Sarah KARASEK, BSc

Statistische Analysen von Fahrzeugreaktions- und Gleislagemessungen

Graphische Methoden zur Identifikation relevanter Variablen zur Beschreibung einer Zielgröße

MASTERARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

eingereicht an der

Technischen Universität Graz

Betreuer:
Univ.-Prof. Dipl.-Ing. Dr.techn. Ernst STADLOBER

Institut für Statistik

Graz, Juni 2016

EIDESSTATTLICHE ERKLÄRUNG

AFFIDAVIT

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit identisch.

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Datum/Date	Unterschrift/Signature

Diese Arbeit entstand am VIRTUAL VEHICLE Research Center in Graz, Österreich. Die Autoren bedanken sich für die Förderung im Rahmen des COMET K2 - Competence Centers for Excellent Technologies Programms des Österreichischen Bundesministeriums für Verkehr, Innovation und Technologie (bmvit), des Österreichischen Bundesministeriums für Wissenschaft, Forschung und Wirtschaft (bmwfw), der Österreichischen Forschungsförderungsgesellschaft mbH (FFG), des Landes Steiermark sowie der Steirischen Wirtschaftsförderung (SFG).

Ebenfalls danken wir den unterstützenden Firmen und Projektpartnern Siemens AG, ÖBB Infrastruktur AG, DB Netz AG und SBB AG Infrastruktur sowie der Technischen Universität Graz.

Kurzfassung

Wer schon einmal eine Zugfahrt bestritten hat, der weiß, wie unangenehm es sein kann, wenn der Zug während der Fahrt plötzlich ruckt. In dieser Arbeit werden Ursachen für solche unerwarteten Fahrzeugreaktionen analysiert. Als Basis der Untersuchung dienen Messdaten einer Bahnstrecke, die mit drei verschiedenen Fahrzeugtypen (Lok, Reisezugwagen, Güterwagen) jeweils drei mal befahren wurde. Für die Analyse werden zunächst Abschnitte im Messsignal identifiziert, in denen ungewöhnliche Werte auftreten könnten. Neben Methoden der Regressionsanalyse werden für diese Identifikation Quantile, sowie die Cook-Distanz verwendet. Als mögliche Ursachen werden verschiedene Bauwerke, Messfahrten, Fahrzeuge und Modellvariablen betrachtet. Dabei werden vor allem graphische Methoden gewählt um die Interpretation einem breiten Publikum, und nicht nur Statistikexperten, zu ermöglichen. Diese Arbeit liefert eine Schritt-für-Schritt Anleitung, wie die betrachteten vermuteten Gründe für unerwartete Fahrzeugreaktionen bestätigt oder ausgeschlossen werden können.

Abstract

Anyone who has ever joined a journey by train, who knows how uncomfortable it can be when the train jerks suddenly while driving. In this work causes of such unexpected vehicle reactions are analyzed. The basis of the investigation are measured data of a railway line, which has been cruised with three different types of vehicle (locomotive, passenger coaches, freight wagons) - three times each. At first sections are identified in the measurement signal for the analysis, in which unusual values may occur. Among methods of regression analysis we used quantiles and Cook's distance for this identification. Various structures, test runs, vehicle types and model variables are considered as possible causes. Mainly graphical methods are chosen to enable a wide audience, and not only statistics experts, an easy way of interpretation. This work provides step-by-step instructions how to confirm or rule out the considered presumed causes of unexpected vehicle reactions.

Danksagung

An dieser Stelle möchte ich mich bei allen bedanken, die mich nicht nur bei der Fertigung dieser Masterarbeit, sondern auch während der gesamten Studienzeit unterstützt haben.

Ein besonderer Dank gebührt Herrn Prof. Stadlober, der nicht nur diese Arbeit betreut hat, sondern mir auch während der letzten Jahre mit seinem Fachwissen und guten Ratschlägen zur Seite stand. Nicht zu vergessen Herrn Prof. Friedl, der auch immer ein offenes Ohr für mich und meine statistischen Fragen hatte.
Ich weiß das Engagement von Ihnen beiden sehr zu schätzen - vielen Dank.

Zudem möchte ich meinen Eltern danken, die immer hinter meinen Entscheidungen stehen und mir während des Studiums den Rücken gestärkt haben. Sie, mein Partner, meine ganze Familie und Freunde gaben mir wichtigen Halt über meine gesamte Studienzeit hinweg.

Ein herzliches Dankeschön geht auch an meine Studienkollegin und Freundin Doris, die mich durch unsere Gespräche nicht nur in fachlicher Hinsicht gestärkt hat, sondern auch eine wichtige persönliche Stütze war.

Zu guter Letzt möchte ich mich bei meinen Betreuern Herrn Luber und Herrn Fuchs, sowie dem Railteam des Virtuellen Fahrzeugs bedanken - ohne euch wäre diese Arbeit nicht zustande gekommen. Danke für lehrreiche und spannende sechs Monate bei euch.

Inhaltsverzeichnis

1 Einleitung

1.1 Motivation und Ziel der Arbeit

Seit der Entwicklung der Eisenbahn im 19. Jahrhundert und den ersten Personenzügen etwas später, wurde das Schienennetzwerk bis zur heutigen Zeit deutlich weiter entwickelt und ausgebaut. Wurde das Zugfahren früher als Luxus bezeichnet, erfreut es sich schon seit längerem großer Beliebtheit und der Zug avancierte zu einem der Hauptverkehrsmittel. Der Grund dafür liegt nicht nur an der raschen und unkomplizierten, sondern auch bequemen Reiseart. Doch die meisten Passagiere beachten diesen Komfort im Schienenfahrzeug wohl kaum, bis unerwartete Fahrzeugreaktionen sie aus ihrer Entspannung reißen. Dabei fragen sich die wenigsten, wie es zu solchen Unannehmlichkeiten kommen kann. Während einer Zugfahrt wirken verschiedene Kräfte sowohl auf das Fahrzeug, als auch auf die Schienen. Durch verschiedene Umstände werden diese Kräfte in manchen Situationen derart verstärkt, dass die Passagiere sie als unangenehm wahrnehmen. In dieser Arbeit sollen genau diese Ursachen für unerwartete Fahrzeugreaktionen identifiziert werden. Diese Informationen dienen einerseits der Sicherheit, da starke physikalische Kräfte zur Abnützung der Schienen beitragen, aber auch der Verbesserung des Fahrkomforts.
Im Wesentlichen soll diese Arbeit einen Wegweiser darstellen, wie unerwartete Fahrzeugreaktionen identifiziert und anschließend analysiert werden können. Für die Ursachenforschung werden einige mögliche Einflussgrößen wie Bauwerke und verschiedene Fahrzeugtypen betrachtet. Dafür werden vor allem graphische Methoden eingesetzt, um die Interpretation auch für Nicht-Mathematiker nachvollziehbar zu gestalten. Die notwendigen statistischen Grundlagen werden dabei an geeigneten Stellen in den Text einfließen.

1.2 Aufbau der Arbeit

Die vorliegende Arbeit lässt sich nach einer kurzen Erläuterung der verwendeten Software in mehrere Kapitel gliedern. In Kapitel 2 werden neben den physikalischen

Grundlagen, die Datenbasis, sowie relevante Variablen beschrieben. Außerdem wird auf die verwendete Methode zur Datenselektion eingegangen. Im anschließenden Kapitel 3 werden die statistischen Zusammenhänge der Variablen untersucht und anschließend passende Regressionsmodelle berechnet, sowie die Modellannahmen überprüft. Kapitel 4 beschäftigt sich schließlich mit der Identifikation von bestimmten Messabschnitten. Dafür werden zuerst zwei allgemeine und in weiterer Folge zwei differenzierte Ansätze, welche schlussendlich verwendet wurden, beschrieben. Anschließend werden die identifierten Abschnitte in Kapitel 5 analysiert und deren mögliche Ursachen untersucht. Das letzte Kapitel rundet diese Arbeit mit einer Zusammenfassung ab.

1.3 Programmiersprache R

Für diese Arbeit wurde das Statistikprogramm R verwendet, welches im Jahr 1993 von Ross Ihaka und Robert Gentleman ins Leben gerufen wurde. In [Ligges, 2005, S. 1] lässt sich folgende Beschreibung finden:

> Bei R handelt es sich um eine *Open Source* Software[1] und eine hochflexible Programmiersprache und -umgebung für (statistische) Datenanalyse und Grafiken, die auch Mittel zum Technologie- und Methodentransfer, etwa mit Hilfe von Zusatzpaketen, bereitstellt. Es gibt Datenzugriffsmechanismen für Textdateien, Binärdaten, R Workspaces, Datensätze anderer Statistiksoftware und Datenbanken.

[1] Open Source Software: Software, deren Quellcode frei erhältlich ist.

2 Grundlagen

Bevor mit den statistischen Analysen begonnen wird, steht eine kurze physikalische Einleitung im Vordergrund, in der die relevante Schienentheorie erklärt wird. Danach werden die verwendeten Datensätze beschrieben und auf die verschiedenen Methoden der Datenselektion zur Erstellung der Datenbasis eingegangen. Anschließend werden einige besonders relevante Variablen für die Analyse des Fahrsystems genauer untersucht.

2.1 Das Fahrzeug/Fahrweg-System

Die Fahrzeug/Fahrweg-Dynamik wird neben dem Rad/Schiene-Kontakt maßgeblich von den Gleislagefehlern bestimmt. Gleislagefehler beschreiben die Lageabweichung der Schiene zu einem theoretischen Soll-Zustand. Zur Beschreibung der Gleislagefehler werden zwei unterschiedliche Koordinatensysteme verwendet: das Gleislage- und das Schienenkoordinatensystem. Beim Gleislagekoordinatensystem beziehen sich die Gleislagefehler auf die Gleismitte, während im Schienenlagekoordiantensystem die Abweichungen der beiden Schienen zur Soll-Lage beschrieben werden.
Die Gleisgeometrieparameter abhängig vom Weg `s` im Gleislagekoordinatensystem lauten

- der Längshöhenfehler `z(s)` der Gleismitte,
- der Richtungsfehler `y(s)` der Gleismitte,
- der Querhöhenfehler `d(s)` der linken und rechten Schiene zueinander, sowie
- der Spurweitenfehler `g(s)`,

und im Schienenlagekoordinatensystem

- der Längshöhenfehler `zL(s)` der linken Schiene,
- der Längshöhenfehler `zR(s)` der rechten Schiene,

- der Richtungsfehler `yL(s)` der linken Schiene und
- der Richtungsfehler `yR(s)` der rechten Schiene.

Diese Parameter sind in Abbildung 2.1 dargestellt, wobei bis auf δ = `d` dieselben Bezeichnungen gelten.

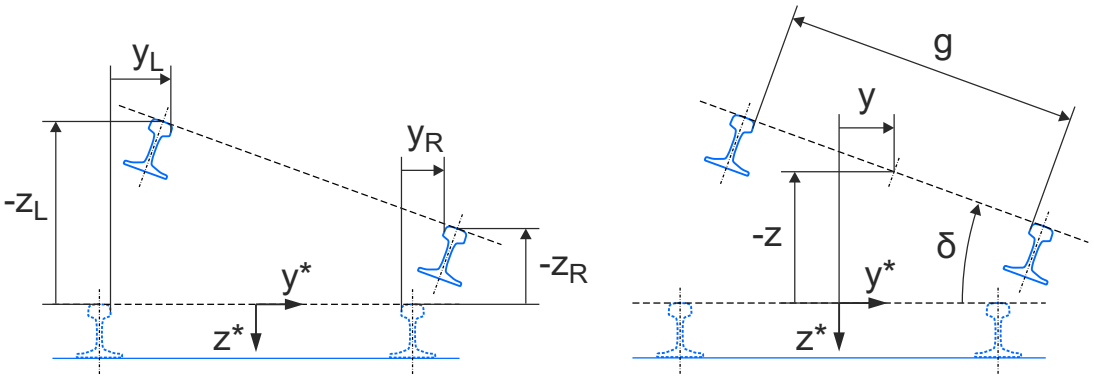

Abbildung 2.1: Gleislageabweichungen [Luber, 2011, S. 6]

Neben den Gleisgeometrieparametern werden drei weitere Größen erhoben,

- der Mittelwert der Krümmung `Ch_mean`,
- der Mittelwert der Überhöhung `U_mean`, sowie
- die Geschwindigkeit `v_mean`.

Die folgenden drei physikalischen Kräfte bilden die Grundlage für die betrachteten acht Zielgrößen und lauten

- `Q11` = vertikale Kraft auf das rechte Rad des ersten Radsatzes,
- `Q12` = vertikale Kraft auf das linke Rad des ersten Radsatzes,
- `sY1` = Summe der lateralen Kräfte auf den ersten Radsatz.

In Abbildung 2.2 werden diese Kräfte dargestellt.

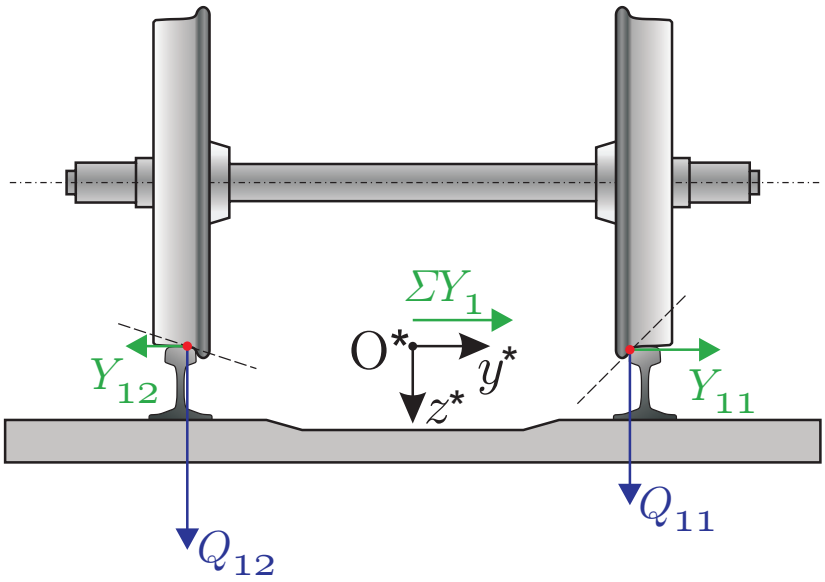

Abbildung 2.2: Darstellung der relevanten physikalischen Kräfte [Luber, 2011, S. 18]

Aus diesen drei Größen werden nach der Norm EN14363 (vgl. [ÖNORM, 2005]) für jeden der betrachteten Abschnitte die 99.85 % -, sowie 0.15 %- Perzentilwerte berechnet.

Zusätzlich wurden die Kräfte auf zwei Arten betrachtet:

1. statischer und dynamischer Anteil (= langwellig)
2. dynamischer Anteil (= kurzwellig).

Der langwellige Anteil, kommt hauptsächlich von der Trassierung bzw. der Strecke. Das Fahrzeug-Verhalten wird dabei von der Quasistatik bestimmt. Der kurzwellige Anteil geht auf die Gleislagefehler zurück. Das Fahrzeug-Verhalten wird hier von der Dynamik bestimmt.

Abbildung 2.3 erläutert, wie die dynamischen und quasistatischen Anteile berechnet werden. Die quasistatischen Größen wie `X_max_h1`, der 0.15 %-Perzentilwert, und `X_max_h2`, der 99.85 %-Perzentilwert, werden ausgehend von der Nulllinie berechnet, während die dynamischen Werte, wie etwa `X_dyn_max_h2` vom berechneten Mittelwert ausgehen.

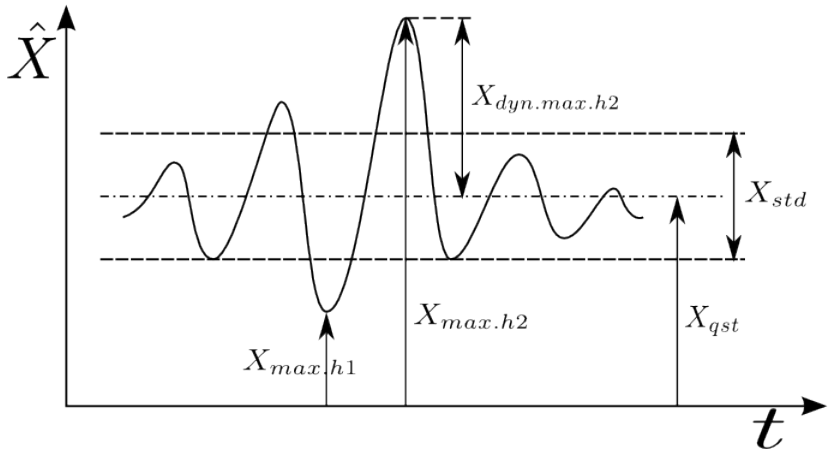

Abbildung 2.3: Definition der quasistatischen und dynamischen Größen

Die acht Zielgrößen werden aus den drei gemessenen Kräften `Q11`, `Q12` und `sY1` wie in Tabelle 2.1 zusammengefasst abgeleitet:

Tabelle 2.1: Ableitung der Zielgrößen aus den originalen Messdaten

- `Q11_max_h2` = 99.85 %-Perzentilwert von `Q11`
- `Q12_max_h2` = 99.85 %-Perzentilwert von `Q12`
- `Q11dyn_max_h2` = 99.85 %-Perzentilwert von `Q11_dyn`
- `Q12dyn_max_h2` = 99.85 %-Perzentilwert von `Q12_dyn`
- `sY1_rms` = quadratischer Mittelwert von `sY1_mean2m`
- `sY1_maxperc` = Maximum aus `sY1_maxperc_h1` und `sY1_maxperc_h2`
- `sY1dyn_rms` = quadratischer Mittelwert von `sY1_mean2m_dyn`
- `sY1dyn_maxperc` = Maximum aus `sY1dyn_maxperc_h1` und `sY1dyn_maxperc_h2`

Wobei `sY1_maxperc_h1` = 0.15 %-Perzentilwert von `sY1_mean2m` und `sY1_maxperc_h2` = 99.85 %-Perzentilwert von `sY1_mean2m` gilt. Die entsprechenden dynamischen Werte berechnen sich analog.

2.2 Datensatz

Als Grundlage für diese Arbeit dienen Messdaten, die am 26.10.2010 (Vormittag) auf der etwa 15 km langen Bahnstrecke von Geislingen nach Westerstetten erhoben wurden. Diese Strecke ist Teil der Filstalbahn - einer Bahnstrecke im deutschen Bundesland Baden-Württemberg von Stuttgart nach Ulm. Auf der Filstalbahn findet man für gewöhnlich alle Arten von Zügen, unter anderem Regionalbahnen und Interregio-Express-Züge.

Für die Aufzeichnungen wurden drei verschiedene Fahrzeuge in einer bestimmten Reihenfolge aneinander gehängt. Neben einer Lok (= $FZ1$) wurden dabei noch ein Reisezugwagen (= $FZ2$) und ein Güterwagen (= $FZ5$) verwendet. Diese Fahrzeuge unterscheiden sich nicht nur durch ihren Aufbau, sondern auch durch einige mechanische Größen, wie etwa Gewicht und Steifigkeit.

Während der Fahrt zeichnete ein Messgerät sämtliche Größen zur Gleislage und Trassierung, sowie Betriebsbedingungen und physikalische Kräfte auf. Diese Aufzeichnungen erfolgten dabei alle 16 cm. In dieser Arbeit wird von der gesamten Strecke nur ein etwa 11 km (= 73 % der gesamten Strecke) langer Abschnitt verwendet. Dieser Entscheidung liegt die Tatsache zugrunde, dass in dem gewählten Abschnitt kein Messsensor ausgefallen ist und somit alle relevanten Größen erhoben werden konnten.

Da das Datenset des reinen Signals mit 68115 Datenpunkten zu fein für die Analysen ist, werden bestimmte Bereiche aus diesem Datensatz ausgewählt, die in weiterer Folge für die Analysen verwendet werden. In Abschnitt 2.3 wird die dafür verwendete Selektionsmethode erläutert, mit welcher schlussendlich der endgültige Datensatz von 115 Datenpunkten entstehen konnte. Jeder dieser Punkte stellt dabei einen Signalabschnitt einer Länge von 75 m dar.

Zusätzlich zur Variable `TestVehicle`, welche den Fahrzeugtyp angibt, wurden drei weitere kategorische Variablen erhoben. `Switch_cat`, `Bridge_cat` und `Station_cat` geben an, ob sich in einem 75 m Abschnitt eine Weiche, eine Brücke oder ein Bahnhof befindet.

Eine weitere Variable, der große Aufmerksamkeit geschenkt wird, beinhaltet Informationen über das Streckenlayout. Diese kategorische Variable beschreibt demnach die Form der Schiene.

Zu Beginn erfolgt die Einteilung in eine der folgenden fünf Kategorien:

- Gerade
- Linksübergang
- Rechtsübergang
- Linksbogen
- Rechtsbogen.

Abbildung 2.4 stellt eine Skizze der Einteilung einer Strecke in Gerade - Rechtsübergang - Rechtsbogen dar.

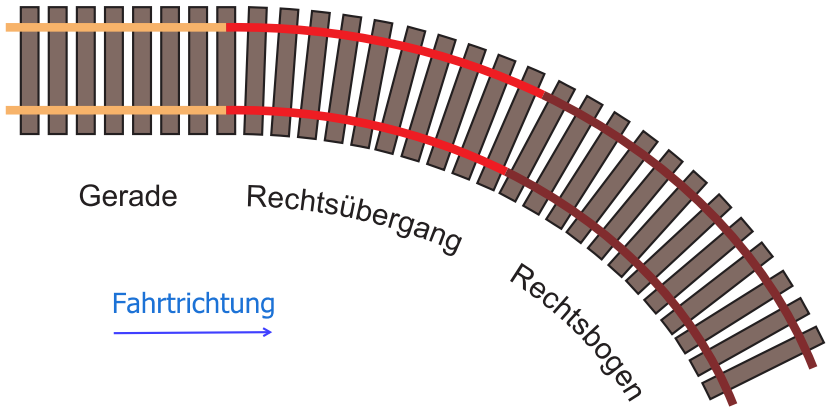

Abbildung 2.4: Skizze einiger Streckenlayoutelemente

Die Variable `LayoutElements_cat`, welche diese Kategorien enthält, wird jedoch in weiterer Folge abgeändert (siehe Abschnitt 2.4).

Ein wichtiger Punkt in Bezug auf ein statistisches Experiment ist dessen Wiederholbarkeit. Diese ist hier gegeben, da die Strecke am gleichen Tag im selben Zugverband drei mal hintereinander befahren wurde, sodass insgesamt neun Sets (Kombinationen aus: Messfahrten 41, 42, 43, Fahrzeug 1, 2, 5) zu jeweils 115 Datenpunkten vorliegen. Hier werden jedoch nur die detaillierten Ergebnisse für den Datensatz gezeigt, der Messungen der Messfahrt 43 und des Fahrzeugtyps 2 (Reisezugwagen) enthält, da die Ergebnisse aller Fahrzeug/Messfahrt-Kombinationen den Rahmen dieser Arbeit sprengen würden.

In Tabelle 2.2 sind alle verwendeten Variablen, sowie deren Definition, übersichtlich zusammengefasst.

Tabelle 2.2: Beschreibung der Variablen

	Variable	Einheit	Beschreibung
Info	s	m	Weg
	km_mean	km	Kilometerstand
Input	z_max	mm	Längshöhenlage (Maximum)
	z_std	mm	Längshöhenlage (Standardabweichung)
	y_max	mm	Richtungslage (Maximum)
	y_std	mm	Richtungslage (Standardabweichung)
	d_max	deg	Querhöhe (Maximum)
	d_std	deg	Querhöhe (Standardabweichung)
	g_max	mm	Spurweitenfehler (Maximum)
	g_std	mm	Spurweitenfehler (Standardabweichung)
	zL_max	mm	Längshöhenlage links (Maximum)
	zL_std	mm	Längshöhenlage links (Standardabweichung)
	zR_max	mm	Längshöhenlage rechts (Maximum)
	zR_std	mm	Längshöhenlage rechts (Standardabweichung)
	yL_max	mm	Richtungslage links (Maximum)
	yL_std	mm	Richtungslage links (Standardabweichung)
	yR_max	mm	Richtungslage rechts (Maximum)
	yR_std	mm	Richtungslage rechts (Standardabweichung)
	Ch_mean	1/m	Mittelwert der Krümmung in einem Abschnitt
	U_mean	m	Mittelwert der Überhöhung in einem Abschnitt
	v_mean	km/h	Mittelwert der Geschwindigkeit in einem Abschnitt
	LayoutElements_cat	-	Streckenführungselemente (ursprünglich 13 Levels)
Zusatz	TestVehicle	-	Binäre Variable Fahrzeugtyp
	Switch_cat	-	Binäre Variable Weiche
	Bridge_cat	-	Binäre Variable Brücke
	Station_cat	-	Kategorische Variable Bahnhof
Zielgröße	Q11_max_h2	kN	vertikale Kraft auf das rechte Rad des ersten Radsatzes; 99.85%-Perzentilwert
	Q12_max_h2	kN	vertikale Kraft auf das linke Rad des ersten Radsatzes; 99.85%-Perzentilwert
	sY1_rms	kN	Summe der lateralen Kräfte auf den ersten Radsatz (statischer und dynamischer Anteil); quadratischer Mittelwert
	sY1_maxperc	kN	Summe der lateralen Kräfte auf den ersten Radsatz (statischer und dynamischer Anteil); Max aus sY1_maxperc_h1 und sY1_maxperc_h2
	Q11dyn_max_h2	kN	dynamischer Wert von Q11_max_h2
	Q12dyn_max_h2	kN	dynamischer Wert von Q12_max_h2
	sY1dyn_rms	kN	dynamischer Wert von sY1_rms
	sY1dyn_maxperc	kN	dynamischer Wert von sY1_maxperc

2.3 Datenselektionsmethoden

Wie bereits zuvor erwähnt, sollen die zu analysierenden Daten mittels geeigneter Methode aus dem gemessenen Signal ausgewählt werden. Dafür wurden zwei Methoden ausprobiert, welche in weiterer Folge erläutert werden.
Bei beiden Methoden geht es darum, Streckenabschnitte auszuwählen, welche eine Länge von 75 m aufweisen. Da der Abstand zwischen zwei Datenpunkten 16 cm beträgt, werden für einen Abschnitt etwa 468 Datenpunkte benötigt.

2.3.1 Methode 1

Zuerst wurde folgende Methode verwendet, um aus dem gemessenen Signal einen brauchbaren Datensatz zu generieren:

1. von links beginnend wird das Signal in 75 m Abschnitte eingeteilt, wobei alle Datenpunkte *in einem Abschnitt dasselbe Streckenlayout* (Gerade, Linksübergang, Rechtsübergang, Linksbogen, Rechtsbogen) haben müssen.

Befinden sich nach einem gewählten Abschnitt noch Datenpunkte desselben Layouts, die jedoch nicht die erforderliche Gesamtlänge von 75 m erreichen, so werden diese weggeschnitten und nicht weiter berücksichtigt.

Diese Vorgehensweise lässt sich mithilfe von Abbildung 2.5 graphisch veranschaulichen.

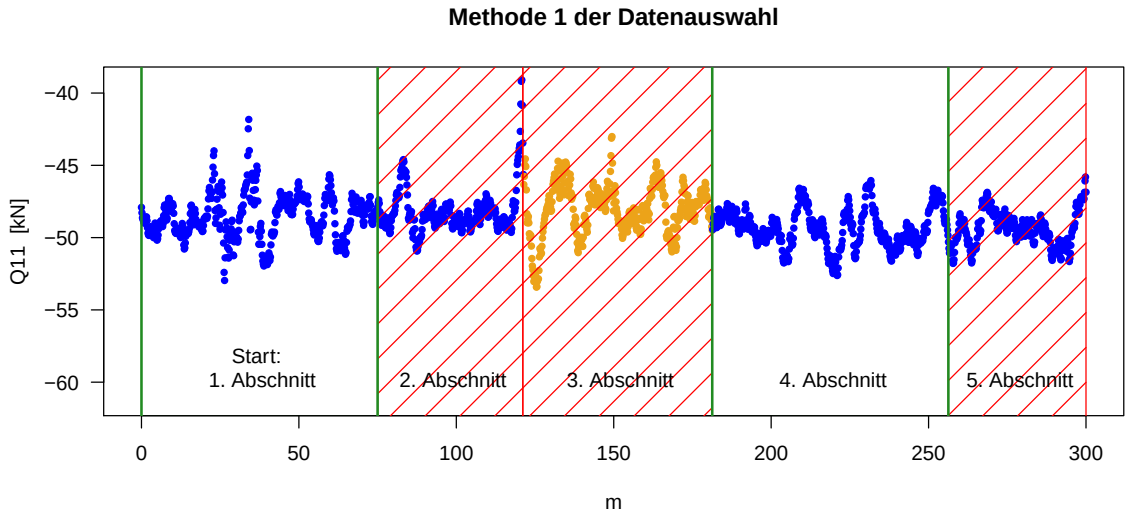

Abbildung 2.5: Beispiel für Methode 1 mit 75 m Abschnitten

2.3.2 Methode 2

Die Selektion der Abschnitte erfolgt bei der zweiten Methode etwas komplizierter und läuft wie folgt ab:

1. suche von links beginnend Stellen, an denen sich das Streckenlayout ändert: dies ist das Zentrum eines neuen Abschnittes, falls die erforderliche Gesamtlänge erreicht wird (solche Abschnitte müssen genau zwei verschiedene Streckenlayouts beinhalten)

2. die restlichen Daten werden erneut von links beginnend in 75 m Abschnitte eingeteilt (in diesen Abschnitten kommt jeweils nur ein Streckenlayout vor).

Erneut werden aufeinanderfolgende Datenpunkte, die keine 75 m Abschnitte bilden können, weggeschnitten.

Eine Skizze dieser Methode stellt Abbildung 2.6 dar.

