ifid Schriftenreihe
Beiträge zu IT-Management & Digitalisierung

Band 4

Rüdiger Buchkremer (Hrsg.)

Big Data im Bobsport

~

Scherge, Babula, Krämer, Mittelmann, Domurath, Klug, Naumann, Herhold, Gerhards, Richard, Brückner, Geier, Woobin, Malu, Benson, Simons, Sommer, Wurm, Merklinger, Friedrich, Nowicki, Foric, Buchkremer

FOM Hochschule
ifid
Institut für IT-Management & Digitalisierung
der FOM University of Applied Sciences

Matthias Scherge, Jan Babula, Jan Krämer, Maximilian Mittelmann, Tim Domurath, Marius Klug, Tom Naumann, Fabian Herhold, Dennis Gerhards, Helena Richard, Oliver Brückner, Artur Geier, Lee Woobin, Felicitas Malu, Silas Benson, Philipp Simons, Sarina Sommer, Nina Wurm, Konstantin Merklinger, Carl-Christoph Friedrich, Thomas Nowicki, Elvisa Foric, Rüdiger Buchkremer

Big Data im Bobsport

ifid Schriftenreihe der FOM, Band 4
Beiträge zu IT-Management & Digitalisierung

Essen 2025

ISBN (Print) 978-3-89275-396-4 ISSN (Print) 2699-562X
ISBN (eBook) 978-3-89275-397-1 ISSN (eBook) 2699-5638

Dieses Werk wird herausgegeben vom ifid Institut für IT-Management & Digitalisierung
der FOM Hochschule für Oekonomie & Management gGmbH

Verlag:
MA Akademie Verlags- und Druck-Gesellschaft mbH, Leimkugelstraße 6, 45141 Essen
info@mav-verlag.de

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Rüdiger Buchkremer (Hrsg.)

Big Data im Bobsport

Matthias Scherge, Jan Babula, Jan Krämer, Maximilian Mittelmann, Tim Domurath, Marius Klug, Tom Naumann, Fabian Herhold, Dennis Gerhards, Helena Richard, Oliver Brückner, Artur Geier, Lee Woobin, Felicitas Malu, Silas Benson, Philipp Simons, Sarina Sommer, Nina Wurm, Konstantin Merklinger, Carl-Christoph Friedrich, Thomas Nowicki, Elvisa Foric, Rüdiger Buchkremer

Autorenkontakt:

matthias.scherge@iwm.fraunhofer.de

ruediger.buchkremer@fom.de

Vorwort

Dieses Buch stellt die Ergebnisse einer groß angelegten Studie zur Analyse von Datensätzen aus dem Bobsport vor, die mehrere Millionen Datenpunkte umfasst. Die Analyse erfolgte mit modernen Werkzeugen der künstlichen Intelligenz und des maschinellen Lernens, was es ermöglichte, tiefgehende Einblicke in die komplexen Datenstrukturen zu gewinnen und Muster zu erkennen.

Ein besonderer Schwerpunkt dieser Arbeit liegt auf der detaillierten Darstellung der mathematischen Werkzeuge und der datentechnischen Herangehensweise. Durch die Anwendung von Data-Mining-Methoden wurden umfassende Analysen der vorhandenen Fachliteratur der letzten Jahrzehnte durchgeführt, insbesondere in Bezug auf die Reibung zwischen Kufe und Eis – ein entscheidender Faktor im Bobsport.

Ein innovativer Bestandteil der Studie war die videotechnische Auswertung der Fahrlinien des Bobs in der Bahn. Durch die Generierung von Heatmaps konnten unterschiedliche Fahrweisen präzise unterschieden und visualisiert werden. Diese Visualisierungen zeigen klar den Zusammenhang zwischen Geschwindigkeit und Fahrlinie und erlauben eine detaillierte Prüfung der Korrelation von Laufzeiten in einzelnen Bahnabschnitten mit der Gesamtlaufzeit.

Die Ergebnisse der Analyse offenbarten klare Muster in den Datensätzen und führten zu Empfehlungen hinsichtlich tribologischer Vorteile. Diese Erkenntnisse könnten nicht nur die Leistung im Bobsport erheblich verbessern, sondern auch als Grundlage für gezieltes Data Mining in anderen Sportdisziplinen dienen, die über umfangreiche Sensorik bei Training und Wettkampf verfügen.

Ein weiteres Highlight dieser Arbeit sind die entwickelten, äußerst instruktiven Datenvisualisierungen. Sie bieten eine anschauliche Darstellung der komplexen Zusammenhänge und tragen wesentlich dazu bei, die gewonnenen Erkenntnisse verständlich zu vermitteln.

Diese Studie stellt einen bedeutenden Fortschritt im Bereich der sportlichen Leistungsanalytik dar und zeigt, wie moderne Technologien genutzt werden können, um sportliche Aktivitäten tiefgehender und präziser zu verstehen. Wir hoffen, dass die Ergebnisse dieser Arbeit nicht nur im Bobsport, sondern auch in anderen Bereichen des Sports und darüber hinaus Anwendung finden werden.

Karlsruhe, im Mai 2025

Prof. Dr.-Ing. habil. Matthias Scherge

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

Formelverzeichnis

1 Einleitung

1.1 Sportliche Rahmenbedingungen

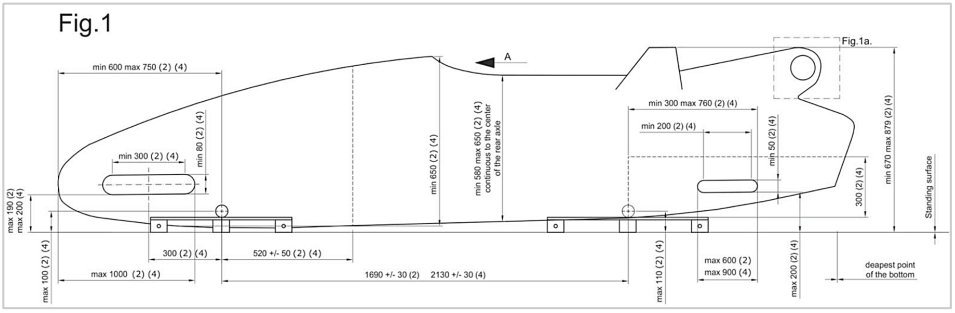
Die zum Ende des 19. Jahrhunderts in St. Moritz begründete Freizeitaktivität des Schlittenfahrens etablierte sich Mitte des 20. Jahrhunderts als Hochleistungssport. Bob und Rennrodeln gelten seit den olympischen Winterspielen 1924 als etablierte Disziplinen und werden seither hinsichtlich des Sportgeräts wie auch der Rahmenbedingungen zunehmend reglementiert. Weltweit sind bis zum heutigen Tag 17 offizielle Wettkampfstätten vorhanden und jährlich Austragungsort der Wettkämpfe im Rennrodel-, Skeleton- und Bobsport (vgl. IBSF, 2022). Der Bobsport ist dabei eine der schnellsten Wintersportarten, bei der die Laufzeitdifferenzen sehr gering sind und häufig im hundertstel Sekundenbereich liegen (vgl. Ubbens *et al.*, 2016, S.92). Aus diesem Grund ist diese Sportart stark vom Einsatz von Technologie abhängig, um eine Optimierung der Fahrt und einen entscheidenden Vorteil gegenüber der Konkurrenz zu erreichen (vgl. Dabnichki, 2015, S. 436).

Bobschlitten haben vier Kufen, zwei vordere und zwei hintere. Bei den Männern gibt es Zweier- und Vierer-Teams, bei den Frauen das Zweier-Team und den Monobob. Das zulässige Mindestgewicht beträgt 170 kg für Zweier- und 210 kg für Viererbobs. Das Höchstgewicht für den Zweierbob beträgt 390 kg und für den Viererbob 630 kg. Die Schlitten verfügen über eine Lenkung und eine Kufenaufhängung, die zusammen mit der Form und der Aerodynamik des Schlittens streng geregelt sind. Für die Kufen ist nur eine Stahlsorte der schweizerischen Firma Kohler zulässig, wobei nachträgliche Änderungen der Materialeigenschaften, beispielsweise durch Erhitzen, verboten sind. Lediglich die Kufengeometrie darf angepasst werden (Ulmer, 2009, S.1f.). Infolgedessen ist das Grundmaterial entsprechend stets identisch.

Die International Bobsleigh & Skeleton Federation (im Folgenden als IBSF) schreibt für alle Wettbewerbe die Regeln vor. Diese Regeln dienen in erster Linie der Sicherheit der Athletinnen und Athleten, verringern aber auch die Möglichkeit eines unfairen Wettbewerbs, da größere Teams mehr Ressourcen für die Entwicklung und Verbesserung des Bobs zur Verfügung haben (vgl. Ubbens *et al.*, 2016, S. 92f.). Unter anderem wird von der IBSF festgelegt, dass Bobbahnen ca. 1,5 km lang sein sollen und eine Höhendifferenz von ca. 120 m nicht überschreiten sollten. Dabei soll das durchschnittliche Gefälle bei 8,6 Prozent liegen und die Bahn eine maximale Steigung von 15 Prozent besitzen (vgl. IBSF (20219), S. 37ff.). Abbildung 1 zeigt, wie der Bobschlitten durch die IBSF reguliert ist und

diese Standards für einen fairen Wettbewerb eingehalten werden müssen (vgl. Dabnichki, 2015, S. 436f.).

Abbildung 1: Regelungen für den Bobschlitten



Quelle: IBSF (2019), S. 52.

Der IBSF benennt insgesamt 17 Wettkampfstandorte in Europa, Asien und Nordamerika zur Austragung von Skeleton- und Bobwettkämpfen. Die jeweiligen Sportstätten unterscheiden sich in Länge, Anzahl der Kurven und Steigung.

Das Bahnreglement des IBSF bestimmt die Richtlinien für Standortnähe, Lage, Bahnverlauf und benötigte Infrastruktur. So darf eine Bahn die Fahrlinien nicht zu stark beschränken und Beschleunigungen von 5g dürfen maximal zwei Sekunden am Stück anhalten. Ebenso ist für ausreichend Sonnenschutz, Beleuchtung sowie An- und Auslaufstrecke zu sorgen (vgl. IBSF, 2019).

Eine Bobbahn ist in drei Hauptabschnitte unterteilt: Start, Fahrt und Ziel. Das Team beginnt mit dem Schieben des Schlittens, während die Uhr durch eine Fotozelle, 15 m von der Startlinie entfernt, ausgelöst wird. Nach dem Ladepunkt, wenn die gesamte Mannschaft in den Schlitten eingestiegen ist, hängt das Ergebnis vollständig von den Fähigkeiten des Piloten und vor allem von der Leistung des Schlittens ab. Die Geschwindigkeit kann bis zu 150 km/h betragen (vgl. IBSF, 2019, S. 38ff.).

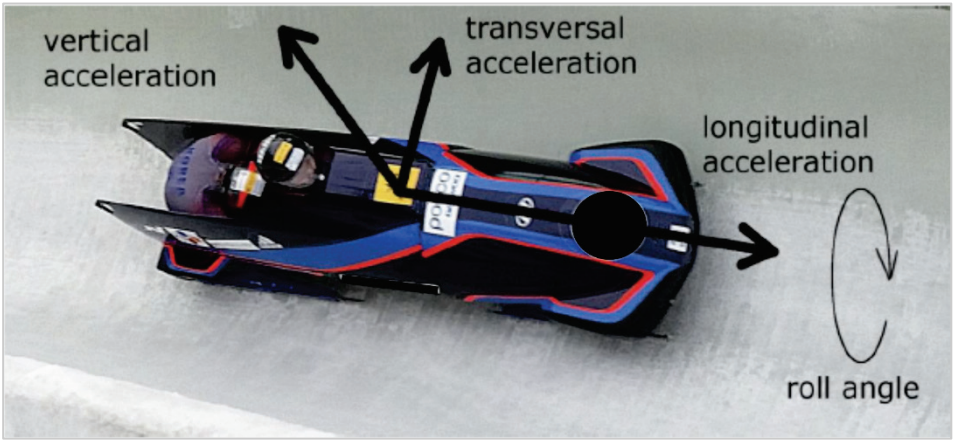
1.2 Physikalische Grundlagen

Für eine Annäherung an die Thematik ist ein Verständnis der physikalischen Grundlagen wichtig, um anschließend aus diesen die Ansätze zur Fahrtzeitoptimierung ableiten zu können. Während der Fahrt eines Bobschlittens wirken drei

relevante Kräfte, welche die Fahrtzeit beeinflussen: die Schwerkraft, welche wiederum die Beschleunigung des Bobs bedingt; der Luftwiderstand des Bobs und des Teams und die Reibung der Kufen auf dem Eis.

Die Schwerkraft als erste relevante Größe sorgt neben der Startbeschleunigung für zusätzliche Beschleunigung und Reibung auf der abschüssigen Strecke (Poirier *et al.*, 2011, S. 2 ff.). Die durch die Schwerkraft induzierte Beschleunigung steigt hierbei mit zusätzlichem Gesamtgewicht des Bobs und der Fahrerinnen und Fahrer an (Dumm *et al.*, 2006, S. 103). Daraus lässt sich ableiten, dass aus isolierter Betrachtung ein möglichst hohes Gewicht unter Berücksichtigung der im Wettkampf zulässigen Obergrenzen wünschenswert ist. In der Literatur wurde geschätzt, dass die Schwerkraft für 84 Prozent der Beschleunigung in Zweierbobfahrten und für 79 Prozent der Beschleunigung in Viererbobfahrten verantwortlich ist (Brüggemann *et al.*, 1997, S. 103).

Abbildung 2: Beschleunigung auf den Bob



Quelle: Scherge (2021), S. 8.

