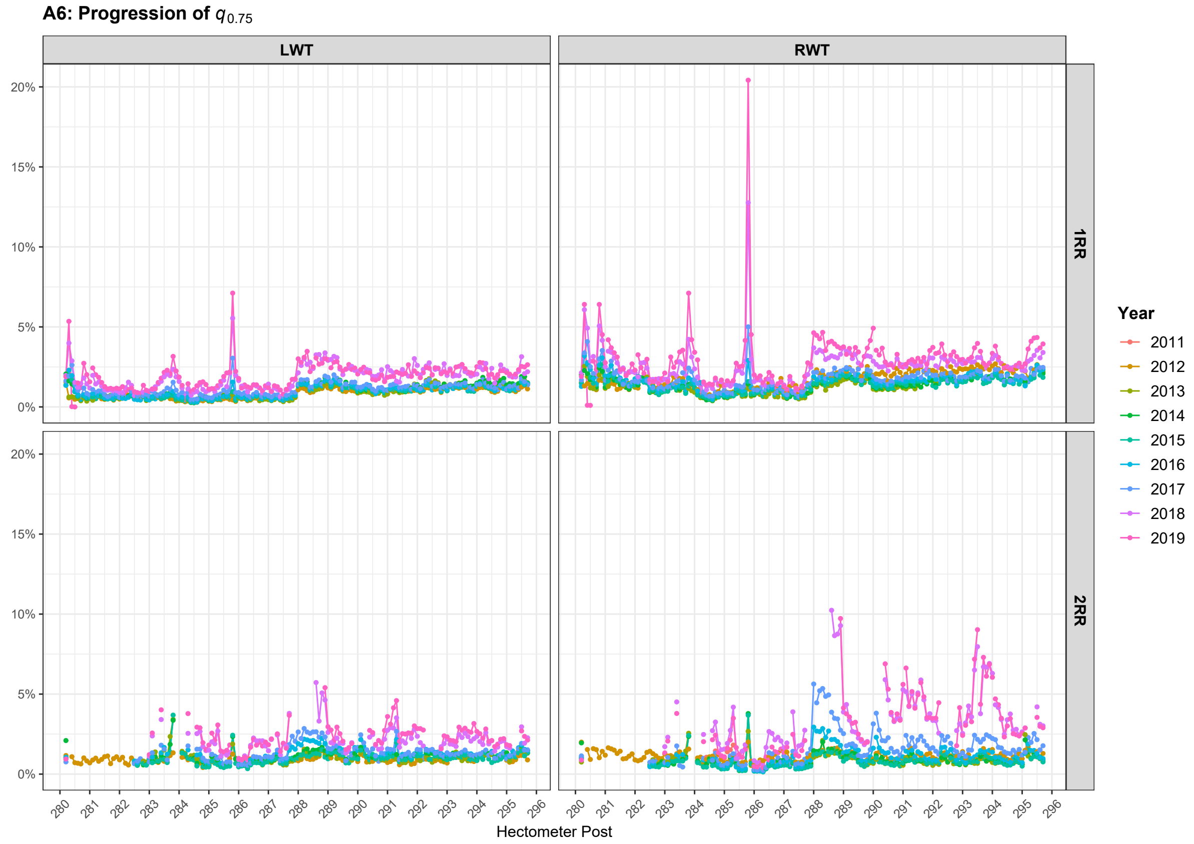

Figure 10.13: Sample empirical $q_{0.75}$ values per year per wheel track and per lane plotted for the road sections of $W = A6$.

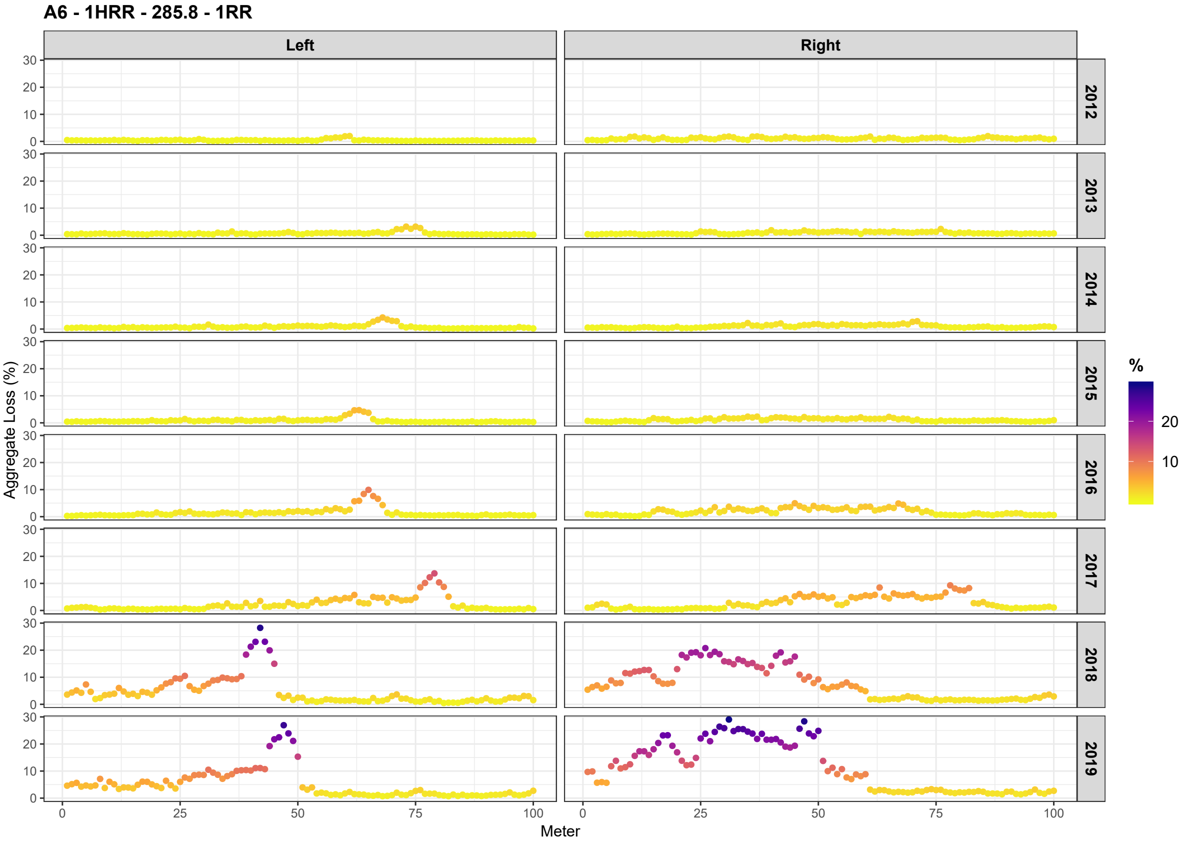

Figure 10.14: Aggregate loss values per year per wheel track plotted for $W = A6, S = 1\text{HRR}^{1\text{RR}}_{285.8}$.