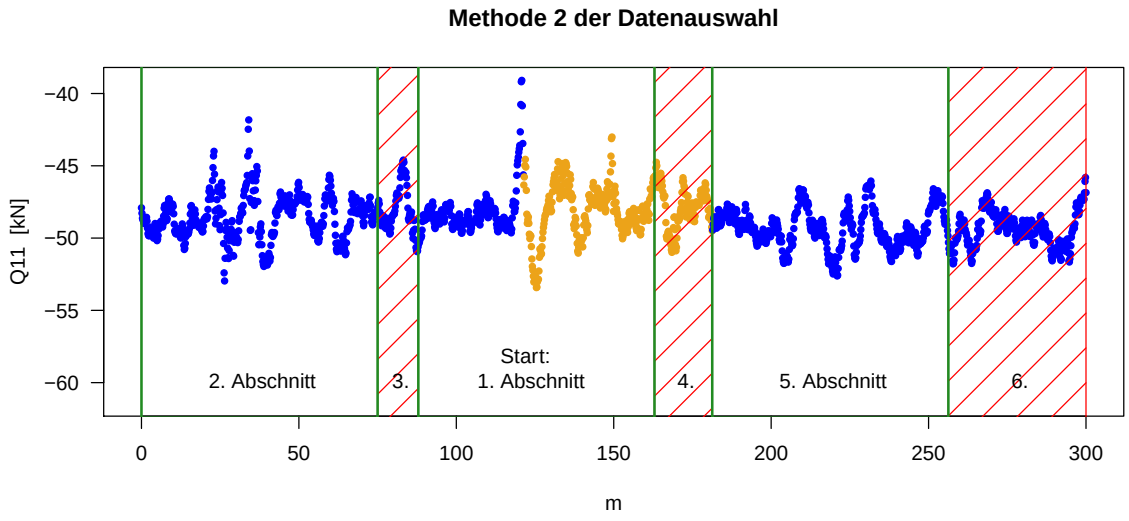

Abbildung 2.6: Beispiel für Methode 2 mit 75 m Abschnitten

2.3.3 Vergleich der beiden Methoden

Die erste Methode hat den unbestrittenen Vorteil, dass sie sehr einfach umzusetzen ist. Ein großer Nachteil besteht darin, dass die Abschnitte nach dem Streckenlayout eingeteilt werden. Es wird jedoch vermutet, dass sich die Zielgrößen im Übergang von einem zum nächsten Streckenlayout stark ändern bzw. anpassen. Dies wird bei Methode 1 völlig außer acht gelassen. Bei Methode 2 werden aber genau diese Übergänge zuerst

ausgewählt und sind somit im neuen Datensatz enthalten. Aufgrund dieser Überlegungen greifen wir in dieser Arbeit auf Methode 2 zurück. Dadurch entstehen neue Kategorien für jene Variable, die das Streckenlayout enthält, welche in Abschnitt 2.4 detailliert beschrieben werden.

2.4 Variable `Layout`

Durch die Datenselektionsmethode 2 (vgl. Abschnitt 2.3) werden nicht nur Abschnitte mit demselben Streckenlayout ausgewählt, sondern auch viele, die zwei verschiedene Layouts beinhalten - die sogenannten Übergänge. So fallen zum Beispiel Abschnitte, in denen das Layout von einer Geraden zu einem Linksübergang wechselt, in die neue Kategorie `Str_TraL`, wobei `Str` für Gerade (engl. *Straight*) und `TraL` für Linksübergang (engl. *Transition left*) steht.

Jene Abschnitte, in denen nur ein Streckenlayout auftritt, erhalten keine neuen Kategorien, sondern behalten ihre ursprünglichen bei.

Für die modifizierte Einteilung des Streckenlayouts wird eine neue Variable generiert - `LayoutElements_cat`. Diese kategorische Variable realisiert nach der besprochenen Datenselektionsmethode 2 in einer von 13 Kategorien, wie in Tabelle 2.3 dargestellt wird. Andere Kategorien als jene in der Tabelle angegeben sind nicht relevant und sollten aus dem Datensatz entfernt werden, sofern sie fälschlicherweise generiert wurden. So ergibt etwa die Kategorie `TraL_CurR` wenig Sinn, da auf einen Linksübergang keine Rechtskurve, sondern nur eine Linkskurve oder eine Gerade folgen kann.

Tabelle 2.3: 13 Kategorien der Variable `LayoutElements_cat` nach der Datenselektion mit Methode 2

`LayoutElements_cat`	Beschreibung
CurL	Linksbogen
CurL_TraL	Übergang von einem Linksbogen zu einem Linksübergang
CurR	Rechtsbogen
CurR_TraR	Übergang von einem Rechtsbogen zu einem Rechtsübergang
Str	Gerade
Str_TraL	Übergang von einer Geraden zu einem Linksübergang
Str_TraR	Übergang von einer Geraden zu einem Rechtsübergang
TraL	Linksübergang
TraL_CurL	Übergang von einem Linksübergang zu einem Linksbogen
TraL_Str	Übergang von einem Linksübergang zu einer Geraden
TraR	Rechtsübergang
TraR_CurR	Übergang von einem Rechtsübergang zu einem Rechtsbogen
TraR_Str	Übergang von einem Rechtsübergang zu einer Geraden

Die Variable `LayoutElements_cat` lässt sich weiters in fünf Hauptkategorien und acht Nebenkategorien, welche die Abschnitte mit Übergängen bezeichnen, einteilen.
Im Folgenden wird die Häufigkeitsverteilung durch ein Stabdiagramm dargestellt (siehe Abbildung 2.7).

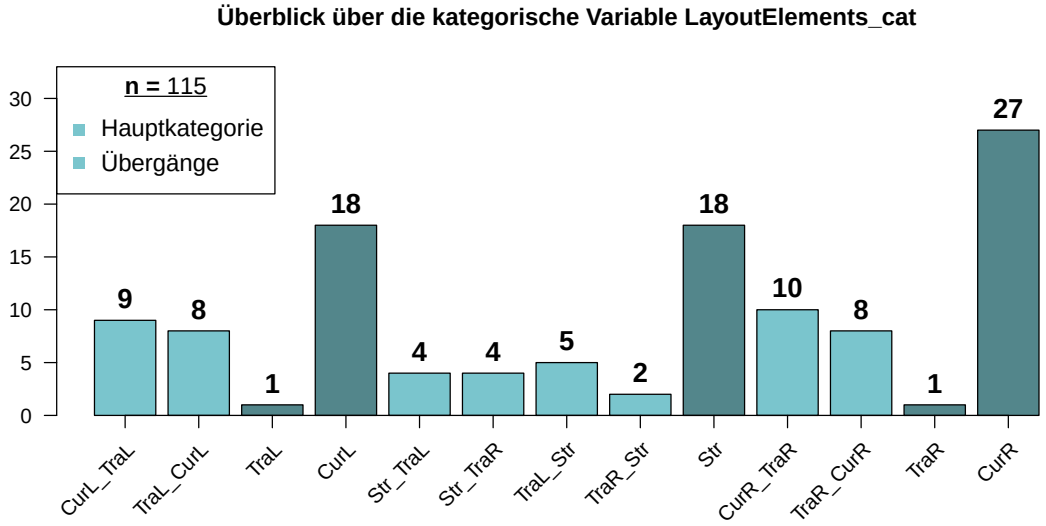

Abbildung 2.7: Stabdiagramm der Variable `LayoutElements_cat`

Man erkennt, dass die Strecke aus sehr vielen Rechtsbögen (27), aber auch einigen Linksbögen (18) und Geraden (18) besteht. Einige Kategorien sind jedoch so dünn besetzt, dass eine neue Einteilung der Klassen zu empfehlen ist. Daher wird eine neue Variable `Layout` erzeugt, in der die Kategorien der Variable `LayoutElements_cat`, welche die Übergänge bezeichnen, zusammengefasst werden. Neben den fünf Hauptkategorien gibt es damit nur mehr drei weitere Kategorien, wie in Tabelle 2.4 zusammengefasst.

Tabelle 2.4: 8 Kategorien der Variable `Layout`

`Layout`	Beschreibung
Ueber_L	Zusammenfassung von CurL_TraL, TraL_CurL
TraL	Linksübergang
CurL	Linksbogen
Ueber_Str	Zusammenfassung von Str_TraL, Str_TraR, TraL_Str, TraR_Str
Str	Gerade
Ueber_R	Zusammenfassung von CurR_TraR, TraR_CurR
TraR	Rechtsübergang
CurR	Rechtsbogen

Diese neue Variable `Layout` realisiert nun einigermaßen gleichmäßig in acht Kategorien, wobei `TraL` und `TraR` mit jeweils nur einer Beobachtung eine Ausnahme bilden (siehe Abbildung 2.8).

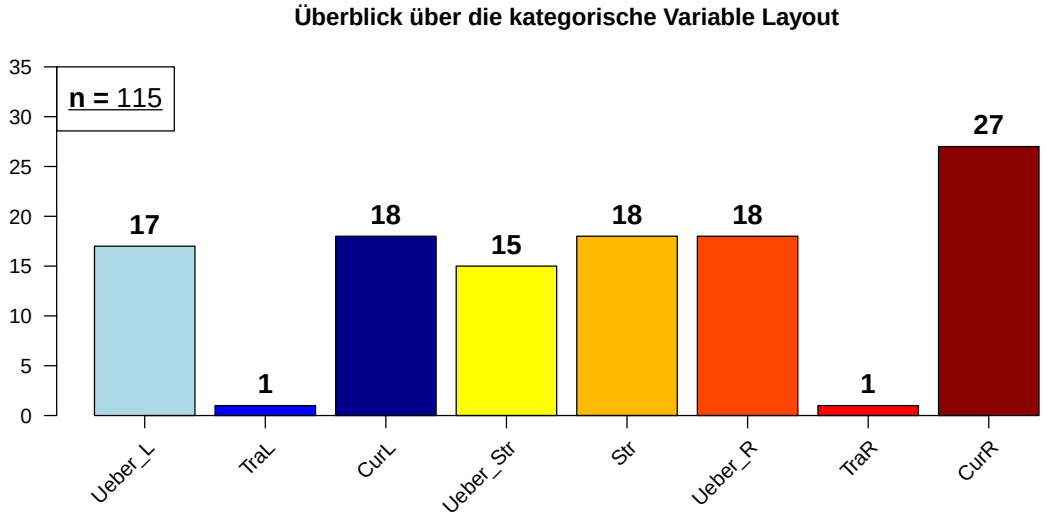

Abbildung 2.8: Stabdiagramm der Variable `Layout`

2.5 Variable `TestVehicle`

Wie bereits zuvor erwähnt, wurde die Teststrecke mit drei verschiedenen Fahrzeugen befahren, für die aufgrund ihrer verschiedenen Bauweisen Unterschiede in den Messungen zu erwarten sind. Aus diesem Grund wird ein Blick auf die Zielgrößen abhängig vom Testfahrzeug geworfen, um einen Überblick über diverse Gemeinsamkeiten zu erhalten. Die Abbildungen 2.9 - 2.11 zeigen den Verlauf der Zielgröße `Q11_max_h2` für Messfahrt 43 für alle drei Fahrzeuge. Auf den ersten Blick lassen sich unmittelbar große Unterschiede in der Skala erkennen. Während sich die Werte für die Fahrzeuge 2 (Reisezugwagen) und 5 (Güterwagen) sehr ähnlich sind, weichen jene für Fahrzeug 1 (Lok) deutlich davon ab.

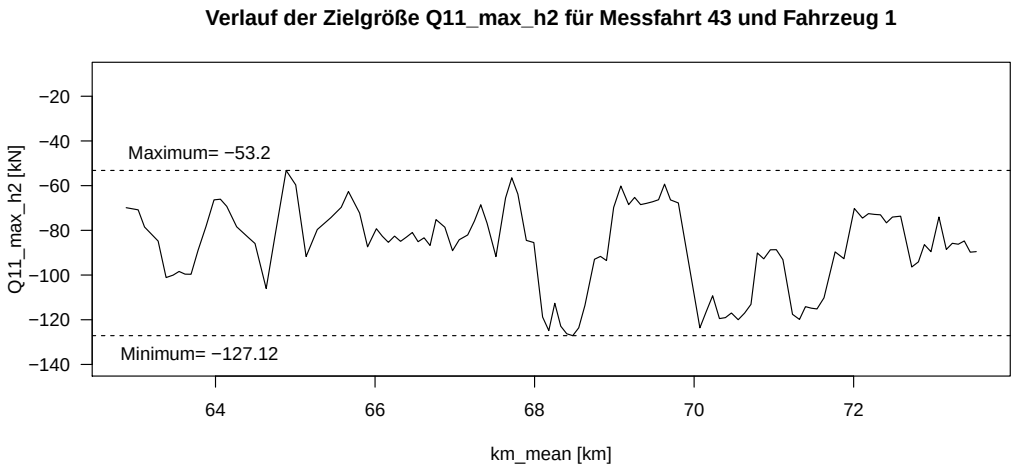

Abbildung 2.9: Verlauf der Zielgröße `Q11_max_h2` für Lok

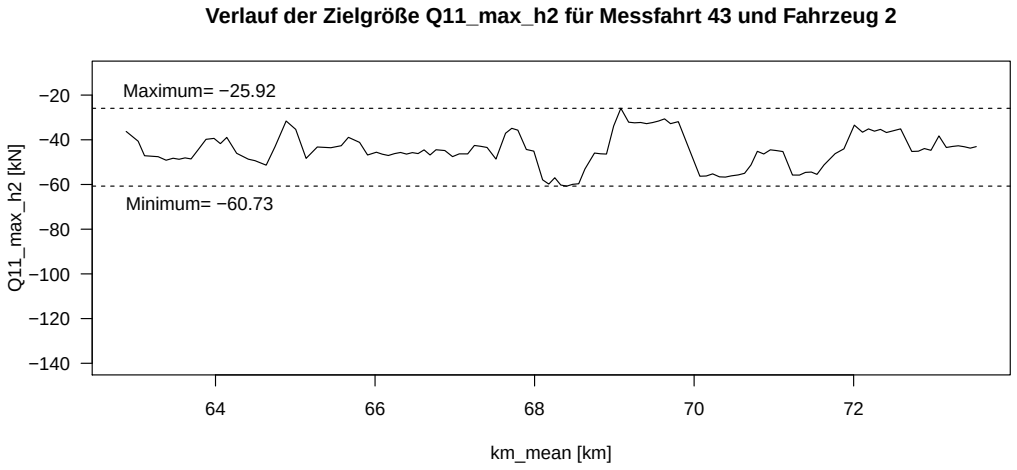

Abbildung 2.10: Verlauf der Zielgröße `Q11_max_h2` für Reisezugwagen

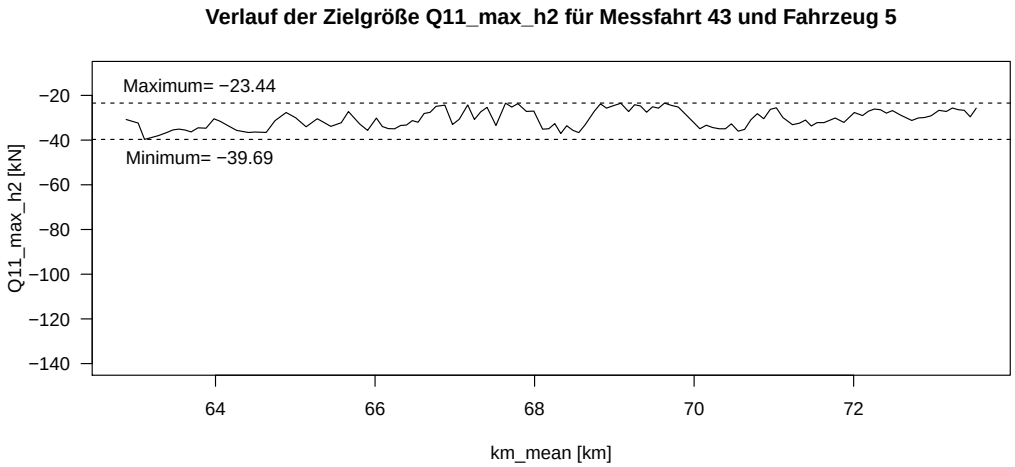

Abbildung 2.11: Verlauf der Zielgröße `Q11_max_h2` für Güterwagen

2.6 Wiederholungsfahrten

Neben den verschiedenen Fahrzeugtypen ist eine Unterscheidung nach der Messfahrt möglich. Da die Teststrecke insgesamt drei mal befahren wurde, können die Fahrten miteinander verglichen werden. Vor den Analysen wird deshalb ein Blick auf den Verlauf der Zielgrößen bezüglich der drei Messfahrten geworfen. Die Abbildungen 2.12 - 2.14 zeigen den Verlauf der Zielgröße `Q11_max_h2` für jede Messfahrt von Fahrzeugtyp 1 (Lok). Während sich die Wertebereiche für alle Messfahrten sehr ähnlich sind, scheint es bei Kilometer 69 einige Unterschiede zwischen Messfahrt 42 und den anderen beiden Fahrten zu geben.

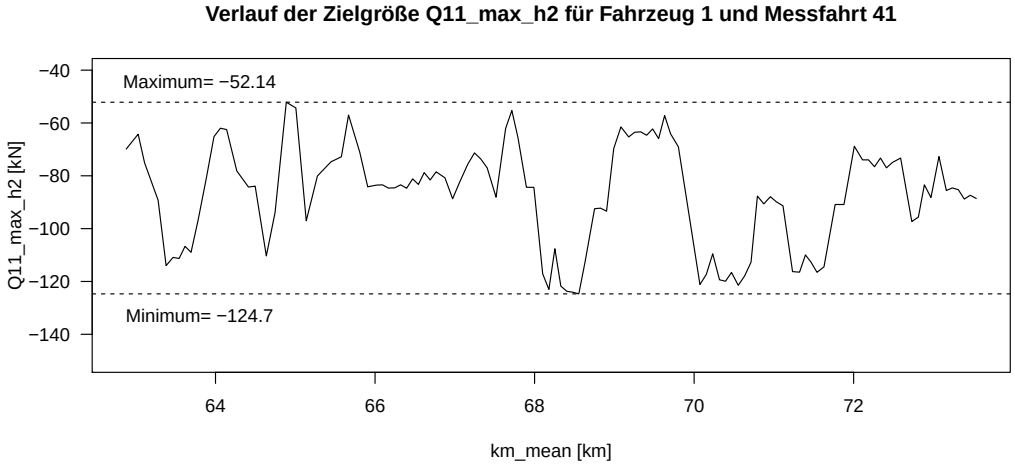

Abbildung 2.12: Verlauf der Zielgröße `Q11_max_h2` für Messfahrt 41 mit Lok

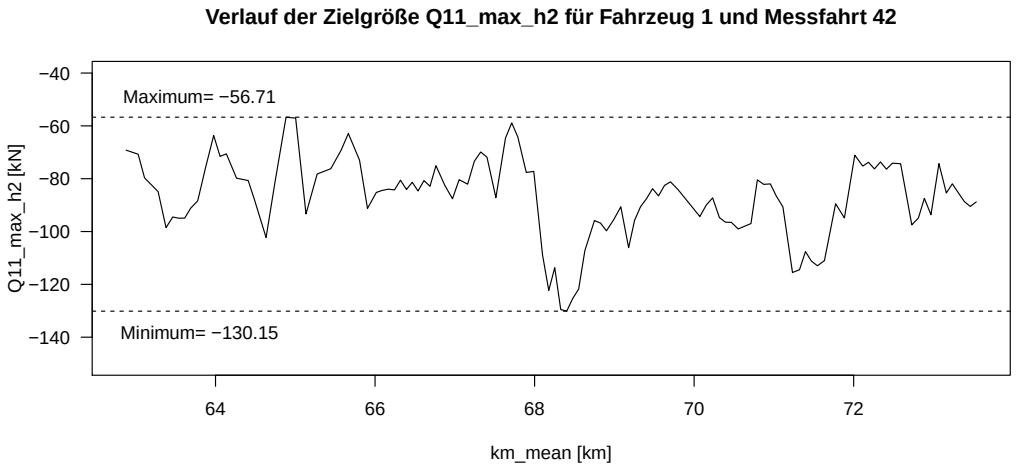

Abbildung 2.13: Verlauf der Zielgröße `Q11_max_h2` für Messfahrt 42 mit Lok

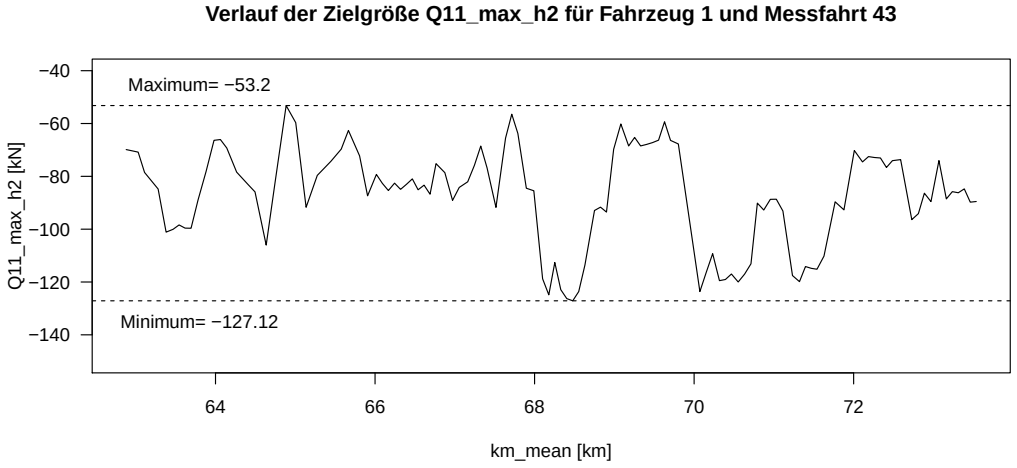

Abbildung 2.14: Verlauf der Zielgröße `Q11_max_h2` für Messfahrt 43 mit Lok

3 Modellbildung

In diesem Kapitel werden die Zusammenhänge der einzelnen Variablen untersucht, um anschließend ein einfaches multiples lineares Regressionsmodell aufzustellen. Zudem müssen die Modellannahmen überprüft werden, in dem diverse Residuenplots betrachtet werden.
Das gefundene Modell soll als Grundlage für weitere Analysen dienen. Dabei wird der Fokus nicht darauf gelegt, ein Modell zu finden, welches die Zielgröße perfekt beschreibt. Es soll vielmehr ein möglichst einfaches lineares Modell gefunden werden. Ein optimales Modell zur Beschreibung der Zielgrößen kann mithilfe einer aufwendigeren Mehrkörpersimulation gefunden werden (vgl. [Luber, 2011]).

3.1 Korrelationskoeffizient

Bevor ein Regressionsmodell aufgestellt wird, sollte die Beziehung zwischen den Variablen analysiert werden. Um den **linearen** Zusammenhang zwischen den Variablen, bzw. insbesondere zwischen Zielgröße und Inputvariablen, zu bewerten, stehen einige Maße zur Verfügung. Diese schätzen die theoretische Korrelation, die unter anderem in [Büning und Trenkler, 1994, S. 218–219] wie folgt definiert wird:

Für zwei Zufallsvariablen X und Y ist der Korrelationskoeffizient

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \tag{3.1}$$

eine geeignete Maßzahl zur Charakterisierung eines linearen Zusammenhangs. So gilt bekanntlich

a) $-1 \leq \rho \leq +1$

b) Die Unabhängigkeit von X und$Y$ impliziert $\rho = 0$. Die Umkehrung gilt im Allgemeinen nicht, wohl aber:

c) Sind X und Y normalverteilt und ist $\rho = 0$, so gilt: X und Y sind unabhängig.

d) $|\rho| = 1$ genau dann, wenn $Y = cX + d$ (mit Wahrscheinlichkeit 1), wobei $c \neq 0$ und d Konstanten sind.

Die am häufigsten verwendeten Schätzer für die Korrelation sind die beiden Maße von Pearson und Spearman, die in [Fahrmeir *et al.*, 2010, S. 139 & 144] so definiert werden:

Korrelationskoeffizient nach Pearson

Der *Bravais-Pearson-Korrelationskoeffizient* ergibt sich aus den Daten (x_i, y_i), $i = 1, \ldots, n$, durch

$$r_P = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}.} \tag{3.2}$$

Wertebereich: $-1 \leq r_P \leq 1$

$r_P > 0$ positive Korrelation, gleichsinniger linearer Zusammenhang

Tendenz: Werte (x_i, y_i) um eine Gerade positiver Steigung liegend

$r_P < 0$ negative Korrelation, gegensinniger linearer Zusammenhang,

Tendenz: Werte (x_i, y_i) um eine Gerade negativer Steigung liegend

$r_P = 0$ keine Korrelation, unkorreliert, kein linearer Zusammenhang

Korrelationskoeffizient nach Spearman

Der *Korrelationskoeffizient nach Spearman* ist definiert durch

$$r_{SP} = \frac{\sum(rg(x_i) - \overline{rg}_X)(rg(y_i) - \overline{rg}_Y)}{\sqrt{\sum(rg(x_i) - \overline{rg}_X)^2 \sum(rg(y_i) - \overline{rg}_Y)^2},} \tag{3.3}$$

Wertebereich: $-1 \leq r_{SP} \leq 1$

$r_{SP} > 0$ gleichsinniger positiver Zusammenhang,

Tendenz: x groß $\Leftrightarrow$ y groß, x klein $\Leftrightarrow$ y klein

$r_{SP} < 0$ gegensinniger negativer Zusammenhang,

Tendenz: x groß $\Leftrightarrow$ y klein, x klein $\Leftrightarrow$ y groß

$r_{SP} \approx 0$ kein monotoner Zusammenhang

Während der Pearson-Korrelationskoeffizient die Differenzen von Beobachtungen und Mittelwerten verwendet, werden beim Korrelationskoeffizient nach Spearman die Ränge

ermittelt, wobei $rg(x_i)$ den Rang der Beobachtung in der Datenreihe $(x_1, \ldots, x_n)$ angibt. $rg(x_i) = j$ bedeutet, dass x_i das j-größte Element in der Datenreihe ist. $\overline{rg}_X$ bezeichnet das arithmetische Mittel $\overline{rg}_X = \frac{1}{n} \sum_{i=1}^{n} rg(x_i)$.
Dadurch entsteht jedoch ein gewisser Informationsverlust im Falle von stetigen Merkmalen, die in unserer Arbeit analysiert werden. Daher wird im Folgenden der Korrelationskoeffizient nach Pearson verwendet.
In Abbildungen 3.1 - 3.4 sind in den unteren Dreiecksmatrizen Scatterplots der jeweiligen Variablen ersichtlich, während in der rechten oberen Matrix der Korrelationskoeffizient nach Pearson berechnet und größenabhängig eingetragen wurde. Zu Gunsten der Lesbarkeit wurde zusätzlich eine farbliche Interpretation gewählt. Je näher der Korrelationskoeffizient bei +1 liegt, desto mehr passt sich der Hintergrund des Panels der Farbe rot an. Je näher der Wert bei −1 liegt, desto mehr geht die Farbe ins Blau über. Für die Fahrzeug/Messfahrt-Kombination Lok/Messfahrt 43 und die Zielgröße `Q11_max_h2` können damit einige lineare Zusammenhänge identifiziert werden. So weisen einige Inputgrößen wie `z_max` und `z_std` untereinander einen linearen Zusammenhang auf, was meist auf den Vergleich von Maximum und Standardabweichung derselben Variable zurückzuführen ist. Aber auch der Zusammenhang zwischen der Längshöhenlage links (`zL_max`) und rechts (`zR_max`) ist physikalisch erklärbar.