Der Luftwiderstand variiert während der Fahrt. Bei höheren Geschwindigkeiten wird der Bremseffekt durch den Luftwiderstand auf ungefähr 60 Prozent geschätzt, während der Reibung des Schlittens auf dem Eis 40 Prozent zuzuordnen sind (Dabnichki, 2015, S. 439). Diese Zahlen sind stark von der Aerodynamik abhängig und variieren entsprechend je nach Bauweise. Das konkrete Fahrverhalten in den Kurven mit einer starken Annäherung an die Bande kann womöglich den Luftwiderstand erhöhen und damit den Schlitten abbremsen, wenngleich die-

ser Effekt schwierig zu quantifizieren ist (Poirier, 2011, S. 12). Aufgrund der Fahrlinie ist dies jedoch ein Aspekt, der vermutlich eher auf den geraden Teilen der Strecke Relevanz hat, weswegen es empfehlenswert ist, auf diesen Abschnitten eine möglichst mittige Fahrlinie anzustreben. In Kurven könnte dieser zusätzliche Luftwiderstand nahe der Bande bei sehr hohen oder sehr niedrigen Rollwinkeln eine mögliche Einflussgröße sein.

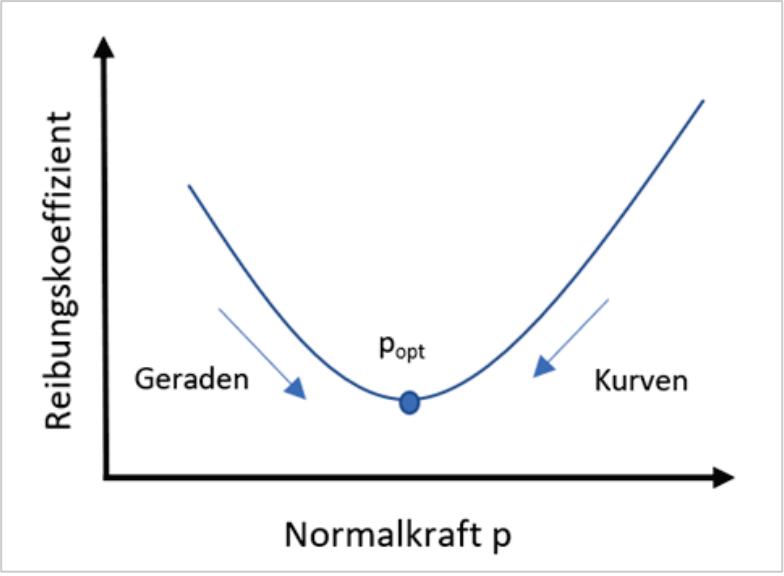
Bei der Reibung, als dritter relevanter Kraft, wechselwirken die Kufen des Bobs mit der Eisfläche und beeinflussen darüber die Beschleunigung und damit die Gesamtgeschwindigkeit. Maß für diese Wechselwirkung ist der Reibungskoeffizient als Verhältnis der Reibungskraft zur Anpresskraft zweier Körper (Sivamani *et al.*, 2003, S. 227). Ein niedrigerer Reibungskoeffizient bedeutet besseres Gleiten, wodurch höhere Geschwindigkeiten realisiert werden können (Hainzlmaier, 2005, S. 62). Die Ursache der ultraniedrigen Reibung ist eine nanometer- bis mikrometerdünne Wasserschicht zwischen Kufe und Eis. *Frictional Heating* gilt als die relevanteste Ursache für die Wasserschicht. Auch in der Literatur wird *Frictional Heating* als dominanter Faktor bei der Entstehung der Wasserschicht gesehen (Dabnichki 2015, S. 439; Irbe und Gross, 2020, S. 3).

Der Effekt, gemäß dem eine höhere Geschwindigkeit und Druck (größere Reibleistungsdichte) zu einer verbesserten Beschleunigung führen, hat jedoch auch Grenzen und ist nichtlinear. Bei Wartungsarbeiten auf der Bobbahn am Königssee sind in den Kurven Fahrrillen aufgefallen, welche durch den hohen Druck ins Eis gepresst wurden. Die Tiefe dieser Schäden stieg hierbei mit der Temperatur an (Hainzlmaier, 2005, S. 55 ff.). Dies deckt sich mit Untersuchungen zur Verringerung der Eishärte bei ansteigenden Temperaturen. Bei diesen Tests wurde eine Stahlkugel aus unterschiedlichen Höhen und bei unterschiedlichen Eistemperaturen auf die Eisschicht fallen gelassen und Tiefe sowie Breite des entstehenden Schadens gemessen (Poirier *et al.*, 2011, S. 130).

Ein zu hoher Druck kann allerdings den Wasserfilm zwischen Eis und Kufe nach außen pressen, wodurch der Reibungskoeffizient wieder ansteigt. Es entsteht zusätzliche Reibungshitze, die Schäden im Eis verursacht (Hainzlmaier, 2005, S. 61f.). Da solche Fahrrillen nicht oder nur kaum auf geraden Streckenabschnitten gefunden wurden, geht Hainzlmaier davon aus, dass der Druck auf geraden Streckenabschnitten unter dem Optimum für den Reibungskoeffizienten liegt und in Kurven tendenziell darüber. An dieser Stelle sei jedoch angemerkt, dass diese Theorie im Kontrast zu den Ergebnissen von Scherge *et al.* steht, welche für steigende Normalkraft einen nahezu konstanten Reibungskoeffizienten festgestellt haben. Da der Reibungskoeffizient dort mit ansteigender Normalkraft bei -5 Grad

Celsius nahezu konstant blieb und die Spannweite der Normalkraft explizit die Belastung in Kurven simulieren sollte, wird daher davon ausgegangen, dass der von Hainzlmaier postulierte Bremseffekt kaum oder gegebenenfalls nur bei Temperaturen nahe Null relevant sein könnte und nicht so stark ausgeprägt ist, wie seine Abbildung suggeriert.

Abbildung 3: Reibungskoeffizient zu Normalkraft



Quelle: in Anlehnung an Hainzlmaier 2005, S. 62.

1.3 Modellierungen

Im Jahr 1995 konstruierten Zhang *et al.* ein Modell unter Verwendung mehrerer Algorithmen zur mathematischen Bestimmung des optimalen Lenkverhaltens im Bobsport. Sie widmeten sich damit dem Optimierungsproblem einer möglichst effizienten Fahrweise am Beispiel der Olympiabahn in Lillehammer. Ihre Ergebnisse beinhalten insbesondere die Darstellung des Zusammenhangs zwischen optimaler Fahrlinie und minimaler Lenkbewegungen in verschiedenen Sektoren der Bahn (vgl. Zhang, Y.L. *et al.*, 1995).

Die im Jahr 1997 von Brüggemann *et al.* durchgeführte Untersuchung der Fahrdynamik verschiedener Bobschlitten im Eiskanal der 17. Olympischen Winter-

spiele in Lillehammer bestätigen die Erkenntnisse von Zhang *et al.* (1995). Angewandte statistische Analysen der Läufe führte sie zu der Erkenntnis, dass insbesondere der Start eine wichtige Determinante für die Zieleinfahrtszeit ist. Eine saubere Fahrweise in Kurven und auf Geraden ist nach ihren Ergebnissen zwar positiv für die Verringerung der Zieleinfahrtszeit, jedoch ist die Startgeschwindigkeit ihr signifikantester Parameter (Brüggemann, G.-P. *et al.*, 1997).

Zanoletti *et al.* (2006) nutzten eine Median-Split Technik, um die Weltcups im Skeleton der Saison 2003 bis 2004 zu analysieren. Mit manuellen Analysen zur Physis der Sportler und Sportlerinnen verglichen Sie die Sportarten Skeleton, Bob und Rennrodeln. Sie untersuchten den Zusammenhang zwischen Startgeschwindigkeit und Zieleinfahrtszeit unter Betrachtung der Anatomie und Erfahrung der Athletinnen und Athleten. Sie bestätigen eine signifikante Abhängigkeit der Zieleinfahrtszeit von der Startzeit.

Im Bereich der Tribologie analysieren Kietzig *et al.* (2010a) die Gleiteigenschaften verschiedener Materialen unter mehreren Umweltbedingungen. Dabei identifizierten sie insbesondere die Wichtigkeit einer geringen thermischen Leitfähigkeit des Kufenmaterials bei geringen Geschwindigkeiten und einen signifikant höheren Einfluss des Wasserfilms bei höheren Geschwindigkeiten. Ähnlich dazu entwickelten Schleinitz *et al.* ein Modell zur Ermittlung des Reibungskoeffizienten am Beispiel des Zweierbobschlittens. Sie nutzten einen sigmoidalen Zusammenhang zwischen Anpressdruck und Reibung. Positiv wirkte sich bei steigendem Druck die Bildung eines Wasserfilms aus. Bei zu hohem Anpressdruck, bei dem der Wasserfilm durch Auspressen zu dünn wird, nimmt die Reibung wieder zu (Vgl. Kietzig *et al.*, 2010a; Irbe *et al.*, 2021; Schleinitz *et al.*, 2022).

Rempfler und Glocker erstellten 2016 eine realitätsnahe Simulation der olympischen Strecke in den Whistler Mountains in Kanada. Mit ihrem Modell, welche eine Berechnung des Reibungskoeffizienten, das Lenkverhalten und die Bobposition enthielt, erreichten sie realitätsnahe Verhältnisse (Vgl. Rempfler & Glocker, 2016.).

1.4 Datenaufzeichnung in Winterberg

Die in dieser Arbeit untersuchten Daten wurden auf der Bahn in Winterberg gemessen. Der datengenerierende Prozess wird durch diese Arbeit nicht behandelt. Die Sportstätte am Standort Winterberg, welche im Jahr 1977 als befestigte Eisbahn eröffnet wurde, ist 1.330 Meter lang und beginnt auf einer Höhe von 760 Metern. In der Abfahrt werden über den Verlauf von 15 Kurven insgesamt 110

Höhenmeter zurückgelegt, wodurch ein durchschnittliches Gefälle von 9° und ein maximales von 15° erreicht werden. Seit ihrer Eröffnung wurde die Wettkampfstätte mehrfach umgebaut und erweitert, um die jeweils gültigen Wettkampfbedingungen und Sicherheitsvorgaben zu erfüllen (vgl. IBSF, 2019) .

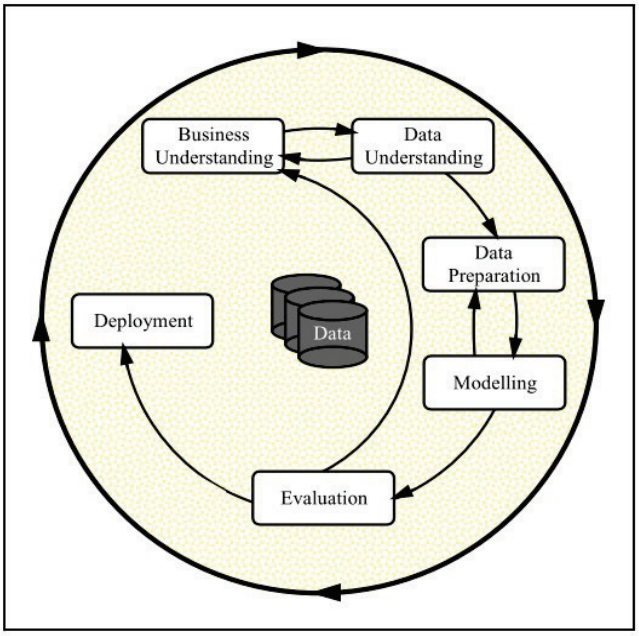
Auf Basis umfassender Messdaten konnten sowohl Fahrten im Monobob, Zweierbob als auch Viererbob analysiert werden. Zusätzlich wurden auch Videodaten ausgewertet. Ziel war es zum einen, das Fahrverhalten zu analysieren und zum anderen zu validieren, in welchem Ausmaß das Lenken die Zeiten determiniert. Zum Zwecke der Gegenüberstellung sind sowohl Trainingsdaten wie auch Fahrten der in Winterberg ausgetragenen Weltmeisterschaft ins Verhältnis gesetzt worden.

Im erweiterten Projektkontext wurde der Zusammenhang von Fahrlinie und Laufzeit in spezifischen Bahnabschnitten untersucht. Partiell wurden auch kinetische Änderungen der Bewegungsgrößen in die Betrachtungen einbezogen. Die Big Data Analyse erfolgte mittels Modellvariationen und divergent strukturierten Parameterwerten.

1.5 Datentechnische Vorgehensweise

Alle hier vorgestellten Analysen basierten auf dem Cross Industry Standard Process für Data Mining (CRISP-DM) nach Wirth und Hipp (2000). Dabei kam ein mehrphasiger, zyklischer Prozess eines strukturierten Data Minings zur Anwendung. Diese Methode strukturiert die Aufbereitung des Wissens. Dafür wurden insgesamt sechs Phasen definiert (vgl. Wirth & Hipp, 2000).

Abbildung 4: Prinzip des CRISP-DM



Quelle: Entnommen aus Wirth & Hipp, 2000.

Die erste Phase des CRISP-DM schreibt die Erarbeitung eines angemessenen Verständnisses der zugrundeliegenden Prozesse und Anforderungen vor. Unter Business Unterstanding wird in dieser Arbeit in einem ersten Schritt die Durchführung von fachlichen Gesprächen mit Fachspezialistinnen und -spezialisten (Wissenschaftlerinnen und Wissenschaftler, Trainerinnen und Trainer, Athletinnen und Athleten) verstanden. Aufbauend auf dieser thematischen Einführung wurde eine strukturierte Literaturrecherche in verschiedenen Literaturbibliotheken durchgeführt. Für die inhaltliche Bearbeitung ist die Literaturrecherche ausschlaggebend. Der fachliche Austausch wird nicht als Expertengespräch, sondern als fachliche Diskussionsbasis und zur Definition der Untersuchungsziele verstanden.