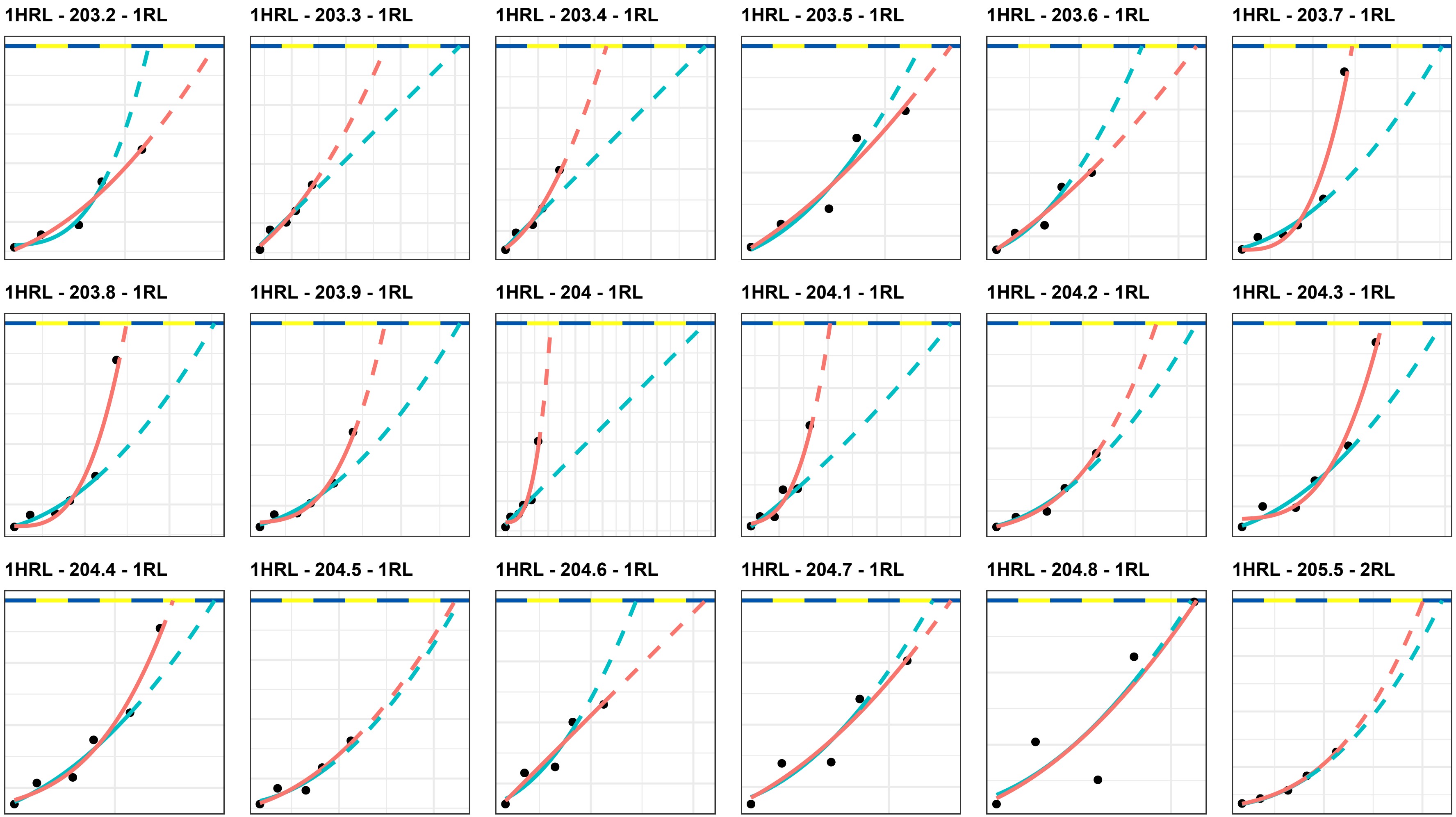

Figure 10.15: Polynomial extrapolations of $q_{0.75}$ estimates plotted for $W = A50$ and its road sections based on both LWT and RWT. The blue curves represent the fits with $n-1$ $q_{0.75}$ values, while the red curves use n $q_{0.75}$ values; all curves are fit using P-splines of order 4. The dashed curves are the extrapolations based on the solid curves. The dashed horizontal lines represent the RWS threshold.

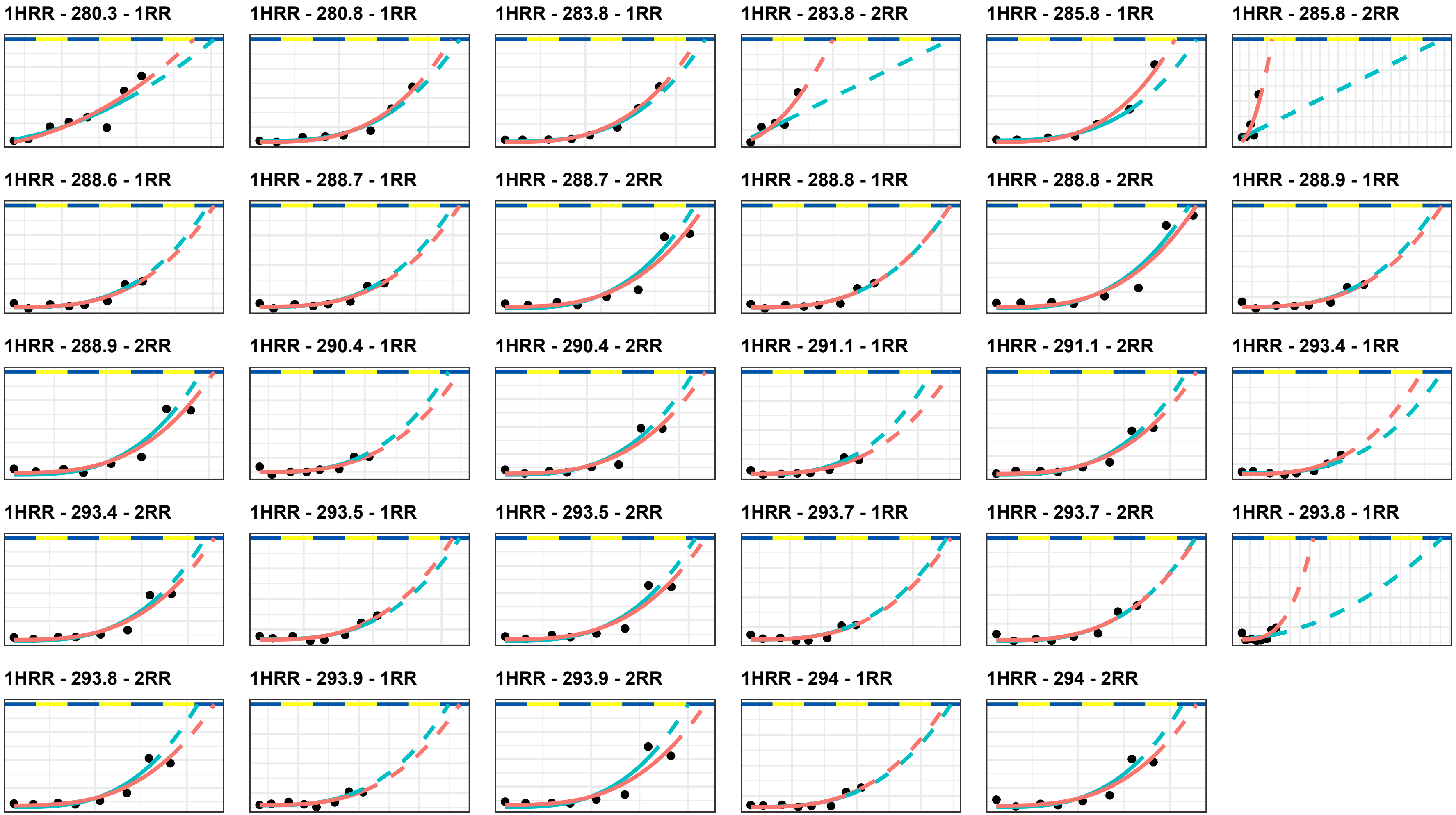

Figure 10.16: Polynomial extrapolations of $q_{0.75}$ estimates plotted for $W = A6$ and some of its road sections based on both LWT and RWT. The blue curves represent the fits with $n-1$ $q_{0.75}$ values, while the red curves use n $q_{0.75}$ values; all curves are fit using P-splines of order 4. The dashed curves are the extrapolations based on the solid curves. The dashed horizontal lines represent the RWS threshold.

10.3. Emerging Problems

After having expanded the concept to multiple roads and their sections, we have seen problems emerge in different orders of magnitude. The following interrelated problems are apparent from § 10.2, in increasing order of complexity/resolution:

1. With the current approach, predictions using data from both wheel tracks could be less consistent than when using either individually.
2. There are not many road sections satisfying our conditions mentioned earlier and in Chapter 6.
3. The consistency between the $n-1$ prediction and the n prediction is extremely dependent on the progressed n-th value of $q_{0.75}$.
4. The general pattern of convexity is lacking in such sense that a prediction of surpassing $q_{0.75} = 10$ is complicated.

We will attempt to provide possible methods to overcome these problems. For each problem a clarification can be given and we will proceed to do so.

For the first problem one should recognise that the current approach of using both wheel tracks depends on the consistency in pattern of consecutive $q_{0.75}$ estimates for the mixed data. Our interpretation as of now joins the LWT and RWT data and regards it as one. Another method we suggest is similar to the 'averaging' mentioned before, but then in terms of the threshold surpassing date. If the LWT and RWT separately truly are more consistent in consecutive $q_{0.75}$ estimates, modelling these should then be done separately which gives two distinct[6] threshold surpassing dates. If it is desired to base the maintenance on both wheel tracks, a naive but efficient estimated threshold surpassing date would lie exactly in between the two TSDs from the LWT and RWT. However, Geurt Jongbloed points out that the minimum of the two TSDs might be more natural to choose for.

The second problem is frankly entirely dependent on the current amount of data we have. As time progresses and more DOS-LCMS data is available, more road sections will be available for proper analysis. That alone is the reason why this problem could initially be considered the least complicated. However, we can understand if the reader thinks this not a fulfilling suggested solution due to its impracticality. For another more practical suggestion one really needs to wonder why there are not enough consecutive $q_{0.75}$ in the first place. We believe it can generally be agreed upon that for road sections which have undergone maintenance, there is not much we can try. The other cause of not enough consecutive $q_{0.75}$ values comes from the fact that we are trying to estimate the date on which $q_{0.75} = 10$, which renders estimates greater than 10 useless: one can opt to decide to perform maintenance in the succeeding year of observing the estimate being larger than 10. Combined with the results we have seen in § 10.2 it naturally implies that often the jump in $q_{0.75}$ was hardly possible to predict given the known sequence of $q_{0.75}$. This in turn could mean that $q_{0.75} = 10$ is too low of a threshold to predict under the current conditions[7] — assuming the lack of convexity in preceding estimates. Recall that this threshold was adhered to as these were the official ones, but an in-depth analysis of the norm using DOS-LCMS data has not been performed yet. This is a reason to consider analysing which norm with respect to the gathered DOS-LCMS data is actually being abode by.

As for the third problem — even though it is stating the obvious — deserves to be mentioned separately in potentially practical use for Rijkswaterstaat. If hypothetically some road section in 2020 is prognosed due for maintenance in 2030, whereas in 2021 the prognosis admits that maintenance is required in 2022 — or worse, 2021 — it is imaginably extremely complicated to plan

[6] Not in general, but it is highly unlikely for these two end up as the exact same extrapolated date.

[7] Official documents specifically condition on a 100m road section, of which we want to highlight the fact that they do not mention a road segment of arbitrary length in general.

around such cases for an executive agency such as Rijkswaterstaat. The solution to this, however, can only be realised if the threshold turns out to be higher than $q_{0.75} = 10$ providing more flexibility in potentially unfavourable behaviour in the DOS-LCMS data, or if the pattern in convexity is more apparent in $q_{0.75}$ estimates. The latter is clearly directly related to problem four which we will soon discuss, but could also be resolved by providing more data to estimate $q_{0.75}$. The DOS-LCMS data as of now is gathered on a yearly basis, but doing so biannually (twice a year) could significantly help in recognising noise in the data.