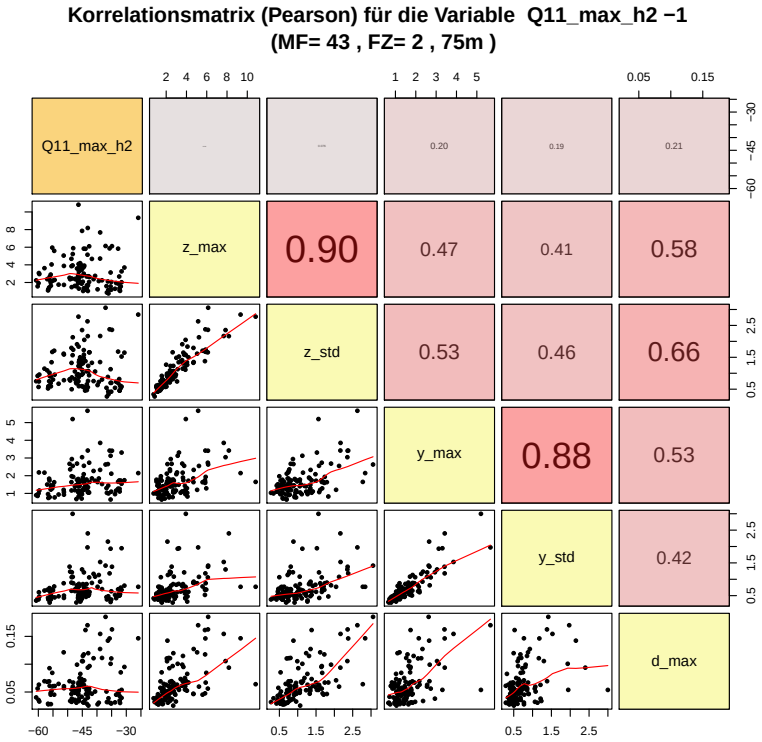

Abbildung 3.1: Korrelationsmatrix mit Scatterplots und Korrelationskoeffizient nach Pearson für die Variable `Q11_max_h2` im Detail - Teil 1

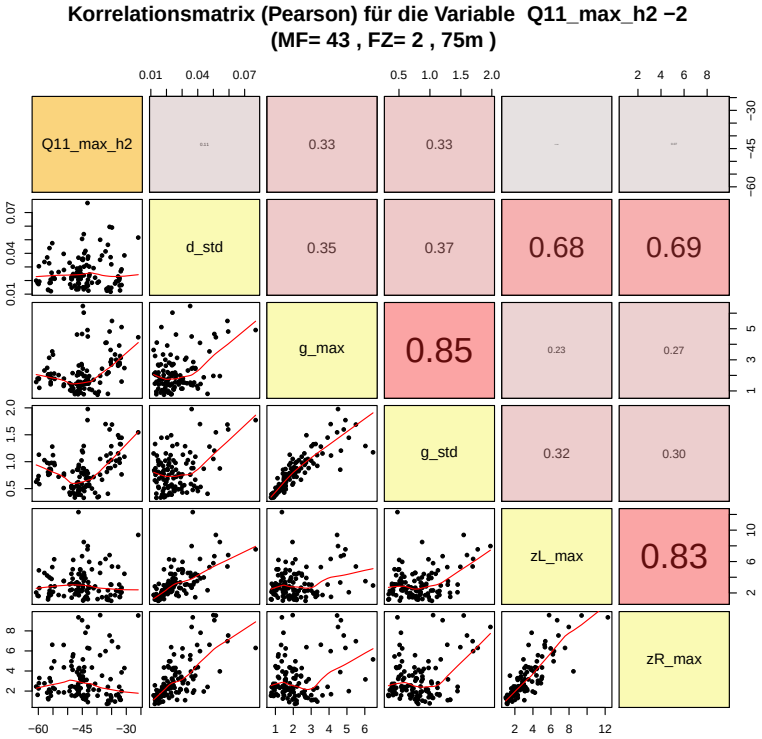

Abbildung 3.2: Korrelationsmatrix mit Scatterplots und Korrelationskoeffizient nach Pearson für die Variable `Q11_max_h2` im Detail - Teil 2

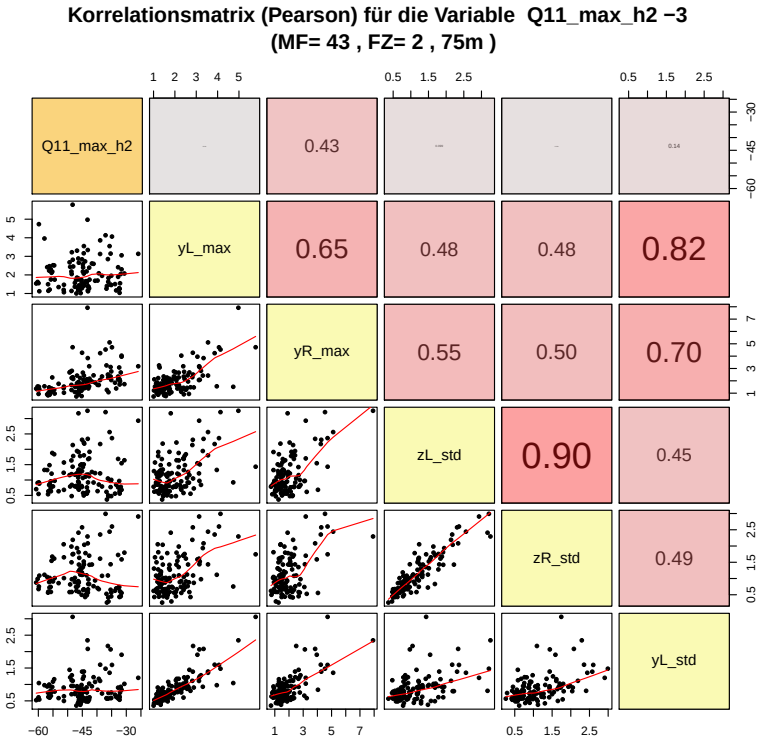

Abbildung 3.3: Korrelationsmatrix mit Scatterplots und Korrelationskoeffizient nach Pearson für die Variable `Q11_max_h2` im Detail - Teil 3

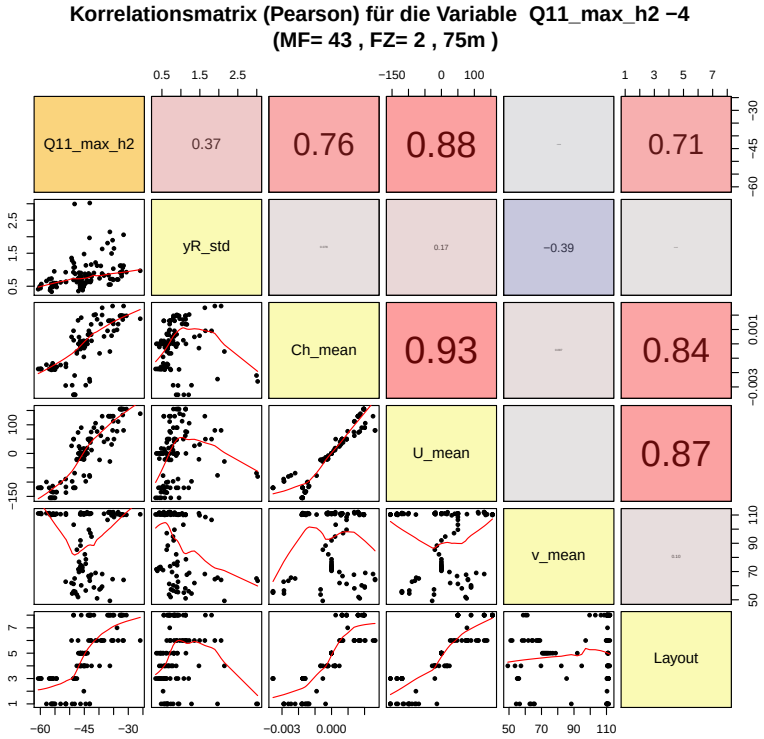

Abbildung 3.4: Korrelationsmatrix mit Scatterplots und Korrelationskoeffizient nach Pearson für die Variable `Q11_max_h2` im Detail - Teil 4

Interessanter ist jedoch die Frage, welche Inputgrößen einen linearen Zusammenhang zur Zielgröße `Q11_max_h2` aufweisen. Dafür wird ein Blick auf die erste Zeile in jeder der vier Graphiken in den Abbildungen 3.1 - 3.4 geworfen. Diese erste Betrachtung identifiziert neben der Variable `Layout` auch noch Krümmung (`Ch_mean`) und Überhöhung (`U_mean`) als mögliche relevante erklärende Größen um `Q11_max_h2` in einem linearen Modell zu beschreiben. Diese Variablen scheinen einen positiven linearen Zusammenhang zur Zielgröße aufzuweisen, da der Korrelationskoeffizient nach Pearson in diesen Fällen über 0.7 liegt. Das bedeutet, dass die Zielgröße ansteigt, wenn die entsprechenden Inputvariablen vergrößert werden - und umgekehrt. Auffällig ist hierbei jedoch, dass diese drei Inputgrößen auch untereinander linear voneinander abzuhängen scheinen.

Abbildung 3.5 liefert zusätzlich einen Überblick über die linearen Zusammenhänge zwischen der Zielgröße `Q11_max_h2` und allen möglichen Inputvariablen. Korrelationskoeffizienten, die betragsmäßig über einem Wert von 0.7 liegen, wurden zusätzlich als solche gekennzeichnet und auf die Anführung der einzelnen Größen verzichtet.

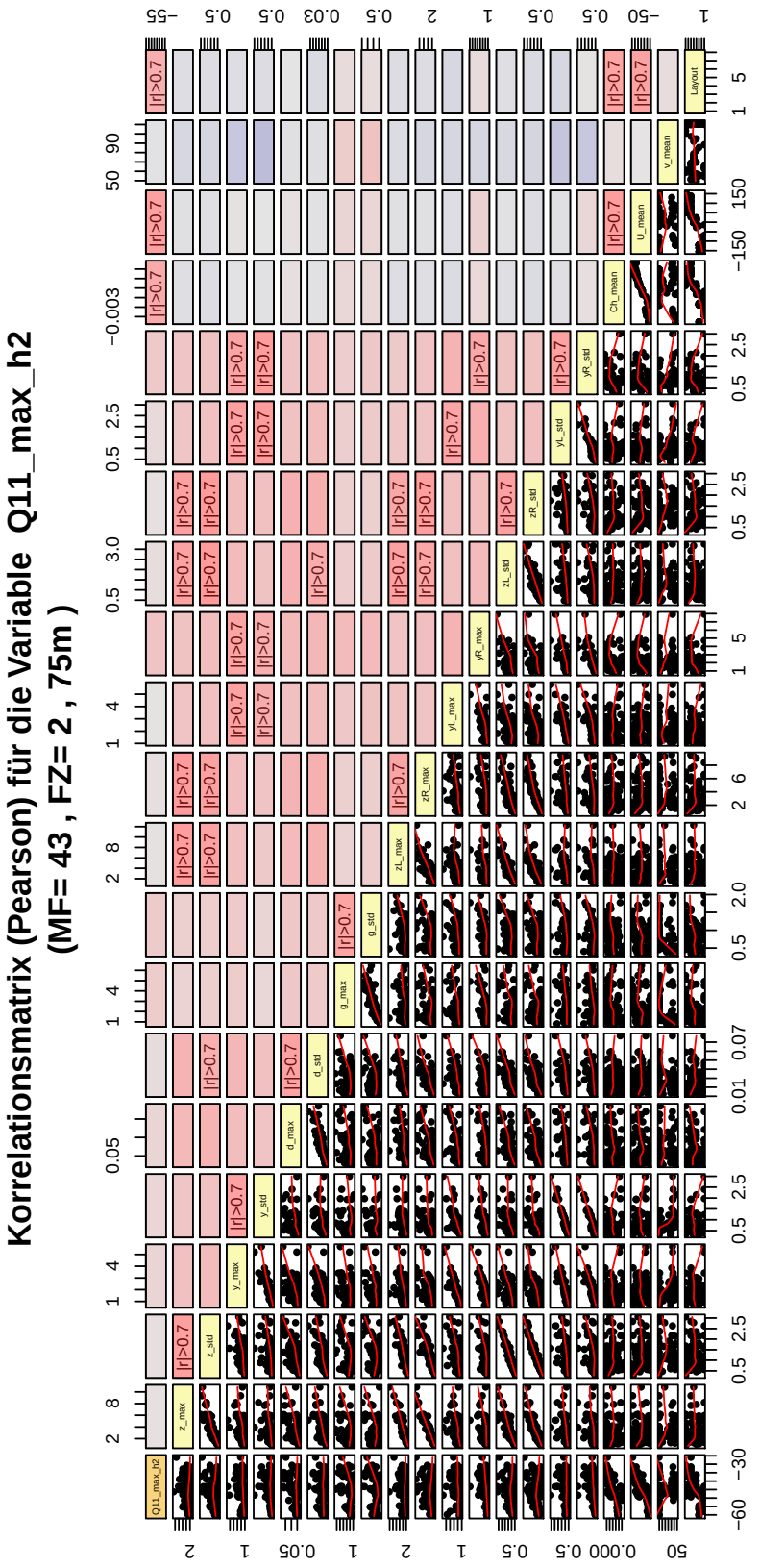

Abbildung 3.5: Korrelationsmatrix mit Scatterplots für die Variable `Q11_max_h2`, Messfahrt 43, Reisezugwagen

3.2 Kreuzkorrelation

Da es möglich ist, dass die Zielgröße zeitverzögert durch eine erklärende Variable beeinflusst wird, ist es sinnvoll, die Kreuzkorrelation zu betrachten, die genau dies aufdeckt. Eine formale Definition lässt sich beispielsweise in [Wei, 1990, S. 296] finden:

Die theoretische **Kreuzkorrelationsfunktion** (CCF) für zwei Zeitreihen $\{X_t\}_{t\in\mathbb{Z}}$ und $\{Y_t\}_{t\in\mathbb{Z}}$ lautet:

$$\rho_{xy}(k) = \frac{\gamma_{xy}(k)}{\sigma_x \sigma_y}, \tag{3.4}$$

für $k = 0, \pm 1, \pm 2, \ldots$, wobei σ_x und σ_y die Standardabweichungen von X_t und Y_t bezeichnen und $\gamma_{xy}(k)$ für die Kreuzkovarianzfunktion steht. Diese ist definiert durch:

$$\gamma_{xy}(k) = \mathbb{E}[(X_t - \mu_X)][(Y_{t+k} - \mu_Y)], \tag{3.5}$$

mit den Erwartungswerten $\mu_X = \mathbb{E}(X_t)$ und $\mu_Y = \mathbb{E}(Y_t)$.

Die empirische Kreuzkorrelationsfunktion für die Daten $(x_1, \ldots, x_n)$ und $(y_1, \ldots y_n)$ dient als Schätzung für $\rho_{xy}(k)$ und ist definiert als

$$\hat{\rho}_{xy}(k) = \frac{\hat{\gamma}_{xy}(k)}{\hat{\sigma}_x \hat{\sigma}_y}, \tag{3.6}$$

mit

$$\hat{\gamma}_{xy}(k) = \begin{cases} \frac{1}{n}\sum_{t=1}^{n-k}(x_t - \overline{x})(y_{t+k} - \overline{y}), & k \geq 0, \\ \frac{1}{n}\sum_{t=1-k}^{n}(x_t - \overline{x})(y_{t+k} - \overline{y}), & k < 0, \end{cases} \tag{3.7}$$

$$\hat{\sigma}_x = \sqrt{\hat{\gamma}_{xx}(0)}, \qquad \hat{\sigma}_y = \sqrt{\hat{\gamma}_{yy}(0)}, \tag{3.8}$$

und $\overline{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$ und $\overline{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$ bezeichnen die Mittelwerte von x_t und y_t.

Die Definition eines stochastischen Prozesses (und speziell einer Zeitreihe), sowie weiterführende Informationen und theoretische Grundlagen lassen sich in [Wei, 1990] nachlesen.
Zwei ausgewählte Graphiken zur Darstellung der Kreuzkorrelation sind in Abbildung 3.6 ersichtlich. In der linken Graphik ist die größte Spitze (engl. *Peak*) bei `Lag 0` erkennbar,

sodass kein zeitlich verschobener Einfluss belegt werden kann. In der rechten Graphik liegt der höchste Wert jedoch etwa bei `Lag 48`. Dies weist darauf hin, dass die größte Korrelation zwischen der Zielgröße `Q11_max_h2` und `g_max` dann vorliegt, wenn `g_max` um 48 Zeiteinheiten verschoben wird. Dieses Ergebnis ist jedoch nicht relevant, da es keine physikalische Erklärung für diesen verschobenen Einfluss gibt. Dies würde tatsächlich bedeuten, dass die Zielgröße erst nach $48 \times 0.16 = 7.68$ Metern von `g_max` beeinflusst wird, da zwischen den Punkten 0.16 m liegen.

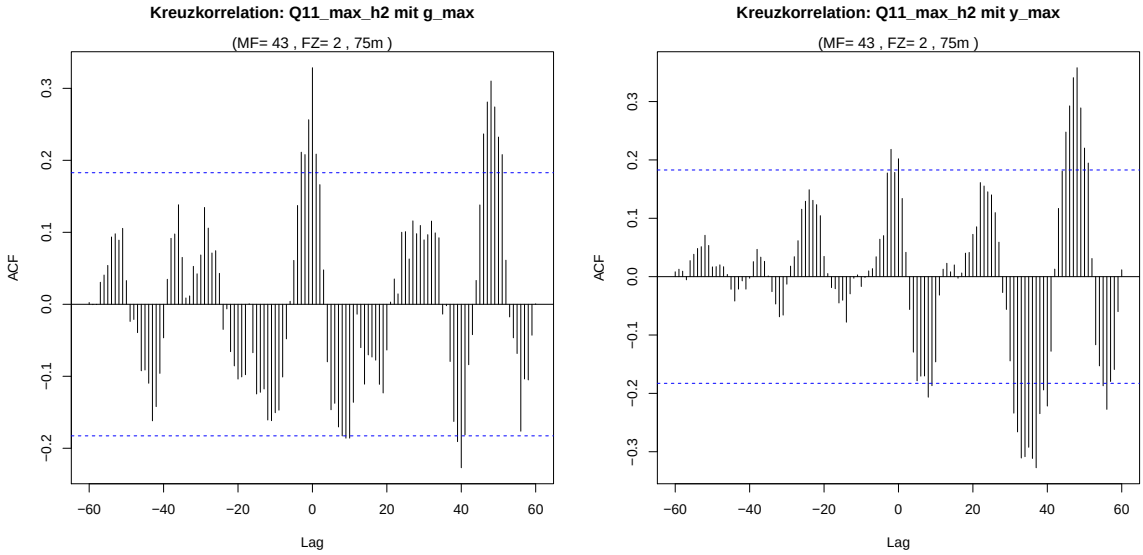

Abbildung 3.6: Kreuzkorrelation der Zielgröße `Q11_max_h2` mit `g_max` bzw. `y_max`

Wichtiger ist jedoch die Betrachtung von nur wenigen Lags wie etwa in Abbildung 3.7.

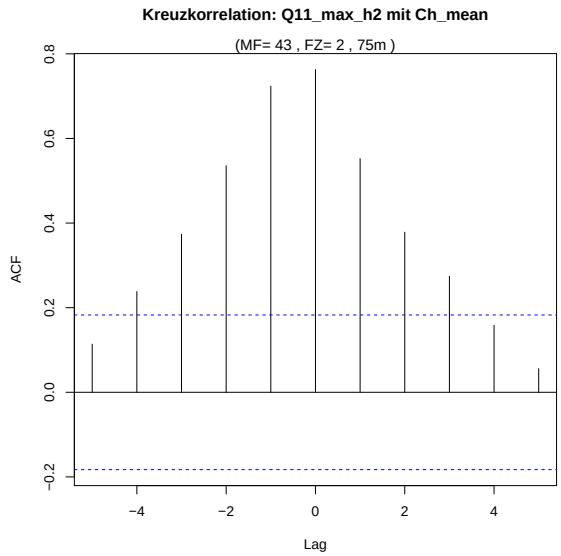

Abbildung 3.7: Kreuzkorrelation der Zielgröße `Q11_max_h2` mit `Ch_mean` bis Lag 5

So kann untersucht werden, ob es Inputgrößen gibt, die mit einer **kurzen** Verzögerung (< 1 Meter) Einfluss auf die Zielgröße ausüben. Anhand der entsprechenden Graphiken lassen sich jedoch keine relevanten zeitlichen Verschiebungen für dieses Beispiel feststellen.

3.3 Modellselektion

Um ein lineares Regressionsmodell aufzustellen gibt es mehrere Vorgehensweisen. Bevor diese beschrieben werden, muss zuerst noch eine mathematisch korrekte Definition für das klassische lineare Regressionsmodell gefunden werden. In [Fahrmeir *et al.*, 2009, S. 29] wird dieses wie folgt beschrieben:

Klassisches lineares Regressionsmodell:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i, \qquad i = 1, \ldots, n, \tag{3.9}$$

mit der metrischen Zufallsvariable Y und den metrischen oder binär kodierten kategorialen Regressoren $x_1, \ldots, x_k$. Die Fehler $\epsilon_1, \ldots, \epsilon_n$ sind unabhängig und identisch verteilt mit

$$\mathbb{E}(\epsilon_i) = 0, \quad \mathrm{Var}(\epsilon_i) = \sigma^2. \tag{3.10}$$

Eine alternative Modellschreibweise lautet

$$Y = X\beta + \epsilon \tag{3.11}$$

mit der Designmatrix $X_{n \times k+1}$, dem p-dimensionalen Parametervektor $\beta = (\beta_0, \ldots, \beta_k)^T$ und dem n-dimensionalen Fehlervektor $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$.

Die Schätzung $\hat{\beta}$ für den Parametervektor β lässt sich mithilfe der Methode der kleinsten Quadrate berechnen. Dafür wird die Fehlerquadratsumme

$$SSE(\beta) = (Y - X\beta)^T (Y - X\beta) \tag{3.12}$$

bzgl. β minimiert. Dadurch erhalten wir schlussendlich $\hat{\beta} = (X^T X)^{-1} X^T Y$.

Unter der Annahme der Normalverteilung der Response

$$Y \sim N(X\beta, \sigma^2 I_n) \tag{3.13}$$

mit der $n \times n$ Einheitsmatrix I_n, gilt für den Schätzer

$$\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1}). \tag{3.14}$$

Um ein passendes Regressionsmodell $\hat{Y} = X\hat{\beta}$ unter allen Modellen mit Linearkombinationen aller Teilmengen der Regressoren $(x_1, \ldots, x_k)$ zu finden, könnte man alle möglichen Modelle berechnen und sich schlussendlich für eines davon entscheiden. Da dieser Aufwand aufgrund der Anzahl der Teilmengen $\left(\sum_{i=0}^{k} \binom{k}{i} = 2^k\right)$ in den meisten Fällen enorm wäre und in keiner Relation zum Nutzen steht, ist es deutlich empfehlenswerter, eine automatische Modellselektion mithilfe eines Statistikprogramms durchzuführen. Dafür gibt es unter anderem folgende Möglichkeiten, wie auch in [Fahrmeir *et al.*, 2009, S. 163 ff] beschrieben wird:

1. *Vorwärts-Selektion*
   Hier wird normalerweise mit einem Modell gestartet, welches nur den Intercept $\hat{\beta}_0$ enthält (Intercept-only-Modell). In jedem Schritt wird nun sukzessive eine weitere x-Variable in das Modell aufgenommen und zwar jene, die die größte Reduktion des gewählten Informationskriteriums nach sich zieht. Das Verfahren wird beendet, falls sich das Informationskriterium durch die Hinzunahme einer Variable nicht mehr verkleinern lässt.

2. *Rückwärts-Selektion*
   Bei dieser Methode wird von einem vollen Modell mit allen x-Variablen ausgegangen. Schritt für Schritt wird nun jene x-Variable aus dem Modell eliminiert, welche durch ihr Entfernen die größte Reduktion des Informationskriteriums bewirkt.

3. *Schrittweise-Selektion*
   Eine Kombination aus Vorwärts-Selektion und Rückwärts-Selektion liefert die Schrittweise-Selektion. Dabei wird in jedem Schritt entweder eine x-Variable hinzugenommen, oder aber auch eine entfernt.

Da die Umsetzung aller drei Methoden in R kein Problem darstellt, wurde in dieser Arbeit die aufwendigere Schrittweise-Selektion verwendet.

Zuvor muss jedoch ein geeignetes Informationskriterium ausgewählt werden, dessen Minimierung zum Modellvorschlag führt. Zwei der bekanntesten Informationskriterien sind in [Fahrmeir *et al.*, 2009, S. 161–162] wie folgt definiert:

Akaike Informationskriterium (AIC)

$$AIC = -2 \cdot l(\hat{\beta}, \hat{\sigma}^2) + 2|k+1|, \tag{3.15}$$

wobei $l(\hat{\beta}, \hat{\sigma}^2)$ der maximale Wert der Log-Likelihood ist, d.h. wenn in die Log-Likelihood die Maximum-Likelihood-Schätzer $\hat{\beta}$ und $\hat{\sigma}^2$ eingesetzt werden. k bezeichnet hierbei die Anzahl der Variablen im Modell, sodass $|k+1|$ die Gesamtanzahl der zu schätzenden Parameter (inkl. Intercept) darstellt.

Bayes'sche Informationskriterium (BIC)

$$BIC = -2 \cdot l(\hat{\beta}, \hat{\sigma}^2) + \log(n)|k|, \tag{3.16}$$

wobei $l(\hat{\beta}, \hat{\sigma}^2)$ wie zuvor den maximalen Wert der Log-Likelihood beschreibt und auch die Maximum-Likelihood-Schätzer sind dieselben wie beim AIC. Der Strafterm $\log(n)$ beinhaltet jedoch nun auch den Stichprobenumfang n.

Obwohl sich die beiden Informationskriterien sehr ähneln, sind sie dennoch unterschiedlich motiviert, wie in [Fahrmeir *et al.*, 2009, Ab. 3.6] nachgelesen werden kann.
Der offensichtlichste formale Unterschied zwischen AIC und BIC liegt in der Bestrafung der Variablenanzahl. Während die Anzahl der Variablen beim AIC mit der festen Zahl 2 bestraft wird, verwendet BIC eine Variante, die von der Anzahl der Beobachtungen n abhängt. AIC als Gütekriterium wird daher größere Modelle vorschlagen, während das BIC eher Modelle mit weniger Variablen bevorzugt.
Da für diese Arbeit ein möglichst einfaches lineares Modell benötigt wird, wird hier das Bayes'sche Informationskriterium verwendet. Die Umsetzung in R benötigt nur folgenden kurzen Code:

Auflistung 3.1: Modellselektion in R

```
1 mod_full <- lm(Zielgröße ~ ., data=Datensatz)
2 mod_bic <- step(mod_full, direction = "both", k=log(length(Zielgröße)))
```

Dabei bezeichnet `mod_full` ein Modell, in dem alle möglichen Variablen enthalten sind. Mithilfe der Funktion `step` wird die Modellselektion durchgeführt. Durch die Definition von `direction = "both"` wird die Schrittweise-Selektion verwendet. Weiters

legt `k=log(length(Zielgröße))` das BIC als Informationskriterium fest.
Für den betrachteten Fall (Reisezugwagen für Messfahrt 43) wird für die Zielgröße `Q11_max_h2` folgendes Modell vorgeschlagen:

$$Q11\_max\_h2 \sim z\_std + y\_std + zR\_max + yL\_max + zL\_std + zR\_std + yL\_std + yR\_std + Ch\_mean + U\_mean.$$

Um mehr über das Modell zu erfahren, wird in R üblicherweise die `summary` betrachtet, wie Auflistung 3.2 zeigt.
Wie an den p-Werten und „*“-Markierungen ersichtlich, sind alle Variablen in diesem Modell statistisch signifikant. Genau so, wie es durch die Modellselektion sein sollte.

Auflistung 3.2: `summary` des Modells in R

```
> summary(mod_bic)
Call:
lm(formula = Zielgröße ~ z_std + y_std + zR_max + yL_max + zL_std +
    zR_std + yL_std + yR_std + Ch_mean + U_mean, data = Datensatz)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0704 -1.2771  0.1116  1.3934  7.5186

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.996e+01  8.502e-01 -58.767  < 2e-16 ***
z_std       -2.495e+01  9.914e+00  -2.516 0.013387 *
y_std       -1.241e+01  3.593e+00  -3.454 0.000801 ***
zR_max      -7.577e-01  3.320e-01  -2.282 0.024504 *
yL_max      -1.658e+00  5.886e-01  -2.816 0.005812 **
zL_std       1.192e+01  4.953e+00   2.407 0.017832 *
zR_std       1.650e+01  5.434e+00   3.037 0.003020 **
yL_std       9.080e+00  3.038e+00   2.989 0.003494 **
yR_std       8.053e+00  2.324e+00   3.464 0.000773 ***
Ch_mean     -1.907e+03  5.172e+02  -3.687 0.000363 ***
U_mean       9.820e-02  9.053e-03  10.848  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.877 on 104 degrees of freedom
Multiple R-squared:  0.8748,    Adjusted R-squared:  0.8628
F-statistic: 72.67 on 10 and 104 DF,  p-value: < 2.2e-16
```

Die `summary` gibt außerdem das Bestimmtheitsmaß für das Modell an. In [Ugarte *et al.*, 2008, S. 586–587] wird dieses Maß R^2 beschrieben als eines, das den Anteil der Variabilität der Zielgröße Y angibt, der durch das lineare Regressionsmodell erklärt werden kann:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}, \tag{3.17}$$

mit

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \ldots\ldots \text{Variation der Modellschätzungen}$$
$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \ldots\ldots \text{Variation der Residuen}$$
$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2 \ldots\ldots \text{Variation der Zielgröße}$$

wobei $SST = SSR + SSE$ gilt.

Durch Hinzufügen weiterer Modellvariablen x wird das Bestimmtheitsmaß R^2 jedoch immer größer, unabhängig davon, ob die neuen Variablen zur Beschreibung der Zielgröße tatsächlich relevant sind. Dies liegt daran, dass SSE nicht kleiner werden kann, wenn weitere Variablen in das Modell aufgenommen werden und SST konstant ist für die betrachtete Zielgröße Y. Daher wird eine Korrektur benötigt, die die Anzahl der Modellvariablen berücksichtigt und nicht automatisch das volle Modell mit allen möglichen Variablen bevorzugt.

Das geeignete Maß dafür ist das **adjustierte** R^2:

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-p}}{\frac{SST}{n-1}} = 1 - \frac{MSE}{MST}, \tag{3.18}$$

mit der Anzahl der Beobachtungen n und der Anzahl der unabhängigen Variablen p.

Aufgrund der Division durch den jeweiligen Freiheitsgrad der Quadratsummen wird die Bestrafung von zu komplexen Modellen ermöglicht. Daher wird in dieser Arbeit der Fokus auf das adjustierte R^2 gelegt.
Für dieses Beispiel wurde ein adjustierte R^2 von 0.8628 berechnet. Dies gibt an, dass knapp 90% der Varianz der Zielgröße durch das Modell beschrieben werden kann. Dieses Modell scheint demnach auf den ersten Blick gut geeignet zu sein, um die Zielgröße `Q11_max_h2` zu beschreiben. Dennoch muss zweifellos geprüft werden, ob die entsprechenden Modellannahmen erfüllt sind, bevor dieses Modell verwendet wird.

3.4 Überprüfung der Modellannahmen

Für das klassische lineare Regressionsmodell müssen einige Annahmen erfüllt sein, wie bereits in (3.10) beschrieben. Zusätzlich wird noch angenommen, dass die Störgrößen normalverteilt sind, womit folgt:

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2). \tag{3.19}$$

Da die Störgrößen ϵ_i nicht beobachtbar sind, werden stattdessen ihre Schätzer, die sogenannten Residuen $\hat{\epsilon}_i = y_i - \hat{y}_i$, verwendet, um die Modellannahmen zu prüfen. Dabei bezeichnet $\hat{y}_i = \widehat{\mathbb{E}(y_i)} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}$ mit $i = 1, \ldots, n$ den Schätzwert von y_i. Folgen die nicht-beobachtbaren Fehler ϵ_i einer Normalverteilung, so muss dies auch für die Schätzungen $\hat{\epsilon}_i$ gelten.
Mithilfe von Residuenplots werden die Annahmen graphisch überprüft. Zusätzlich können Hypothesentests zur Unterstützung verwendet werden.
Um die ersten beiden Annahmen, Erwartungswert gleich Null und homoskedastische Varianzen, zu kontrollieren, werden die durch das Modell geschätzten Werte gegen die Residuen geplottet (siehe linke Graphik in Abbildung 3.8).

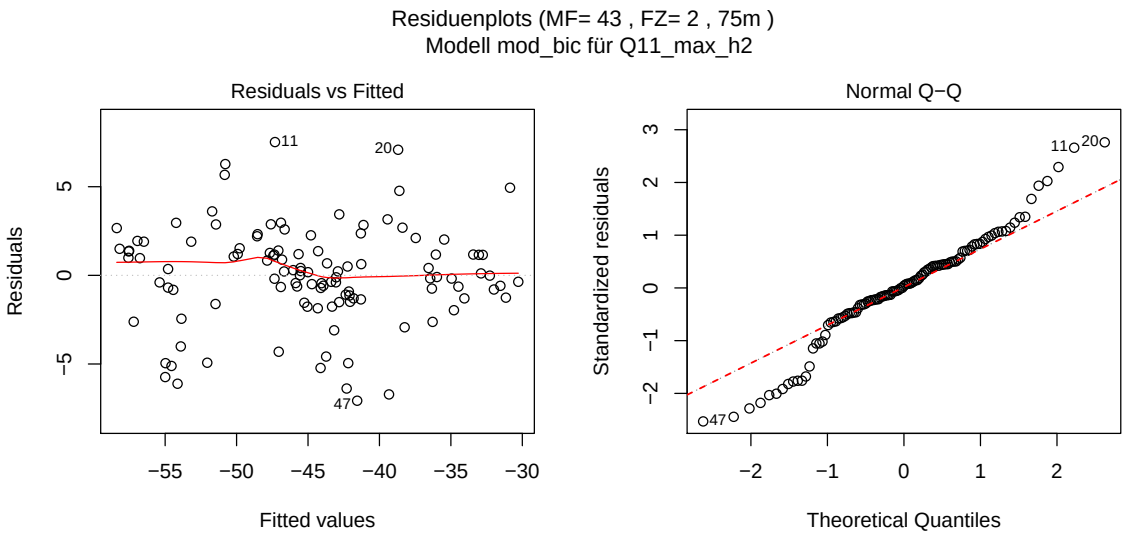

Abbildung 3.8: Residuenplots zur Überprüfung der Varianzhomogenität und Normalverteilungsannahme

Die Residuen weisen keine offensichtliche Struktur auf und scheinen annähernd zufällig um 0 zu schwanken, ohne ein sichtbares Muster zu bilden. Damit können die ersten beiden Modellannahmen als erfüllt angesehen werden. Zur Prüfung der Normalverteilung wird ein Quantil-Quantil-Plot (QQ-Plot) verwendet, in welchem die theoretischen

Normalverteilungsquantile gegen die geordneten standardisierten Residuen aufgetragen werden.
Die dafür notwendige Definition der Quantile lässt sich zum Beispiel in [Fahrmeir *et al.*, 2010, S. 287] finden:

p-Quantil

Für $0 < p < 1$ ist das *p-Quantil* x_p die Zahl auf der x-Achse, für die

$$F(x_p) = p \tag{3.20}$$

gilt. Für streng monotone Verteilungsfunktionen $F(x)$ sind die p-Quantile eindeutig bestimmt. Der Median x_{med} entspricht dem 0.5-Quantil $x_{0.5}$.