Mit der Phase des Data Understandings führen Wirth und Hipp die Einarbeitung in die fachlichen Hintergründe der Daten als zweitem Schritt des CRISP-DM auf. Sie empfehlen einen iterativen Austausch zwischen den ersten beiden Phasen, um gleichermaßen die benötigten fachlichen Hintergründe sowohl aus fachlicher Sicht als auch aus Sicht der Daten bewerten zu können. Dabei orientiert sich

diese Arbeit an den technischen Voraussetzungen des Versuchsaufbaus, welcher die Daten generiert und beginnt mit einer explorativen Analyse der verfügbaren Daten (vgl. Wirth & Hipp, 2000).

Nach Wirth und Hipp wird anschließend zur Phase des Data Understandings die Phase Data Preparation begonnen. Die dritte Phase des CRISP-DM setzt die Erkenntnisse der ersten beiden Phasen in eine fachliche Datenaufbereitung um. Dabei werden insbesondere gefundene Datenqualitätsprobleme mit angemessenen Maßnahmen behandelt. Die Phase der Data Preparation wird iterativ mit der vierten Phase des Modellings durchgeführt. Wirth und Hipp beschreiben hierbei die Modellentwicklung zur Bearbeitung der Forschungsfragen und Hypothesen. Die dabei entstehenden Ergebnisse werden in der fünften Phase, der Evaluation, ausgewertet und anschließend werden die Erkenntnisse als Eingang zur Erweiterung des Business Understandings verwendet. Auf die Phase des Deployment verzichtet diese Arbeit und führt die notwendigen Erkenntnisse in das Projekt zurück (vgl. Wirth & Hipp, 2000).

2 Literaturrecherche

2.1 Zielsetzung

Für wissenschaftliche Literatur im Bereich des Bobsports gilt, wie für sämtliche Fragestellungen und Disziplinen, dass die Menge der publizierten Ausarbeitungen einen Umfang erreicht, der durch konventionelle Methoden nicht umfassend berücksichtigt werden kann. Werden etwa Langzeittrends oder eine Übersicht auf der Metaebene angestrebt, ist oft langjährige Fachexpertise nötig. Die vorliegende Ausarbeitung nutzt das Konzept des „doppelten Trichters der Künstlichen Intelligenz“, um sich einen Überblick über die Datenflut zu verschaffen.

Der „doppelte Trichter der Künstlichen Intelligenz“ wurde im Mai 2020 von Rüdiger Buchkremer *et al.* (2019a, 2018b) vorgestellt und ist somit ein recht neuer Ansatz, der durch Anwendungen, wie die vorliegende und daraus gewonnene Praxiserfahrung, weiterentwickelt werden kann. Im Rahmen dieser Methodik werden Textanalyse- und Visualisierungstechniken angewendet, um das Ziel (Extraktion von Wissen aus einem Datenkorpus) zu erreichen und die eingesetzte Suchtaxonomie aufzuwerten.

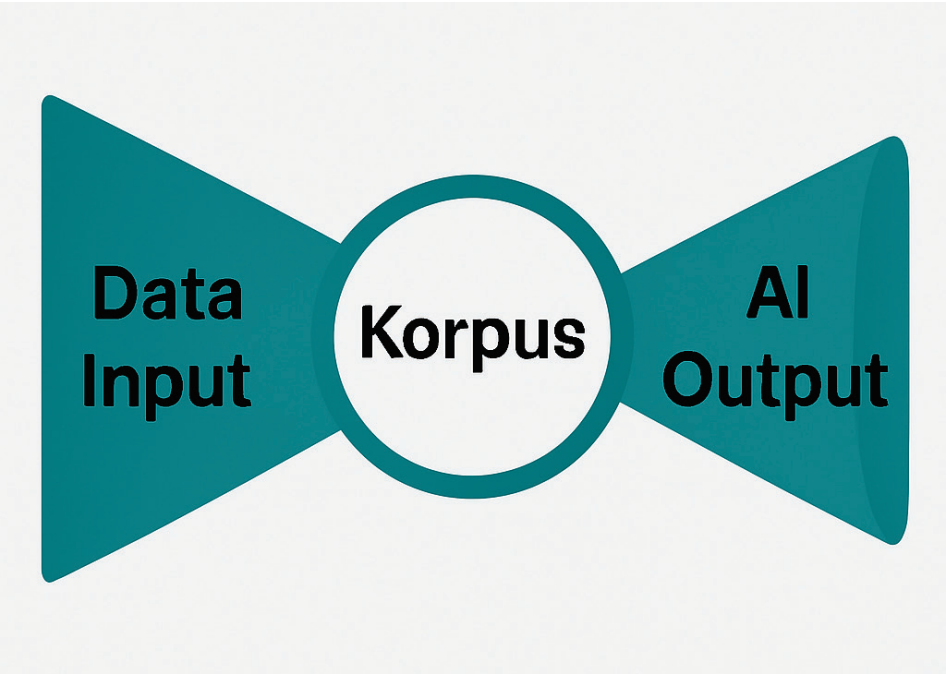
Die genaue Funktionsweise dieser Methodik soll im weiteren Verlauf verdeutlicht werden. Um dies zu erleichtern, beruht auch der formelle Aufbau dieser Ausarbeitung auf der strukturellen Abfolge des „doppelten Trichters der Künstlichen Intelligenz“.

Innerhalb des Ergebnisteils soll eine spezifische Auswahl an Fachliteratur zum Thema Bobsport ausgewertet und visualisiert werden. Die abschließend in der Literatur identifizierten Entwicklungen sollen dabei helfen, weitere Fachartikel und die bisher gewonnenen Erkenntnisse, anhand des Kontextes zum Entstehungszeitpunkt besser einzuordnen.

2.2 Vorgehensweise und Aufbau der Literaturrecherche

Die Vorgehensweise und der Aufbau der Literaturrecherche sind in Abbildung 5 visualisiert.

Abbildung 5: Der doppelte Trichter der Künstlichen Intelligenz



Quelle: vgl. Buchkremer *et al.*, 2019.

So soll in diesem Kapitel eine kurze Übersicht der Teilschritte und im jeweiligen Unterkapitel eine theoretische Einführung erfolgen. Eingangs wird der Prozess, durch welchen die zu verarbeitenden Daten gewonnen werden, ausführlich beschrieben. Daran anschließend wird ein erstes Preprocessing durchgeführt. Um die Taxonomie in einem späteren Arbeitsschritt aufzuwerten, wird ein Clustering und eine Latent Dirichlet Allocation (LDA) durchgeführt. So wird es möglich, Verbindungen zwischen Themen sowie Knoten und Kanten, die für eine Netzwerkanalyse benötigt werden, zu identifizieren. Der nächste Abschnitt ist erneut dem Bereich der Datenvorverarbeitung zuzuordnen. Es werden eine Reihe an Bereinigungen, die im Bereich des „Natural Language Processing“ (NLP) üblich sind, Anwendung finden. Beispiele sind die Entfernung von Stopwords, Duplikaten und von Zeichen ohne inhaltlichen Mehrwert wie Hypertext Markup Language Text (HTMLTags) oder Nummern. Somit ergibt sich ein vorläufiger Datenkorpus der als Grundlage für alle in dieser Arbeit eingesetzten Methoden dienen soll. Ab dieser Stelle ergibt sich eine Abweichung zum „doppelten Trichter der Künstlichen Intelligenz“. Die innerhalb dieser Methodik folgenden Schritte der Evolution

sowie Topic-, Trend- und prädiktiven Analyse werden bei dem durchgeführten Vorgehen nicht Bestandteil der Ausarbeitung sein.

2.3 Literaturrecherche und Datenaufbereitung

Ziel der Literaturrecherche war es, möglichst viele wissenschaftliche Arbeiten, Artikel oder sonstige Texte mit Bezug zum Thema Bobsport aus den zur Verfügung stehenden Quellen zu extrahieren. Für den Datenkorpus wurden drei unterschiedliche Quellen verwendet, welche nun genauer betrachtet werden sollen. Zunächst wurden alle gängigen Internetportale wie

- Institute of Electrical and Electronics Engineers (IEEE),
- SpringerLink,
- Ebsco und
- Google Scholar

mit Keywords zum Thema Bobsport durchsucht und relevante Artikel extrahiert. Um die Artikel verwalten zu können, wurde das Tool Mendeley verwendet, welches mehrere Vorteile bietet. Zum Beispiel können die Texte in diversen Formaten exportiert und so weiterverarbeitet werden. Für die Extraktion der Publikationen sind allerdings nicht alle Attribute relevant, die spätere Auswertung wird nur anhand der Abstracts und dem Veröffentlichungsdatum der jeweiligen Publikationen erfolgen. Wie Buchkremer *et al.* (2019a, 2019b) gezeigt haben, verfügen die Abstracts über die höchste Informationsdichte und lassen sich daher am besten auswerten.

2.3.1 Suchtaxonomie

Bei der Suche nach geeigneter Literatur ist es essenziell, die richtigen Schlüsselworte zu verwenden, da Portale wie IEEE, Ebsco, SpringerLink oder Google Scholar über umfangreiche Bibliotheken verfügen und sehr viele Artikel bei der Suche aufgelistet werden. Eine einfache Recherche mit dem Suchbegriff Bobsport ist daher nicht zielführend und lieferte in keinem der Portale ein zufriedenstellendes Ergebnis. Oftmals musste manuell geprüft werden, welche Artikel eine Relevanz für dieses Projekt haben, da sehr viele Publikationen aus fachfremden Bereichen gefunden wurden. Die meisten Publikationen sind in englischer Sprache formuliert bzw. selbst nicht englischsprachige Artikel besitzen meist ein englischsprachiges Abstract, daher müssen auch die Suchbegriffe englischsprachig sein. Bei der Recherche genügte es nicht, nur nach Keywords wie

„bob“ oder „bobsleigh“ zu suchen, es mussten auch weitere Suchbegriffe ergänzt werden, um möglichst viele Publikationen zu erhalten. So wurden dem Bobsport verwandte Sportarten wie „Skeleton“ und „Luge“ oder andere relevante Themen wie „surface“, „speed“ oder „winter“ hinzugefügt. Mit diesem Pool von Suchbegriffen und den daraus gebildeten Variationen wurden die Portale durchsucht, um eine breit gefächerte Datenbasis zu erhalten.

Die final angewandte Suchtaxonomie und die Anzahl der gefundenen Publikationen stellen sich wie folgt dar:

Tabelle 1: Wordcloud aus Suchwörtern in Abstracts generiert. Die Größe der Wörter repräsentiert die Häufigkeit des Erscheinens

Plattform	Suchtaxonomie	Anzahl Treffer
Ebsco	"Bobsleigh" OR "bobsled"	9.499
Ebsco	"Bobsleigh" OR"bobsled" AND "Speed"	6.065
Ebsco	"Bobsleigh" OR"bobsled" AND "Surface"	4.932
Ebsco	"Luge" AND "sport" AND "winter"	2.347
Ebsco	"Skeleton" AND "sport" AND "winter"	10.679
IEEE	"Bobsleigh" OR"bobsled" AND "Speed"	34
IEEE	"Bobsleigh" OR"bobsled" AND "Surface"	23
Scholar	"Bobsleigh" OR"bobsled" AND "Speed"	3.850
Scholar	"Bobsleigh" OR"bobsled" AND "Surface	2.860
Springer	"Bobsleigh" OR "bobsled" AND "Speed"	255
Springer	"Skeleton" AND "sport" AND "winter"	3.108
Springer	"Luge" AND "sport" AND "winter"	1.288

2.3.2 Data Preprocessing

Nachdem die wissenschaftlichen Artikel aus den Portalen identifiziert wurden, müssen diese an einem zentralen Ort gespeichert bzw. verwaltetet werden. Mendeley bietet eine Übersicht über die gesammelten Publikationen und es können auch eventuell fehlende Attribute in den Publikationen ergänzt werden. Oftmals war es nach dem manuellen Import von Artikeln notwendig, Attribute wie Autorennamen oder das Veröffentlichungsdatum zu ergänzen. Da die Daten aus mehreren Quellen stammen, war eine manuelle Vervollständigung oftmals notwendig. Nachdem die Literatur im comma-separated-values Format (csv-Format) vorlag,

musste diese nach Mendeley überführt werden. Zunächst wurden die Digital Object Identifier System Nummern (DOI Nummer) aus den verwendeten csv-Dateien gefiltert, um damit in Mendeley einen neuen Eintrag anzulegen. Oftmals war es jedoch notwendig, noch eine Internetrecherche für die DOI Nummer durchzuführen, da Mendeley nicht automatisch die Abstracts ergänzt. Nachdem die Artikel und die Metadaten vervollständig wurden, war die initiale Vorverarbeitung abgeschlossen.

Im nächsten Schritt wurden die Daten exportiert. Mendeley bietet hierfür eine Funktion an, um Artikel in diversen Formaten zu exportieren. Dabei ist es jedoch nicht möglich, nur einzelne Attribute zu verwenden, es werden immer alle Attribute, die zu einem Paper zur Verfügung stehen, exportiert. In einem Zwischenschritt wird der Mendeley-Export in ein csv-Format konvertiert, so können die Daten leichter in der späteren Auswertung weiterverarbeitet werden.