The fourth and final problem we have observed is in one sense the most complex and in another sense quite straightforward given our current knowledge. A first suggestion would be to always include a fixed point $q_{0.75} < 1$ on the date of the most recent pavement rehabilitation, in order to enhance the convexity in the pattern. This is justifiable, as the aggregate loss progresses relatively slow when the asphalt was just resurfaced as seen in Figure 10.8. The idea adds one $q_{0.75}$ value and could have resulted in more road sections accessible for analysis. The only potential problem with this suggestion is that it assumes proper administration, and this can imaginably be hard to register correctly for such a great amount of road sections. The second suggestion is a rerun of what we have done for the $q_{0.75}$ estimates, but on individual meter level. Indeed, it is possible to also fit smooth convex monotone curve fits on a 1m — or technically a $1/100^{\text{th}}$ length — level. Initially we planned to circumvent the noise in data by considering the complete 100 aggregate loss values per year and consider its progression in time. However, from the results of $W = A44$ in particular, it appears that even that sometimes does not find a way around the not monotonically increasing pattern. By doing so, we are forcing the behaviour we seek in regard of individual meter level, but should be wary of how this tends to average aggregate loss values, which in turn can downplay the severity of ravelling.

As the last problem is such an influential one, we will show what the results are of putting monotonic increasing and convexity constraints on a 1m level for $W \in \{A44, A50, A6\}$. Figure 10.17 shows the effect of this on $W = A44$ and the first thing to notice is that ΔPRL appears to be much smaller relative to the prior values in Table 10.6. If this were to be true in general, the third problem (large values for ΔPRL) would be partially solved. Do note that for $W = A44$, $S = 1\text{HRR}_{2.5}^{1\text{RR}}$, we see a failure of the polynomial extrapolation with all n $q_{0.75}$ estimates. This was not unexpected as we already warned for this potential problem in Chapter 9, but in cases like these we could opt for the linear extrapolation. Table A.3 in Appendix A provides the exact values of Figure 10.17 similar to Table 10.6. For the remaining roads, the visualisations are given in Figures 10.18 and 10.19 while the exact values are provided in Tables B.3 and C.3 in Appendices B and C. Figures 10.17 to 10.19 should be compared to Figures 10.9, 10.15 and 10.16 on Pages 76, 84 and 85.

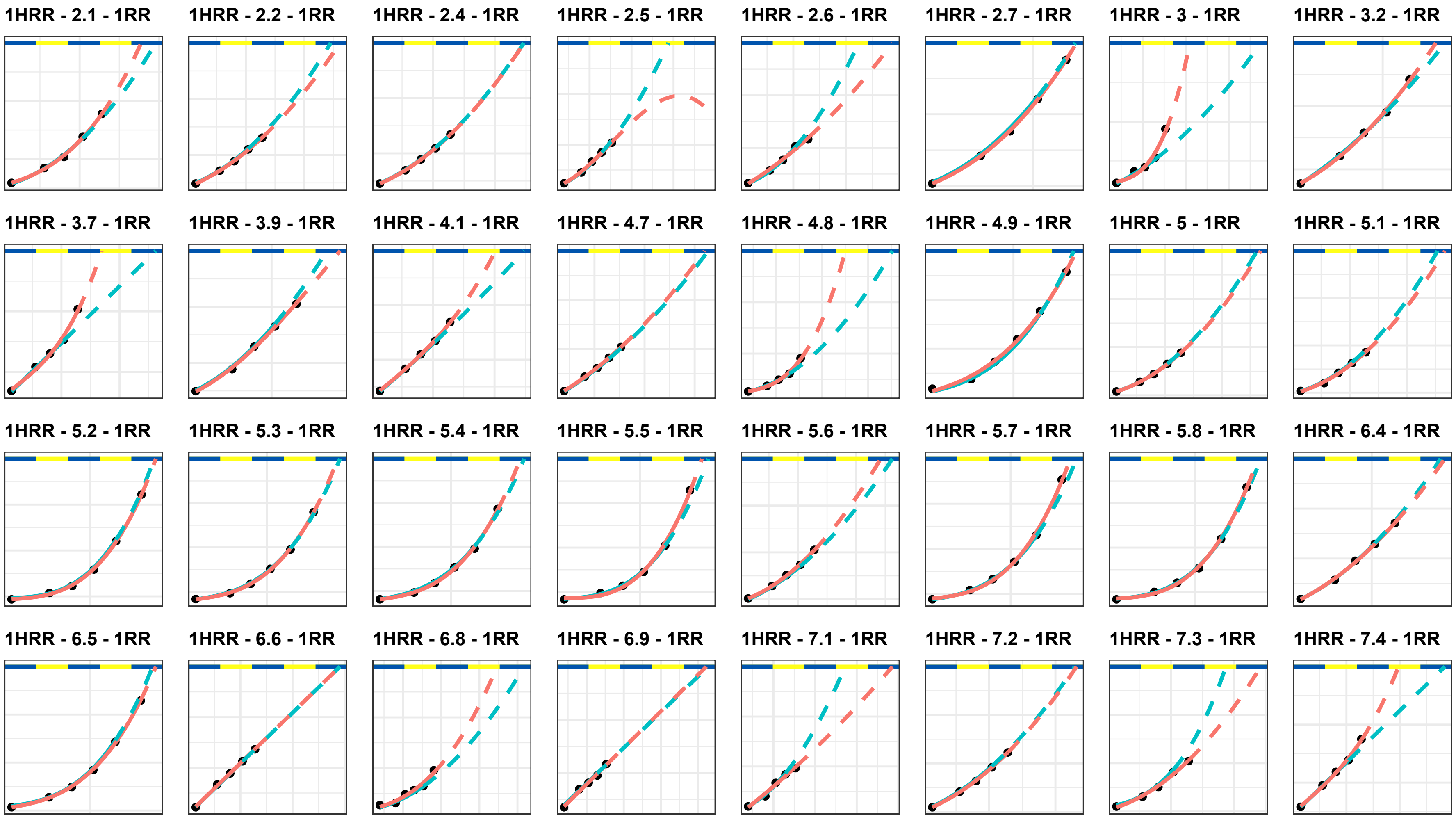

Figure 10.17: Polynomial extrapolations of $q_{0.75}$ estimates plotted for W = $A44$ and its road sections based on both LWT and RWT, where the data was transformed on 1 meter level under the constraints of monotonic increasing and convex. The blue curves represent the fits with $n-1$ $q_{0.75}$ values, while the red curves use n $q_{0.75}$ values; all curves are fit using P-splines of order 4. The dashed curves are the extrapolations based on the solid curves. The dashed horizontal lines represent the RWS threshold.