Die standardisierten Residuen ergeben sich aus dem Quotienten der Residuen und ihrer geschätzten Standardabweichung (vgl. [Fahrmeir *et al.*, 2009, S. 110]):

Standardisierte Residuen

Die standardisierten Residuen sind definiert durch

$$r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}, \tag{3.21}$$

wobei h_{ii} das i-te Diagonalelement der Prädiktionsmatrix $H = X(X^T X)^{-1} X^T$ bezeichnet.

Der Normal-Q-Q-Plot in Abbildung 3.8 deutet auf eine gute Anpassung an die Normalverteilung im Zentrum hin, da die Punkte sehr nahe an der Referenzlinie liegen. Wären die Residuen jedoch tatsächlich normalverteilt, so dürften die Ränder (Tails) der Verteilung nicht so stark von der Referenz abweichen, wie es hier ersichtlich ist. Die Verletzung der Normalverteilungsannahme kann zusätzlich noch mit einem Hypothesentest bestätigt werden. Einer der gängigsten Tests dafür ist der Shapiro-Wilk-Test. Dieser Signifikanztest prüft, ob die Stichprobe aus einer normalverteilten Population stammt. Für die Definition der Teststatistik muss zuvor ein anderer Begriff eingeführt werden, wie in [Friedl, 2005, S. 8] angeführt:

Bezeichne $g : \mathbb{R}^n \to \mathbb{R}^n$ die Abbildung $g(x_1, \ldots, x_n) = (x_{(1)}, \ldots, x_{(n)})$ mit $x_{(1)} \leq \ldots \leq x_{(n)}$. Dann heißt $x_{(\cdot)} = (x_{(1)}, \ldots, x_{(n)})$ **geordnete Stichprobe** zu $x = (x_1, \ldots, x_n)$,

$X_{(\cdot)} = (X_{(1)}, \ldots, X_{(n)})$ die geordnete Statistik (**Ordnungsstatistik**) und $X_{(i)}$ die i-te geordnete Statistik.
Damit lässt sich die Teststatistik des Shapiro-Wilk Tests nach [Friedl, 2005, S. 26] wie folgt definieren:

$$W = \frac{[\sum_{i=1}^{n} a_i X_{(i)}]^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2,} \tag{3.22}$$

mit Koeffizienten a_i, welche in Form von Tabellen vorliegen oder approximiert werden.

In der Teststatistik wird der Schätzer der Varianz unter der Normalverteilung mit jener Varianz verglichen, die aus der Stichprobe berechnet wird. Stammt die Stichprobe tatsächlich aus einer Normalverteilung, so wird dieser Quotient nahe bei 1 liegen. Für starke Abweichungen hingegen nahe bei 0. Die Teststatistik wird mit einem kritischen Wert verglichen, welcher vom Stichprobenumfang abhängt und in eigenen Tabellen vorliegt. Ist die Teststatistik dabei signifikant kleiner als der kritische Wert, so wird die Nullhypothese, und somit die Annahme der Normalverteilung, verworfen und die Stichprobe gilt nicht als normalverteilt.
Eine alternative Interpretation verwendet den p-Wert, welcher in R automatisch berechnet wird. Der p-Wert gibt hier die Wahrscheinlichkeit an, dass die Stichprobe gezogen wird, falls sie aus einer Normalverteilung stammt. Ist dieser Wert kleiner als ein gewähltes α (häufig 5%), so wird die Nullhypothese verworfen.

Für das ausgewählte Beispiel ergibt sich mit R folgender Output:

Auflistung 3.3: Shapiro-Wilk Test in R

```
1 > shapiro.test(mod_bic$residuals)
2
3         Shapiro-Wilk normality test
4
5 data:  mod_bic$residuals
6 W = 0.9676, p-value = 0.00686
```

Mit einem kleinen p-Wert von 0.00686 verwirft der Test die Annahme, dass die betrachteten Residuen aus einer Normalverteilung stammen.

Um die Verteilung der Residuen besser beurteilen zu können, werden graphische Hilfsmittel verwendet. Der Boxplot der Residuen (Abbildung 3.9) lässt einige Ausreißer

erkennen, allerdings lässt sich auf den ersten Blick keine signifikante Schiefe der Verteilung feststellen.

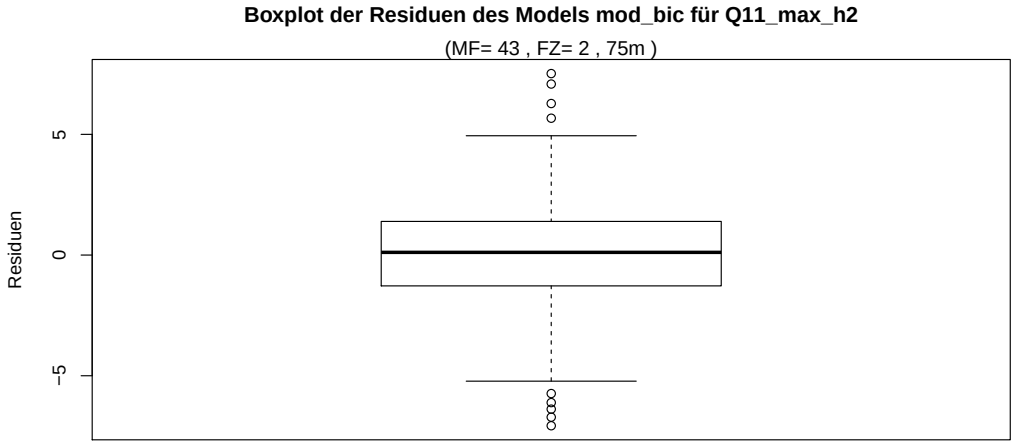

Abbildung 3.9: Boxplot der Residuen

Um einen detaillierten Eindruck über die Verteilung der Residuen zu erhalten, bietet sich ein Histogramm mit geschätzter Dichtefunktion an, wie in Abbildung 3.10 visualisiert wird. Die zwei Gipfel in der Verteilung bestätigen dabei die Abweichung von der Normalverteilung.

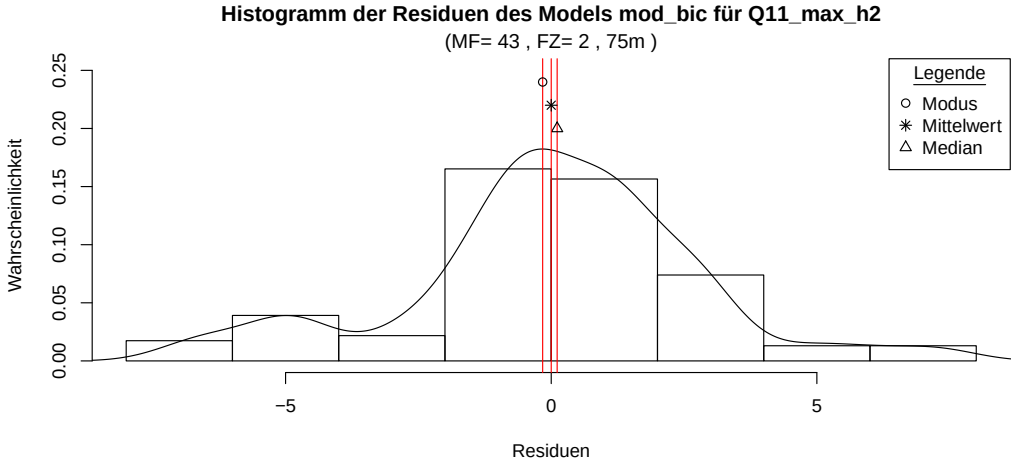

Abbildung 3.10: Histogramm der Residuen

Mit der Berechnung von Modus (= `argmax` der Verteilungsdichte), Mittelwert und Median der Verteilung, liefert die Lageregel, wie in [Fahrmeir *et al.*, 2010, S. 292] beschrieben, eine „Faustregel“, nach der die Verteilung kategorisiert werden kann:

Lageregel

Symmetrische unimodale Verteilung:	$x_{mod} = x_{med} = \mathbb{E}(X)$
Rechtsschiefe Verteilung:	$x_{mod} < x_{med} < \mathbb{E}(X)$
Linksschiefe Verteilung:	$x_{mod} > x_{med} > \mathbb{E}(X)$

mit Modus x_{mod}, Median x_{med} und Erwartungswert $\mathbb{E}(X)$.

Für diesen Fall ergibt sich unter der Verwendung des Mittelwertes $\hat{x}_{mean}$ als Schätzer für $\mathbb{E}(X)$:

$$\begin{aligned} \hat{x}_{mod} &= -0.16 \\ \hat{x}_{med} &= \quad 0.11 \\ \hat{x}_{mean} &= \quad 6.67e-17 \end{aligned}$$

und somit

$$\hat{x}_{mod} < \hat{x}_{mean} < \hat{x}_{med},$$

wodurch die Lageregel nicht anwendbar ist. Dies liegt daran, dass hier keine 1-gipflige Verteilung vorliegt.

Die Verletzung der Normalverteilungsannahme sollte in weiterer Folge berücksichtigt und auf robuste, nichtparametrische Methoden ausgewichen werden.

3.5 Zusammenfassung der Modelle

Anschließend wurde für jede der acht Zielgrößen für jede der neun Fahrzeug/Messfahrt-Kombinationen ein individuelles Modell generiert (also insgesamt $8 \times 9 = 72$ Modelle). Diese werden für die Zielgröße `Q11_max_h2` in Tabelle 3.1 dargestellt und können somit leicht untereinander verglichen werden.
Dabei fällt auf, dass die Modelle in den meisten Fällen aus mehr als drei Variablen bestehen, aber durchaus gewisse Ähnlichkeiten aufweisen. Das adjustierte R^2 soll zusätzlich aufzeigen, wie gut die berechneten Modelle tatsächlich sind. Dadurch lässt sich erkennen, dass die Regressionsmodelle für Messfahrt 42 nicht so passend zu sein scheinen, wie dies für die anderen beiden Messfahrten der Fall ist.
Erneut soll hier darauf hingewiesen werden, dass die Diagnoseplots in den meisten Fällen eine schiefe Verteilung der Residuen aufgedeckt haben. Diese Problematik muss in weiterer Folge auf jeden Fall berücksichtigt werden.

Tabelle 3.1: Zusammenfassung der Parameter der BIC- Modelle für die Zielgröße `Q11_max_h2`

Messfahrt → Fahrzeug ↓	MF 041	MF 042	MF 043
I (Lok)	(z_max, z_std, zL_std, zR_std, Ch_mean, U_mean, v_mean) $R^2_{adj} = 0.9151$ $SE = 5.635$	(z_std, y_std, g_std, zL_max, zR_std, Ch_mean) $R^2_{adj} = 0.5552$ $SE = 9.936$	(z_max, y_std, zL_std, zR_std, U_mean, v_mean) $R^2_{adj} = 0.8966$ $SE = 5.954$
II (RZW)	(y_max, yL_std, U_mean) $R^2_{adj} = 0.6926$ $SE = 4.201$	(U_mean) $R^2_{adj} = 0.1334$ $SE = 5.154$	(z_std, y_std, zR_max, yL_max, zL_std, zR_std, yL_std, yR_std, Ch_mean, U_mean) $R^2_{adj} = 0.8628$ $SE = 2.877$
V (GW)	(zL_std, U_mean, v_mean, zL_max) $R^2_{adj} = 0.7377$ $SE = 2.168$	(y_std, zR_max, zR_std, U_mean, v_mean) $R^2_{adj} = 0.6029$ $SE = 2.361$	(y_std, g_max, g_std, zL_std, U_mean, v_mean) $R^2_{adj} = 0.7262$ $SE = 2.155$

4 Identifikation der Signalwerte

Hat man ein Regressionsmodell wie in Kapitel 3 gefunden, werden nun jene Datenpunkte identifiziert, die vom Modell nicht gut genug geschätzt werden. Dafür werden die Residuen bzw. die standardisierten Residuen betrachtet. Anschließend wird in Kapitel 5 analysiert, aus welchen Gründen diese Werte als auffällig eingestuft wurden.
Die Definitionen für den statistischen Begriff „Ausreißer“ unterscheiden sich in der Literatur, bezeichnen jedoch immer einen Datenpunkt, der in gewisser Art und Weise auffällig wirkt. In [Hawkins, 1980, S. 1] findet man folgende Erklärung:

> The intuitive definition of an outlier would be 'an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism'. An inspection of a sample containing outliers would show up such characteristics as large gaps between 'outlying' and 'inlying' observations and the deviation between the outliers and the group of inliers, as measured on some suitable standardized scale.

Es werden zuerst ein allgemeiner, sowie ein parametrischer Ansatz zur Identifikation von extremen Werten vorgestellt, wobei letzterer jedoch nur unter Annahme der Normalverteilung gültig ist. Aus diesem Grund werden noch zwei weitere Methoden beschrieben, welche ohne direkte Verteilungsannahmen auskommen.

4.1 Allgemeiner Ansatz

Ein allgemeiner Ansatz definiert jene Datenpunkte als auffällige Werte, die außerhalb einer bestimmten Anzahl an Standardabweichungen vom Mittelwert entfernt liegen. Dieser Ansatz (vgl. [Czitrom und Spagon, 1987, S. 301–302]) setzt dabei keine Normalverteilung voraus, sondern gilt auch approximativ, falls sich die Häufigkeitsverteilung durch eine Normalverteilung approximieren lässt, und lautet formal:

$$[\overline{x} - k\hat{\sigma}_x \quad ; \quad \overline{x} + k\hat{\sigma}_x]\,, \tag{4.1}$$

mit einer Konstanten k und der geschätzten Standardabweichung $\hat{\sigma}_x$ von x.

Die Konstante k wird dabei üblicherweise auf 2 oder auch 3 gesetzt ([Czitrom und Spagon, 1987, S. 302]). Letzteres wird oft auch als „3-Sigma-Regel“ bezeichnet.
Diese Methode ist zwar einfach umzusetzen, ist jedoch in den meisten Fällen zu ungenau.

4.2 Parametrischer Ansatz

Unter der Normalverteilungsannahme können extreme Werte mithilfe der Normalverteilungsquantile identifiziert werden, wie etwa in [Behnke, 2006, S. 300]. Dafür muss zu Beginn festgelegt werden, welcher Anteil von Datenpunkten erwartet wird.
$(1-\alpha)$ % der Daten $(x_1, \ldots, x_n)$ liegen dann innerhalb folgender Grenzen:

$$\left[\overline{x} - z_{1-\frac{\alpha}{2}}\hat{\sigma}_x \quad ; \quad \overline{x} + z_{1-\frac{\alpha}{2}}\hat{\sigma}_x\right], \tag{4.2}$$

mit dem Mittelwert $\overline{x}$ als Schätzer für den Erwartungswert, dem Normalverteilungsquantil $z_{1-\frac{\alpha}{2}}$ und dem Schätzer der Standardabweichung $\hat{\sigma}_x$.

Werte, die außerhalb dieses Konfidenzintervalls liegen, werden als Signalwerte bezeichnet. Da die Zielgrößen aus den betrachteten Datensätzen jedoch keiner Normalverteilung folgen, sollte auf nichtparametrische Methoden ausgewichen werden, wie sie in Abschnitt 4.3 und 4.4 beschrieben werden.

Die Tatsache, dass die Verteilung der (standardisierten) Residuen nicht nur von der (Standard-) Normalverteilung abweicht, sondern auch allgemein von einer symmetrischen Verteilung, lässt sich zudem rechnerisch nachweisen. Unter einer symmetrischen Verteilung müsste zum Beispiel das 95% Quantil mit dem Negativen des 5% Quantil übereinstimmen, da unter diesen Umständen beide gleich weit vom Erwartungswert entfernt sind.
Für die Zielgröße `Q11_max_h2` (Reisezugwagen, Messfahrt 43) ergibt sich jedoch (gerundet auf zwei Dezimalstellen):

$$q_{0.95} = 1.45$$
$$q_{0.05} = -1.94$$

wobei für die Quantile der Standardnormalverteilung folgt:

$$z_{0.95} = 1.64$$
$$z_{0.05} = -1.64.$$

Das bestätigt die fehlende Symmetrie der Residuenverteilung bzw. in unserem Fall eine linksschiefe Verteilung.

4.3 Quantil-Signalwerte

Um nicht weiterhin fälschlicherweise die Normalverteilung annehmen zu müssen, werden die empirischen Quantile der Residuen für die Grenzen verwendet.
In dieser Arbeit wird $\alpha = 10\%$ benützt, sodass die 95% und 5% Quantile betrachtet werden. Damit folgt für die Grenzen:

$$\left[\overline{r} - q_{\frac{\alpha}{2}}\sigma_{r_i} \quad ; \quad \overline{r} + q_{1-\frac{\alpha}{2}}\sigma_{r_i}\right], \tag{4.3}$$

mit den standardisierten Residuen r_i, dem Mittelwert dieser Residuen $\overline{r}$, den empirischen Quantilen $q_{\frac{\alpha}{2}}$ bzw. $q_{1-\frac{\alpha}{2}}$ und der Standardabweichung der Residuen σ_{r_i}.

Für den Mittelwert der Residuen gilt $\overline{r} = 0$, unabhängig davon, ob diese standardisiert oder roh betrachtet werden. Die Varianz der standardisierten Residuen berechnet sich durch $Var(r_i) = Var\left(\frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}\right) \approx 1$, wodurch für die Standardabweichung $\sigma_{r_i} \approx 1$ gilt.

Jene standardisierten Residuen, deren Wert die obere Grenze übersteigt bzw. die untere Grenze unterschreitet, zählen zu den extremsten 10% der Daten und werden als Quantil-Signalwerte (= **Q-Signalwerte**) klassifiziert.
In Abbildung 4.1 lassen sich die Q-Signalwerte für die Zielgröße `Q11_max_h2` für Messfahrt 43 mit einem Reisezugwagen erkennen. Dafür werden die durch das Modell geschätzten Werte für die Zielgröße gegen die standardisierten Residuen aufgetragen.

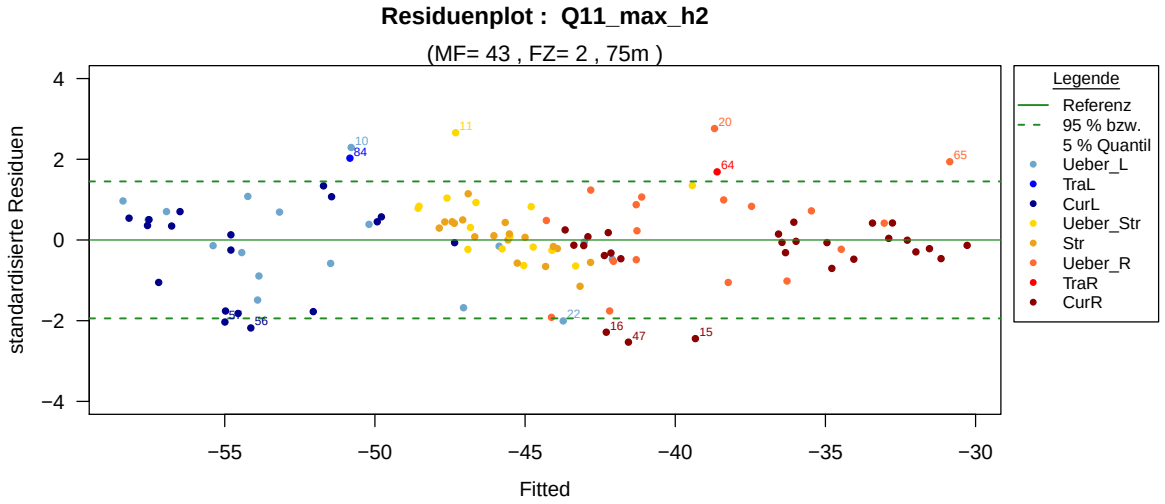

Abbildung 4.1: Identifikation der Q-Signalwerte

Zur Erinnerung: jeder Punkt stellt einen Datenabschnitt von 75 m dar in dem eine der acht Streckenlayoutkategorien auftreten. Daher werden die eingezeichneten Punkte zu-

sätzlich mit einer kategorienabhängigen Farbe versehen. Dadurch sollen Zusammenhänge zwischen den Q-Signalwerten und der Variable `Layout` sichtbar gemacht werden. Zusätzlich wurde die Indexnummer der auffälligen Residuen, und somit der dazugehörigen Abschnitte, angegeben.
In diesem Fall scheint es auf den ersten Blick keine Abhängigkeit vom Streckenlayout zu geben, da Abschnitte aus verschiedenen Kategorien als Q-Signalwerte identifiziert werden. Es lässt sich jedoch erkennen, dass oft zwei aufeinanderfolgende Abschnitte herausstechen, wie etwa die Abschnitte 15 und 16 sowie 64 und 65. Des Weiteren liegen viele standardisierte Residuen nahe der eingezeichneten Grenzen und nur wenige deutlich außerhalb.
Abbildung 4.1 identifiziert die Q-Signalwerte, sagt jedoch nicht viel über deren Position im Datensatz aus. Lediglich die Indexnummer liefert Informationen darüber. Da diese in der Abbildung jedoch nicht immer gut lesbar sind, kann Abbildung 4.2 zur Hilfe herangezogen werden.

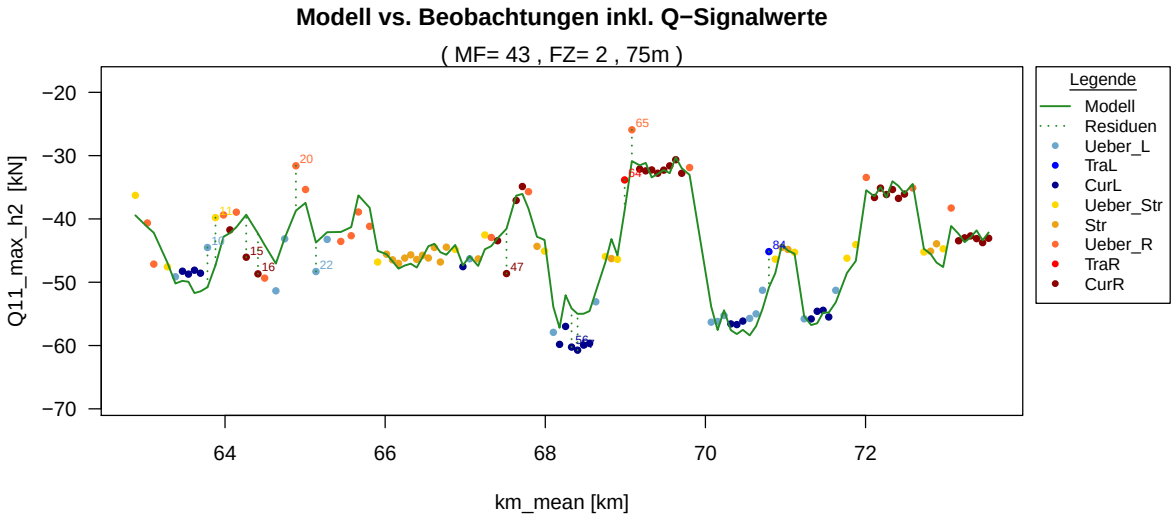

Abbildung 4.2: BIC-Modell inkl. Q-Signalwerte und deren Residuen

In Abbildung 4.2 wird der Verlauf der Zielgröße `Q11_max_h2` über den Kilometerstand `km_mean` geplottet. Die eingezeichneten Punkte die, wie zuvor definiert, 75 m Abschnitte darstellen, sind nach den Kategorien der Variable `Layout` gefärbt. Zusätzlich wird der durch das Regressionsmodell aus Abschnitt 3.3 geschätzte Verlauf der Zielgröße sichtbar gemacht (durchgehende Linie). Weiters sind die Residuen, also die Abweichungen der Modellschätzungen zu den beobachteten Werten, der Q-Signalwerte eingezeichnet (punktierte Linien).

Anhand dieser Darstellung lassen sich nicht nur die Größen der relevanten Residuen darstellen und vergleichen, sondern auch die Position der Q-Signalwerte im Datensatz wird verdeutlicht. Damit lässt sich erkennen, wo das Modell gut schätzt und wo nicht. Besonders auffällig sind in diesem Beispiel die Abschnitte mit den Nummern 20 und 65, die vom Modell unterschätzt werden. In beiden Abschnitten realisiert die Variable `Layout` in der Kategorie `Ueber_R`, wie anhand der farblichen Markierung erkennbar ist. Des Weiteren interessant ist die Schätzung nach Kilometer 68, wo einige aufeinander folgende Abschnitte in einem Linksbogen (`Layout=CurL`) vom Modell überschätzt werden. Zwei dieser Abschnitte (56 und 57) wurden dabei als Q-Signalwerte markiert.

4.3.1 Betrachtung der Position der Q-Signalwerte bzgl. Inputgrößen

Nachdem die extremsten 10 % der Daten als Q-Signalwerte identifiziert wurden, kann man sich die Frage stellen, ob diese Abschnitte nur bezüglich der Zielgröße auffällig sind. So könnte es sein, dass diese auch besonders hohe/niedrige Werte für die verschiedenen Inputgrößen liefern. Um dies herauszufinden, werden die größten/kleinsten Werte der Inputgrößen sichtbar gemacht und den Q-Signalwerten gegenüber gestellt.
In Abbildung 4.3 werden die gemessenen Werte für die Inputgrößen `y_max` und `y_std` gegen den Kilometerstand `km_mean` aufgetragen.
Die horizontale durchgehende Linie bezeichnet den Mittelwert der Inputgrößen, während die punktierten Linien die empirischen Quantile derselben darstellen. In diesem Fall wurden die 90 % und 10 % Quantile verwendet, sodass die extremsten 20 % der Daten außerhalb dieser Grenzen liegen. Der Grund dafür ist schnell erklärt: wären die Grenzen so gesetzt, dass erneut nur 10 % der Daten außerhalb liegen würden, wären das wahrscheinlich zu wenige Punkte, die mit den Q-Signalwerten verglichen werden können. Da es hierbei jedoch nur um einen Vergleich und nicht um eine erneute Definition von Signalwerten geht, sind die hier gewählten Grenzen vorteilhafter.
In der Graphik wurden die zuvor identifizierten Q-Signalwerte als rote Sternsymbole dargestellt, um sie von anderen Datenpunkten unterscheiden zu können. Dadurch können die Abschnitte 20 und 22 leicht als, bezüglich `y_max` und `y_std`, auffällige Q-Signalwerte identifiziert werden. Aber auch die Abschnitte mit den Indexnummern 56 und 57 gehören zu den extremsten 20 % der betrachteten Variablen. Es könnte demnach sein, dass ein Fehler in der Messung von `y_max` und `y_std` oder ein äußerer Einfluss zu einem extremen Messwert geführt hat. Sind diese Variablen im verwendeten Regressionsmodell enthalten, so könnte dies eine Erklärung für den Q-Signalwerte darstellen.

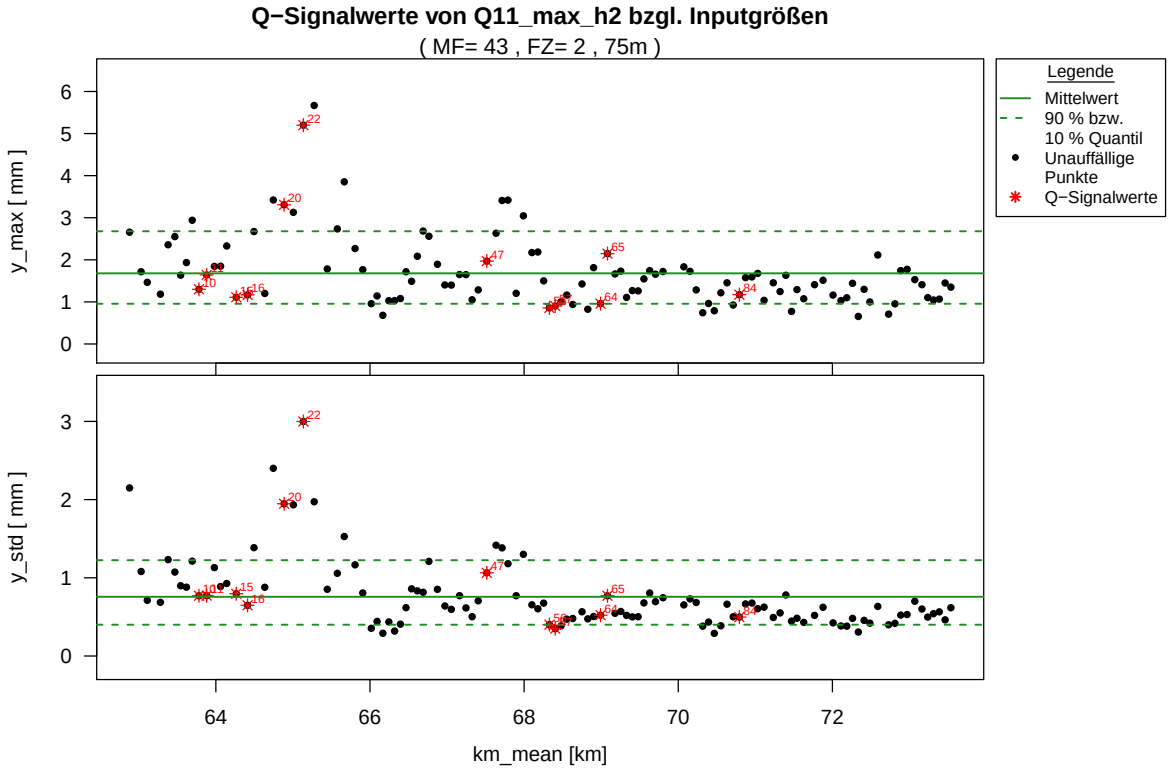

Abbildung 4.3: Inputgrößen `y_max` und `y_std` mit Quantilsgrenzen und Q-Signalwerte

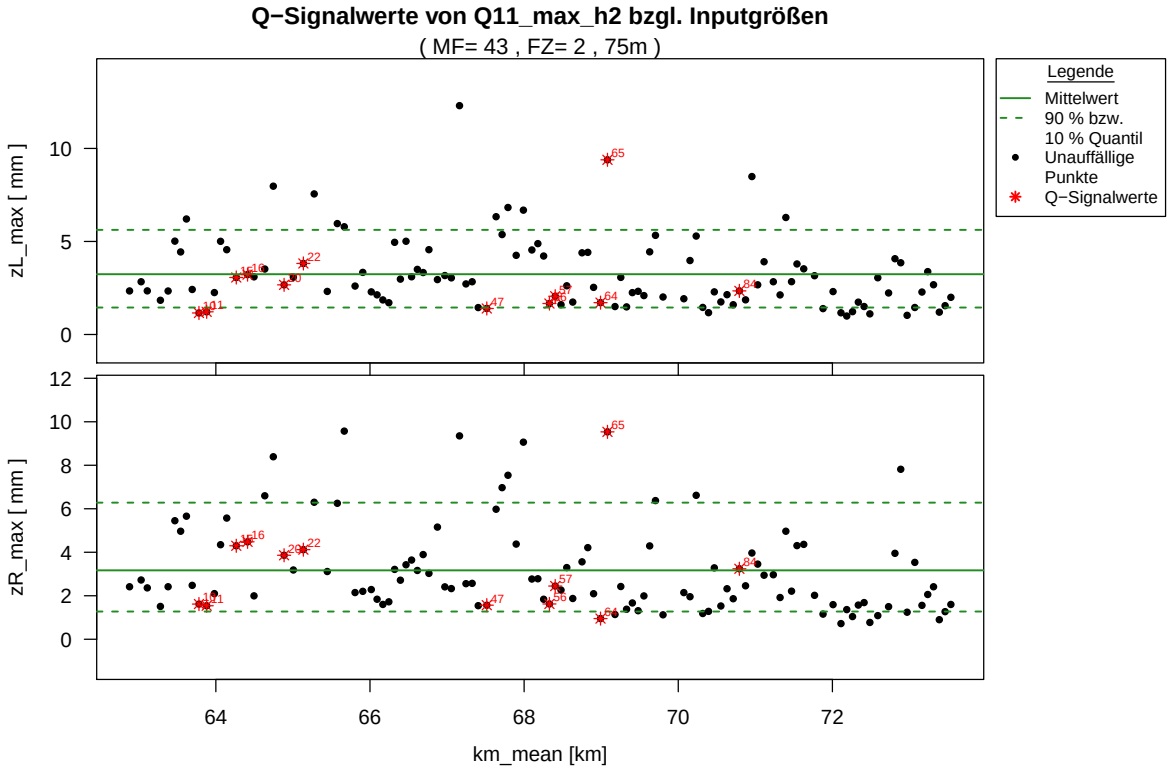

Abbildung 4.4: Inputgrößen `zL_max` und `zR_max` mit Quantilsgrenzen und Q-Signalwerte

Auch bezüglich der Inputgrößen `zL_max` und `zR_max` lassen sich Abschnitte finden, die sowohl als Q-Signalwerte markiert, als auch bezüglich der Inputgrößen auffällig sind. Abbildung 4.4 deckt Abschnitt 65 als solchen auf. Alle anderen Q-Signalwerte befinden sich jedoch innerhalb der Quantilsgrenzen dieser Inputgrößen bzw. knapp außerhalb.