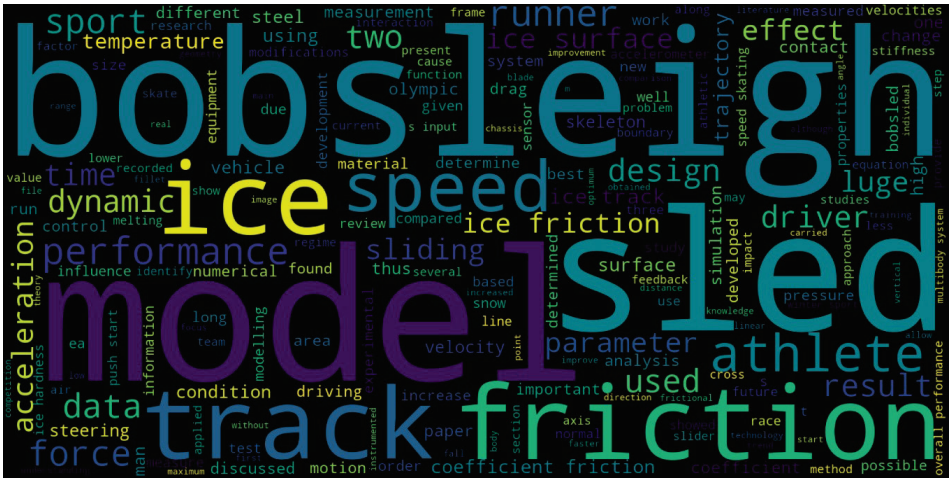
Python bietet die *pandas* Library an, mithilfe derer die Daten nicht nur im csv-Format eingelesen, sondern in das sog. Dataframe Format überführt werden. Ein Dataframe ist dabei eine Darstellung in tabellarischer Form, welche leicht manipuliert werden kann. Da nicht alle eingelesen Elemente für die Auswertung notwendig sind, müssen diese noch gefiltert werden. Für die Analyse sind besonders das Datum der Veröffentlichung sowie das Abstract aus der jeweiligen Veröffentlichung relevant.

2.3.3 Deskriptive Datenanalyse

Nachdem die Daten bereinigt waren, musste sich ein Überblick über diese verschafft werden. Dies ist ein unerlässlicher Schritt, welcher auch die Auswertung vereinfacht. Da die Publikationen in Mendeley verwaltet werden, sind dort beim Import automatisch eine Reihe von Attributen wie Titel, Autoren, Abstract, Jahr der Veröffentlichung, Institut, DOI-Nummer, Journal, ggf. Seitenzahl oder der Verleger ergänzt worden. Für eine Auswertung der Abstracts, im Rahmen dieser Arbeit, sind jedoch nicht alle Attribute relevant und konnten teils vernachlässigt werden. Es wurden nur Publikationen betrachtet, welche ein englischsprachiges Abstract besitzen. Bei den Publikationen handelte es sich fast ausschließlich um Journalartikel, vereinzelt wurden Berichte von Tagungen verwendet. Alle Publikationen wurden in einem Zeitraum von 1991 bis 2022 veröffentlicht. Besonders auffällig war, dass der größte Teil der Publikationen in den 2010er Jahren, also 2010 bis 2019, veröffentlicht wurde.

Im nächsten Schritt konnten, nach kurzer Bearbeitung der Abstracts, quantitative Aussagen über die Wörter in den Abstracts getroffen werden. Zunächst wurden die Abstracts tokenisiert. Darunter wird verstanden, einen Text in eine Liste von Wortbestandteilen und Zeichen zu überführen. Danach kann diese Liste von Stopwords befreit werden, da diese keinen Mehrwert für die Auswertung bieten. Stopwords sind dabei Wörter wie „between“, „we“ oder „some“. Die Python Library *nltk* bietet für diesen Zweck eine Liste von Stopwords, auf welche zurückgegriffen werden kann. Nach dem Entfernen der Stopwords können die relevanten Keywords identifiziert werden und man kann diese für weitere Auswertungen verwenden. Als erster Orientierungspunkt über die vorkommenden Keywords wurde eine Wordcloud gebildet. Mithilfe der Wordcloud Library wurde das Schaubild mit den nachstehenden Befehlen generiert.

Abbildung 6: Wordcloud aus Suchwörtern in Abstracts generiert. Die Größe der Wörter repräsentiert die Häufigkeit des Erscheinens



2.3.4 Vorbereitende Methodiken für eine Topicmap

Nachdem die Daten als Dataframe eingelesen wurden, müssen diese noch bereinigt und transformiert werden, damit sie mithilfe des LDA-Algorithmus zu Topics zusammengefasst werden können. Dazu sind allerdings verschiedene Zwischenschritte notwendig, um die Daten in ein geeignetes Format zu überführen, welches vom LDA Algorithmus akzeptiert wird. Da die Daten aus diversen Quellen stammen und die Vollständigkeit nicht gewährleistet ist, wird geprüft, ob leere Felder sog. NA-Values im Dataframe vorhanden sind. Für die Auswertung der

Abstracts wurden fehlende Einträge mittels der Methode *dropna()* entfernt. Diese wird mithilfe der *subset()*-Methode ergänzt, um nur Einträge aus dem Dataframe zu löschen, welche kein Abstract besitzen. Insgesamt können so alle NA-Values gelöscht werden und es ergibt sich der Ausdruck: *papers.dropna(subset=[‚Abstract'])*.

Anschließend wurden alle Wörter nach lower case formatiert. Dabei wurden alle Großbuchstaben mit Kleinbuchstaben ersetzt. Dieser Schritt ist notwendig, da später auch Stopwords gefiltert werden sollen. Eine Unterscheidung von Groß- und Kleinschreibung wäre auch möglich, war aber umständlich und nicht effizient. Danach wurden Sonderzeichen aus den Artikeln entfernt, da sonst nicht alle Stopwords erkannt und letztendlich das Endergebnis verfälscht wäre. Anschließend konnten die Stopwords herausgefiltert werden. Ohne diesen Schritt würden Wörter wie z. B. „the" oder „and" das Ergebnis verfälschen, da diese besonders oft verwendet werden. Im letzten Schritt wurden die einzelnen Abstracts in eine Liste von Worten umgewandelt. Mithilfe der Python Library *gensim* kann aus den einzelnen Abstracts ein sog. Dictionary erstellt werden. Ein Dictionary wird dazu verwendet, um Paare von Key:Values zu speichern. Das Dictionary speichert hierbei die Keywords und wie oft diese in allen Abstracts vorkommen. Somit ist die notwendige Bereinigung der Daten abgeschlossen. Nun müssen die Daten noch transformiert bzw. zusammengefasst werden, um den LDA-Algorithmus auszuführen zu können. Welches Format dafür notwendig ist wird nun näher erläutert.

Die *gensim* Library bietet die LDA-Methode an, welche drei Parameter verwendet, um das LDA Modell zu erstellen. Zum einen eine Liste, welche aus den einzelnen Wörtern aus allen Abstracts besteht. Zum anderen eine Liste, welche aus den einzelnen Worten der Abstracts besteht und der Anzahl, wie oft diese Worte vorkommen, z. B. [‚ice'→3]. Diese beiden Listen und die Anzahl der Topics wurden nun verwendet, um die LDA -Auswertung zu vollziehen.

2.4 Theoretische Grundlagen: Netzwerkanalyse mit LDA

Im späteren Verlauf der Arbeit soll der untersuchte Datenkorpus visualisiert werden. Es gibt verschiedene Möglichkeiten, eine solche Visualisierung zu ermöglichen, eine dieser ist die „Latent-Dirichlet-Allokation" (LDA). LDA wird als intuitiver Ansatz zur Berechnung der Ähnlichkeit zwischen Texten verwendet, um die jeweiligen Verteilungen der einzelnen Texte über die Themen zu erhalten. Die Grundidee ist, dass die Dokumente als Zufallsmischungen über latente Themen

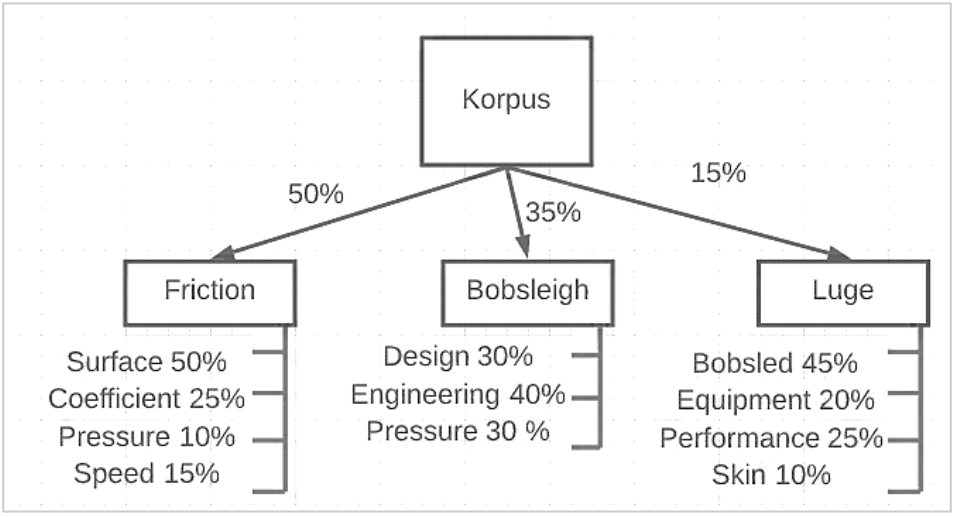
dargestellt werden, wobei ein Thema durch eine Verteilung über Wörter charakterisiert wird. Die Wörter mit den höchsten Wahrscheinlichkeiten in jedem Thema vermitteln in der Regel eine gute Vorstellung davon, welches Thema ausgewählt werden sollte.

Bei diesem Vorgehen handelt es sich um einen Prozess, der Abhängigkeiten zwischen einer Vielzahl von Variablen beschreibt. Es existieren verschiedene Algorithmen, welche die Problemstellung einer zu hohen Menge an Variablen adressieren.

Einige häufig verwendete Algorithmen sind „Gibbs sampling“, „Expectation maximization“ und „Variational Bayes inference“. Sie alle bestimmen LDA-Paramater, die den Datenkorpus und Themen der Texte beschreiben.

Beispielhaft wird im Weiterem ein fiktiver Korpus mit elf verschiedenen Wörtern betrachtet. Es soll angenommen werden, dass dieser Korpus eine Mischung aus drei Themen ist: „Friction“, „Bobsleigh“ und „Luge“. Jedes dieser Themen ist wiederum eine Mischung aus verschiedenen Wortsammlungen. Bei der Erstellung dieses Korpus wird zunächst ein Thema aus der Verteilung ausgewählt, und später wird aus dem ausgewählten Thema ein Wort aus den Wortverteilungen dieses Themas selektiert. Dieses Vorgehen beschreibt die Funktionsweise der LDA und wird in Abbildung 7 dargestellt.

Abbildung 7: LDA Topic Bildung



2.4.1 Clustering-Layout nach Modularität

Wie auf der offiziellen Seite der Open Graph Visualization Platform „Gephi“ ersichtlich ist, kann Modularität als ein Maß für die Struktur von Netzen oder Graphen verwendet werden (Jacomi *et al.*, 2014). Das Maß wurde entwickelt, um die Stärke der Aufteilung eines Netzes in Modulen (auch Gruppen, Cluster oder Gemeinschaften genannt) zu bemessen. Netze mit hoher Modularität haben dichte Verbindungen zwischen den Knoten innerhalb von Modulen, aber spärliche Verbindungen zwischen Knoten in verschiedenen Modulen. Modularität wird häufig in Optimierungsmethoden zur Erkennung von Gemeinschaftsstrukturen in Netzwerken verwendet.

Die gleiche Quelle beschreibt auch, dass sich gezeigt hat, dass die Modularität eine Auflösungsgrenze hat und daher nicht in der Lage ist, kleine Gemeinschaften zu erkennen. Biologische Netzwerke, einschließlich der Gehirne von Tieren, weisen ein hohes Maß an Modularität auf. Tatsächlich kann der rohe Modularitätswert nicht verwendet werden, um zu bestimmen, ob es sich um eine „gute“ Partition handelt, da ein ähnlich hoher Wert auch in einem Zufallsdiagramm erzielt werden kann. Beim Vergleich mit dem traditionellen Clustering-Koeffizienten ergibt sich, sobald die beiden Variablen unabhängig sind, eine Korrelation gleich Null, so dass jede Korrelation, die von Null verschieden ist, als „hoch“ eingestuft wird. Bei der Modularität ist dieser Wert für Zufallsgraphen jedoch nicht Null, sodass es schwierig ist, zu sagen, ob die Modularität „hoch“ und damit eine „gute“ Partition ist.

Da für die Visualisierung der Literaturdaten jedoch die Stärke des Zusammenhangs eine wichtige Kennziffer ist, reicht ein simples Clustering nicht aus, um die angestrebte Netzwerkanalyse durchzuführen.