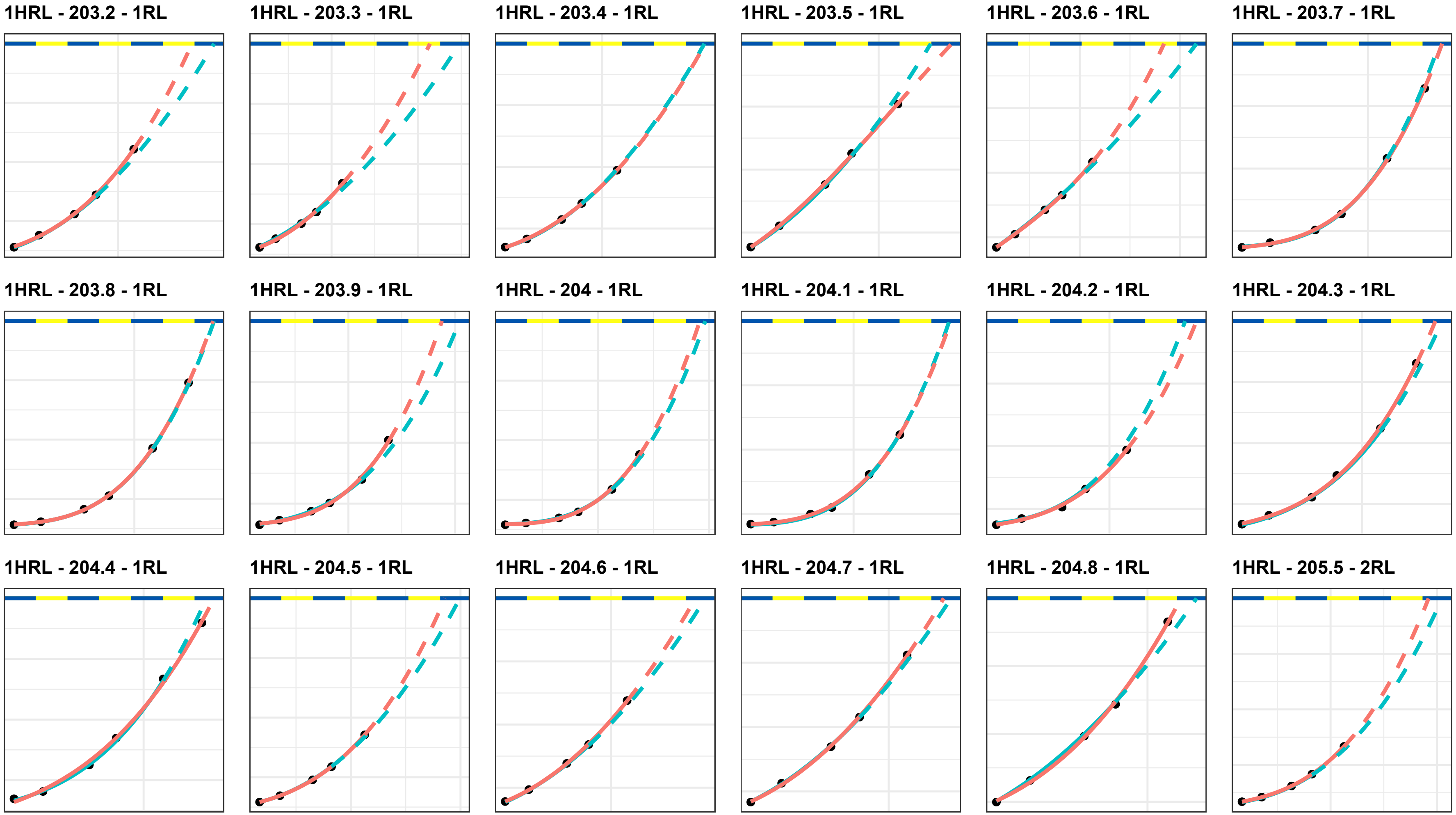

Figure 10.18: Polynomial extrapolations of $q_{0.75}$ estimates plotted for $W = A50$ and its road sections based on both LWT and RWT, where the data was transformed on 1 meter level under the constraints of monotonic increasing and convex. The blue curves represent the fits with $n-1$ $q_{0.75}$ values, while the red curves use n $q_{0.75}$ values; all curves are fit using P-splines of order 4. The dashed curves are the extrapolations based on the solid curves. The dashed horizontal lines represent the RWS threshold.

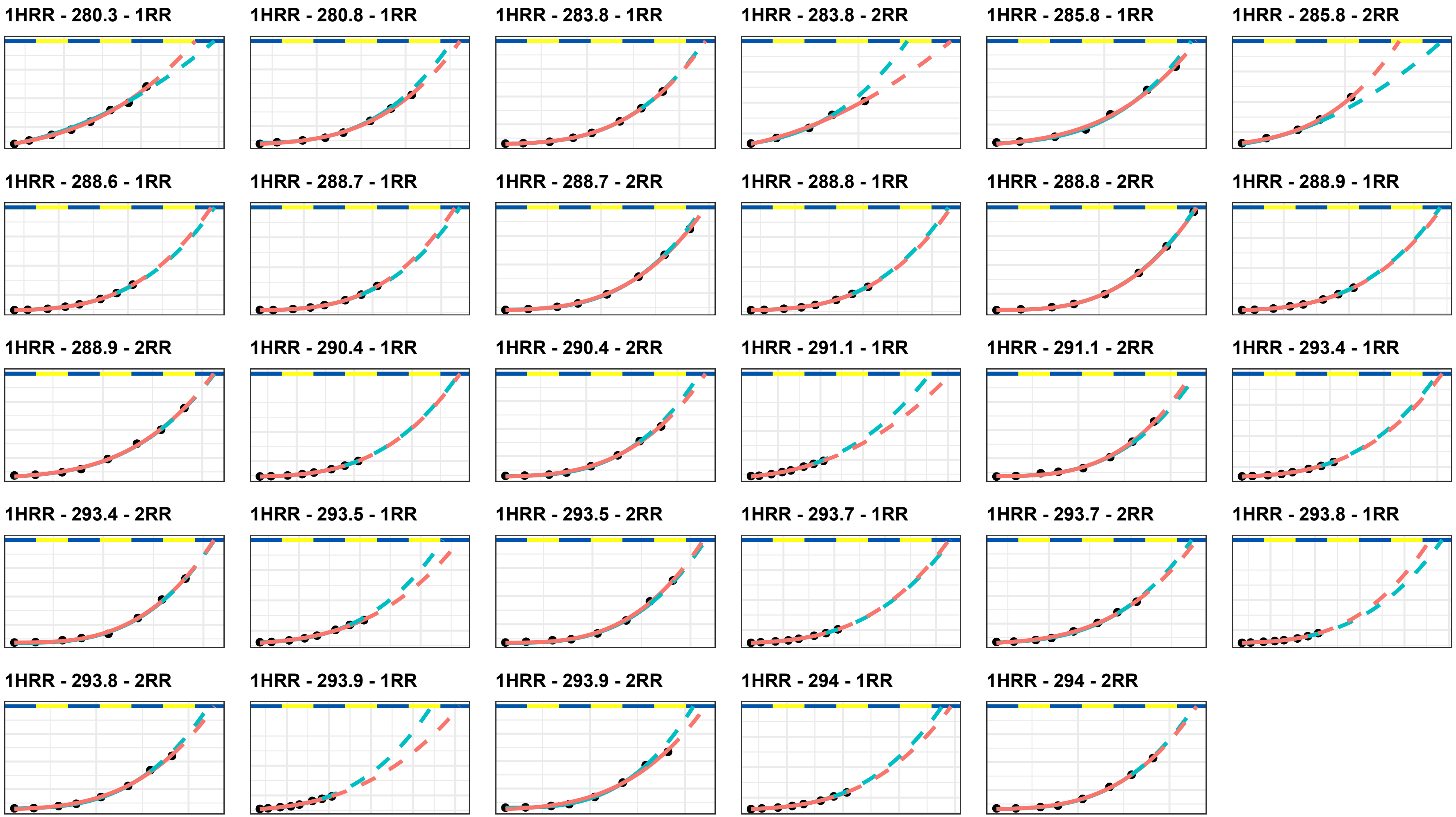

Figure 10.19: Polynomial extrapolations of $q_{0.75}$ estimates plotted for $W = A6$ and its road sections based on both LWT and RWT, where the data was transformed on 1 meter level under the constraints of monotonic increasing and convex. The blue curves represent the fits with $n-1$ $q_{0.75}$ values, while the red curves use n $q_{0.75}$ values; all curves are fit using P-splines of order 4. The dashed curves are the extrapolations based on the solid curves. The dashed horizontal lines represent the RWS threshold.

11 Conclusion and Recommendations

Conclusion

The main question which we wanted to answer was the following:

Main Question

Given DOS-LCMS data over multiple years from a road section, how can we predict its corresponding remaining lifetime?

The lifetime of a road section is currently defined as:

Threshold (Verra et al. [6] and DHV et al. [12])

If more than 25% of a road segment (approximately 100 m) measures an aggregate loss percentage of at least 10%, the corresponding road segment needs maintenance.

To answer the main question, we have constructed two fundamental steps introduced in Chapter 6:

1. Finding the 75th percentile based on the data.
2. Fit the progression of 75th percentiles to a monotonic increasing convex curve.

For each road section it is possible to handle the first step in a parametric or non-parametric method. The parametric approach is appealing in the sense that the quantiles can be computed explicitly, but definitely needs manual (human) inspection if one does not want to rely on distribution fitting algorithms in statistical software packages. For an immediate overview of an entire road it is simply not viable to find unique data transformations per road to fit proper parametric distributions. Additionally the little discrepancy between the parametric and non-parametric estimates of $q_{0.75}$ found for the prototype supports the assumption that finding such parametric distributions is perhaps too time consuming for large scale analysis. Therefore it is advisable to work with non-parametric estimates for their time-efficiency, showing barely any difference in estimates computed using manpower with respect to the prototype.

The second step can be realised using P-splines to construct a function which tries to capture the general pattern admitted from the 75th percentiles found in the first step. Although the constraints of monotonic increasing and convexity can be asked, whether the curve fit is simply linear will eventually be entirely dependent on the pattern of the estimates. Fitting an increasing curve is not difficult, as the most trivial way to fit such a line is by simple linear regression on the estimates. The result of the convexity condition, however, is the part which is extremely dependent on the data. We have seen that if the general pattern does not seem to be convex, the only fit that can be acquired is the straight — though increasing — line. While in theory a linear fit is convex, it does not coincide with the pattern we expect to see for severely ravelled road section in particular. If such a convex pattern is lacking, putting constraints on a meter level proved to be quite helpful. Yet we did not thoroughly analyse the differences in quantile approximations for the roads we applied this to.

For extrapolations we have seen that the predictions are consistent the closer we are to the Rijkswaterstaat threshold. We do have to reiterate that we do not know whether the threshold is correct in terms of being applied in practice, and additionally, our analysis has not cross-validated whether the road sections which had undergone maintenance were in fact due to ravelling. More complex even, is that if a road section was said to be resurfaced due to ravelling, we still cannot conclude that the road section has reached the threshold right before the date of maintenance: for all we know it could have been due to convenience of resurfacing a longer stretch of a road. The complexity lies in the fact that the current threshold was not based on DOS-LCMS data. If such a threshold is decided upon, even if it turns out that the current one is valid, the predictions can also be used as mere indications of overdue maintenance.

Recommendations

Our analysis has led to several options for improvements which could be made for future research which we will state and clarify. These are mainly based on the two general topics:

- optimal alignment,
- threshold recalibration.