Obwohl die Variable `v_mean`, welche die Geschwindigkeit beschreibt, in diesem Beispiel nicht in das Regressionsmodell aufgenommen wurde, wird der Verlauf aufgrund der Struktur dennoch betrachtet. Da der Zugverband zu Beginn eine gewisse Beschleunigungsphase benötigt, steigt die Geschwindigkeit erst über einen relativ großen Zeitraum an und bleibt danach konstant (Anm. Die Bremsphase ist in diesen Datensätzen nicht enthalten.). Dadurch liegen, wie in Abbildung 4.5 ersichtlich, nur jene Abschnitte außerhalb der Quantilsgrenzen, in denen der Zugverband beschleunigt hat und jene, die im oberen Bereich etwas von der konstanten Geschwindigkeit abweichen.

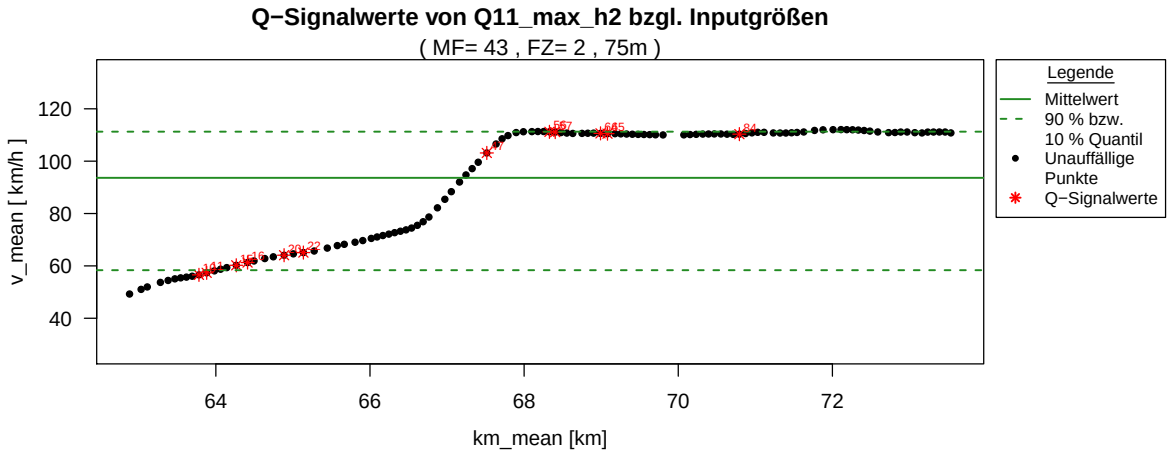

Abbildung 4.5: Inputgröße `v_mean` mit Quantilsgrenzen und Q-Signalwerte

Zu beachten gilt, dass bei der Verwendung der Quantile als Grenzen nur die extremsten Werte identifiziert werden, lokale Extremwerte jedoch unerkannt bleiben, solange sie sich innerhalb der Quantile befinden. Abbildung 4.6 dient zur Veranschaulichung des Problems. Der rot markierte Extremwert fällt nicht auf, wenn nur jene Punkte betrachtet werden, die außerhalb der Quantilsgrenzen liegen.

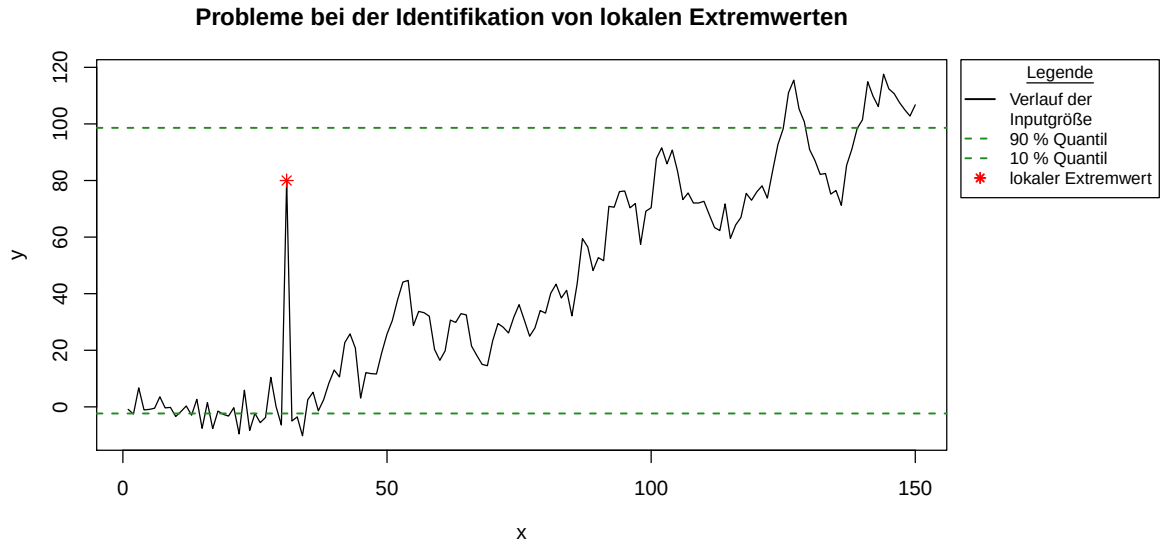

Abbildung 4.6: Probleme bei der Erkennung lokaler Extremwerte

Gerade bei Daten, die einen Trend aufweisen, wie jene in Abbildung 4.6, und somit keine 1-gipflige Dichte, sondern beispielsweise eine 2-gipflige, besitzen, können solche lokalen Extremwerte schnell untergehen, da die Werte zwischen den Gipfeln nicht als Signalwerte erkannt werden, obwohl sie als solche definiert werden sollten. Daher sollte man einen Blick auf die Dichte werfen, und diese zusätzlich einzeichnen. Dies könnte einen Hinweis darüber liefern, dass die Betrachtung der Quantile nicht ausreichend aussagekräftig ist.

Wird in Abbildung 4.3 im Hintergrund zusätzlich die Dichtefunktion eingezeichnet, so lässt sich eine 1-gipflige Verteilung erkennen und die Betrachtung der Quantilsgrenzen scheint ein adäquates Mittel zur Identifikation von Signalwerten bezüglich dieser Inputgrößen (Abbildung 4.7).

Die Dichtefunktion der Variable `v_mean` zeigt jedoch ein anderes Verhalten. In Abbildung 4.8 lässt sich die 2-gipflige Verteilung gut erkennen. Werte, die innerhalb dieser Spitzen liegen, werden nicht als extreme Werte erkannt. In diesem Fall befinden sich jedoch keine Q-Signalwerte in diesem Bereich.

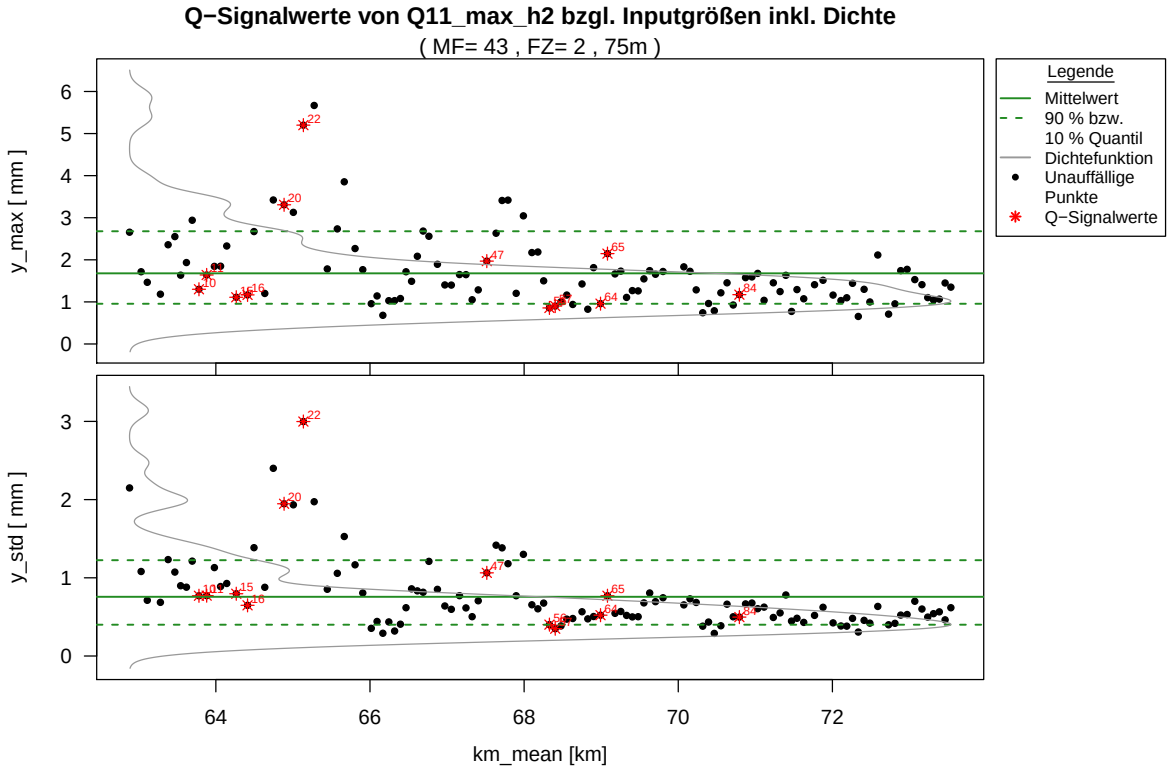

Abbildung 4.7: Inputgrößen `y_max` und `y_std` mit Quantilsgrenzen und Q-Signalwerte inkl. Dichtefunktion

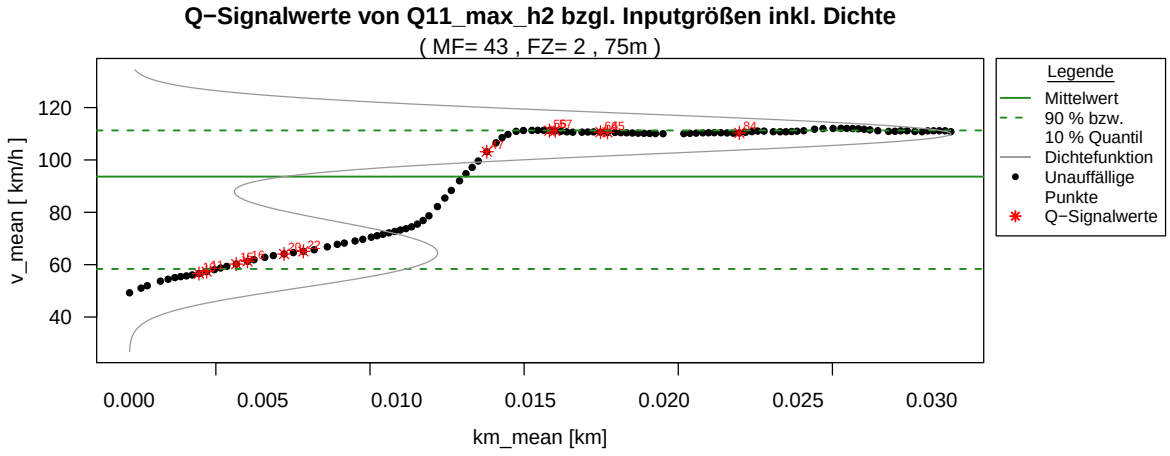

Abbildung 4.8: Inputgröße `v_mean` mit Quantilsgrenzen und Q-Signalwerte inkl. Dichtefunktion

Wurden nun Q-Signalwerte identifiziert, welche auch bezüglich bestimmter Inputgrößen auffällig sind, weiß man jedoch noch nicht, ob diese auch relevant sind, da einige Q-Signalwerte extremer sind als beispielsweise jene, die nur knapp außerhalb der

definierten Grenzen liegen. Daher bietet sich eine Graphik an, in der sich auf einen Blick erkennen lässt, ob die Q-Signalwerte, die auch bezüglich der Inputgrößen auffallen, ein großes Residuum haben und damit besonders schlecht vom Modell geschätzt werden. Abbildung 4.9, in der die standardisierten Residuen gegen die Inputgrößen `y_max` und `y_std` geplottet werden, liefert diesen Überblick, wobei die Datenpunkte erneut nach der Variable `Layout` gefärbt wurden, um zusätzliche Zusammenhänge identifizieren zu können.
Jener Punkt, der Abschnitt 22 repräsentiert, fällt dabei durch sein hohes negatives Residuum auf, da er in der Graphik sehr weit links liegt. Zusätzlich befindet er sich deutlich außerhalb der Quantilsgrenzen der Inputgrößen und sollte daher im Auge behalten werden.

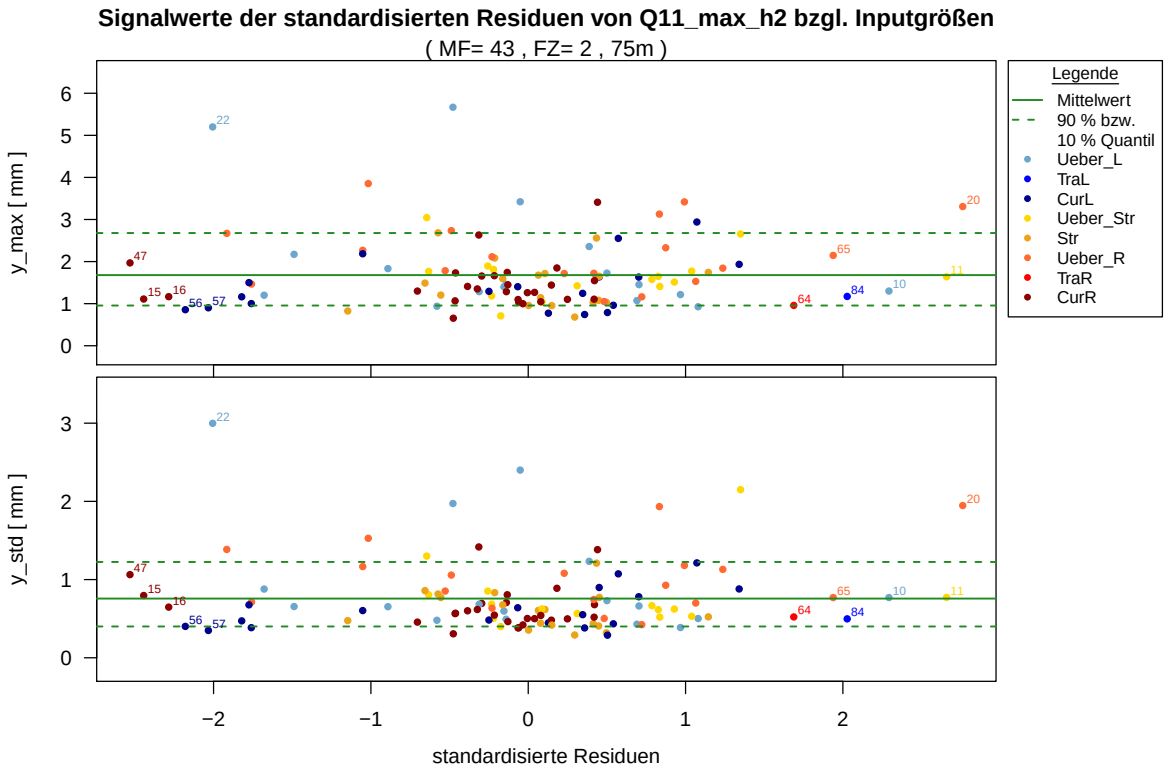

Abbildung 4.9: Standardisierte Residuen gegen die Inputgrößen `y_max` und `y_std` inklusive Quantilsgrenzen und gefärbt bezüglich `Layout`

Da anhand der Graphiken nicht immer eindeutig ersichtlich ist, ob ein Q-Signalwerte inner- oder außerhalb der Grenzen der Inputgrößen liegt und es zudem mühsam ist, alle relevanten Graphiken zu analysieren, kann in R eine Funktion geschrieben werden, die anhand der Daten eine entsprechende Tabelle erstellt (siehe Tabelle 4.10). Mit Hilfe

dieser Tabelle kann auf einen Blick erkannt werden, wo sich die Q-Signalwerte bezüglich der Inputgrößen befinden.
Dafür gibt es folgende Möglichkeiten:

- U = der Q-Signalwert liegt außerhalb der unteren Quantilsgrenze der Inputgröße
- X = der Q-Signalwert liegt innerhalb der Quantilsgrenzen der Inputgröße
- O = der Q-Signalwert liegt außerhalb der oberen Quantilsgrenze der Inputgröße.

Zum Einen sind in dieser Tabelle Werte bedeutungsvoll, die bezüglich vieler Inputgrößen auffallen, wie etwa Abschnitt 65. Andererseits sind auch jene Q-Signalwerte interessant, die sich bezüglich aller Inputgrößen unauffällig verhalten. Dies gilt in diesem Beispiel für die Abschnitte 15 und 84. Diese Q-Signalwerte scheinen sich jedenfalls nicht durch die Inputgrößen erklären zu lassen, sondern dürften auf andere Gründe zurückzuführen sein.

Abbildung 4.10: Tabelle der Position der Q-Signalwerte bezüglich der Inputgrößen

Q−Signalwerte bzgl. Q11_max_h2
(MF= 43 , FZ= 2 , 75m)

	Q−Signal	z_max	z_std	y_max	y_std	d_max	d_std	g_max	g_std	zL_max	zR_max	yL_max	yR_max	zL_std	zR_std	yL_std	yR_std	Ch_mean	U_mean	v_mean
1	10	X	X	X	X	U	U	U	U	U	X	X	X	U	X	X	X	U	X	U
2	11	U	X	X	X	U	X	U	U	U	X	X	X	X	X	X	X	X	X	U
3	15	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
4	16	X	X	X	X	X	X	U	X	X	X	U	X	X	X	X	X	X	X	X
5	20	X	X	O	O	X	X	X	X	X	X	X	O	X	X	O	O	O	X	X
6	22	X	X	O	O	X	X	X	X	X	X	O	O	X	X	O	O	U	X	X
7	47	X	U	X	X	X	X	X	X	U	X	X	X	U	X	X	X	X	X	X
8	56	X	X	U	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
9	57	X	X	U	U	X	X	X	X	X	X	X	X	X	X	U	U	X	X	X
10	64	X	X	X	X	X	X	X	X	X	U	X	X	U	X	X	X	X	X	X
11	65	O	O	X	X	O	O	O	O	O	O	X	X	O	O	X	X	X	O	X
12	84	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Ein Vorteil der Q-Signalwerte besteht in der trivialen Berechnung und der simplen Interpretation. Nichtsdestotrotz muss auf den offensichtlichsten Nachteil hingewiesen werden, der in der willkürlichen Festlegung der Anzahl der Q-Signalwerte liegt. Mit einem Wert für α in (4.3) von 10% werden immer 10% der Zielgröße als Q-Signalwerte definiert (hier: 10% von $115 \approx 12$ Q-Signalwerte). Ein kleinerer Wert für α führt zu einer kleineren Anzahl an Q-Signalwerten.
Aufgrund dieser Einschränkung ist es sinnvoll, eine weitere Methode zur Identifikation von Signalwerten zu betrachten. Dafür bietet sich die Cook-Distanz an, welche in Abschnitt 4.4 eingeführt wird.

4.4 Cook-Signalwerte

Neben den Q-Signalwerten (vgl. Kapitel 4.3) wird eine weitere Methode vorgestellt, um auffällige Datenpunkte als solche zu identifizieren. Dafür wird die Cook-Distanz benötigt, welche der amerikanische Statistiker R. Dennis Cook 1977 einführte. Die Cook-Distanz misst den Effekt, wenn eine Beobachtung aus dem Datensatz gestrichen wird und die kleinste-Quadrate-Schätzung ohne diesen Punkt erfolgt.

Nach einigen Umformungen lautet eine formale Definition der **Cook-Distanz**:

$$D_i = \left[\frac{y_i - x_i^T\hat{\beta}}{\hat{\sigma}\sqrt{1-h_{ii}}}\right]^2 \frac{h_{ii}}{p(1-h_{ii})}, \tag{4.4}$$

mit den Residuen $y_i - x_i^T\hat{\beta} = \hat{\epsilon}$, der geschätzten Standardabweichung $\hat{\sigma}$, dem i-ten Diagonalelement der Hatmatrix h_{ii}, sowie der Anzahl der Parameter p.
Nähere Details lassen sich in [Cook, 1977,] nachlesen.

Als Cook-Signalwerte (= **C-Signalwerte**) werden nun jene Datenpunkte bezeichnet, deren Cook-Distanz größer als eine definierte Grenze ist. Diese ist jedoch nicht einheitlich festgelegt. Während [Fahrmeir *et al.*, 2009, S. 178] als Grenze 0.5 bzw. 1 verwendet, wird in [Bollen und Jackman, 1990] die schärfere Grenze $4/n$, mit dem Stichprobenumfang n, benützt.
In dieser Arbeit wird die Grenze $4/n = 4/115 \approx 0.0348$ angewendet, da diese strenger ist und somit mehr Datenpunkte als C-Signalwerte identifiziert werden. Eine größere, weitere Grenze, die eventuell gar keinen Datenpunkt als C-Signalwerte bezeichnet, würde hier keine Verwendung finden.

Abbildung 4.11 liefert einen Überblick über die berechneten Cook-Distanzen zu jedem Datenpunkt. Dabei wurden jene Werte, die größer als $4/n$ sind, automatisch mit der entsprechenden Indexnummer versehen.
Erneut fallen die Datenpunkte mit den Indexnummern 20 und 22, sowie 65 besonders auf. Diese Abschnitte wurden bereits als Q-Signalwerte identifiziert und ziehen auch bezüglich der Cook-Distanz Aufmerksamkeit auf sich.
In die Berechnung der Cook-Distanz fließt neben dem standardisierten Residuum $\frac{y_i - x_i^T\hat{\beta}}{\hat{\sigma}\sqrt{1-h_{ii}}} = r_i$ (vgl. 3.21) auch die sogenannte Hebelwirkung (engl. *leverage*) der i-ten Beobachtung ein, die mit h_{ii} durch das i-te Diagonalelement der Hatmatrix H gegeben ist.

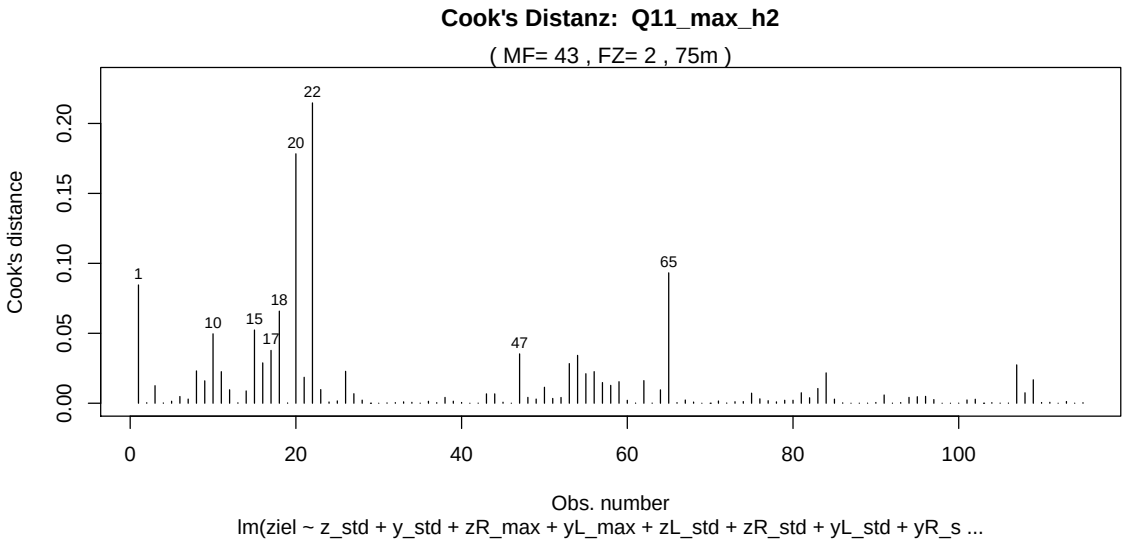

Abbildung 4.11: Cook-Distanzen; 9 Datenpunkte mit $d_i > 0.0348$

Beobachtungen mit einer großen Cook-Distanz haben neben einem hohen Residuum auch eine große Hebelwirkung. Dies lässt sich mit einer Graphik, in der die Hebelwirkung gegen die standardisierten Residuen aufgetragen wird, anschaulich präsentieren (vgl. Abbildung 4.12).

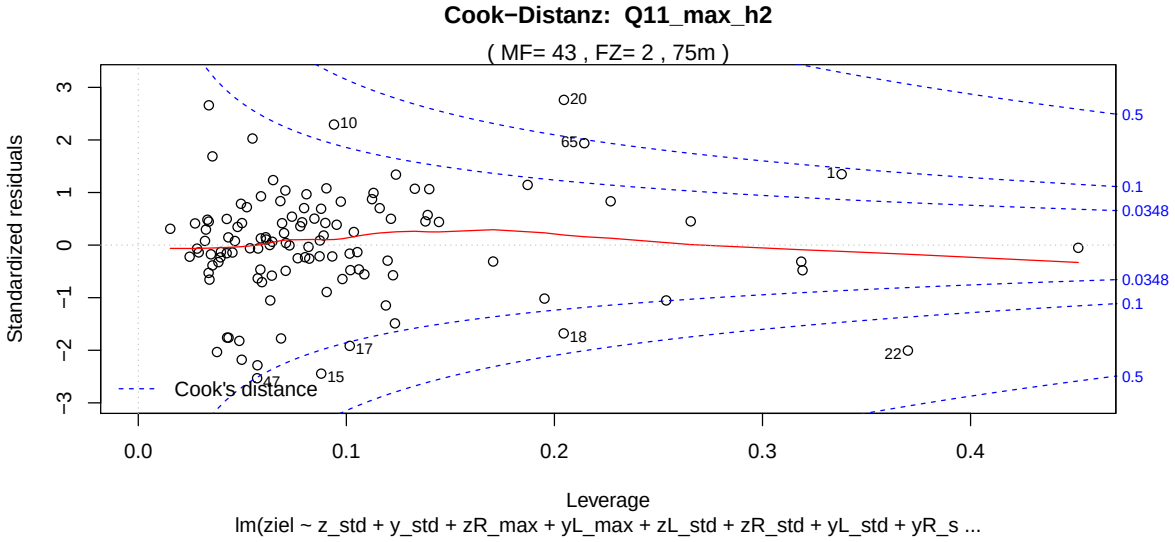

Abbildung 4.12: Plot der Hebelwirkung gegen die standardisierten Residuen

Diese Graphik zeigt, dass ein hohes Residuum alleine meist nicht ausreicht, damit die entsprechende Beobachtung einen starken Einfluss auf das Regressionsmodell darstellt. Eine Kombination aus hohem Residuum und großer Hebelwirkung kann jedoch ein deutlicher Hinweis sein.

In Abbildung 4.12 sind die Grenzen für die Cook-Distanz 0.5, 0.1 und $4/n \approx 0.0348$ als strichlierte Isolinien eingezeichnet. Dies lässt sich in R schnell realisieren, wobei in `Cook_Out` die Anzahl der zu markierenden Punkte angegeben wird:

Auflistung 4.1: Plot der Hebelwirkung gegen die standardisierten Residuen mit verschiedenen Grenzen in R

```
1 plot(mod_bic, which=5, main=paste("Cook's Distanz: ", Q11_max_h2),
2       caption="Modell: mod_bic", id.n = Cook_Out, labels.id = NULL,
3       cook.levels=c(round(4/length(Q11_max_h2),4), 0.1, 0.5, 1))
```

Werden in Abbildung 4.12 jene Quantilsgrenzen eingezeichnet, die zur Entdeckung der Q-Signalwerte verwendet wurden, so können Q-Signalwerte und C-Signalwerte miteinander verglichen werden und jene Beobachtungen identifiziert werden, die mithilfe beider Methoden als auffällig markiert wurden (siehe Abbildung 4.13).