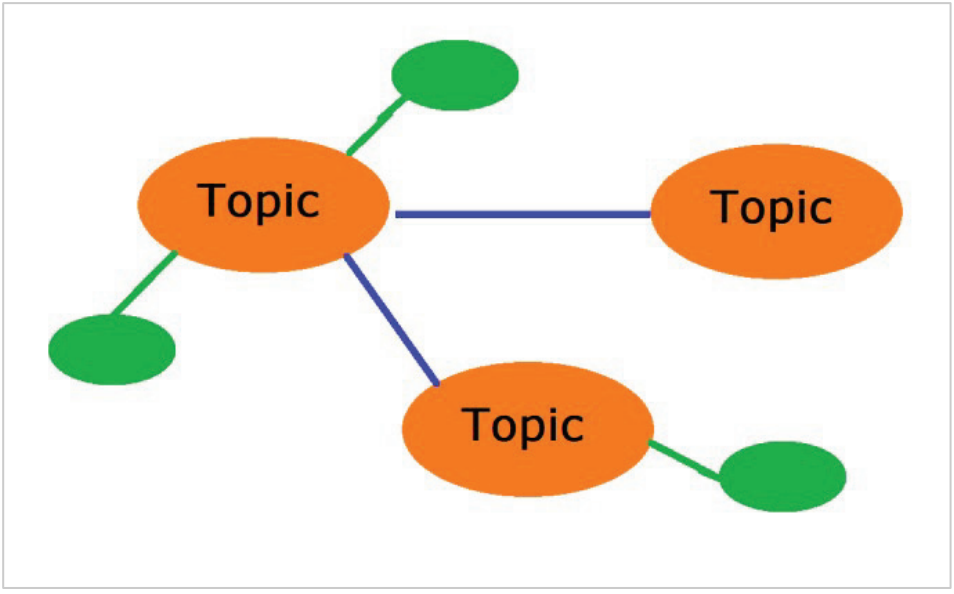
2.4.2 Topic-Map

Da im Verlauf des Projektes verschiedene Auswertungsmethoden angewendet wurden, die alle zur Kategorie Topic Modelling gehören, sollten sich die Visualisierungsmethoden daran orientieren bzw. geeignet sein, um diesen Sachverhalt darzustellen. Eine Möglichkeit, um Wissensstrukturen zu formulieren, vermitteln und darzustellen bietet die sog. „Topic-Map“ (vgl. Qiu & Yuan, 2014). Die „Topic-Map“, die ursprünglich als eine disziplinspezifische Form der Concept Map entwickelt wurde, ist im Allgemeinen eine Darstellung von Konzepten über Informationen und Beziehungen und nicht des Wissens selbst. Themenkarten dienten

der Darstellung von Back-of-the-Book-Indexstrukturen, damit mehrere Indizes aus verschiedenen Quellen zusammengeführt werden können.

Bei einer Darstellung durch eine Topic-Map handelt es sich um eine besondere Form eines Netzwerks, in dem es n-Knoten gibt, welche Topics repräsentieren. Topics können dabei Gegenstände oder Sachverhalte sein. Untereinander sind die Topics miteinander verbunden. Des Weiteren gibt es x-Knoten, welche mit den einzelnen Topics verbunden sind. So lässt sich ein LDA-Modell genau abbilden, da alle Wörter eines Topics an ein Topic gebunden sind. Kanten sind entweder zwischen den Topics gezeichnet oder zwischen den Topics und den zugehörigen Wörtern.

Abbildung 8: Topic-Map



2.5 Analyse und Visualisierung des bereinigten Datenkorpus

Die Gesamtheit der zu analysierenden Inhalte machen den Datenkorpus aus. Diese können in unterschiedlichsten Formen und Größen vorliegen. In jedem Fall enthalten sie jedoch Text sowie in einigen Fällen Metadaten und stellen gleichzeitig eine Reihe von zusammenhängenden Inhalten zu einem bestimmten Themenkomplex dar. In seiner Rohform ist der Datenkorpus häufig strukturlos. Die Inhalte sind kaum gebündelt und können in unterschiedlichen Speicherplätzen

liegen. Die Relation der verschiedenen Dateien und Texte muss erst identifiziert und verifiziert werden.

Wie Buchkremer *et al.* aufzeigen, ist bei der Anwendung des „doppelten Trichters der Künstlichen Intelligenz" die Qualitätssicherung des Datenkorpus essenziell. Hierzu wird der Korpus mit Metadaten angereichert und harmonisiert, indem Duplikate entfernt sowie Fehler bei der Konvertierung und Formatierung der Datensätze bereinigt werden. Als Metadaten können Schlüsselwörter oder Tagging-Informationen, die von den Autorinnen und Autoren bereitgestellt werden oder durch einen Vergleich mit bekannten Taxonomien, Katalogen oder Ontologien entstehen, herangezogen werden.

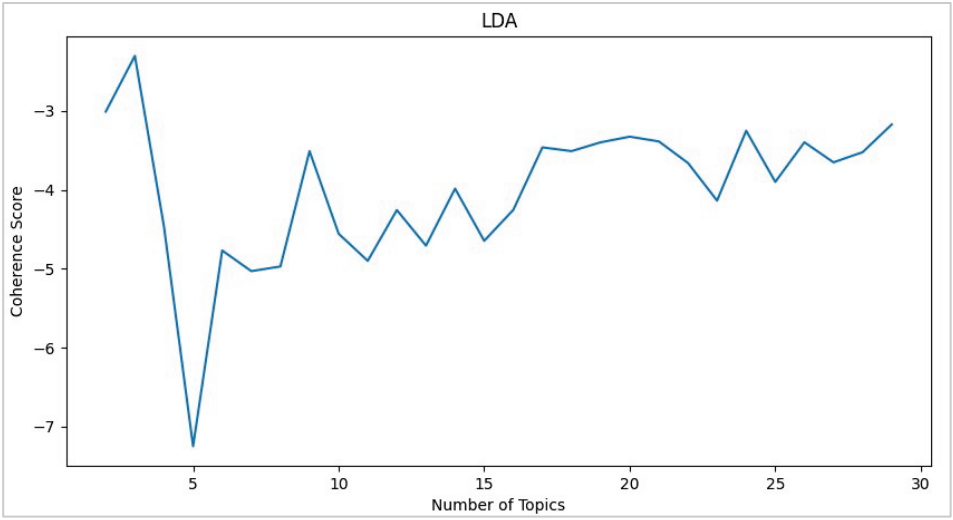
Der vorläufige Datenkorpus ist das Ergebnis aus den durchgeführten Bereinigungen und enthält 521 Dokumente.

2.5.1 LDA-Modell generieren

Bisher konnten die Daten vorbereitet werden, sodass sie nun dem LDA-Algorithmus zugeführt werden können. Es gibt nur noch einen variablen Parameter, welcher der LDA-Methode übergeben werden muss. Dabei handelt es sich um die Anzahl der Topics, welche vom Algorithmus generiert werden sollen. Die richtige Anzahl zu finden, ist keinesfalls ein triviales Problem und die vermeintlich beste Anzahl kann nur durch „Ausprobieren" herausgefunden werden. Die Anzahl automatisiert zu bestimmen, ist Gegenstand der Forschung in den letzten Jahren gewesen. So konnten Röder *et al.* bereits 2015 neue Ansätze definieren und implementieren, um die Anzahl der Topics bestmöglich zu schätzen. Um eine Abschätzung zu geben, gibt es das sog. Coherence Model welches den Coherence Score berechnet, welcher unter Verwendung der LDA Methode zwischen -14 und +14 liegt. Dieses Modell wird von *gensim* bereitgestellt. Um den besten Coherence Score zu bestimmen, wird dazu mehrfach, hier mit 30 Iterationen, jeweils ein LDA-Modell erstellt und jeweils der Coherence Score berechnet. Das LDA-Modell wird mit folgendem Ausdruck berechnet, wobei *k* für den Laufindex von 0 bis 30 steht: *LdaModel(papers_corpus, num_topics=k, id2word=papers_dic)*.

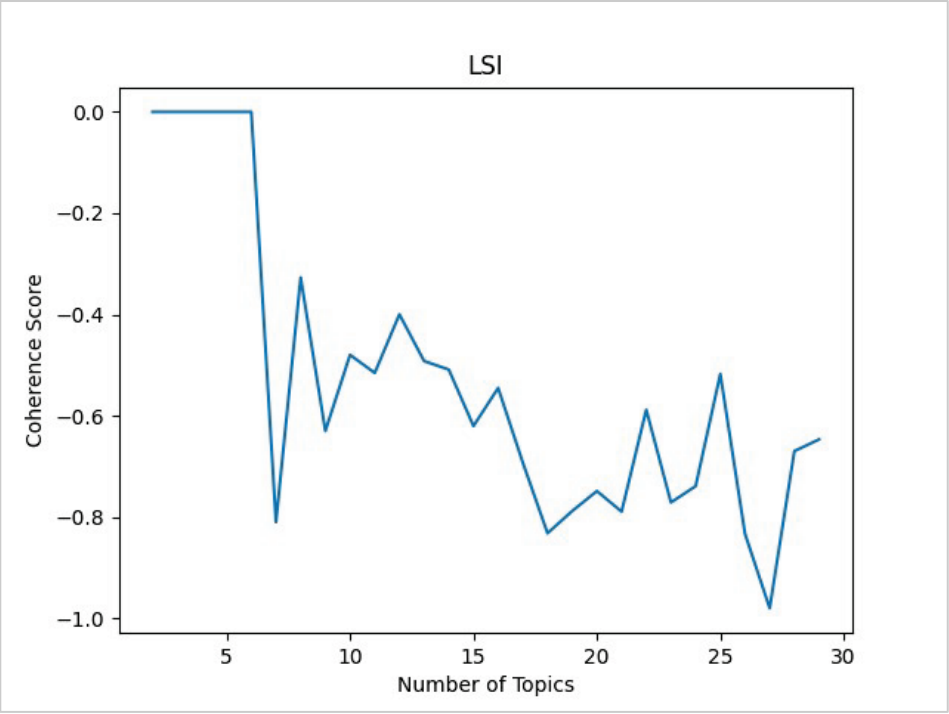
Nach insgesamt 30 durchlaufenden Iterationen und der Berechnung des Coherence Scores, ergibt sich die folgende Grafik. In diesem Fall liegt der höchste Score bei ca. 3,75, welcher 8 Topics entspricht.

Abbildung 9: Coherence Score mit LDA-Modell



Eine weitere Analyse für die optimale Anzahl an Topics kann mittels Latent Semantic Indexing (LSI) Modell vorgenommen werden. Die Ausführung ist nahezu identisch mit der Auswertung der LDA-Analyse und es kann ebenfalls ein Coherence Score berechnet werden. Dieser hat auch einen Peak bei 8 Topics, welcher sich mit dem Ergebnis der LDA-Analyse deckt. Daher werden nun acht Topics für die LDA-Auswertung gewählt.

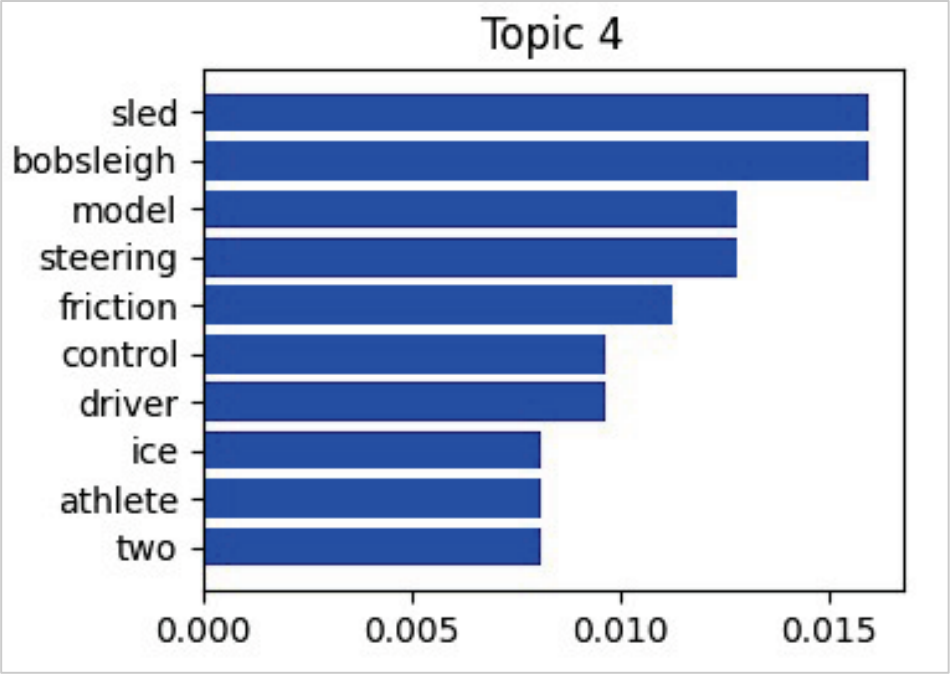
Abbildung 10: Coherence Score mit LSI-Modell



Die optimale Anzahl von acht Topics wurde zweifach verifiziert und für das LDA-Modell verwendet. Der Ablauf des Algorithmus soll nun näher erläutert werden.

Für jedes Topic werden alle Wörter aus dem Dictionary untersucht und geprüft, wie sehr sie zu einem Topic passen. Dabei erhält jedes Wort aus dem Dictionary eine Prozentanzahl basierend darauf, wie gut das Wort zum Topic passt. Da alle Wörter im Dictionary dem Topic zugeordnet werden, summieren sich diese Wahrscheinlichkeiten auf 1. So wird nun für alle Topics vorgegangen. Als Endergebnis erhält man also eine Liste pro Topic mit allen Wörtern und deren Wahrscheinlichkeiten. Ein weiterer Aspekt ist, dass LDA selbst keine Topicnamen vergibt. Diese werden lediglich nummeriert. Anbei wurde hier Topic 4 mit den 10 Worten, mit der höchsten Wahrscheinlichkeit, visualisiert. Dabei kann deutlich erkannt werden, wie sich hier Wörter um einen „Übergriff“ sammeln, beispielweise könnte hier der Topicname „Bobschlittensteuerung bei vereister Oberfläche“ verwendet werden.