Keep in mind that we have provided quite some suggestions in § 10.3 in terms of applying DOS-LCMS data in our framework, which we believe need not be restated here. Prior to the two general recommendations, we do like to point out that potential future research using this framework could be explored in various ways. Unfortunately due to time constraints, these could not be included in the presented work. One could focus on

- aggregate loss on a meter basis: why are some parts of the road section more prone to aggregate loss?
- analysis on differences between constraints and no constraints on a meter basis: although ΔPRL appeared to be favourable in terms of consistency for the former, at what cost was this acquired, that is, are the constraints justifiable?
- analysis of accuracy of the quantile estimates: how accurate are the current approximations of the 75th percentile?
- analysis of accuracy of proposed remaining lifetimes: in line with the previous item, how accurate are the current proposed remaining lifetimes?
- setting up more well-defined rules with the current provided predictions: how do the current predictions fare in comparison to their actual years of maintenance[1]?

Finally we would also like to dedicate a part for the interest of Rijkswaterstaat. That part will conclude the recommendations.

[1] This does remain slightly ambiguous, because performing maintenance on a road section is not guaranteed to be due to its short remaining lifetime.

Optimal Alignment

Before starting the main point, we suggest that the involved parties do not only register the hectometer posts properly to the first decimal, but also to the second decimal. Now it is still ambiguous whether maintenance from hectometer post 2.1 - 2.2 means that this happened exactly on this road section between the two respective driver location signs, or if in fact it ranges from 2.05 - 2.15. If the planned boundaries deviated from the actually used boundaries, it is extremely helpful if this is registered adequately.

Although the non-convex patterns are not necessarily caused by GPS inaccuracies, it still does not help with predictions using our proposed scheme. A particular distressing example can be found in Figure 11.1. In years 2012-2015 there does not seem to be too much inaccuracy in the assignment of the aggregate loss values on a meter basis, although one could argue that in years 2014 and 2015, the values seem to have slightly shifted towards the right compared to 2012. Regardless, that is not the point of Figure 11.1: year 2017 clearly shows high aggregate loss values in the first ±20 meters, while in 2016 it seems that asphalt was completely resurfaced. This is partially true and can be confirmed by Table A.2, where it shows that there was maintenance on lane 1RR of $W = A44$ from hectometer 2.9 - 3.0 supposedly on 2016-07-13. The DOS-LCMS data shows that this date might be faulty, because the DOS-LCMS measurements were supposedly done on 2016-02-27. Nevertheless, the initial high aggregate loss values in 2017 in Figure 11.1 should then coincide with the final meters of hectometer 2.8 - 2.9. Indeed, we cannot find a date of maintenance for this road section and Figure 11.2 shows high aggregate loss values throughout. In particular in year 2018 for Figure 11.2, the final meters yield low aggregate loss values opposed to the other measured values. Now the point is almost complete: assuming that there was a shift in position assignation caused by whichever reason, Figure 11.1 and especially year 2017 shows that the only logical deduction[2] is that for the consecutive road section ranging from hectometers 3.0 - 3.1 will admit low aggregate values in particular for year 2017, but perhaps also for other years. Indeed, as you might expect from this build-up, Figure 11.3 depicts this.

The point to consume is that Figures 11.1 to 11.3 show that the assignment of aggregate loss values per meter are not consistent across the years, which explains why on a meter basis — apart from some noise — the sequence of aggregate loss throughout the years is not increasing. This in turn will also influence quantiles on the entire road section, which skews the predictions. It should be redundant to point out that a $q_{0.75}$ approximation of year 2017 in Figure 11.3 could be one of the reasons why the threshold surpassing date for some road sections changed drastically with respect to the $n - i$ approach; in this case the $q_{0.75}$ prediction based on the data could be too low, exposing why the jump in consecutive years at times seemed rather unforeseen.

These inconsistencies should help with convincing that for proper use of DOS-LCMS data in whichever proposed method — such as the one presented in this thesis but by far not limited to this method — the position assignation should be much more consistent throughout the years. The only reason for seeing ravelling not increasing should be due to standard noise in data measurements or from resurfaced asphalt if the consecutive year shows very minimal ravelling. Therefore we suggest that if the involved parties are going to rely more and more on only the DOS-LCMS data in whichever manner and less on manpower in the form of visual inspections, an excellent investment to start with would be in GPS accuracy. However, if such an investment is difficult to implement directly — which we can imagine as many parties could be involved in such a process — the data could be aligned prior to usage. At worst case, one could plot visuals such as Figures 11.1 to 11.3 and propose to shift the data manually. The best case scenario without fundamentally changing steps in the current DOS-LCMS measurement scheme would be to compute an algorithm or apply some set of rules which align the data: for example, as we have seen in this brief example, an immediate look-up

[2] A small remark to also acknowledge is that maintenance, though registered to be for only one road section, could be extended slightly further into the preceding and succeeding road section.

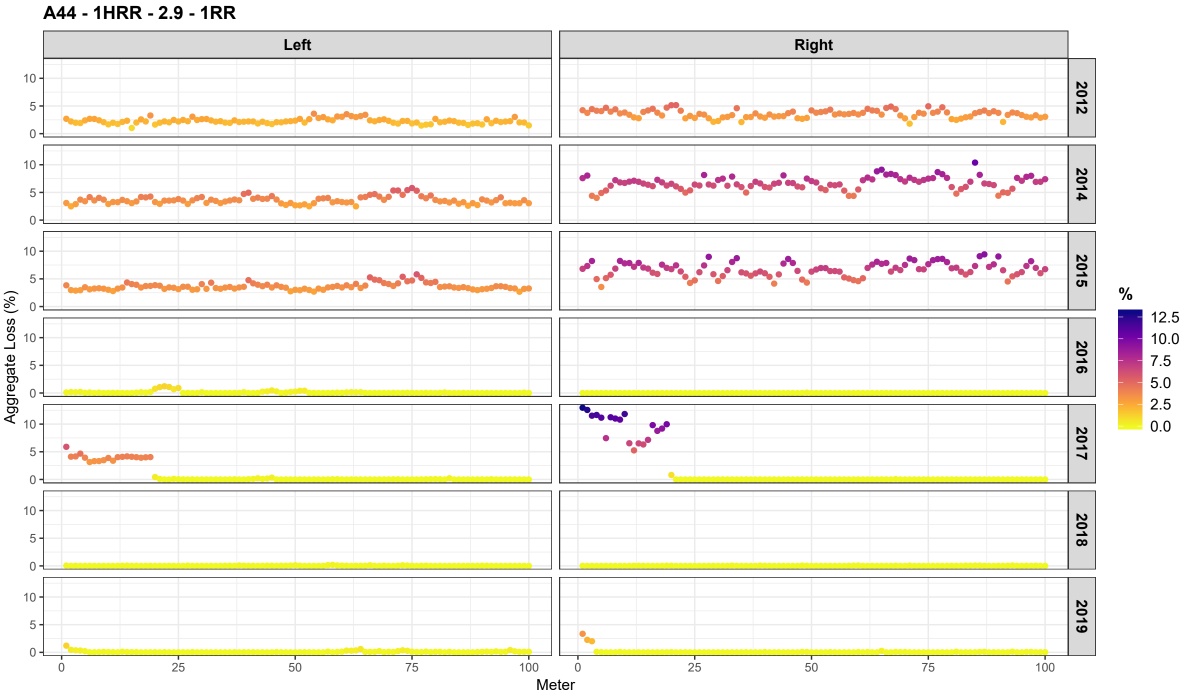

Figure 11.1: Aggregate loss values per year per wheel track plotted for $W = A44, S = 1\text{HRR}_{2.9}^{1\text{RR}}$.

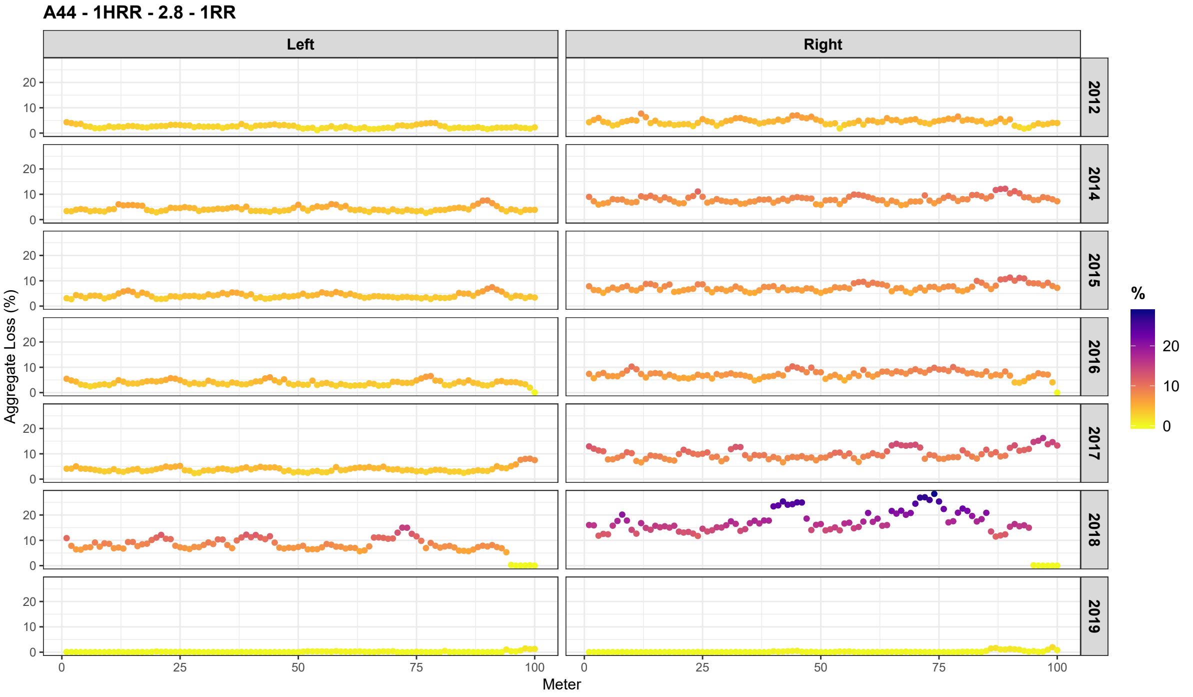

Figure 11.2: Aggregate loss values per year per wheel track plotted for $W = A44, S = 1\text{HRR}_{2.8}^{1\text{RR}}$.