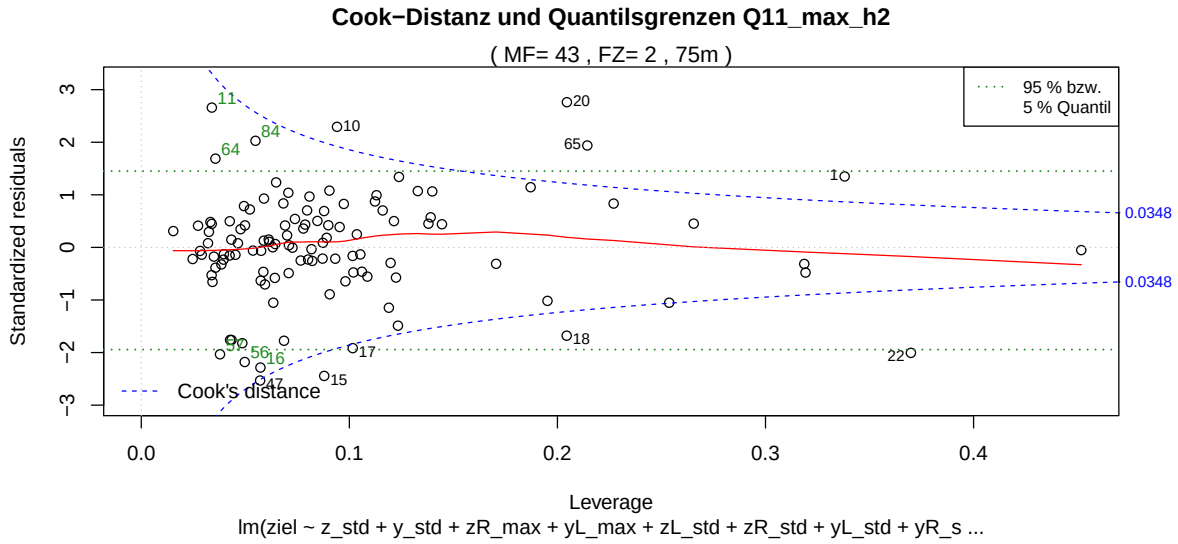

Abbildung 4.13: Plot der Hebelwirkung gegen die standardisierten Residuen mit eingezeichneten Quantilsgrenzen

Abschnitte, die außerhalb der Quantilsgrenzen (punktiert) und gleichzeitig außerhalb der Cook-Grenzen (strichliert) liegen, wurden von beiden Methoden als auffällig bezeichnet. Es gibt jedoch auch Punkte, die nur von einer der beiden Methoden detektiert wurden. Sechs Abschnitte wurden nur mithilfe der Quantile gefunden, wohingegen es nur drei Punkte gibt, die nur durch ihre hohe Cook-Distanz auffallen, jedoch innerhalb der Quantilsgrenzen liegen. Kombiniert man beide Methoden, so lassen sich insgesamt 15 verschiedene Abschnitte identifizieren, die durch mindestens eine der beiden Methoden gefunden wurden.

Auch für die C-Signalwerte wird eine automatische Tabelle erstellt, die für jeden Abschnitt die Cook-Distanz berechnet und farblich markiert, falls diese den Wert $4/n$ übersteigt (vgl. Tabelle 4.1).

Tabelle 4.1: Tabelle der Cook-Distanzen für jeden Abschnitt

Cook−Distanz: Q11_max_h2
(MF= 43 , FZ= 2 , 75m)
Markierung: Distanz > 4/n = 0.034783

Nummer	Distanz	Nummer	Distanz	Nummer	Distanz	Nummer	Distanz
1	0.084551	30	0.000027	59	0.015387	88	0.000067
2	0.000358	31	0.000266	60	0.002100	89	0.000085
3	0.012543	32	0.000435	61	0.000137	90	0.000546
4	0.000202	33	0.000999	62	0.016158	91	0.005888
5	0.001438	34	0.000638	63	0.000112	92	0.000094
6	0.004815	35	0.000067	64	0.009573	93	0.000471
7	0.002961	36	0.001387	65	0.093210	94	0.004182
8	0.023093	37	0.000394	66	0.000433	95	0.004662
9	0.016012	38	0.004164	67	0.002302	96	0.004912
10	0.049619	39	0.001463	68	0.000836	97	0.002605
11	0.022558	40	0.000538	69	0.000011	98	0.000019
12	0.009634	41	0.000012	70	0.000000	99	0.000023
13	0.000296	42	0.000096	71	0.001592	100	0.000089
14	0.008785	43	0.006710	72	0.000190	101	0.002339
15	0.052348	44	0.006700	73	0.001081	102	0.002850
16	0.028832	45	0.000728	74	0.001175	103	0.000009
17	0.037752	46	0.000052	75	0.007199	104	0.000420
18	0.065839	47	0.035280	76	0.003146	105	0.000101
19	0.000196	48	0.004192	77	0.001819	106	0.000133
20	0.178306	49	0.002978	78	0.000986	107	0.027422
21	0.018553	50	0.011419	79	0.002121	108	0.007463
22	0.214685	51	0.003420	80	0.002133	109	0.016752
23	0.009758	52	0.004115	81	0.007482	110	0.000500
24	0.000886	53	0.028334	82	0.003901	111	0.000653
25	0.001656	54	0.034211	83	0.010541	112	0.000019
26	0.022800	55	0.021096	84	0.021698	113	0.001225
27	0.006811	56	0.022590	85	0.002916	114	0.000063
28	0.002217	57	0.014772	86	0.000269	115	0.000381
29	0.000000	58	0.012774	87	0.000026		

Der größte Vorteil dieser Methode liegt darin, dass zwar die Grenzen festgelegt werden, aber keine bestimmte Anzahl an Punkten, welche als C-Signalwerte definiert werden. Dadurch ist eine glaubwürdigere Aussage über die markierten Werte möglich.

5 Analyse der Signalwerte

Nachdem in Kapitel 4 einige Punkte als Signalwerte definiert wurden, soll für möglichst viele von ihnen in Kapitel 5 geklärt werden, warum es sich um auffällige Abschnitte handelt. Zu Beginn werden die gefundenen Signalwerte zusammengefasst, analysiert und die originalen Abschnitte der Länge 75 m betrachtet.
Zuvor wurde gezeigt, dass ein unerwartet hoher bzw. niedriger Wert für eine Inputgröße ausschlaggebend für die Markierung eines Punktes als Signalwert sein kann. In diesem Kapitel wird nun zusätzlich untersucht, ob es noch andere Variablen gibt, die nicht als mögliche Inputgrößen zur Verfügung stehen, und die als Ursache für unerwartete Fahrzeugreaktionen in Frage kommen.
Zum Schluss werden weitere mögliche Ursachen und alternative Herangehensweisen besprochen.

5.1 Zusammenfassung der Signalwerte

Um einen Überblick über alle bisher gesammelten Informationen zu den Signalwerten zu erhalten, kann in R eine automatische Tabelle erstellt werden (vgl. Tabelle 5.1). Diese beinhaltet neben der Abschnittsnummer auch noch die Spalten `h_Gesamt`, `h_Quantile` und `h_Cook`, die angeben, wie oft ein Signalwert insgesamt, alleine mithilfe der Quantile und nur unter Verwendung der Cook Distanz als solcher identifiziert wurde.
Zudem wird der Kilometerstand und das entsprechende Layout des Signalwerts angegeben. Die letzten Spalten zeigen an, bezüglich welcher Zielgrößen ein Signalwert gefunden wurde („X“) oder nicht gefunden wurde („-“). Dabei wurden die Fälle, in denen C-Signalwerte vorkommen, noch zusätzlich farblich markiert.

Anhand der Tabelle lässt sich so zum Beispiel erkennen, dass der Signalwert mit der Indexnummer 1 bei der Betrachtung von allen acht Zielgrößen mit beiden Methoden (4.3 und 4.4) insgesamt 11 mal als Signalwert aufgefallen ist. Davon fünf mal als Q-Signalwert und sechs mal als C-Signalwert.

Der Signalwert mit der Nummer 112 wurde hingegen nur ein einziges Mal gemeldet und das als Q-Signalwert.
Diese Interpretationen lassen sich für alle gefundenen Signalwerte fortführen.

Tabelle 5.1: Zusammenfassung der Signalwerte bzgl. aller Zielgrößen und Markierung der C-Signalwerte

Signalwerte gesamt
(MF= 43 , FZ= 2 , 75m)

	Nr.	h_Gesamt	h_Quantile	h_Cook	km	Layout	Q11_max_h2	Q12_max_h2	sY1_rms	sY1_maxperc	Q11dyn_max_h2	Q12dyn_max_h2	sY1dyn_rms	sY1dyn_maxperc
1	1	11	5	6	62.8802	Ueber_Str	X	X	X	X	X	–	X	–
2	3	2	1	1	63.1103	Ueber_R	–	X	–	–	–	–	–	–
3	4	1	1	0	63.2805	Ueber_Str	–	–	–	X	–	–	–	–
4	5	1	1	0	63.3806	Ueber_L	–	–	–	–	–	–	X	–
5	6	2	1	1	63.4682	CurL	–	–	–	–	X	X	–	–
6	7	2	1	1	63.5432	CurL	–	–	–	–	–	–	X	–
7	8	2	1	1	63.6182	CurL	–	–	–	–	–	–	X	–
8	9	4	2	2	63.6933	CurL	–	X	–	–	–	–	X	–
9	10	3	1	2	63.7810	Ueber_L	X	–	–	–	–	–	X	–
10	11	1	1	0	63.8810	Ueber_Str	X	–	–	–	–	–	–	–
11	13	2	1	1	64.0611	CurR	–	–	–	–	X	–	–	–
12	14	4	2	2	64.1413	Ueber_R	–	–	–	–	X	–	–	X
13	15	2	1	1	64.2650	CurR	X	–	–	–	–	–	–	–
14	16	3	3	0	64.4114	CurR	X	X	–	X	–	–	–	–
15	17	2	1	1	64.4951	Ueber_R	X	–	–	–	–	–	X	–
16	18	6	2	4	64.6352	Ueber_L	X	X	–	X	–	X	–	–
17	19	4	1	3	64.7454	Ueber_L	–	X	–	–	–	X	–	X
18	20	8	4	4	64.8854	Ueber_R	X	X	–	X	–	–	X	–
19	21	4	1	3	65.0056	Ueber_R	–	–	–	X	–	X	X	–
20	22	5	1	4	65.1357	Ueber_L	X	X	–	–	X	X	–	–
21	23	3	1	2	65.2758	Ueber_L	–	–	–	–	–	X	X	–
22	24	1	1	0	65.4459	Ueber_R	–	X	–	–	–	–	–	–
23	26	4	2	2	65.6662	Ueber_R	–	–	X	–	X	X	–	–
24	27	2	2	0	65.8062	Ueber_R	–	X	X	–	–	–	–	–
25	34	1	1	0	66.3946	Str	–	–	–	–	–	X	–	–
26	37	5	3	2	66.6147	Str	–	–	X	–	–	X	X	–
27	38	2	1	1	66.6897	Str	–	–	–	–	X	–	–	–
28	39	1	1	0	66.7646	Str	–	–	–	–	–	X	–	–
29	40	1	1	0	66.8751	Ueber_Str	–	–	–	–	–	–	X	–
30	44	3	2	1	67.2454	Ueber_Str	–	–	–	–	X	–	–	X
31	47	4	2	2	67.5157	CurR	X	–	–	–	–	–	X	–
32	48	3	2	1	67.6361	CurR	–	–	–	–	X	–	–	X
33	49	5	3	2	67.7141	CurR	–	–	X	X	–	X	–	–
34	50	3	1	2	67.7904	Ueber_R	–	–	X	–	X	–	–	–
35	52	8	4	4	67.9906	Ueber_Str	–	–	X	X	–	X	–	X
36	53	2	2	0	68.1006	Ueber_L	–	X	–	X	–	–	–	–
37	55	5	3	2	68.2534	CurL	–	X	–	–	X	–	–	X
38	56	2	2	0	68.3283	CurL	X	–	–	X	–	–	–	–
39	57	1	1	0	68.4034	CurL	X	–	–	–	–	–	–	–
40	59	1	1	0	68.5534	CurL	–	–	X	–	–	–	–	–
41	61	1	1	0	68.7512	Ueber_Str	–	–	X	–	–	–	–	–
42	62	1	1	0	68.8262	Str	–	–	–	–	–	–	–	X
43	64	3	2	1	68.9914	TraR	X	–	X	–	–	–	–	–
44	65	11	5	6	69.0814	Ueber_R	X	–	X	X	X	X	–	X
45	67	1	1	0	69.2542	CurR	–	–	–	–	–	–	–	X
46	73	2	1	1	69.7043	CurR	–	–	–	–	–	X	–	–
47	82	1	1	0	70.6328	Ueber_L	–	X	–	–	–	–	–	–
48	84	2	2	0	70.7930	TraL	X	X	–	–	–	–	–	–
49	86	2	0	2	70.9557	Str	–	–	X	–	–	X	–	–
50	87	2	2	0	71.0306	Str	–	–	X	–	–	–	–	X
51	91	2	1	1	71.3960	CurL	–	X	–	–	–	–	–	–
52	96	1	1	0	71.8783	Ueber_Str	–	–	–	X	–	–	–	–
53	100	2	1	1	72.2611	CurR	–	–	–	–	–	–	–	X
54	101	2	1	1	72.3362	CurR	–	–	X	–	–	–	–	–
55	104	1	1	0	72.5890	Ueber_R	–	–	–	X	–	–	–	–
56	106	1	1	0	72.8117	Str	–	–	–	–	–	–	X	–
57	107	4	2	2	72.8866	Str	–	–	–	–	–	X	–	X
58	109	4	2	2	73.0693	Ueber_R	–	–	–	–	X	–	–	X
59	111	1	1	0	73.2371	CurR	–	–	–	–	X	–	–	–
60	112	1	1	0	73.3121	CurR	–	–	–	–	–	X	–	–

Weiters interessant ist die Summe der gefundenen Signalwerte für jede Zielgröße. Tabelle 5.2 liefert genau diese Information für den betrachteten Datensatz mit Messfahrt 43 und Fahrzeug 2. Die erste Zeile gibt dabei die Summe der gefundenen C-Signalwerte für jede Zielgröße an. Dabei fällt auf, dass diese Anzahl recht unterschiedlich für die verschiedenen Zielgrößen ist. Während für die Zielgröße `sY1_maxperc` nur sechs C-Signalwerte gefunden wurden, war dies für die Zielgröße `Q12dyn_max_h2` mehr als doppelt so oft der Fall. Im Vergleich dazu bleibt die Anzahl der identifizierten Q-Signalwerte aufgrund der Definition über alle Zielgrößen konstant und liegt hier bei 12 (≈ 10 % von $n = 115$).
Die letzte Zeile gibt an, wie viele Signalwerte von beiden Methoden entdeckt wurden. Während beispielsweise für die Zielgröße `sY1dyn_maxperc` fast jeder Signalwert von beiden Methoden als solcher markiert wurde, stimmen für `Q12_max_h2` deutlich weniger C-Signalwerte mit den Q-Signalwerten überein.

Tabelle 5.2: Summe der Signalwerte für jede Zielgröße

Summe der Signalwerte für jede Zielgröße
(MF= 43 , FZ= 2 , 75m)

	Zielgröße	Q11_max_h2	Q12_max_h2	sY1_rms	sY1_maxperc	Q11dyn_max_h2	Q12dyn_max_h2	sY1dyn_rms	sY1dyn_maxperc
1	C−Signalwerte	9	8	10	6	10	13	9	10
2	Q−Signalwerte	12	12	12	12	12	12	12	12
3	C− & Q−Signalwerte	6	5	8	5	8	8	7	9

5.2 Signalausschnitte

Die Indexnummer der Signalwerte stellt nicht die tatsächliche Position im originalen Signaldatensatz dar, da mithilfe einer Datenselektionsmethode (vgl. Abschnitt 2.3) bestimmte Abschnitte ausgewählt wurden. Diese Abschnitte können jedoch im Signaldatensatz markiert werden, um einen Überblick über die Lage der Signalwerte zu erhalten.
Für die Zielgröße `Q11_max_h2` wird dabei das Signal von `Q11` betrachtet, wie in Abbildung 5.1 dargestellt wird. Zusätzlich wurden die C-Signalwerte markiert.
Dabei fällt besonders auf, dass sich die meisten C-Signalwerte innerhalb der ersten drei Kilometer der Strecke befinden und nur ein markierter Abschnitt zwischen dem sechsten und siebten Kilometer liegt.

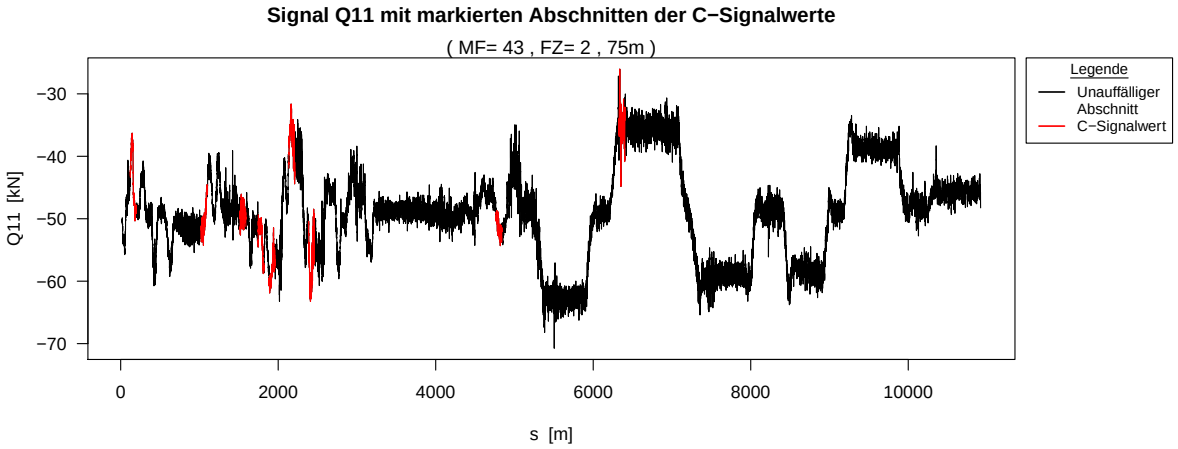

Abbildung 5.1: Verlauf von `Q11` aus dem Signaldatensatz mit markierten C-Signalwerten

Zusätzlich kann Abbildung 5.2 herangezogen werden, um Informationen über das Streckenlayout zu erhalten. Zu beachten ist, dass die Variable `Layout` im Signaldatensatz nur in fünf Kategorien realisiert, da es sich hierbei um die rohen Messdaten handelt und die Einteilung in acht Kategorien erst später durchgeführt wurde (vgl. 2.4).

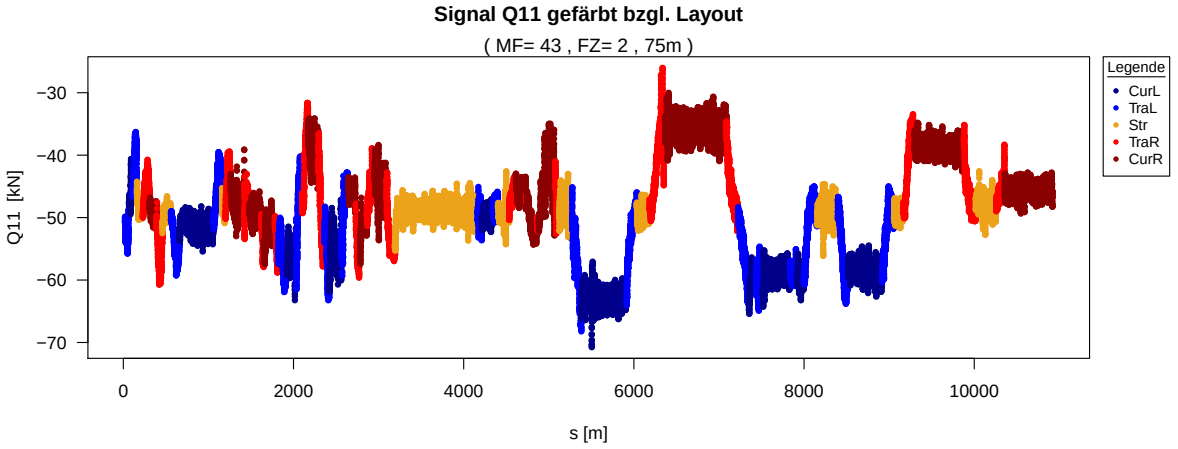

Abbildung 5.2: Verlauf von `Q11` aus dem Signaldatensatz gefärbt bzgl. `Layout`

Durch diese Betrachtung fällt auf, dass die Strecke zu Beginn aus vielen kleineren Bogen besteht und sich das Streckenlayout somit viel öfter ändert als im späteren Verlauf.

Mithilfe Abbildung 5.1 wird zwar sichtbar, wo sich die Abschnitte befinden, die als Signalwerte identifiziert wurden, aber sie liefert keine weiteren Informationen über die Zielgröße in diesem Datenausschnitt oder über den Wert einer anderen Zielgröße in diesem Abschnitt. Aus diesem Grund wird in R für jeden Signalwert der entsprechende

Signalausschnitt geplottet (vgl. Abbildung 5.3). Die Ableitung der Zielgrößen aus den ursprünglichen Messwerten wurde bereits in 2.1 beschrieben.

In den sechs Graphiken in Abbildung 5.3 wird der Verlauf der verschiedenen Messgrößen dargestellt. Dabei werden die Messreihen in den bereits bekannten Farben für das Streckenlayout gefärbt. Zusätzlich wurden Graphiken rot umrandet, falls die für diesen Ausschnitt berechnete Zielgröße als Signalwert identifiziert wurde.

In diesem Beispiel wird der Ausschnitt für den Signalwert mit der Nummer 1 dargestellt. Dieser Abschnitt beinhaltet die Kategorien `TraL` und `Str`, wie anhand der Farben zu erkennen ist. Außerdem wurden fünf der sechs Graphiken rot umrandet. Dies weist darauf hin, dass dieser Abschnitt bezüglich aller Zielgrößen, bis auf `Q12dyn_max_h2`, als Signalwert identifiziert wurde.

Einerseits liefert diese Art der Abbildung einen raschen Überblick darüber, ob ein Abschnitt häufig als Signalwert markiert wurde oder nicht. Andererseits dient sie auch der Überprüfung auf Messfehler. Sollte in einem Abschnitt tatsächlich ein einzelner unerwartet hoher Wert aufgezeichnet worden sein, so kann das mithilfe dieser Abbildungen aufgedeckt werden.

In diesem Beispiel lässt sich erkennen, dass sich die Messwerte mit dem wechselnden Streckenlayout stark verändern, wodurch die Werte der berechneten Zielgrößen nicht in den erwarteten Rahmen fallen und als auffällig markiert werden.
Es finden sich hier jedoch keine Hinweise auf einen Messfehler in diesem Abschnitt des Signals.

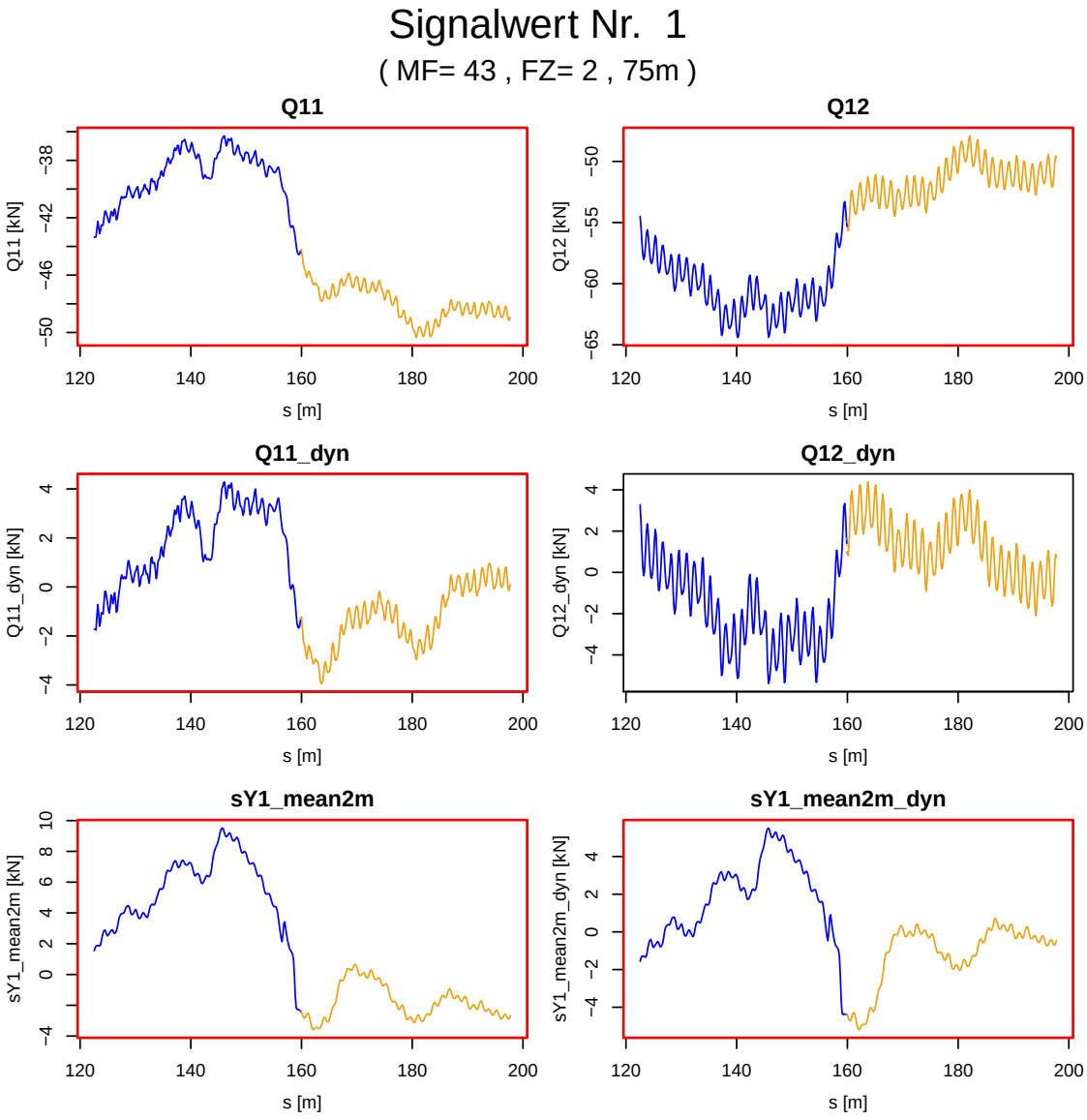

Abbildung 5.3: Signalausschnitte für Signalwert 1 mit Markierungen

5.3 Häufigkeiten der Signalwerte

Um die Anzahl der gefundenen Signalwerte für alle Zielgrößen graphisch darzustellen, bieten sich Balkendiagramme an (vgl. Abbildungen 5.4 und 5.5). Diese liefern Informationen darüber, welche Signalwerte am häufigsten identifiziert wurden und von welcher Methode. Für den betrachteten Datensatz lässt sich erkennen, dass die Abschnitte mit den Indexnummern 1 und 65 sehr oft von beiden Methoden als Signalwerte markiert wurden.

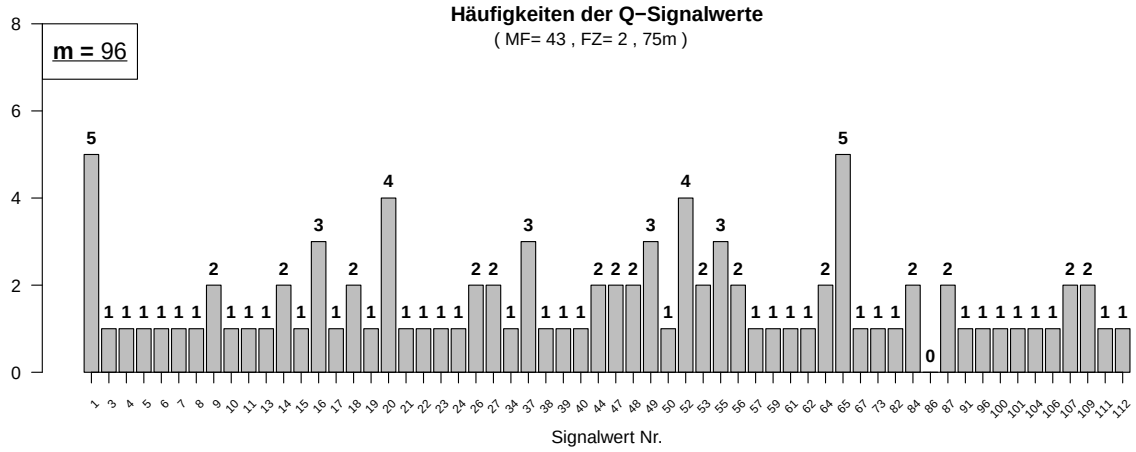

Abbildung 5.4: Häufigkeiten der Q-Signalwerte

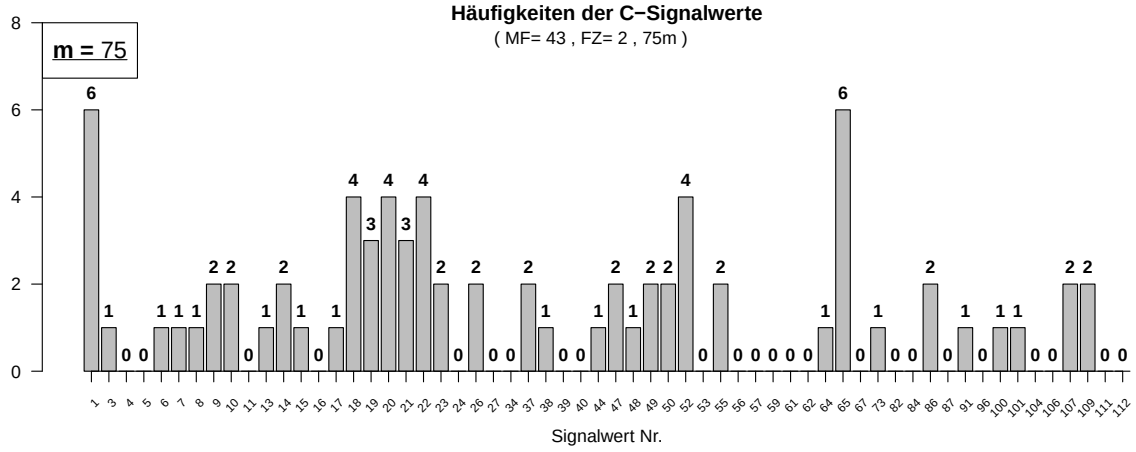

Abbildung 5.5: Häufigkeiten der C-Signalwerte

Jeder Abschnitt kann bezüglich jeder Zielgröße höchstens ein mal pro Methode als Signalwert identifiziert werden. Da für die Methode der Quantile in dieser Arbeit $\alpha = 10\%$ gewählt wurde und der betrachtete Datensatz aus $n = 115$ Datenpunkten

besteht, werden $10\% \times 115 \approx 12$ Abschnitte als Signalwerte markiert. Daher ergibt sich für die Anzahl der Q-Signalwerte exakt $8 \times 12 = 96$. Die Anzahl der C-Signalwerte liegt erfahrungsgemäß niedriger und beläuft sich in diesem Beispiel auf 75.