Abbildung 11: Topic aus LDA



Anschließend wurden die Topics noch in eine Topic-Map überführt. Dazu müssen die Wörter aus den Topics in eine Netzwerkform überführt werden. Diese Umwandelung soll nun näher erläutert werden. Mit den ersten zehn Wörtern aus jedem Topic wird eine Liste gebildet. Danach werden Bigrams und Trigrams aus dieser Liste generiert. Im nächsten Schritt wird berechnet, ob und in welcher Menge diese Bi- und Trigrams in den Abstracts vorkommen. Sollten die Bigrams weiter als 15 Wörter voneinander entfernt sein, wird dies nicht als Verbindung gewertet. Sollten Bigrams keine Verbindung vorweisen, werden diese einfach aus der Liste gelöscht und nicht weiter betrachtet, damit keine einzelnen Elemente in dem Netzwerk dargestellt werden. Als Ergebnis erhält man eine Liste mit zwei Wörtern und eine Angabe, wie oft es eine Übereinstimmung gab. Zuletzt wird die Liste der Python Library *networkx* übergeben. Diese kann mit wenigen Befehlen ein Netzwerk aus den vorher definierten Listen erstellen und exportieren.

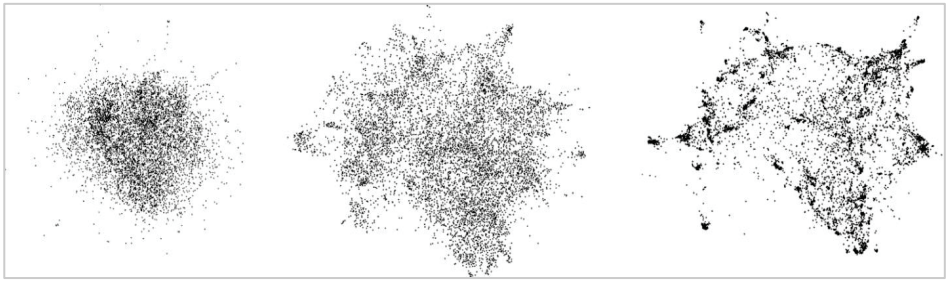
2.5.2 Netzwerkvisualisierung mit *Gephi*

Bei der Gegenüberstellung von Visualisierungsmethoden wie der „Wordcloud“ und einer „Topic-Map“ wird, sobald die gewichtete Anzahl der Wörter in einem Datenkorpus als alleiniges Merkmal zur Analyse genutzt wird, deutlich, dass keine Aussagen über inhaltliche Strukturen möglich sind, für die etwa Semantik oder Kontext berücksichtigt werden muss. Daher wird sich im Weiteren auf der Ebene der Netzwerkanalyse bewegt.

Gephi ist eine open-source Software zur Visualisierung von Netzwerken. *Gephi* bietet verschiedene Methoden und Algorithmen die dazu dienen, die Zusammenhänge zwischen den Themen innerhalb des Datenkorpus zu identifizieren. Eine dieser Methoden ist der in dieser Ausarbeitung verwendete *ForceAtlas2*. Die im Verlauf dieses Kapitels aufbereiteten Visualisierung beruhen auf dieser Methodik.

Jacomy *et al.* (2014) beschreiben *ForceAtlas2* als ein kräfteorientiertes Layout: Es simuliert ein physikalisches System, um ein Netzwerk auf einen Gegenstandsraum abzubilden. Knoten stoßen sich gegenseitig wie geladene Teilchen ab, während Kanten ihre Knoten anziehen wie Federn. Diese Kräfte erzeugen eine Bewegung, die zu einem ausgeglichenen Zustand konvergiert. Mit einer solchen Konfiguration soll die Interpretation der Daten erleichtert werden. Sollte sich im Weiteren auf den Term „Gravitation“ bezogen werden, ist das Layout *ForceAtlas2* referenziert. Abbildung 12 zeigt dieses Layout in verschiedenen Gravitationsstufen.

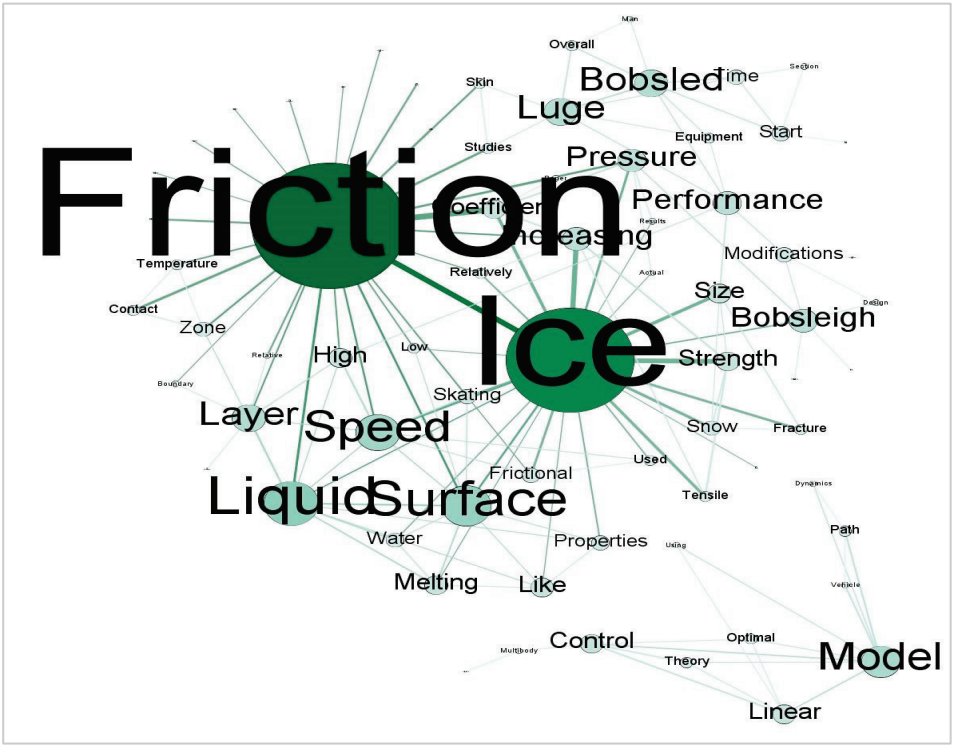
Abbildung 12: Layouts mit verschiedenen Gravitationsstufen



Anhand der zuvor beschriebenen LDA-Methodik und dem gravitationsbasierten Layout *ForceAtlas2* wird der vorläufige Datenkorpus durch folgende Netzwerke visualisiert. Literaturdaten, die im Zeitraum 2015 bis 2022 entstanden, nehmen

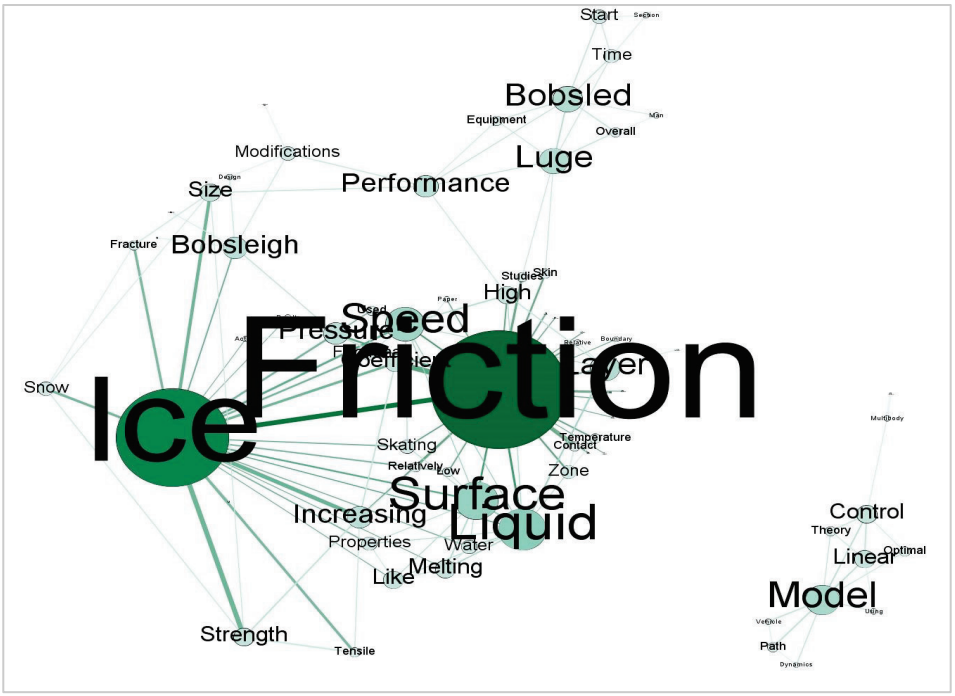
im Datenkorpus einen Anteil von 33 Prozent ein und dass resultierende Netzwerk ist in Abbildung 13 zu sehen.

Abbildung 13: Netzwerk Literaturdaten 2015 bis 2022



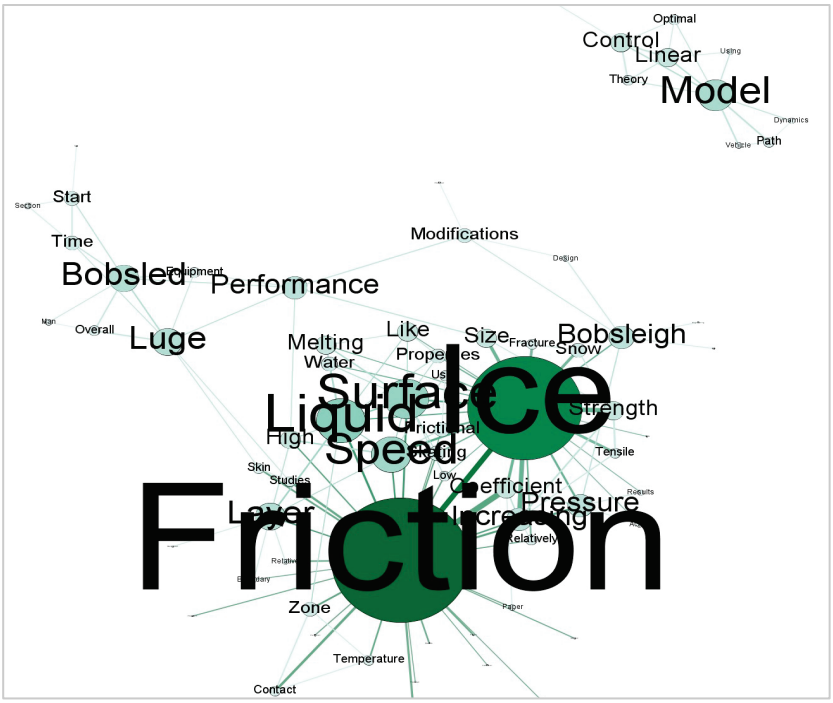
Literaturdaten, die im Zeitraum 2007 bis 2014 entstanden, nehmen im Datenkorpus einen Anteil von 45 Prozent ein und das resultierende Netzwerk ist in Abbildung 14 zu sehen.

Abbildung 14: Netzwerk Literaturdaten 2007 bis 2014



Literaturdaten, die im Zeitraum 1997 bis 2006 entstanden, nehmen im Datenkorpus einen Anteil von 22 Prozent ein und das resultierende Netzwerk ist in Abbildung 15 zu sehen.

Abbildung 15: Netzwerk Literaturdaten 1997 bis 2006



2.6 Fazit

Die gewählte Methodik des „doppelten Trichters der Künstlichen Intelligenz" wurde partiell angewandt, um einen Einblick in die ausgewählte Literatur zu erhalten und die aus dem Datenkorpus gewonnenen Erkenntnisse beruhen auf einem Datenkorpus von 521 Publikationen. Die innerhalb des „doppelten Trichters der Künstlichen Intelligenz" vorgesehenen Maßnahmen zur Datenvorverarbeitung haben einen Großteil des Umfangs dieser Ausarbeitung bestimmt. Aus der ausgewerteten Literatur lässt sich erkennen, dass die „Reduktion von Reibung" das dominierende Thema im Bobsport und in Arbeiten war und ist, deren Fokus auf den Möglichkeiten zur Reduktion der Laufzeit liegt.

Abseits dieser Beständigkeit lässt sich durch die Gravitation in der gewählten Netzwerkanalyse erkennen, dass lineare Modelle zur Erklärung dieser Reduktion nach wie vor Anwendung finden aber keine gesonderte Rolle innerhalb der Publikation einnehmen. Die Nähe von Texten zu linearen Modellen und dem Schwer-

punkt „Reibung/Geschwindigkeit reduzieren“ in Publikationen ab 2015 könnte darauf hindeuten, dass lineare Modelle in neuen Ausarbeitungen eher im Rahmen der Einleitung und Beschreibung der Fragestellung selbst genutzt wurden, während der Schwerpunkt in neuen Ausarbeitungen über lineare Modelle hinausgeht. Bis zum Jahr 2014 haben die Themen Streckendesign und Olympia ein deutliches Wachstum gezeigt, anschließend ist die Entwicklung stagniert. Die olympischen Winterspiele finden seit 1924 in einem Vierjahresturnus statt. 2014, 2018 und 2022 fanden die Olympiaden in Russland, Südkorea und China statt. Da in der durchgeführten Analyse nur englische Texte berücksichtigt wurden und ab 2014 kein Gastgeberland primär Literatur in englischer Sprache produziert, könnte dieser Rückgang durch die Standorte der olympischen Spiele erklärt werden.

Aufgrund der strikten Reglementierungen im Bobsport sind innovative Entwicklungen über einen zeitlichen Verlauf in den Publikationen kaum zu identifizieren. So wurden im Oktober 2018 zwar Änderungen im IBSF Reglement beschlossen, diese umfassten jedoch nicht die Ausrüstung oder Rennverläufe.