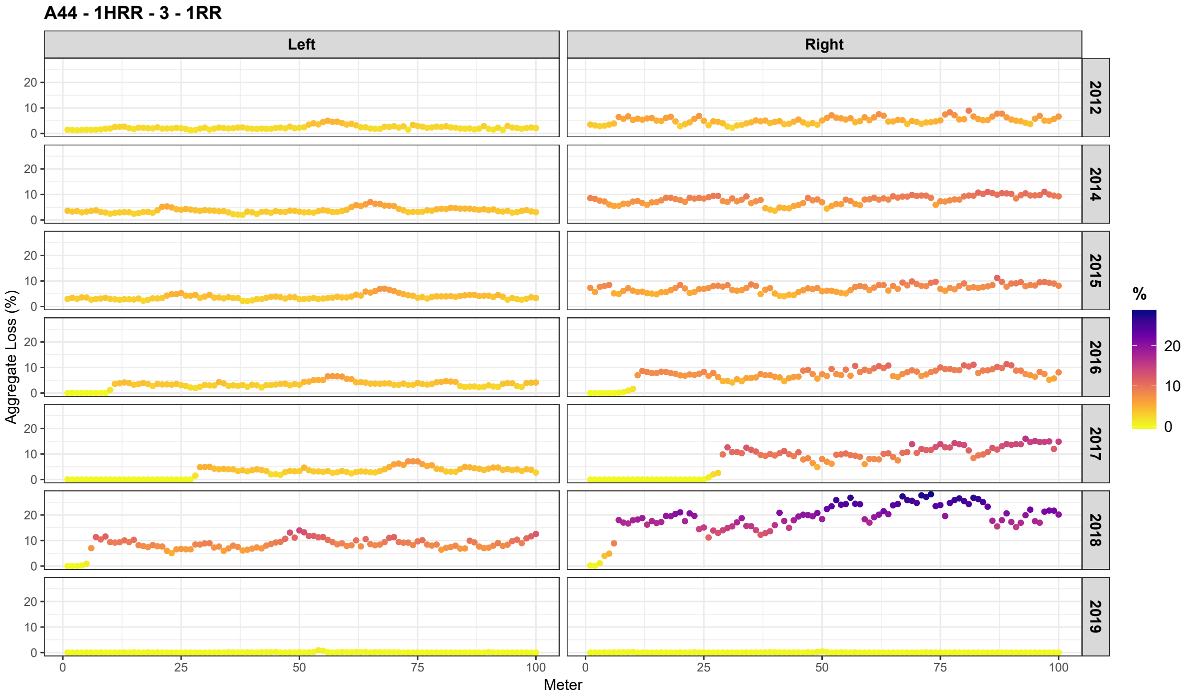

Figure 11.3: Aggregate loss values per year per wheel track plotted for $W = A44, S = 1\text{HRR}^{1\text{RR}}_{3.0}$.

for road maintenance can indicate which values should belong where. Alternative ambitious and innovative ideas could stem from some metric which indicates monotonicity in the desired manner, and perform cross-validation using this measure. A quick but probably naive measure would be to compute the differences in consecutive aggregate loss values along the years for each meter; we can imagine many much more polished and effective measures would exist if this concept is considered more carefully.

Threshold Recalibration

Not unrelated to the previous topic is analysis on determining a new threshold entirely based on DOS-LCMS data. The current threshold indicates when to resurface asphalt per 100 meter road sections. Regardless of being nit-picky on this specific length of exactly 100 meters, one could first of all consider looking at longer stretches of the road. From Tables A.2, B.2 and C.2 in Appendices A to C, maintenance is often performed on a span of more than 100 meters rather than exactly on one road section. Reasons for this may vary, but if in practice the policy is based on a factor such as convenience, it makes sense to also reconsider the threshold to account for more than 100 meters. In turn this will help with more robust estimates of statistics such as the 75th percentile. The downside to this is that if maintenance was actually performed for only one road section while the preceding and succeeding road sections were untouched, it would skew the predictions too.

The other part of the threshold is characterised by the percentile and its corresponding value to consider. We iterate yet again that the threshold was defined using specific values for aggregate loss, even though no data was available to associate these with. Now that the DOS-LCMS scheme is available, it makes sense to validate in how many cases the threshold was actually adhered to. If there is interest in recalibrating or even redefining the threshold, a starting point would be to first validate how often the threshold was lived up to. If this is less than anything tolerable from Rijkswaterstaat standards, say below 80%, then it suggest that the norm should be seriously reconsidered. One idea to deduce new thresholds is by inspecting the data and in particular the stretches of the road which have been resurfaced: one should ask themselves why these parts in specific were resurfaced with respect to ravelling. Parts of which the maintenance was not due to ravelling should logically be excluded. Specifically we can look at the DOS-LCMS measurements 1 year before maintenance was performed. By doing so, we can find patterns in maintenance policy which could in turn be dependent on factors such as the respective district, but more importantly: a new percentile and a new cut-off as opposed to the 75^{th} percentile and 10% as of now can be deduced and be considered the new norm.

Practical Proposals for Rijkswaterstaat

From the current framework in which the data has been explored and the important extrapolation step has been proposed, the research has unfortunately not led to a model in operational use for Rijkswaterstaat[3]. We acknowledge that such a model is extremely interesting for Rijkswaterstaat, but due to some constraints — of which one is our available time — this has not been realised yet. We will attempt to provide a setup for a model which could potentially be used by Rijkswaterstaat for predictions. However, this setup naturally is completely dependent on the DOS-LCMS data for which some serious (quality) demands are required and which we will discuss first.

Demands for DOS-LCMS Data

The framework is entirely data-dependent, and hence it is sensible to be critical about the DOS-LCMS measurements we had available. Here we will reiterate what is still lacking and what exactly may be expected from the data.

A starting requirement for the data is for it to be consistent and stable. Specifically, if the data indicates that over two (consecutive) years there is an increase of aggregate loss of some x%, it should correspond with actual aggregate loss and not to the result of other factors such as:

- a change in algorithms which compute the aggregate loss percentages,
- positional inaccuracies previously discussed with respect to optimal alignment,
- a change in measuring or processing methods.

The first factor essentially means that for the same raw data, a change in algorithms could lead to significantly different aggregate loss percentages. The second factor has been largely discussed just before, but we would like to add that it is also possible to set some ‘business rules’ which allow the disqualification of certain data points. This implies that the used measure for the threshold does not always have to be based on 100 values of aggregate loss percentages, but in general probably less. The third factor is related to the improvements in the measuring methods; in 2018, the resolution of the sensor was increased from 5 mm to 2 mm, which allows for detection of ravelling on finer textures of pavement than only PA. In particular *two layer porous asphalt* (TLPA) is one of the surface types which have finer textures, and has become more and more desired due to their increased noise reduction over regular PA. However, it also implies that we are not certain on the comparability of the aggregate loss of both devices: are they significantly different? We acknowledge that innovation and development is a great initiative, but for a model to work properly and be representative, consistency in data is key.

Severity of Ravelling

It is shown in Chapter 10 that there is little to no use in the predictions for which the ravelling intensity is relatively low, say around 0-3%, because of its sensitivity to deviations. From Chapter 10 a clear threshold for reasonable predictions is not evident — and even dependent on how much we can rely on the computed aggregate loss values — but it appears to become reasonable from 4% onward.

[3] Although this was not the intention of the research, it was an overzealous personal aim initially.

Model Framework

The input for the model would be DOS-LCMS measurements over multiple years satisfying the demands, including potential business rules. The output can greatly vary, but it fundamentally provides a proposed remaining lifetime extrapolated from fitted curves. Using that, other interesting approaches are rendered possible such as

- deducing when the remaining lifetime is 5 years, such that rejuvenating products can be applied appropriately;
- providing overviews of parts of the road which are due for maintenance the earliest.

The actual output really depends on where the interests of Rijkswaterstaat lie.

Appendices

A

A44

Table A.1: Construction dates of $W = A44$ on carriageway 1HRR

HmStart	HmStop	Lane	Date
2.1	7.7	ALL	2002-09-09

Table A.2: Maintenance of $W = A44$ on carriageway 1HRR.