Die Darstellung als Stabdiagramm eignet sich sehr gut für den Vergleich der beiden Methoden, da auf einen Blick erkannt werden kann, ob ein bestimmter Signalwert von beiden Methoden identifiziert wurde und wie oft dies der Fall war.

Zusätzlich können Stabdiagramme bezüglich dem Layout zeigen, welches Streckenlayout am häufigsten unter den Signalwerten vorkommt (vgl. Abbildungen 5.6 und 5.7). Da jeder Abschnitt nur in eine der acht möglichen Layoutkategorien fallen kann, ergeben sich bei 59 Q-Signalwerten und 36 C-Signalwerten genau 59 und 36 Einträge in den Stabdiagrammen bezüglich dem Layout.

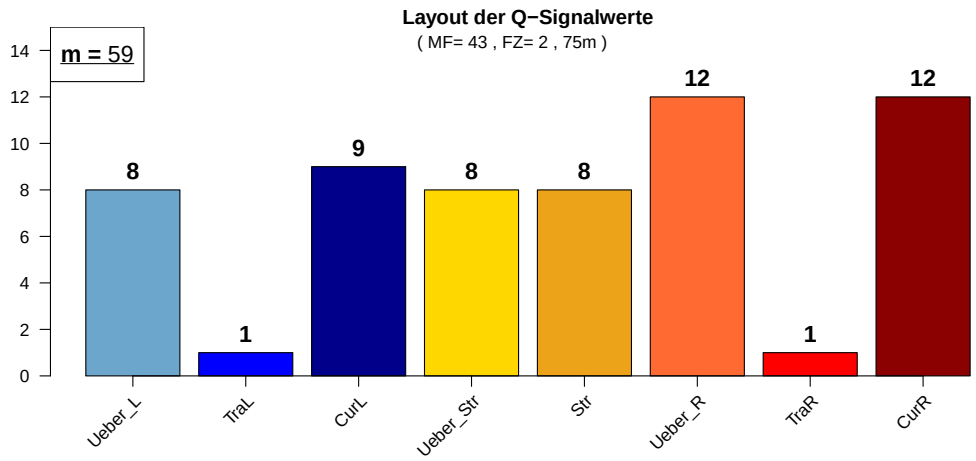

Abbildung 5.6: Häufigkeiten des Streckenlayouts der Q-Signalwerte

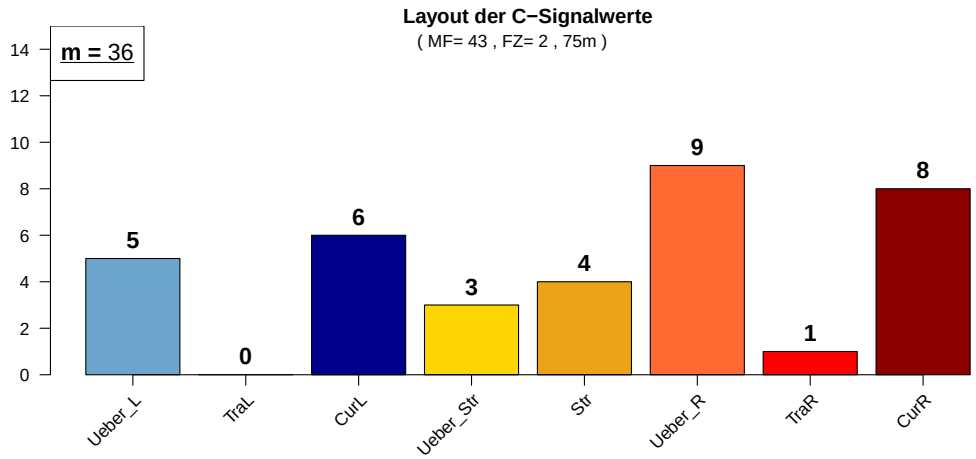

Abbildung 5.7: Häufigkeiten des Streckenlayouts der C-Signalwerte

Die meisten Signalwerte befinden sich in einer Rechtskurve (`CurR`) oder einem Rechtsübergang (`Ueber_R`) und vergleichsweise wenige C-Signalwerte liegen auf einer geraden Strecke (`Str`). Die Verwendung der absoluten Werte kann jedoch irreführend wirken, da kein Bezug zur Gesamtanzahl der vorhandenen Streckenlayouts herstellt wird. Aus diesem Grund werden zusätzlich die Stabdiagramme mit den relativen Anteilen in Prozent betrachtet (vgl. Abbildungen 5.8 und 5.9).

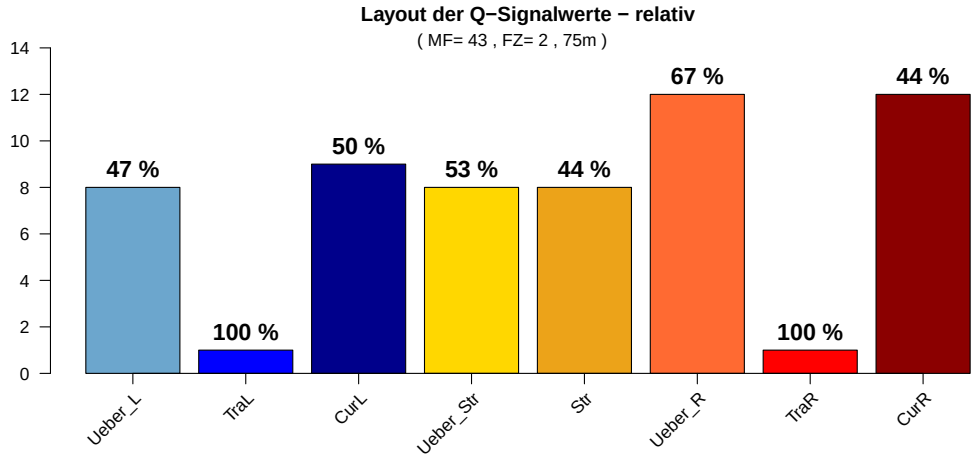

Abbildung 5.8: Häufigkeiten des Streckenlayouts der Q-Signalwerte

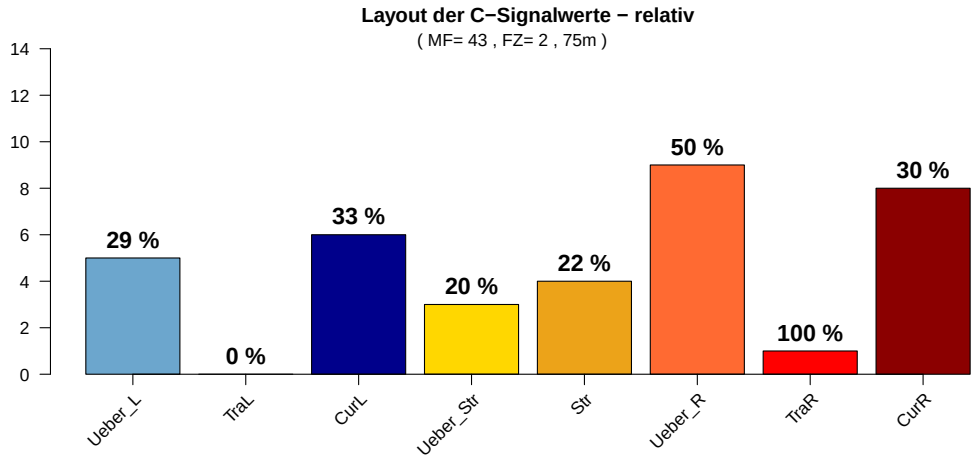

Abbildung 5.9: Häufigkeiten des Streckenlayouts der C-Signalwerte

Anhand der Darstellung der relativen Anteile in Prozent lässt sich erkennen, dass zwar gleich viele (=12) Q-Signalwerte in einer Rechtskurve oder einem Rechtsübergang liegen, die Gesamtanzahl dieser Layoutkategorien jedoch unterschiedlich ist. So wurden die verhältnismäßig meisten Q-Signalwerte in einem Rechtsübergang gefunden - 67 % der Abschnitte wurden bezüglich mindestens einer der acht Zielgrößen als Q-Signalwerte identifiziert. Die Häufigkeiten in den Kategorien `TraL` und `TraR` lassen sich nicht

aussagekräftig interpretieren, da sie mit jeweils nur einem Abschnitt zu dünn besetzt sind.

Da insgesamt weniger C-Signalwerte als Q-Signalwerte gefunden wurden, fallen auch die relativen Häufigkeiten geringer aus. Dennoch wurden auch die meisten C-Signalwerte in einem Rechtsübergang identifiziert - 50 % der Abschnitte in einem Rechtsübergang wurden bezüglich mindestens einer Zielgröße als C-Signalwert markiert.

5.4 Ursache: Bauwerk

Typischerweise beinhaltet eine Bahnstrecke nicht nur einen Start- und Zielbahnhof, sondern auch weitere Bahnhöfe, sowie diverse Brücken und Weichen. Diese Bauwerke könnten die Ursache für einige Signalwerte sein, da ein Fahrzeug möglicherweise anders reagiert, wenn es nicht über einfache Schienen, sondern über diese Bauwerke fährt. Abbildungen 5.10 und 5.11 eignen sich besonders gut dafür, Bauwerke als potentielle Ursache für Signalwerte zu erkennen. Neben dem Verlauf der Zielgröße `Q11_max_h2` wird die Position der zuvor identifizierten Signalwerte markiert. Dabei lassen sich einige Überschneidungen der Q-Signalwerte (grüne Quadrate) mit den C-Signalwerten (rote Kreise) erkennen.

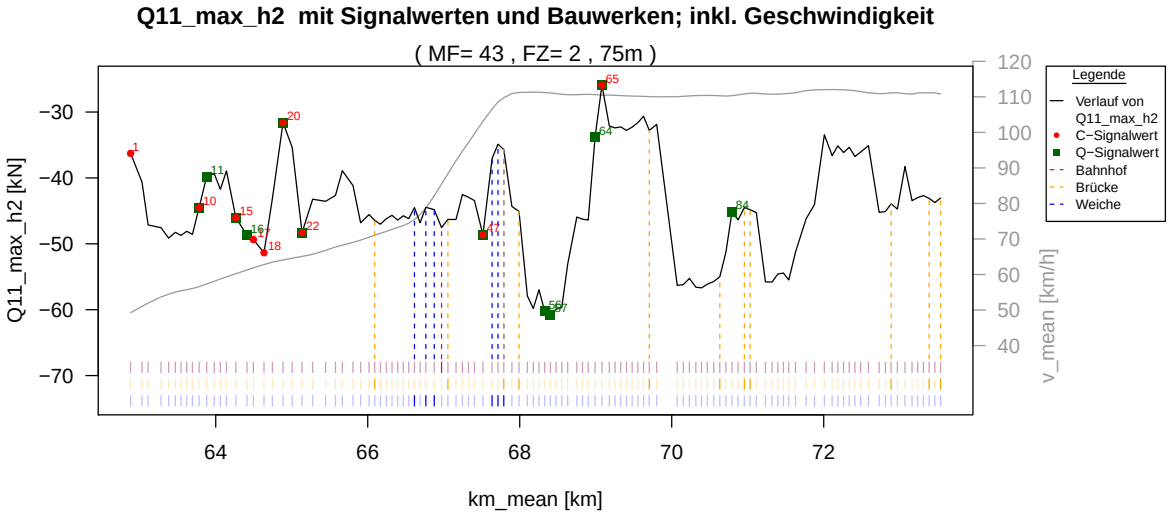

Abbildung 5.10: Verlauf der Zielgröße `Q11_max_h2` mit eingezeichneter Position der Signalwerte und Bauwerke; inkl. Geschwindigkeit

Der untere Teil der Graphik dient zur Erkennung der Bauwerke. Die erste der drei Reihen steht für einen Bahnhof, die zweite für eine Brücke und die letzte schließlich

für eine Weiche auf der Strecke. Befindet sich in einem Streckenabschnitt tatsächlich ein Bauwerk, so wird die strichlierte Linie bis zum Zielgrößenverlauf senkrecht nach oben verlängert. Befindet sich an diesem Berührpunkt ein Signalwert, so könnte das entsprechende Bauwerk ein Grund dafür sein. Zusätzlich wurde in Abbildung 5.10 die Geschwindigkeit im Hintergrund eingezeichnet um diese ergänzende Information bereit zu stellen.
Insgesamt liegen auf dem betrachteten Streckenabschnitt 1 Bahnhof, 11 Brücken und 6 Weichen.

In Abbildung 5.11 wird zusätzlich im Hintergrund statt der Geschwindigkeit die Krümmung `Ch_mean` dargestellt. Dadurch wird erneut der starke Zusammenhang zwischen der Größe der Bögen auf der Strecke und der gefundenen Anzahl an Signalwerten verdeutlicht.

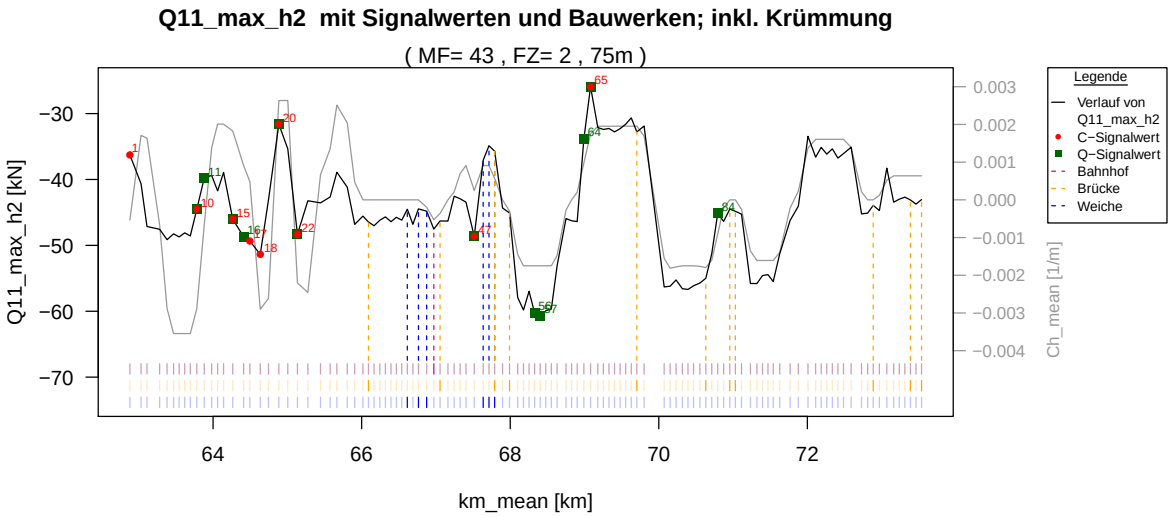

Abbildung 5.11: Verlauf der Zielgröße `Q11_max_h2` mit eingezeichneter Position der Signalwerte und Bauwerke; inkl. Krümmung

Anhand der Graphiken lässt sich erkennen, dass auf dieser Teststrecke nur ein Bahnhof liegt, der sich bei Kilometer 67 befindet. Dieser Abschnitt wurde für die Zielgröße `Q11_max_h2` jedoch nicht als Signalwert identifiziert. Auch die Brücken und Weichen im näheren Umfeld um den Bahnhof scheinen keinen starken Einfluss auf diese Zielgröße zu haben.

Für die Zielgröße `Q11_max_h2` konnte kein Bauwerk als Ursache für einen Signalwert gefunden werden. Für `Q12dyn_max_h2` jedoch beispielsweise schon (vgl. Abbildung 5.12). Auf der Strecke sind fünf Abschnitte zu finden, in denen Weichen vorkommen, wobei

drei dieser Abschnitte als Signalwerte auffallen (Indexnummern 37, 39 und 49). Zwei dieser Signalwerte wurden sogar als C- und Q-Signalwert markiert. Des weiteren sind vier Abschnitte, in denen Brücken vorkommen, als Signalwerte identifiziert worden (Indexnummern 52, 73, 86 und 107). Dies legt die Vermutung nahe, dass die Bauwerke einen starken Einfluss auf die dynamischen Zielgrößen ausüben könnten.

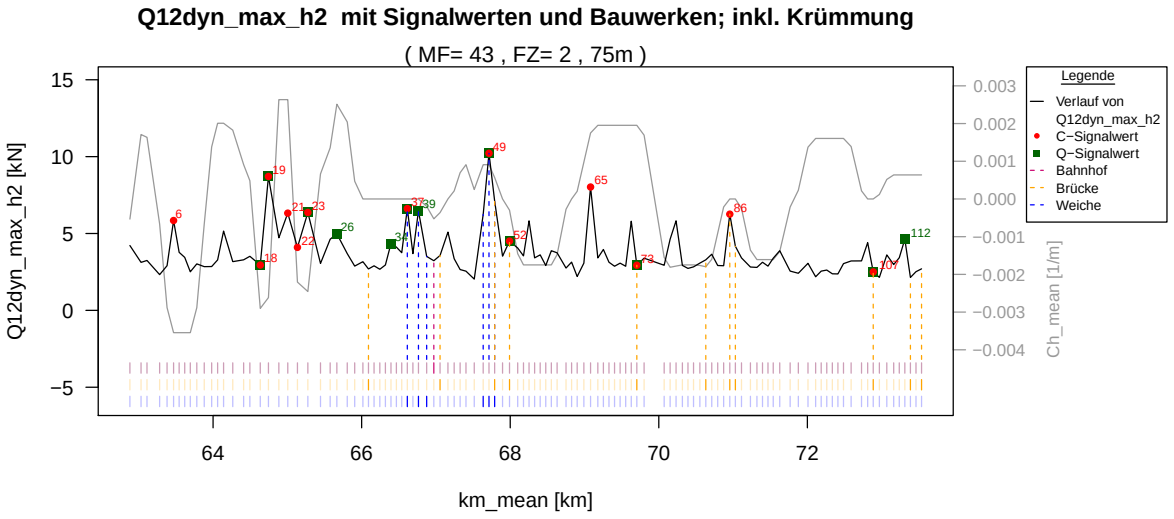

Abbildung 5.12: Verlauf der Zielgröße `Q12dyn_max_h2` mit eingezeichneter Position der Signalwerte und Bauwerke; inkl. Krümmung

5.5 Ursache: Messfahrten

Eine weitere mögliche Ursache für einige Signalwerte könnte in den unterschiedlichen Messfahrten zu finden sein. Falls sich die drei Messfahrten in manchen Abschnitten stark voneinander unterscheiden und dort zudem Signalwerte auftreten, könnte dies ein Hinweis auf eine Unregelmäßigkeit in einer Messfahrt sein. Um dies herauszufinden wird der Verlauf der Zielgröße für alle drei Messfahrten in einer Graphik dargestellt.

Da es zu unübersichtlich wäre, alle drei Fahrzeugtypen gleichzeitig zu berücksichtigen, erfolgt eine Aufteilung in drei Abbildungen (vgl. Abbildungen 5.13 , 5.14 und 5.16), mit denen die Messfahrten miteinander verglichen werden können.

Abbildung 5.13 zeigt den Verlauf der Zielgröße `Q11_max_h2` (durchgehende Linien) für alle drei Messfahrten 41, 42 und 43 für den Fahrzeugtyp 1. Zusätzlich sind die Positionen der C-Signalwerte für jede der Messfahrten in unterschiedlichen Symbolen dargestellt.

Außerdem wurden die jeweiligen Geschwindigkeiten pro Messfahrt als strichlierte Linien eingezeichnet.

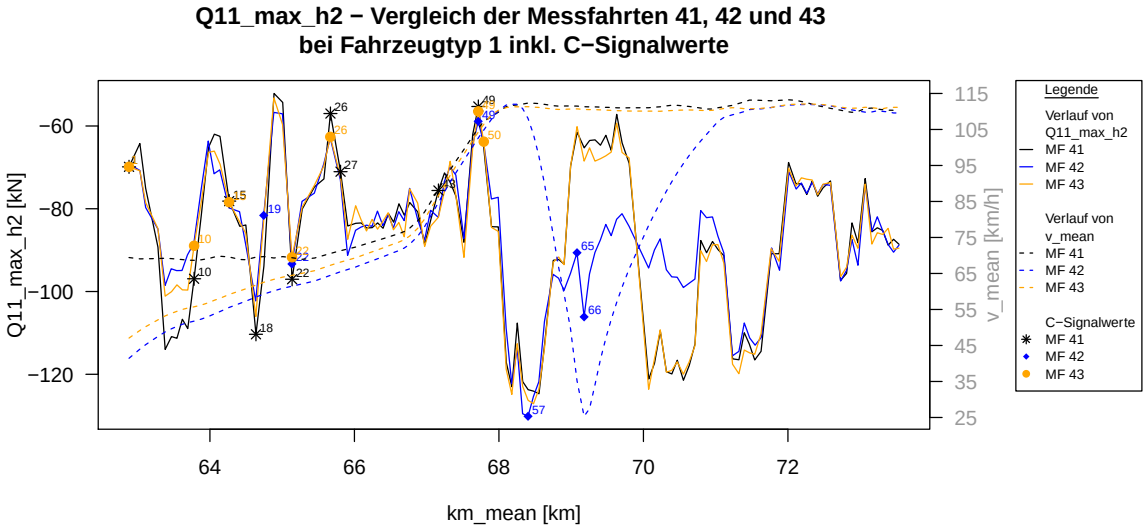

Abbildung 5.13: Verlauf der Zielgröße `Q11_max_h2` für Messfahrten 41, 42 und 43 für die Lok inkl. Markierung der C-Signalwerte

Besonders auffällig in dieser Graphik ist der Geschwindigkeitsabfall um Kilometer 69 für Messfahrt 42, während die Geschwindigkeit in den anderen beiden Messfahrten ab Kilometer 68 konstant bleibt. Im Bereich des Abfalls wurden für Messfahrt 42 zwei C-Signalwerte mit den Indexnummern 65 und 66 identifiziert. Da diese nur in dieser Messfahrt aufgefallen sind, liegt es nahe, den Geschwindigkeitsabfall dafür verantwortlich zu machen.

Dieselbe Graphik für den Reisezugwagen zeigt besonders viele C-Signalwerte hintereinander beziehungsweise in unmittelbarer Nähe zueinander (vgl. Abbildung 5.14).
Die Signalwerte mit den Nummern 15, 17, 18, 20 und 22 für Messfahrt 43 liegen sehr nahe beieinander. Da diese Abschnitte in den anderen Messfahrten nicht aufgefallen sind, könnte auch dies ein Hinweis auf eine Unregelmäßigkeit in der Messfahrt sein.

Auch für Messfahrt 42 können viele aufeinander folgende Signalwerte lokalisiert werden. Dies fällt besonders zwischen Kilometerstand 68 und 70 auf. Da in diesem Bereich jedoch der zuvor erkannte Geschwindigkeitsabfall liegt, könnten diese Signalwerte daher resultieren.

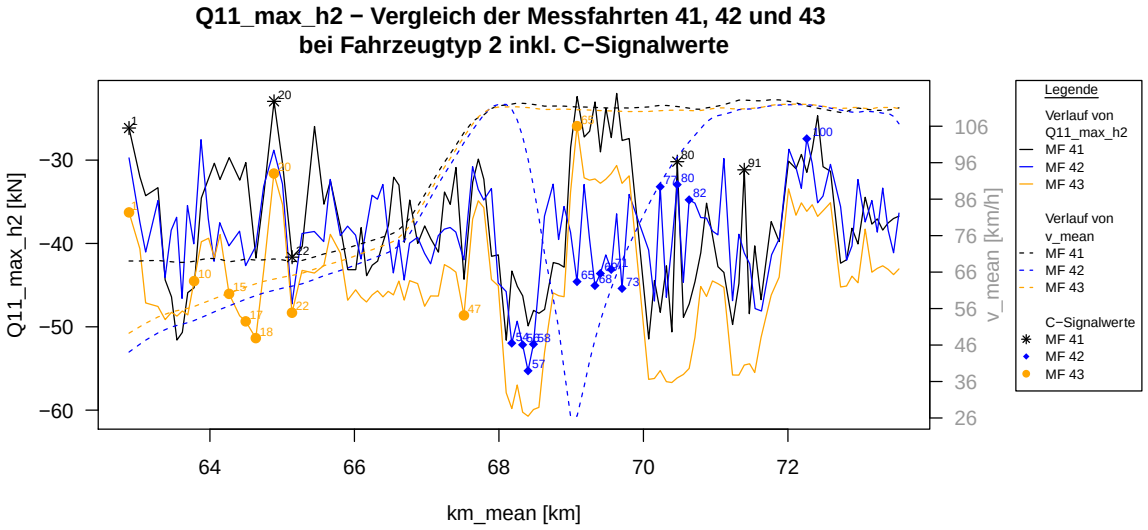

Abbildung 5.14: Verlauf der Zielgröße `Q11_max_h2` für Messfahrten 41, 42 und 43 für den Reisezugwagen inkl. Markierung der C-Signalwerte

Des Weiteren fällt jedoch auf, dass sich die Verläufe der Zielgröße `Q11_max_h2` für den Reisezugwagen sehr stark bezüglich der Messfahrt unterscheiden. Liegen die Linien für die Lok oft direkt übereinander, so lassen sich in Abbildung 5.14 große Skalenunterschiede erkennen. Da es jedoch schwierig zu beurteilen ist, welche der Messfahrten aus der Reihe tanzt, wird der Verlauf der anderen Zielgrößen zum Vergleich herangezogen. So liefert beispielsweise Abbildung 5.15 weitere wichtige Informationen.

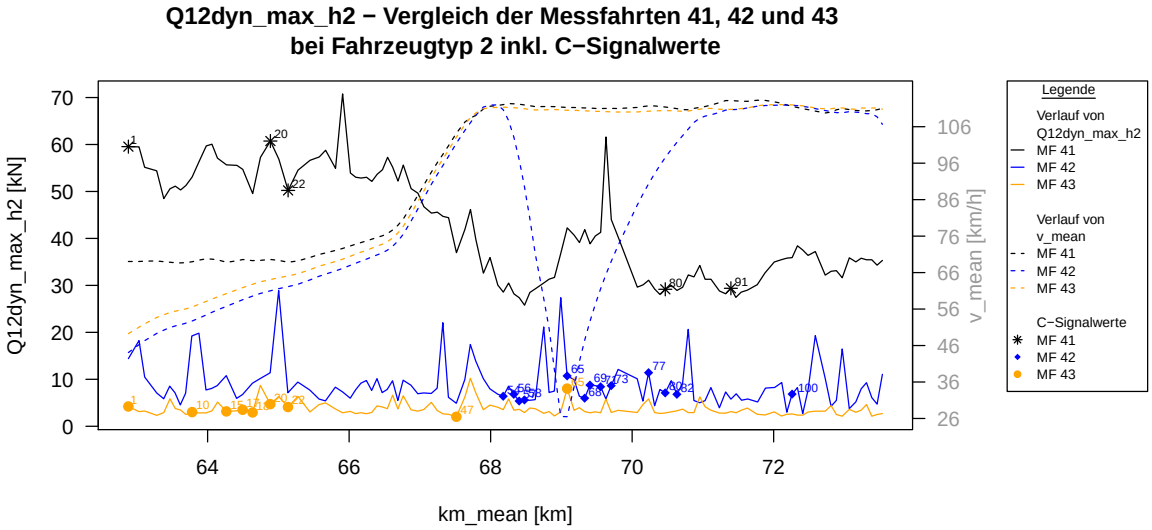

Abbildung 5.15: Verlauf der Zielgröße `Q12_max_h2` für Messfahrten 41, 42 und 43 für den Reisezugwagen inkl. Markierung der C-Signalwerte

Der Verlauf der Zielgröße `Q12dyn_max_h2` (vgl. Abbildung 5.15) für den Reisezugwagen zeigt ein deutlich unterschiedliches Verhalten für die Messfahrten. Während sich die Wertebereiche für die Messfahrten 42 und 43 stark ähneln, passt die Messung für Messfahrt 41 überhaupt nicht ins Bild. Da diese Unterschiede auch bezüglich der anderen Zielgrößen so extrem sichtbar werden, muss abgeklärt werden, ob die Messungen einen systematischen Fehler aufweisen. Dieser Shift in den Daten für den Reisezugwagen in Messfahrt 41 schließt einen weiteren Vergleich mit den anderen Messfahrten eigentlich aus, da es sich somit nicht um vergleichbare Wiederholungen handelt. Hier wird jedoch bewusst entschieden, die Daten weiter zu behandeln um aufzuzeigen, dass diese Schwierigkeiten auch in weiterer Folge sichtbar werden und zu nicht-interpretierbaren Ergebnissen führen werden.

In Abbildung 5.16, in welcher der Güterwagen unter die Lupe genommen wird, zeigen sich einige Signalwerte in Messfahrt 42, die mit hoher Wahrscheinlichkeit der abfallenden Geschwindigkeit zugeordnet werden können. Weiters wird sichtbar, dass manche Signalwerte in jeder der drei Messfahrten identifiziert wurden. Dies ist hier für die Abschnitte mit den Nummern 1, 19, 22, 26 und 39 der Fall. Die Ursache für diese Signalwerte scheint also nicht in den unterschiedlichen Messfahrten zu finden sein.
Die Begründung für den Signalwert mit der Nummer 47 scheint jedoch schon eher an der Messfahrt zu liegen. Dieser Abschnitt wurde nur in Messfahrt 43 gefunden und scheint zudem ein lokales Minimum darzustellen.