3 Fahrlinienanalyse auf Basis von Videodaten

3.1 Zielsetzung

Sportanalysen, insbesondere in Sportarten mit einer hohen medialen Aufmerksamkeit, fordern immer neue Ansätze der Betrachtung und Leistungsanalyse. Dazu gehört unter anderem die gezielte Videoanalyse und die damit verbundene Bildverarbeitung mit dem Ziel, neue Erkenntnisse in das Training einfließen zu lassen (Schlipsing *et al.*, 2013, S. 235). Laut Liebermann *et al.* im *Journal of Sport Sciences* aus dem Jahr 2002, haben die Fortschritte in der Informationstechnologie zu erweitertem und verbessertem Feedback bei Sportlerinnen und Sportlern geführt (Liebermann *et al.*, 2002, S. 756).

Die Videoanalyse ist laut Stein *et al.* (2018) verstärkt im Fußball vertreten. Daher haben sie, basierend auf Videoaufnahmen aus dem Fußball, Bewegungsdaten abgeleitet, analysiert und in Form von Wärmebildkarten (engl.: Heatmaps) visualisiert. Diese analytischen Visualisierungen werden mithilfe von Computer Vision Techniken in das Video eingebaut und erlauben somit eine Analyse des Fußballspiels (freie Räume, Bewegungen der Spieler etc.) im Kontext des Originalvideos. (Stein *et al.*, 2018, S. 13-15) Huang *et al.* (2019) wenden diese Technik auf den Tennissport an und haben mithilfe von Convolutional Neural Networks (CNN) und Deconvolutional Neural Networks (DeconvNet) aus aufeinanderfolgenden Bildern eine Heatmap der Position des Balls visualisiert. Mit einem F1-Score von 98,2 Prozent kann das trainierte Modell die Position sehr schneller und kleiner Objekte, wie im Tennis oder Badminton üblich, erkennen. (Huang *et al.*, 2019) Auch Koshkina *et al.* (2021) nutzten CNNs, um aus Videos von Eishockeyspielen wertvolle Visualisierungen und Statistiken für Trainerinnen und Trainer abzuleiten. Dazu werden die Videos einer stationären Kamera mithilfe eines CNNs analysiert, indem die Personen auf dem Eis erkannt und in Spieler und Schiedsrichter segmentiert werden. Im Nachgang werden dann, mithilfe des K-Means Algorithmus, Clusterzentren der zwei Teams geschätzt, in welchen die Spieler, basierend auf der Distanz des Spielers zu den Zentren, zugeordnet werden. Die daraus generierte Heatmap basiert auf 800 bis 900 Bildern pro Spiel und ermöglicht es, Trainern und Spielern die Verteilung des gegnerischen Teams im Spiel genauer analysieren und verstehen zu können (Koshkina, Pidaparthy & Elder, 2021). Des Weiteren haben auch Tora *et al.* (2027) mithilfe CNNs Videos aus dem Eishockeysport analysiert. Im Nachgang an die Analyse der Videos und der Extraktion der notwendigen Features aus den Bildern erfolgte mithilfe des LSTM-Algorithmus eine Klassifikation des Puck-Besitzes in fünf Events. Auch in

diesem Fall zielten Tora *et al.* (2017) auf die Nutzung der Ergebnisse im Spielkontext ab, um strategische Konzepte und einzelne Spieler konkreter zu analysieren (Tora *et al.*, 2017).

Laut Seidenschwarz *et al.* (2020) ist die Unterstützung der Trainer zur Analyse von Sportlern durch eine benutzerfreundliche Lösung bedeutend. Daher haben sie ein Videoanalysetool zur Analyse der taktischen Leistung von Fußballern entwickelt. Ziel war es, ein User Interface zu entwickeln, welches für eine Zielgruppe mit geringer IT-Affinität bestimmt ist (Seidenschwarz *et al.*, 2020). Kanth *et al.* (2018) wenden Ähnliches auf den Stabhochsprung an. Im *Journal of Image and Graphics* haben sie mithilfe der Bibliothek Open CV, die Schrittlänge, Geschwindigkeit sowie Kontaktzeit von Fuß und Boden bestimmt. Ein von ihnen entwickeltes GUI unterstützt bei der Videoanalyse (Kanth *et al.*, 2018). Des Weiteren beschreiben Kim und Um (2017) ein User Interface zur Analyse von Eishockeyspielen. Die Darstellung der Ergebnisse der Videoanalyse erfolgt in Form von Heatmaps und Diagrammen (Kim & Um, 2017).

Durch den Fortschritt in der Informationstechnologie, insbesondere in den Computer Vision Technologien, ergeben sich für diverse Sportarten neue Möglichkeiten, um Spielzüge, Fahrweisen oder auch Bewegungsabläufe gezielt zu analysieren und zu optimieren. Im Bobsport werden neben einer normalen Sichtung der Videos die Videodaten noch nicht weiter ausgewertet. Die Videoaufnahmen ermöglichen jedoch grundsätzlich eine Analyse der Fahrlinie der Bobs und bieten somit Potenzial als weitere Datenquelle für den Trainer und die Piloten. Aufgrund der begrenzten finanziellen und zeitlichen Ressourcen im heutigen olympischen Bobsport konnten derartige Analysen bis dato nicht intensiv durchgeführt werden.

Ziel dieses Kapitels ist es daher, auf Basis von bereitgestellten Videos die Fragen zu beantworten, ob aus dem vorliegenden Videomaterial Erkenntnisse über die Fahrlinien der Bobfahrer abgeleitet werden können und wenn ja, wie die maschinelle Verarbeitung der Daten gelingen kann.

3.2 Data Understanding/Datenkorpus

Für die Beantwortung oben genannter Fragen wurden Videoaufzeichnungen aus der Eisarena Winterberg bereitgestellt. Insgesamt standen 115 Aufzeichnungen von Zweier- und Viererbobfahrten aus der fest installierten Kameraanlage zur Verfügung. Die Videoeigenschaften werden beispielhaft anhand der Aufzeichnung der Zweierbobfahrt von Laura Nolte vom 12.12.2021 erläutert.

Abbildung 16: Beispielhafter Ausschnitt aus Aufzeichnung von Zweierbob L. Nolte vom 12.12.2021

Die Videos wurden im Format MP4 (MPEG-4) mit einer Auflösung von 1280 × 720 Pixeln aufgezeichnet. Die Bildwiederholrate der Videos betrug dabei 50 Frames pro Sekunde. Die Kameraanlage, aus welcher die Videos stammen, ist mit 26 Kameras entlang der Bobbahn ausgestattet. Die Kameras sind an den insgesamt 14 Kurven sowie Verbindungsabschnitten angebracht.

Grundsätzlich lassen sich die Kameraperspektiven in statische und dynamische Aufnahmen unterteilen. Die Kameras K1, K3, K4, K6, K8, K11, K12 und K13 haben statische Perspektiven, während die Kameras K2, K5, K7, K9 und K14 mit Schwenkvorrichtungen ausgestattet sind und somit dynamische Perspektiven bieten. Die dynamischen Kameras verfolgen die Bobs bei der Durchfahrt durch lange Kurven, für welche sonst diverse statische Kameras benötigt würden. Abbildung 17 zeigt beispielhaft eine dynamische Kameraperspektive, Abbildung 18 eine statische Kameraperspektive.

Abbildung 17: Dynamische Kameraperspektive in Kurve 7

Abbildung 18: Statische Kameraperspektive in Kurve 11

Die zur Verfügung stehenden 115 Videoaufzeichnungen sind Zusammenschnitte der 26 Kameraperspektiven, welche mit Einblendungen der jeweiligen Bobs sowie gemessenen Zeiten an den jeweiligen Lichtschranken angereichert werden. Für die Analyse der Fahrlinie in einer bestimmten Kurve müssen diese Videos demnach in die einzelnen Kameraperspektiven aufgeteilt werden, um diese wiederrum über verschiedene Fahrten hinweg vergleichen zu können. Für die Aufteilung der Kameraperspektiven wurde ein automatischer Szenenerkennungsalgorithmus angewendet.

Für die Analyse wurden ausschließliche statische Kameraperspektiven verwendet, da nur bei diesen Perspektiven die Generierung einer Heatmap technisch möglich ist.

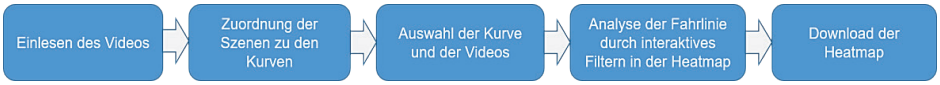
3.3 Vorgehen und Implementierung

Dieses Kapitel widmet sich der Beschreibung des Entwicklungsvorgehens und der Implementierung der Applikation zur Fahrlinienanalyse der Bobs. Im Entwicklungsvorgehen werden wichtige Thematiken der Usability, User Experience und Nutzerzentrierung beleuchtet, die für die Erstellung der Prototypen berücksichtigt werden müssen. Die auf dem Prototyp basierenden implementierten Benutzeroberflächen und deren Funktionalitäten, Aspekte der Datenhaltung und der Ausführbarkeit der Applikation werden in folgenden Kapiteln dargestellt. Usability, User Experience und Nutzerzentrierung sind wichtige Erfolgsfaktoren bei der Erstellung von Applikationen für den Endnutzer. Meist werden potenzielle Nutzende nicht in den Entwicklungsprozess neu zu entwickelnder Software mit einbezogen und müssen sich notgedrungen an die Gegebenheiten anpassen. Applikationen, denen die Akzeptanz der Nutzenden verwehrt bleiben, sind in den meisten Fällen zum Scheitern verurteilt. Nur wenn das Produkt auf die Bedürfnisse der Nutzenden ausgelegt ist und sie das System intuitiv bedienen können, wird es als positives und unterstützendes System wahrgenommen.

3.4 Aufbau des Graphical User Interface

Das hier vorgestellte Graphical User Interface wurde nutzerzentriert entwickelt, sodass es ohne tiefergehende technische Kenntnisse genutzt werden kann. Der Aufbau des GUI sieht den in Abbildung 19 dargestellten User Flow vor.

Abbildung 19: User Flow des User Interface



Im ersten Schritt liest der Nutzende ein neues Video aus seinem Dateiexplorer ein und vergibt einen Videonamen. Das Video wird nach Einlesung in seine einzelnen Szenen geteilt.

Im nächsten Schritt erfolgt eine manuelle Zuordnung der Szenen zu den Kurven durch den Nutzenden. Mit Klick auf den Videonamen und den Button *Szenen bearbeiten* wird eine Tabelle geöffnet, die mit der korrekten Kurve auf Basis der Zeitinformationen zu den jeweiligen Szenenabschnitten gefüllt wird.

Die Anwendung kann auch ohne das Einlesen des Videos genutzt werden. Dazu kann der Nutzende auf die bereits eingelesenen Videos zurückgreifen. Nach Zuordnung der Szenen zu den Kurven wählt der Nutzende, neben den zu analysierenden Videos, unter Punkt 3 im Drop Down Menü eine Kurve zur Analyse aus und erstellt über den gleichnamigen Button die Heatmap. Wie im User Flow in Abbildung 19 zu erkennen, hat der Nutzende nun die Möglichkeit, die Fahrlinien der Bobs zu vergleichen und Erkenntnisse abzuleiten. Dabei unterstützt die interaktive Filtermöglichkeit, mit der die einzelnen Fahrlinien ein- bzw. ausgeblendet werden können. Für den weiteren Gebrauch ist es zudem möglich, sich die generierte Heatmap herunterzuladen.

Zur Erstellung des User Interface wurde auf die Pythonbibliothek *PySimpleGui* zurückgegriffen. Der Aufbau der Startseite ist durch die Nummerierung der Schrittabfolge für den Nutzenden einfach nachzuvollziehen. Der sichtbare freie Bereich zwischen den linken und rechten Steuerungselementen dient zur Darstellung der Heatmap.

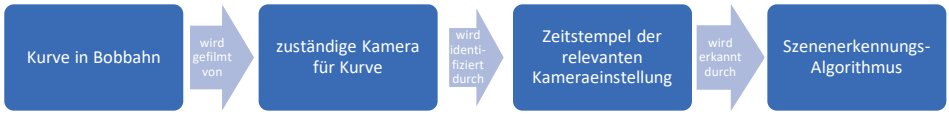
3.4.1 Einlesen und Verarbeitung der Videos

Um die Funktionalität des Videoeinlesens für die Nutzenden bereitzustellen, wurde mithilfe von PySimpleGUI ein entsprechender Lesedialog erstellt. Im Dialog wird der Nutzende aufgefordert einen Dateipfad zur Videodatei anzugeben. Dies kann er über einen gesonderten Dialog zur Navigation des Dateisystems durchführen. Die Funktionalität zum Durchsuchen des Dateisystems wird ihm über die in PySimpleGUI integrierte Funktion *.FileBrowse()* bereitgestellt („Python GUIs for Humans“, o. J.).

Für die weitere Verarbeitung zu einer Heatmap müssen die zu analysierenden Videodaten entsprechend vorverarbeitet werden. Um der Anwenderin bzw. dem Anwender dafür einen möglichst großen Komfort zu bieten, soll die Generierung der Fahrlinienanalyse anhand der Auswahl einer Kurve der Bobbahn gesteuert werden können. Da diese Informationen jedoch nicht in den Videorohdaten vorliegen, muss eine Zuordnung der Kurve auf die zuständige Kamera vorgenommen werden. Die zuständige Kameraperspektive muss dann wiederrum im Video isoliert werden.