HmStart	HmStop	Lane	Date	Surface
6.0	6.1	ALL	2009-06-01	PA
2.1	2.9	2RR	2015-10-01	PEA
2.9	5.5	2RR	2016-07-13	PEA
5.5	6.7	2RR	2015-10-01	PEA
6.7	7.6	2RR	2016-07-13	PEA
2.9	3.0	1RR	2016-07-13	PEA
3.3	3.4	1RR	2016-07-13	PEA
4.4	4.7	1RR	2016-07-13	PEA
6.9	7.6	2RR	2016-07-13	PEA
6.0	6.2	2RR	2018-04-09	PA
2.1	2.3	ALL	2018-09-25	PEA
2.3	6.9	1RR	2018-09-25	PEA

HmStart	HmStop	Lane	Date	Surface
2.3	5.9	2RR	2018-09-25	PEA
5.9	6.2	2RR	2018-01-10	PA
6.2	6.9	2RR	2018-09-25	PEA

Table A.3: Lifetime predictions for $W = A44$ and its road sections using kernel approximations and polynomial extrapolations, where the data is constraint to be increasing and convex on a 1m level. n denotes the number of consecutive $q_{0.75}$ values used for the calculation, PRL = Proposed Remaining Lifetime (days relative to the date of the last percentile used), ΔPRL = difference in PRL between prediction n and $n-1$ (negative implies that prediction $n-1$ was $|\Delta\text{PRL}|$ further back in time.

C.way	Hm	Lane	n	$q_{0.75}$	Age	Date	TSD	PRL	ΔPRL
1HRR	2.1	1RR	5	6.95	14.44	2017-02-14	2019-02-08	724	-270
1HRR	2.2	1RR	5	5.76	14.44	2017-02-14	2022-06-15	1947	233
1HRR	2.4	1RR	5	5.86	14.44	2017-02-14	2021-11-16	1736	6
1HRR	2.5	1RR	5	5.09	14.44	2017-02-14	2026-04-24	3356	1389
1HRR	2.6	1RR	5	5.42	14.44	2017-02-14	2023-07-10	2337	934
1HRR	2.7	1RR	5	9.40	14.44	2017-02-14	2017-06-24	130	16
1HRR	3.0	1RR	5	7.27	14.44	2017-02-14	2019-05-25	830	-2407
1HRR	3.2	1RR	5	8.55	14.44	2017-02-14	2018-03-22	401	-139
1HRR	3.7	1RR	5	7.59	14.44	2017-02-14	2018-10-07	600	-1363
1HRR	3.9	1RR	5	7.63	14.44	2017-02-14	2019-02-04	720	219
1HRR	4.1	1RR	5	6.73	14.44	2017-02-14	2020-01-24	1074	-677
1HRR	4.7	1RR	5	5.09	14.44	2017-02-14	2023-10-22	2441	-104
1HRR	4.8	1RR	5	4.77	14.44	2017-02-14	2021-01-05	1421	-1493
1HRR	4.9	1RR	6	8.93	15.56	2018-03-27	2018-09-02	159	43
1HRR	5.0	1RR	5	4.71	14.44	2017-02-14	2022-10-02	2056	95
1HRR	5.1	1RR	5	4.63	14.44	2017-02-14	2022-08-28	2021	187
1HRR	5.2	1RR	6	8.05	15.56	2018-03-27	2018-11-09	227	27
1HRR	5.3	1RR	6	6.99	15.56	2018-03-27	2019-05-28	427	-37
1HRR	5.4	1RR	6	7.15	15.56	2018-03-27	2019-04-30	399	-66
1HRR	5.5	1RR	6	8.22	15.56	2018-03-27	2018-10-10	197	-106

C.way	**Hm**	**Lane**	n	$q_{0.75}$	**Age**	**Date**	**TSD**	**PRL**	**ΔPRL**
1HRR	5.6	1RR	5	5.14	14.44	2017-02-14	2021-09-10	1669	-296
1HRR	5.7	1RR	6	8.84	15.56	2018-03-27	2018-08-13	139	-99
1HRR	5.8	1RR	6	8.39	15.56	2018-03-27	2018-09-24	181	-43
1HRR	6.4	1RR	5	6.77	14.44	2017-02-14	2019-07-16	882	77
1HRR	6.5	1RR	6	8.23	15.56	2018-03-27	2018-11-26	244	52
1HRR	6.6	1RR	5	5.91	14.44	2017-02-14	2023-08-20	2378	-4
1HRR	6.8	1RR	6	4.90	15.56	2018-03-27	2024-11-05	2415	-1042
1HRR	6.9	1RR	5	5.42	14.44	2017-02-14	2027-11-24	3935	-77
1HRR	7.1	1RR	5	4.92	14.44	2017-02-14	2026-06-08	3401	1663
1HRR	7.2	1RR	5	5.57	14.44	2017-02-14	2021-04-11	1517	40
1HRR	7.3	1RR	5	5.11	14.44	2017-02-14	2021-09-04	1663	802
1HRR	7.4	1RR	5	6.87	14.44	2017-02-14	2019-11-26	1015	-1268

B

A50

Table B.1: Construction dates of $W = A50$ on carriageway 1HRL

HmStart	HmStop	Lane	Date
139.9	140.3	ALL	2010-03-18
140.3	141.6	ALL	2012-11-08
141.6	142.2	ALL	2013-04-16
142.2	142.9	ALL	2012-03-30
142.9	146.7	ALL	2013-03-19
146.7	148.4	ALL	2012-09-17
139.9	140.3	ALL	2006-06-04
140.3	146.7	ALL	2013-05-07
146.7	148.5	ALL	2012-09-17
205.4	205.5	1RL	2007-11-01
205.4	205.5	2RL	2002-07-12
204.8	205.4	ALL	2002-01-01
203.2	204.8	ALL	2002-12-31
203.1	203.2	ALL	2003-01-01
202.9	203.1	1RL	2003-01-01
202.9	203.1	2RL	2002-07-12

Table B.2: Maintenance of $W = A50$ on carriageway 1HRL.

HmStart	HmStop	Lane	Date	Surface
205.5	205.4	2RL	2017-11-01	PA
205.4	204.8	ALL	2014-08-14	PA
203.6	203.1	1RL	2018-02-02	PA
204.8	203.6	1RL	2018-11-29	PA
204.8	203.1	2RL	2016-11-30	PA

Table B.3: Lifetime predictions for $W = A50$ and its road sections using kernel approximations and polynomial extrapolations, where the data is constraint to be increasing and convex on a 1m level. n denotes the number of consecutive $q_{0.75}$ values used for the calculation, PRL = Proposed Remaining Lifetime (days relative to the date of the last percentile used), ΔPRL = difference in PRL between prediction n and $n-1$ (negative implies that prediction $n-1$ was |ΔPRL| further back in time.

C.way	Hm	Lane	n	$q_{0.75}$	Age	Date	TSD	PRL	ΔPRL
1HRL	203.2	1RL	5	5.53	14.52	2017-07-04	2019-05-16	681	-259
1HRL	203.3	1RL	5	4.20	14.52	2017-07-04	2021-07-22	1479	-502
1HRL	203.4	1RL	5	4.86	14.52	2017-07-04	2020-07-14	1106	11
1HRL	203.5	1RL	5	7.60	14.52	2017-07-04	2018-11-19	503	191
1HRL	203.6	1RL	5	5.42	14.52	2017-07-04	2020-05-12	1043	-472
1HRL	203.7	1RL	6	8.21	15.51	2018-07-02	2018-12-17	168	27
1HRL	203.8	1RL	6	7.40	15.51	2018-07-02	2019-03-06	247	-11
1HRL	203.9	1RL	6	5.11	15.51	2018-07-02	2020-07-04	733	-243
1HRL	204.0	1RL	6	4.40	15.51	2018-07-02	2020-08-31	791	-72
1HRL	204.1	1RL	6	5.58	15.51	2018-07-02	2020-02-26	604	19
1HRL	204.2	1RL	5	4.85	14.52	2017-07-04	2019-07-26	752	122
1HRL	204.3	1RL	6	8.27	15.51	2018-07-02	2019-01-11	193	-68
1HRL	204.4	1RL	6	8.99	15.51	2018-07-02	2018-10-24	114	72
1HRL	204.5	1RL	5	4.28	14.52	2017-07-04	2020-06-09	1071	-195
1HRL	204.6	1RL	5	5.95	14.52	2017-07-04	2019-08-28	785	-113
1HRL	204.7	1RL	5	7.80	14.46	2017-06-12	2018-05-01	323	-63

C.way	**Hm**	**Lane**	n	$q_{0.75}$	**Age**	**Date**	**TSD**	**PRL**	**ΔPRL**
1HRL	204.8	1RL	5	9.14	14.46	2017-06-12	2017-09-27	107	-125
1HRL	205.5	1RL	7	1.20	11.82	2019-08-25	2032-10-10	4795	883
1HRL	203.0	2RL	7	3.40	17.03	2019-07-18	2068-08-03	17914	7222
1HRL	203.1	2RL	7	2.82	17.03	2019-07-18	2033-06-26	5092	-12937
1HRL	205.5	2RL	5	3.33	14.99	2017-07-04	2020-09-07	1161	-186

C

A6

Table C.1: Construction dates of $W = A6$ on carriageway 1HRR

HmStart	HmStop	Lane	Date
288	280.2	ALL	2005-11-15
295.8	288	ALL	2005-08-31

Table C.2: Maintenance of $W = A6$ on carriageway 1HRR.