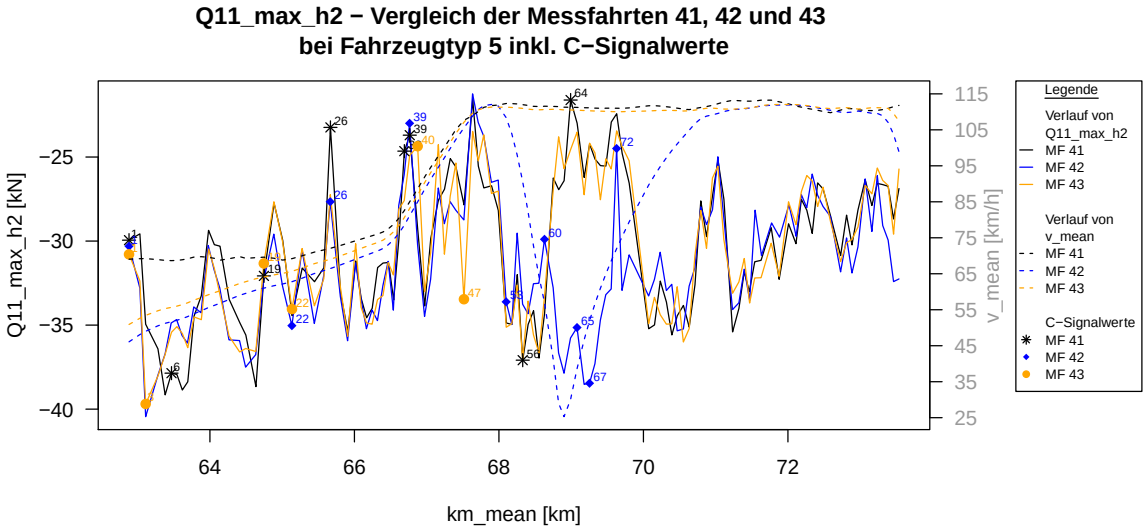

Abbildung 5.16: Verlauf der Zielgröße `Q11_max_h2` für Messfahrten 41, 42 und 43 für den Güterwagen inkl. Markierung der C-Signalwerte

Allgemein fällt auf, dass sich die beiden Messfahrten 41 und 43, zumindest für die Fahrzeugtypen 1 und 5, sehr ähnlich sind. Diese lassen sich demnach gut miteinander vergleichen und die Unterschiede in den identifizierten Signalwerten haben eine hohe Aussagekraft.

5.6 Ursache: Fahrzeuge

Neben den verschiedenen Messfahrten könnten auch die Fahrzeugtypen aufgrund ihrer unterschiedlichen Beschaffenheit eine Ursache für einige Signalwerte sein. Um dies herauszufinden werden Abbildungen betrachtet, in denen der Verlauf der Zielgröße für alle Fahrzeugtypen für eine bestimmte Messfahrt dargestellt wird (vgl. Abbildungen 5.17, 5.18 und 5.19). Erneut wurden die Positionen der C-Signalwerte markiert und die Geschwindigkeiten als strichlierte Linien eingezeichnet. Diese Linien verlaufen nahezu parallel, da die Strecke in einem Zugverband befahren wurde und nicht mit jedem Fahrzeug gesondert. Anhand dieser Graphiken sollen Gemeinsamkeiten und Unterschiede zwischen den Fahrzeugtypen untersucht werden.

In Abbildung 5.17 wird der Verlauf der Zielgröße `Q11_max_h2` für Fahrzeugtypen 1, 2 und 5 für Messfahrt 41 dargestellt.

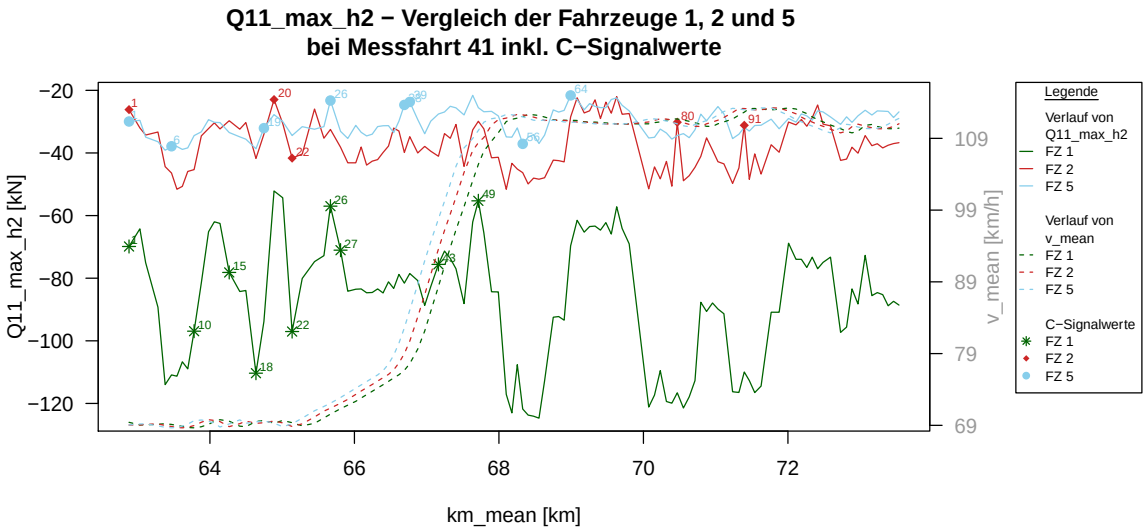

Abbildung 5.17: Verlauf der Zielgröße `Q11_max_h2` für Fahrzeugtypen 1, 2 und 5 für Messfahrt 41 inkl. Markierung der C-Signalwerte

Dabei fällt sofort der ähnliche Wertebereich für die Fahrzeuge 2 und 5 auf, während für Fahrzeugtyp 1 deutlich größere Werte für die Zielgröße `Q11_max_h2` gemessen wurden.

Dennoch unterscheiden sich die gefundenen Signalwerte für Fahrzeugtyp 2 und 5 deutlich, sodass davon ausgegangen werden kann, dass diese Fahrzeuge einen starken Einfluss auf die Zielgröße ausüben.

Abbildung 5.18 zeigt denselben Verlauf für die Messfahrt 42. Hier wird deutlich, dass Fahrzeugtyp 2 und 5 sehr stark auf den Geschwindigkeitsabfall reagieren und viele C-Signalwerte in diesen Bereich fallen.

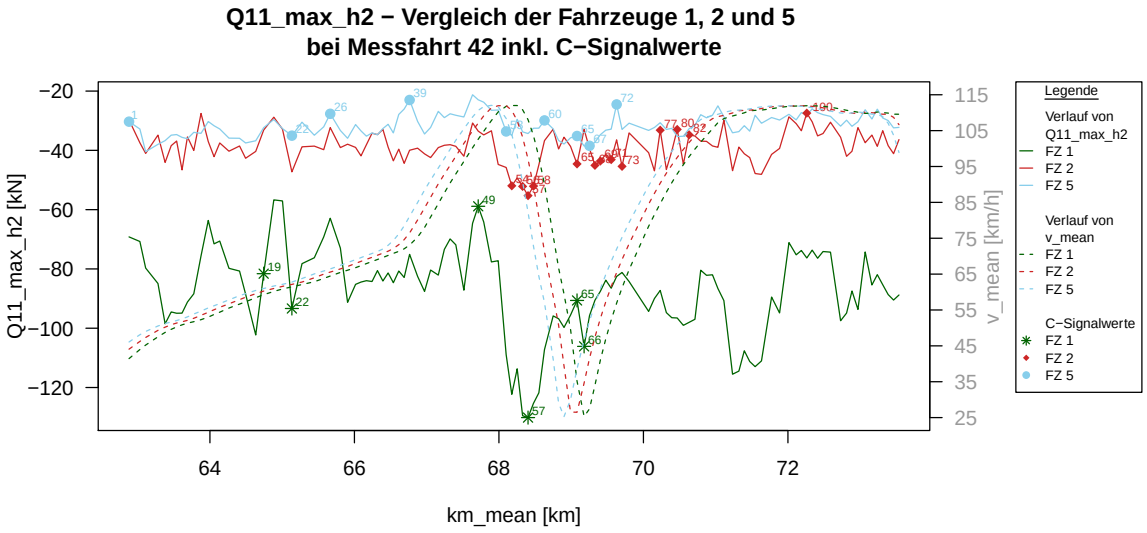

Abbildung 5.18: Verlauf der Zielgröße `Q11_max_h2` für Fahrzeugtypen 1, 2 und 5 für Messfahrt 42 inkl. Markierung der C-Signalwerte

Im Gegensatz dazu scheint Fahrzeugtyp 1 unverkennbar schwächer auf die Geschwindigkeit zu reagieren, da nur drei Signalwerte in diesem Umfeld identifiziert wurden. Diese Robustheit könnte sich auf das höhere Gewicht der Lok zurückführen lassen.

In Abbildung 5.19 wird schlussendlich der Verlauf der Zielgröße für Messfahrt 43 dargestellt. Hier wird noch einmal deutlich, dass sich die meisten C-Signalwerte in der ersten Hälfte der Messstrecke befinden, während sich nach Kilometer 68 nur mehr ein einziger auffälliger Abschnitt finden lässt. Dieses Verhalten lässt sich für alle drei Fahrzeugtypen erkennen, sodass der Grund dafür nicht die Fahrzeugwahl zu sein scheint. Die unterschiedlichen C-Signalwerte wären jedoch sehr wohl ein Hinweis für diese Ursache.

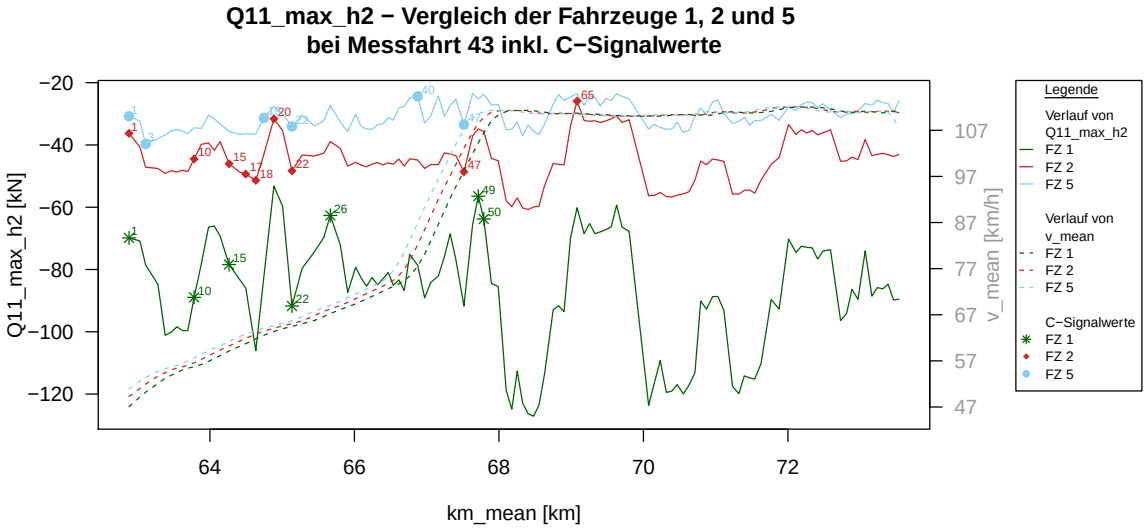

Abbildung 5.19: Verlauf der Zielgröße `Q11_max_h2` für Fahrzeugtypen 1, 2 und 5 für Messfahrt 43 inkl. Markierung der C-Signalwerte

Da es insgesamt nur wenige Abschnitte gibt, die bezüglich aller Fahrzeugtypen gleichzeitig als Signalwerte identifiziert wurden, spricht dies dafür, dass die Fahrzeugwahl einen Einfluss auf die Zielgröße und damit die C-Signalwerte hat.

5.7 Ursache: Variablen

Auch wenn sich einige Signalwerte mithilfe der Bauwerke, unterschiedlichen Messfahrten und Fahrzeugtypen erklären lassen, darf nicht außer Acht gelassen werden, dass die Modelle für die acht Zielgrößen für jede der Fahrzeug/Messfahrt-Kombinationen meistens unterschiedliche Variablen beinhalten. Das macht einen Vergleich schwieriger. Aus diesem Grund werden für jeweils eine Zielgröße und ein Fahrzeug einheitliche Modelle berechnet. Dadurch lassen sich die Ergebnisse für die verschiedenen Messfahrten besser vergleichen. Auch einheitliche Modelle über alle Fahrzeugtypen wären möglich, falls für die verschiedenen Fahrzeuge die selben Variablen relevant sind. Durch die Parameterschätzungen lassen sich dann die unterschiedlich starken Einflüsse der Variablen erkennen.

In den Tabellen 5.3 und 5.4 werden die gemeinsamen Modelle und die entsprechenden Parameterschätzungen für die Messfahrten 41 und 43 zusammengefasst. Messfahrt 42 wird hier nicht mehr berücksichtigt, da ein Modell über die gesamte Strecke aufgrund des Geschwindigkeitsabfalles nicht sinnvoll wäre. Man könnte jedoch eine dummy-Variable einführen, die dieses Ereignis berücksichtigt.

Jedes dargestellte Modell beinhaltet maximal drei Variablen und wurden so erstellt, dass es ein möglichst hohes adjustiertes R^2 generiert.

In Tabelle 5.3, in der Modelle für die quasistatischen Zielgrößen dargestellt werden, lassen sich durchwegs sehr ähnliche Parameterschätzungen für Variablen der einheitlichen Modelle für die beiden Messfahrten 41 und 43 erkennen. Außerdem unterscheiden sich die Werte des Gütekriteriums für die jeweiligen Modelle kaum voneinander. Dies weist darauf hin, dass die Modelle für beide Messfahrten ähnliche Ergebnisse liefern und die verwendeten Variablen unabhängig von der Messfahrt relevant sind.

Die Modelle für den Reisezugwagen in Messfahrt 41 müssen jedoch gesondert betrachtet werden, da zuvor ein Shift in den Daten aufgefallen ist (vgl. Abschnitt 5.5). Anhand der Modelle für die quasistatischen Größen sind nur wenige Auffälligkeiten erkennbar, die erst durch genaueres Hinsehen sichtbar werden. Die meisten Parameterschätzungen sind ähnlich jenen für die Messfahrt 43. Einzelne Schätzungen weisen jedoch auf Unstimmigkeiten hin: Im Modell für die Zielgröße `sY1_maxperc` liegt die Parameterschätzung in Messfahrt 43 für den Intercept bei 2.5898 und für die Variable `Ch_mean` bei -371.6838. Im Gegensatz dazu lauten die Schätzungen in Messfahrt 43 42.7442 bzw. -877.1711. Diese Unterschiede werden für die dynamischen Zielgrößen jedoch um einiges deutlicher.

Die Tabelle der einheitlichen Modelle für die dynamischen Größen (5.4) zeigt deutliche Divergenzen: Die Modelle weisen sowohl für die Lok, als auch den Güterwagen große Ähnlichkeiten zwischen den Messfahrten 41 und 43 auf. Sowohl die Parameterschätzungen, als auch das adjustierte R^2 und der Standardfehler scheinen jeweils vergleichbar zu sein.
Die Modelle für den Reisezugwagen weisen jedoch große Unterschiede auf. Aufgrund des zuvor identifizierten systematischen Fehlers ist die Aussagekraft dieser Modelle natürlich in Frage zu stellen und ein normaler Vergleich nicht zulässig. Wir haben uns jedoch bewusst für die Berechnung der Modelle entschieden, um zu zeigen, wie sich ein solcher Fehler auswirkt. Anhand des Modells für die Zielgröße `Q11dyn_max_h2` sollen die Unterschiede aufgezeigt werden, die einen stutzig machen sollten:

- Parameterschätzung: die Schätzungen unterscheiden sich stärker, als man erwarten würde - Schätzung für `d_max` liegt bei 6.8595 bzw. 25.2900.
- Adjustiertes R^2: auch hier werden große Unterscheide (0.0953 zu 0.5884) sichtbar
- Standardfehler: liegt für die Messfahrt 41 bei 3.6620 - im Vergleich zum Fehler der Messfahrt 43= 0.9598.

Die Messungen für den Reisezugwagen für Messfahrt 41 müssen demnach unbedingt überprüft werden.
Die Modelle für Lok und Güterwagen für die Messfahrten 41 und 43 scheinen jedoch durchaus vergleichbar zu sein.

Tabelle 5.3: Tabelle der einheitlichen Modelle für alle Zielgrößen - Teil 1

Fahrzeug → Zielgröße ↓	I (Lok)		II (RZW)		V (GW)	
Q11_max_h2	**MF 41**	**MF 43**	**MF 41**	**MF 43**	**MF 41**	**MF 43**
	Intercept −80.8384	−80.5859	Intercept −41.7056	−49.2954	Intercept −43.3393	−42.7109
	U_mean 0.1836	0.1721	U_mean 0.0626	0.0707	U_mean 0.0325	0.0274
	v_mean −0.1319	−0.1451	g_std 2.5376	2.9199	v_mean 0.1077	0.0993
	y_std 7.7092	8.3292	y_std 2.6218	2.7355	z_std 2.2348	2.5229
	R^2_{adj} 0.8733	0.8717	R^2_{adj} 0.6810	0.8246	R^2_{adj} 0.7227	0.6973
	SE 6.8850	6.6340	SE 4.2790	3.2520	SE 2.2290	2.2660
Q12_max_h2	**MF 41**	**MF 43**	**MF 41**	**MF 43**	**MF 41**	**MF 43**
	Intercept −50.3193	−60.4433	Intercept −3.2914	−39.4400	Intercept −38.2852	−39.3891
	U_mean −0.2021	−0.1815	U_mean −0.1481	−0.1175	U_mean −0.0313	−0.0289
	v_mean −0.4238	−0.3401	v_mean −0.6583	−0.0670	v_mean 0.0619	0.0701
	d_std 310.9039	339.1263	Ch_mean 2662.2667	2943.0000	d_std 105.4701	114.4134
	R^2_{adj} 0.8786	0.8380	R^2_{adj} 0.8812	0.8005	R^2_{adj} 0.6740	0.6305
	SE 7.8470	8.5820	SE 5.7540	3.553	SE 2.3600	2.5720
sY1_rms	**MF 41**	**MF 43**	**MF 41**	**MF 43**	**MF 41**	**MF 43**
	Intercept −1.6374	−0.8929	Intercept 5.6631	−0.0391	Intercept −2.5536	−1.9062
	g_std 2.0632	2.2530	g_std 0.7358	1.0192	v_mean 0.0483	0.0429
	v_mean 0.0211	0.0130	v_mean −0.0452	0.0006	z_std 0.6901	0.5984
	d_std 36.6791	28.4481	d_std 6.3328	10.5785		
	R^2_{adj} 0.5866	0.6072	R^2_{adj} 0.8314	0.5562	R^2_{adj} 0.4632	0.4922
	SE 0.9313	0.8724	SE 0.3644	0.3736	SE 0.9683	0.9484
sY1_maxperc	**MF 41**	**MF 43**	**MF 41**	**MF 43**	**MF 41**	**MF 43**
	Intercept 3.5970	2.2270	Intercept 42.7442	2.5898	Intercept −4.4068	−3.9970
	Ch_mean −2526.5640	−2383.1680	Ch_mean −877.1711	−371.6838	Ch_mean −200.3358	−247.4000
	g_std 17.5270	19.4380	v_mean −0.0479	0.0276	v_mean 0.1055	0.1023
	y_std 6.8250	5.7610	y_std 3.1685	4.8576	g_std 3.3636	3.2530
	R^2_{adj} 0.3187	0.3656	R^2_{adj} 0.3128	0.2267	R^2_{adj} 0.6459	0.7035
	SE 11.2000	10.4200	SE 3.6150	3.5690	SE 1.9140	1.8650

Tabelle 5.4: Tabelle der einheitlichen Modelle für alle Zielgrößen - Teil 2

Fahrzeug → / Zielgröße ↓	I (Lok)			II (RZW)			V (GW)		
Q11dyn_max_h2		**MF 41**	**MF 43**		**MF 41**	**MF 43**		**MF 41**	**MF 43**
	Intercept	3.3539	2.2862	Intercept	6.6912	0.6839	Intercept	−3.4523	−3.9758
	d_max	42.4215	34.6228	d_max	6.8595	25.2900	v_mean	0.1149	0.1192
	g_std	2.8572	3.7190	g_std	1.4436	1.2750	z_std	2.0502	2.2475
	z_std	2.3988	2.5147	U_mean	−0.0138	0.0008			
	R^2_{adj}	0.5485	0.5789	R^2_{adj}	0.0953	0.5884	R^2_{adj}	0.5786	0.6641
	SE	2.8370	2.7000	SE	3.6620	0.9598	SE	1.8930	1.9160
Q12dyn_max_h2		**MF 41**	**MF 43**		**MF 41**	**MF 43**		**MF 41**	**MF 43**
	Intercept	2.7819	1.9973	Intercept	92.5703	1.9698	Intercept	−2.7747	−3.6896
	g_max	1.1245	1.4389	g_std	5.7462	1.6599	g_max	0.6856	0.5476
	d_std	119.4945	109.6872	v_mean	−0.5705	−0.0114	v_mean	0.0855	0.0895
	z_std	2.2820	2.4246	z_std	−0.1063	1.2849	z_std	1.8667	2.5654
	R^2_{adj}	0.4897	0.5329	R^2_{adj}	0.8205	0.5910	R^2_{adj}	0.6099	0.7342
	SE	3.2580	3.2360	SE	4.7280	0.9003	SE	1.7480	1.5720
sY1dyn_rms		**MF 41**	**MF 43**		**MF 41**	**MF 43**		**MF 41**	**MF 43**
	Intercept	0.2772	0.0986	Intercept	28.7006	0.7228	Intercept	0.2369	0.1054
	Ch_mean	−54.9099	−46.9028	Ch_mean	247.2360	−39.3852	Ch_mean	−243.8000	−164.0000
	g_std	0.7847	0.3006	v_mean	−0.1814	−0.0035	U_mean	0.0029	0.00189
	y_std	0.5108	0.6420	y_std	0.0848	0.5190	y_std	0.3847	0.3607
	R^2_{adj}	0.4820	0.5911	R^2_{adj}	0.9773	0.4868	R^2_{adj}	0.6953	0.5862
	SE	0.4019	0.2732	SE	0.5229	0.2955	SE	0.1630	0.1718
sY1dyn_maxperc		**MF 41**	**MF 43**		**MF 41**	**MF 43**		**MF 41**	**MF 43**
	Intercept	−1.3122	−1.5155	Intercept	11.1514	1.7921	Intercept	−3.8148	−3.3479
	d_std	143.4529	146.5715	d_std	51.6531	58.1318	z_std	1.6836	1.7757
	g_std	7.6770	7.4336	g_std	2.3675	3.1006	v_mean	0.0827	0.0785
	U_mean	−0.0094	−0.0064	v_mean	−0.0747	−0.0243			
	R^2_{adj}	0.6630	0.7139	R^2_{adj}	0.3218	0.6121	R^2_{adj}	0.4544	0.5194
	SE	2.5620	2.2570	SE	2.3210	1.1240	SE	1.7930	1.7600

In weiterer Folge könnten für diese einheitlichen Modelle erneut Signalwerte identifiziert werden und diese mit den zuvor gefundenen verglichen werden. Übereinstimmende Signalwerte sprechen dabei gegen die Annahme, dass die gewählten Variablen einen starken Einfluss auf die Identifikation der Signalwerte hat. Unterscheiden sich die gefundenen Signalwerte jedoch deutlich, so sind diese stark von der Variablenwahl abhängig.

5.8 Weitere mögliche Ursachen für Signalwerte

In dieser Arbeit wurden neben den Bauwerken, die verschiedenen Fahrzeugtypen und Messfahrten, sowie die Wahl der Variablen in Betracht gezogen. In keinem Fall wird dabei Anspruch auf Vollständigkeit dieser Liste erhoben, da noch viele weitere Faktoren einen wesentlichen Einfluss auf die Zielgrößen ausüben können. Unter anderem könnten die Wetterverhältnisse entscheidend sein, oder aber auch der Zustand der Gleise und des Fahrzeugs. Weitere mögliche Faktoren müssten mit den Schienenverantwortlichen in Erfahrung gebracht und die entsprechenden Daten erhoben werden.

Abgesehen von der Ursachenforschung sollte zudem berücksichtigt werden, dass ein lineares Modell eine zu strikte Einschränkung für diese Daten darstellen könnte und ein komplexeres Modell eine bessere Alternative darstellen würde. Des Weiteren gilt immer zu beachten, dass bei Messdaten immer wieder Schwankungen auftreten, die zufälliger Natur sind.

Dennoch kann diese erarbeitete Vorgehensweise einen Aufschluss über auffällige Messabschnitte liefern und diese in vielen Fällen auch erklären.

6 Zusammenfassung

Ziel der vorliegenden Arbeit war es, einen Leitfaden für die Identifikation und Analyse von Messabschnitten, in denen unerwartete Fahrzeugreaktionen auftreten, bereit zu stellen. Dieser Wegweiser soll vor allem für Nicht-Mathematiker verständlich und anwendbar sein.

Die Strecke, auf der die Messdaten erhoben wurden, wurde mit drei verschiedenen Schienenfahrzeugen (Lok, Reisezugwagen, Güterwagen) jeweils dreimal befahren, sodass neun unterschiedliche Fahrzeug/Messfahrt-Kombinationen die Datenbasis bilden.

Zu Beginn wurde nach einem kurzen Einführungskapitel (Kapitel 1) und der Beschreibung der physikalischen Kräfte und des Datensatzes (Kapitel 2), die Grundlage für die nachfolgenden Analysen definiert: das Regressionsmodell (Kapitel 3). Nach der explorativen Datenanalyse, in der die Zusammenhänge der Variablen untersucht werden, wird eine Methode zur Erstellung des Regressionsmodells mithilfe des R-Befehls *step* aufgezeigt (3.3). Da jedoch für jedes Modell gewisse Annahmen getroffen werden, die nicht immer erfüllt sind, müssen diese, wie in Kapitel 3.4 überprüft werden. Eine dieser Voraussetzungen ist die Normalverteilung der Residuen, die für die meisten Modelle in dieser Arbeit jedoch verletzt wurde. Dieser Verstoß wurde in weiterer Folge berücksichtigt, sodass in Kapitel 4, in dem die Abschnitte identifiziert wurden, welche in gewisser Weise auffällig sind, auf nichtparametrische Methoden zurückgegriffen wurde.
Die Identifikation von Abschnitten erfolgte auf zwei Arten: mittels Quantilen (4.3) und mithilfe der Cook-Distanz (4.4). Die markierten Abschnitte wurden, je nach verwendeter Methode, Q-Signalwerte bzw. C-Signalwerte genannt und in Kapitel 5 genauer unter die Lupe genommen. In diesem letzten Kapitel ging es darum, die Ursache für die Identifikation der Signalwerte zu finden. Dies erfolgte durch graphische Methoden, die einfach zu interpretieren sein sollten.
Als erste mögliche Ursache wurden die Bauwerke untersucht (Kapitel 5.4). Während die quasistatischen Kräfte selten von den Bauwerken Bahnhof, Brücke und Weiche beeinflusst werden, scheinen einige Signalwerte für dynamische Kräfte tatsächlich im Bereich von Bauwerken zu liegen.

Anschließend wurde untersucht, ob die Signalwerte auf unterschiedliche Messfahrten zurückzuführen sind (Kapitel 5.5). Durch diese Betrachtungen fiel auf, dass sich eine der Messfahrten in einem Teilbereich gravierend von den anderen unterscheidet und Signalwerte in diesem Gebiet eventuell dadurch verursacht wurden.
Des Weiteren wurden die drei verwendeten Fahrzeugtypen als Ursache für die Signalwerte in Betracht gezogen (Kapitel 5.6). Diese Analysen zeigten einen großen Unterschied in den Fahrzeugreaktionen bezüglich der Fahrzeugtypen. Während zwei Fahrzeugtypen (Reisezugwagen, Güterwagen) ähnliche Ergebnisse lieferten, schienen die Werte der Zielgrößen für die Lok in einem anderen Bereich zu liegen. Einige Signalwerte ließen sich demnach durch die verschiedenen Fahrzeugtypen erklären.
In Kapitel 5.7 wurden schließlich einheitliche Modelle für zwei der Messfahrten berechnet, da bis zu diesem Zeitpunkt individuelle Modelle verwendet wurden. Der Vergleich der Parameterschätzungen liefert Informationen darüber, welche Variablen für die Zielgrößen bezüglich beider Messfahrten besonders relevant sind.
Zum Ende wurden noch weitere mögliche Ursachen für Signalwerte aufgezählt, die in weiteren Analysen mit einbezogen werden könnten.

Insgesamt lässt sich hieraus der Schluss ziehen, dass die Ursache für einige auffällige Abschnitte mithilfe graphischer Methoden aufgedeckt werden konnte und somit wichtige Informationen für weitere Forschungen Richtung Fahrkomfort und Schienenverschleiß liefern.

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[Behnke, 2006] Behnke, J. (2006). *Grundlagen der statistischen Datenanalyse*. VS Verlag für Sozialwissenschaften, Wiesbaden.

[Büning und Trenkler, 1994] Büning, H., und Trenkler, G. (1994). *Nichtparametrische statistische Methoden*. 2. Auflage. Walter de Gruyter, Berlin New York.

[Bollen und Jackman, 1990] Bollen, K.A., und Jackman, R. (1990). Regression diagnostics: an expository treatment of outliers and influential cases. *In:* Fox, J., und Scott Long, J. (eds), *Modern Methods of Data Analysis*. SAGE Publications, Inc, London.

[Cook, 1977] Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics*, **19**, 15–18.

[Czitrom und Spagon, 1987] Czitrom, V., und Spagon, P. D. (1987). *Statistical Case Studies for Industrial Process Improvement*. ASA-SIAM Series on Statistics and Applied Probability. Society for Industrial Mathematics, USA.

[Devore und Farnum, 2004] Devore, J., und Farnum, N. (2004). *Applied Statistics for Engineers and Scientists*. 2. Auflage. Duxbury, USA.

[Fahrmeir *et al.*, 2009] Fahrmeir, L., Kneib, T., und Lang, S. (2009). *Regression - Modelle, Methoden und Anwendungen*. 2. Auflage. Springer, Berlin Heidelberg.

[Fahrmeir *et al.*, 2010] Fahrmeir, L., Künstler, R., Pigeit, I., und Tutz, G. (2010). *Statistik - Der Weg zur Datenanalyse*. 7. Auflage. Springer, Berlin Heidelberg.

[Friedl, 2005] Friedl, H. (2005). *Computerstatistik*. Vorlesungsskript. Institut für Statistik, Graz.

[Hawkins, 1980] Hawkins, D. M. (1980). *Identification of Outliers*. 1. Auflage. SPRINGER-SCIENCE+BUSINESS MEDIA, B.V., Niederlande.

[Ligges, 2005] Ligges, U. (2005). *Programmieren mit R*. 3. Auflage. Springer, Heidelberg.

[Luber, 2011] Luber, B. (2011). *Methode zur Bewertung von Gleislageabweichungen auf Basis von Fahrzeugreaktionen*. Dissertation, Fakultät für Maschinenbau und Wirtschaftsingenieurwesen der Technischen Universität Graz.

[ÖNORM, 2005] ÖNORM (2005). *EN 14363 - Fahrtechnische Prüfung für die Zulassung von Eisenbahnfahrzeugen - Prüfung des Fahrverhaltens und staionäre Versuche*.

[Stadlober, 2011] Stadlober, E. (2011). *Angewandte Statistik*. Vorlesungsskript. Institut für Statistik, Graz.

[Ugarte *et al.*, 2008] Ugarte, M.D., Militino, A.F., und Arnholt, A.T. (2008). *Probability and Statistics with R*. 1. Auflage. CRC Press, Taylor & Francis Group, USA.

[Wei, 1990] Wei, W.W.S. (1990). *Time series analysis*. 1. Auflage. Addison-Wesley Publishing Company, USA.