HmStart	HmStop	Lane	Date	Surface
280.4	280.3	2RR	2012-12-31	PA
282.5	280.4	2RR	2013-06-13	PA
284.1	283.9	2RR	2012-12-31	PA
280.5	280.3	2RR	2016-10-21	PA
280.8	280.6	2RR	2016-10-01	PA
282.1	282	2RR	2018-08-20	PA
282.4	282.3	2RR	2018-06-26	PA
282.8	282.5	2RR	2018-10-20	PA
283	282.8	2RR	2016-10-01	PA
283.4	283.2	2RR	2016-10-01	PA
283.7	283.5	2RR	2018-08-20	PA

HmStart	HmStop	Lane	Date	Surface
284	283.7	2RR	2016-10-01	PA
284.3	284.1	2RR	2016-07-28	PA
284.6	284.4	2RR	2016-10-01	PA
286	285.8	2RR	2016-10-01	PA
287.3	287.2	2RR	2018-08-20	PA
288.6	287.8	2RR	2018-09-05	PADI
288.9	288.6	2RR	2019-05-22	PA
290.4	289.9	2RR	2018-09-05	PA
292.8	292.3	2RR	2018-09-04	PA
295.5	295.2	2RR	2018-09-04	PA

Table C.3: Lifetime predictions for $W = A6$ and a sample of its road sections using kernel approximations and polynomial extrapolations, where the data is constraint to be increasing and convex on a 1m level. n denotes the number of consecutive $q_{0.75}$ values used for the calculation, PRL = Proposed Remaining Lifetime (days relative to the date of the last percentile used), ΔPRL = difference in PRL between prediction n and $n-1$ (negative implies that prediction $n-1$ was |ΔPRL| further back in time.

C.way	Hm	Lane	n	$q_{0.75}$	Age	Date	TSD	PRL	ΔPRL
1HRR	280.3	1RR	8	6.03	13.41	2019-04-12	2021-10-14	916	-364
1HRR	280.8	1RR	8	5.25	13.41	2019-04-12	2021-06-13	793	130
1HRR	283.8	1RR	8	5.45	13.41	2019-04-12	2021-02-18	678	20
1HRR	283.8	2RR	5	5.13	10.34	2016-03-16	2019-01-21	1041	527
1HRR	285.8	1RR	7	7.70	12.48	2018-05-07	2019-01-14	252	63
1HRR	285.8	2RR	5	5.38	10.34	2016-03-16	2017-11-13	607	-538
1HRR	288.6	1RR	8	3.38	13.62	2019-04-12	2023-10-17	1649	-70
1HRR	288.7	1RR	8	3.45	13.62	2019-04-12	2023-10-21	1653	-103
1HRR	288.7	2RR	8	8.11	13.63	2019-04-14	2019-11-11	211	35
1HRR	288.8	1RR	8	3.32	13.62	2019-04-12	2024-02-10	1765	59
1HRR	288.8	2RR	8	9.60	13.63	2019-04-14	2019-05-20	36	10
1HRR	288.9	1RR	8	3.39	13.62	2019-04-12	2024-09-13	1981	41
1HRR	288.9	2RR	8	6.93	13.63	2019-04-14	2020-06-17	430	-12

C.way	**Hm**	**Lane**	n	$q_{0.75}$	**Age**	**Date**	**TSD**	**PRL**	**ΔPRL**
1HRR	290.4	1RR	8	2.48	13.62	2019-04-12	2026-05-07	2582	22
1HRR	290.4	2RR	8	5.21	13.63	2019-04-14	2021-03-27	713	91
1HRR	291.1	1RR	8	2.41	13.62	2019-04-12	2031-04-22	4393	724
1HRR	291.1	2RR	8	5.76	13.63	2019-04-14	2020-12-01	597	-76
1HRR	293.4	1RR	8	2.90	13.62	2019-04-12	2027-05-09	2949	110
1HRR	293.4	2RR	8	6.59	13.63	2019-04-14	2020-06-03	416	-7
1HRR	293.5	1RR	8	3.36	13.62	2019-04-12	2025-07-31	2302	404
1HRR	293.5	2RR	8	6.39	13.63	2019-04-14	2020-07-09	452	-34
1HRR	293.7	1RR	8	2.56	13.62	2019-04-12	2028-01-03	3188	-60
1HRR	293.7	2RR	8	4.47	13.63	2019-04-14	2022-03-17	1068	87
1HRR	293.8	1RR	8	2.25	13.62	2019-04-12	2029-04-16	3657	-396
1HRR	293.8	2RR	8	5.52	13.63	2019-04-14	2021-02-14	672	94
1HRR	293.9	1RR	8	2.42	13.62	2019-04-12	2031-06-12	4444	987
1HRR	293.9	2RR	8	5.85	13.63	2019-04-14	2020-11-10	576	187
1HRR	294.0	1RR	8	2.93	13.62	2019-04-12	2026-09-14	2712	232
1HRR	294.0	2RR	8	5.37	13.63	2019-04-14	2021-03-08	694	32

Bibliography

[1] *Meerjarenplanning verhardingsonderhoud 2019-2025* (Ministerie van Infrastructuur en Waterstaat, Rijkswaterstaat, Grote Projecten en Onderhoud, 2020).

[2] J.-F. Hébert, *Pavemetrics | Laser Crack Measurement System (LCMS),* (2020), library Catalog: www.pavemetrics.com.

[3] W. Van Aalst, G. Derksen, P. Schackmann, F. Bouman, and P. Paffen, *Automated Raveling Inspection and Maintenance Planning on Porous Asphalt in the Netherlands,* International Symposium Non-Destructive Testing in Civil Engineering (NDTCE 2015) , 3 (2015).

[4] W. Van Aalst, G. Derksen, P. Schackmann, E. Zwier, F. Bouman, and P. Paffen, *Automated road survey and pavement management on porous asphalt,* in *Life-Cycle of Engineering Systems*, edited by J. Bakker, D. M. Frangopol, and K. v. Breugel (CRC Press, 2016) 1st ed., pp. 2082–2085.

[5] G. Leegwater, B. Luiten, R. Krans, M. Koole, and E. Van Osch, *LAM: Levensduurvoorspellings-AsfaltModel vanuit data en fysische modellen,* (2020).

[6] N. Verra, M. Van den Bol, and B. Gaarkeuken, *De levensduur van ZOAB: gemiddelde levensduurbepaling op basis van MJPO-2003,* (2003).

[7] G. Derksen, *Levensduur ZOAB,* (2014).

[8] B. Ebrahimi, H. Wallbaum, K. Svensson, and D. Gryteselv, *Estimation of Norwegian Asphalt Surfacing Lifetimes Using Survival Analysis Coupled with Road Spatial Data,* Journal of Transportation Engineering, Part B: Pavements **145** (2019), 10.1061/JPEODX.0000115.

[9] E. L. Kaplan and P. Meier, *Nonparametric Estimation from Incomplete Observations,* Journal of the American Statistical Association **53**, 457 (1958).

[10] L. Wasserman, *All of Nonparametric Statistics*, Springer Texts in Statistics (Springer-Verlag, New York, 2006).

[11] D. R. Cox, *Regression Models and Life-Tables,* Journal of the Royal Statistical Society: Series B (Methodological) **34**, 187 (1972).

[12] DHV, Unihorn, KOAC-NPC, D. V. en Scheepvaart, and P. Antes, *De beoordeling van rafeling en scheurvorming in de MJPV,* (2009).

[13] W. Van Aalst, *DOS: beknopte analyse ijkvakken en toetsingsmetingen 2019,* (2019).

[14] J. Driessen, R. Landwier, Y. Verhoeven, and R. Verwoerd, *Beschrijvende Plaatsaanduiding Systematiek BPS*, DWW No. 039 (Ministerie van Verkeer en Waterstaat, Rijkswaterstaat, Dienst Weg- en Waterbouwkunde, 2005).

[15] E. Rahm and H. H. Do, *Data Cleaning: Problems and Current Approaches,* , 11 (2000).

[16] *Meerjarenplanning verhardingen (MJPV),* (2007).

[17] T. W. Anderson and D. A. Darling, *Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes,* Annals of Mathematical Statistics **23**, 193 (1952).

[18] S. S. Shapiro and M. B. Wilk, *An analysis of variance test for normality (complete samples),* Biometrika **52**, 591 (1965).

[19] N. M. Razali and Y. B. Wah, *Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests,* Journal of Statistical Modeling and Analytics **2**, 21 (2011).

[20] D. Freedman and P. Diaconis, *On the histogram as a density estimator:L2 theory,* Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **57**, 453 (1981).

[21] A. W. v. d. Vaart, *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press, Cambridge, 1998).

[22] W. Q. Meeker and L. A. Escobar, *Statistical Methods for Reliability Data*, Wiley Series in Probability and Statistics (Wiley, New York, 1998).

[23] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, 2nd ed., Wiley Series in Probability and Mathematical Statistics (Wiley, New York, 1994).

[24] A. L. Kolmogorov, *Sulla determinazione empirica di una legge di distribuzione,* Giornale dell'Istituto Italiano degli Attuari **4**, 83 (1933).

[25] T. J. Ypma, *Historical Development of the Newton–Raphson Method,* SIAM Review **37**, 531 (1995).

[26] P. Heidelberger and P. A. W. Lewis, *Quantile Estimation in Dependent Sequences,* Operations Research **32**, 185 (1984).

[27] N. Pya and S. N. Wood, *Shape constrained additive models,* Statistics and Computing **25**, 543 (2015).

[28] J. O. Ramsay, G. Hooker, and S. Graves, *Functional Data Analysis with R and MATLAB*, Use R! (Springer-Verlag, New York, 2009).

[29] C. De Boor, *A Practical Guide to Splines* (Springer, New York, 2001).

[30] J. Ramsay and B. W. Silverman, *Functional Data Analysis*, 2nd ed., Springer Series in Statistics (Springer-Verlag, New York, 2005).

[31] P. H. C. Eilers and B. D. Marx, *Flexible Smoothing with B-splines and Penalties,* Statistical Science **11**, 89 (1996).

[32] N. Pya, *scam: Shape Constrained Additive Models,* (2020).