Eine manuelle Identifikation der Schnitte im Video, welche den Wechsel von einer Kameraperspektive zu einer weiteren anzeigen, ist eine zeitaufwändige und ressourcenintensive Arbeit, welche in anderen Kontexten teilweise noch heute manuell durchgeführt wird, beispielsweise um Nachrichtensendungen zu teilen, sodass die Konsumentin bzw. der Konsument die Möglichkeit erhält, nur einzelne Nachrichtenszenen zu betrachten (Cui *et al.*, 2017). Um diese Zuordnung für die Anwenderin bzw. den Anwender komfortabler zu gestalten, wurde das Video nach dem Einlesen vorverarbeitet und durch Anwendung eines Szenenerkennungsalgorithmus automatisch in einzelne Szenen unterteilt. In Abbildung 20 wird dieser Prozess grafisch dargestellt.

Abbildung 20: Prozessschritte zur Zuordnung der Videoaufnahmen zu ausgewählten Kurven



Grundsätzlich verwenden alle Szenenerkennungsalgorithmen hierbei ähnliche Vorgehensweisen. Die hohe Anzahl an Informationen in einem Video (Farbwerte der Pixel, Helligkeit, Kontrast, Länge, Auflösung, etc.) wird auf eine kleine Menge an Informationen, sogenannte Features, reduziert und aus den einzelnen Frames extrahiert und verarbeitet. (Cotsaces *et al.*, 2006, S. 30) Diese extrahierten Features können dann zwischen aufeinanderfolgenden Frames verglichen werden, um eine Änderung festzustellen. Da die Metriken, welche die extrahierten Features abbilden, jedoch hochsensibel sind und bei jedem Framewechsel variieren, muss neben der Auswahl der Features auch ein Schwellwert angegeben werden, ab welchem die Änderungen in den Metriken groß genug sind, um einen Szenenwechsel (Wechsel der Kameraperspektive) zu markieren.

Je nachdem, welche Art von Szenenwechsel vorliegt, können jedoch auch diese Schwellwerte schwanken. Grundsätzlich lassen sich vier Hauptarten von Szenenwechseln unterscheiden. Bei *Cut changes* (a) finden abrupte Wechsel einer Szene statt. Bei solchen Wechseln gehört der letzte Frame zur vorherigen Szene und der nächste Frame zur neuen Szene. Bei *Dissolve changes* (b) findet eine sichtbare Überlappung verschiedener Kameraperspektiven statt, welche sich über eine größere Menge an Frames hinzieht. Sogenannte *Fade changes* (c) sind ähnlich zu Dissolve changes, jedoch überlagern die aufeinanderfolgenden Szenen sich nicht direkt, sondern es findet ein fade-out auf ein schwarzes Bild statt, woraufhin das schwarze Bild wieder auf die nächste Szene wechselt (fade-in). Ebenfalls über mehrere Frames hinweg finden *wipe changes* (d) statt. Diese Szenenwechsel bewegen die vorherige und nächste Szene in horizontaler oder vertikaler Richtung aus dem Bild heraus bzw. in das Bild herein, sodass für eine kleine Menge an Frames die beiden Szenen in Teilen nebeneinander angezeigt werden.

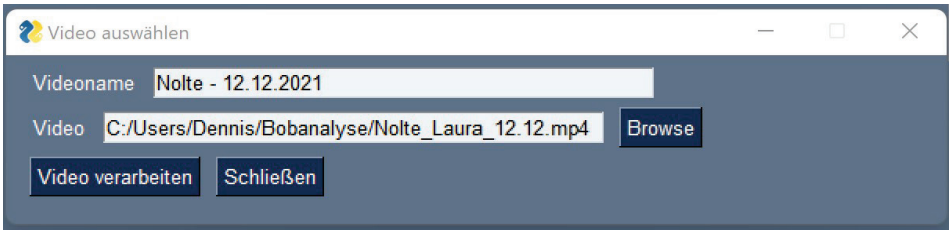
Wie in Abbildung 21 gezeigt, werden in den vorliegenden Videodaten ausschließlich Wipe changes verwendet, um die unterschiedlichen Kameraperspektiven zu verbinden und den Bob entlang der Bahn zu verfolgen. Für die spätere Anwendung der Detektoren wurde darauf geachtet, dass die Entscheidung über einen Szenenwechsel nicht ausschließlich auf Basis eines Frames getroffen wird, welcher einen Wipe change zeigt.

Abbildung 21: Szenenwechsel mit Wipe change

Die automatische Szenenerkennung wurde in dieser Arbeit über Open-Source Python Package *PySceneDetect* realisiert. Das Package wurde erstmalig 2014 vom Entwickler Brandon Castellano veröffentlicht und wird bis heute weiterentwickelt (Castellano, 2014). Die in dieser Arbeit verwendete Version v0.5.6.1 wurde im Oktober 2021 veröffentlicht. Das Paket kann über das Python Command-Line-Interface oder über den direkten Methodenaufruf angesteuert werden. Für die Einbindung in das GUI wurde der direkte Methodenaufruf gewählt und implementiert.

Wenn der Benutzer ein Video einliest und der *Video verarbeiten* Button geklickt wird, dann wird die Methode .find_scenes() aufgerufen, um das Video einzulesen und die Szenen zu extrahieren. Dabei wird vorab geprüft, ob ein entsprechender Videoname sowie Dateipfad angegeben wurde. Falls nicht, wird eine Fehlermeldung ausgegeben. Abbildung 22 zeigt hierzu den Videoauswahldialog.

Abbildung 22: GUI – Videoauswahldialog



Die find_scenes Methode erhält als Input drei Parameter: Den Videopfad (*video_path*), einen booleschen Wert, welcher angibt, ob ein Analysedokument über die Szenenextraktion erstellt werden soll (*export_stats)* sowie ein Video Manager-Objekt, welches eine entsprechende Bibliothek enthält, um Videos verschiedenen Typs programmatisch einzulesen und die einzelnen Frames der Videos zu untersuchen (*video_manager*). Für die Erkennung der Szenen stehen im Paket drei verschiedene Detektor-Algorithmen zur Verfügung:

Threshold Detector

Der Threshold Detector (deutsch: Schwellwertdetektor) basiert auf Helligkeit und Sättigungsfeatures, welche aus den einzelnen Frames berechnet werden. Hierzu wird für jeden Pixel im Video der jeweilige Helligkeits- und Sättigungswert, welcher wiederrum zu einem Gesamtdurchschnitt des gesamten Frames, bestehend aus 1280×780 Pixeln, konsolidiert wird, genutzt. Auf diese Weise ergibt sich eine Gleitkommazahl zwischen 0 und 255, welche den durchschnittlichen Helligkeits-/Sättigungswert des Frames angibt. Der Threshold Detector legt jedoch die Annahme zugrunde, dass ein Szenenwechsel immer durch die Einblendung eines schwarzen Bildes eingeleitet wird, daher vergleicht der Algorithmus die berechneten Durchschnittswerte stets gegen die Werte eines schwarzen Bildes und markiert einen Szenenwechsel nur dann, wenn die Werte übereinstimmen – im Video also ein vollständig schwarzer Frame angezeigt wird. [„Scene Detection Algorithms“]

Content-Aware Detector

Der Content- Aware Detector (deutsch: inhaltssensitiver Detektor) berücksichtigt weitere Features, indem die RGB-Werte der Pixel in den HSV-Farbraum konvertiert werden. Im HSV-Farbraum werden die Eigenschaften der Farbwerte (Hue), Farbsättigung (Saturation) und des Hellwertes (Value) berücksichtigt. Auf diese Weise werden die Farben im Bild stärker berücksichtigt und es kann ein Vergleich zwischen verschiedenen Frames durchgeführt werden. [„Scene Detection Algorithms“]

Adaptive Content Detector

Der Adaptive Content Detector (deutsch: anpassender inhaltssensitiver Detektor) weist dieselbe Funktionsweise wie der Content Detector auf, verwendet jedoch zum Framevergleich einen rollierenden Durchschnitt von mehreren Frames anstatt jeden Frame einzeln mit seinem Nachfolger zu vergleichen. [„Scene Detection Algorithms“]

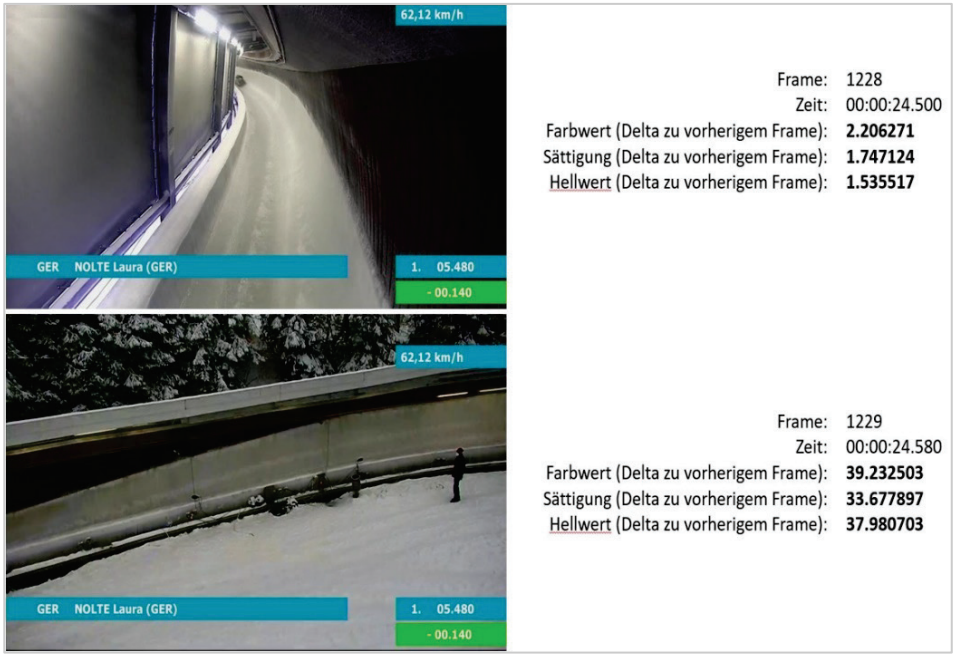
Für die Realisierung der Szenenerkennung der Videodaten wurde der Content Aware Detector verwendet, da ein Wechsel der Kameraperspektive unmittelbar aufeinanderfolgend passiert. Die *.find_scenes()* Methode übergibt dem Szenenmanager-Objekt den ContentDetector und wendet daraufhin die *detect_scenes* Methode an, um die Szenen zu extrahieren.

Wie bereits in der allgemeinen Einführung zu Szenenerkennungsalgorithmen beschrieben, ist es für alle Detektoren notwendig zu definieren, bei welchen Schwellwerten die Veränderung einer Messmetrik als Szenenwechsel erkannt werden soll. Bedenkt man die zugrundeliegenden Einflussgrößen (Farbwerte, Helligkeit, Sättigung), so wird deutlich, dass ein solcher Schwellwert nicht nur von Video zu Video, sondern auch von Szene zu Szene unterschiedlich hoch sein kann. Auch Daudpota *et al.* (2019) sehen im Festlegen der korrekten Schwellwerte eines der größten Probleme der klassischen Szenenerkennungsalgorithmen und haben festgestellt, dass diese Schwellwerte nicht unabhängig der Dateninputs festgelegt werden können (Daudpota *et al.*, 2019, S. 1415). Mittlerweile haben sich unterschiedliche Vorgehen zur Festlegung dieser Schwellwerte etabliert, welche primär darauf abzielen, die Schwellwerte möglichst dynamisch und unter Berücksichtigung des Dateninputs zu berechnen. Cotsaces *et al.* (2006) unterschieden zwischen statischen, (static thresholding), anpassenden (adaptive tresholding), probabilistischen (probabilistic thresholding) Festlegungen der Schwellwerte, sowie der Verwendung von vortrainieren Modellen (trained classifier) (Cotsaces *et al.*, 2006, S. 31).

Im Rahmen der Implementierung wurden verschiedene Methoden zur Festlegung des Schwellwertes getestet. Da in der aktuellen Ausbaustufe des Programms lediglich statische Kurven mit festinstallierten Kameras untersucht wurden, hat die statische Festlegung der Schwellwerte bereits zufriedenstellende Ergebnisse erzeugt. Die Analyse dessen wurde dabei über eine Log-Datei des *SceneDetect* Packages durchgeführt. In der exportierbaren stats.csv Datei, können die Metriken für Farbwert, Farbsättigung und Hellwert je Frame eingesehen werden und mit den Bildern der Frames verglichen werden. In Abbildung 23 sind die

Werte und Bilder kombiniert dargestellt und es wird deutlich, wie sich ein Szenenwechsel in den Daten ausdrückt. Für die Verarbeitung der vorliegenden Videodaten wurde nach Prüfung mehrerer Videos von unterschiedlichen Aufnahmetagen ein Schwellwert von 27 festgelegt. Zusätzlich wurde festgelegt, dass Szenen erst dann als solche erkannt werden sollen, wenn diese mindestens 15 Frames – also 0,3 Sekunden – lang sind.

Abbildung 23: Vergleich der HSV-Werte bei einem Szenenwechsel



Die Rückgabe der Methode ist eine Liste der identifizierten Szenen, welche der Nutzerin bzw. dem Nutzer tabellarisch mit Zeitstempeln und weiteren Attributen angezeigt werden können.

Für die Zuordnung zu einer entsprechenden Kurve wurde der Liste noch eine Spalte *Kurve* hinzugefügt, welche die Nutzerin bzw. der Nutzer in der Anzeige manuell ändern kann. Aufgrund der implementierten Datenspeicherung muss diese Aufgabe von der Benutzerin bzw. dem Benutzer pro analysierter Fahrt nur einmalig durchgeführt werden.

Im nächsten Kapitel werden die Daten der extrahierten Szenen und zugehörigen Kurven verwendet, um die eigentliche Fahrlinienanalyse zu generieren.

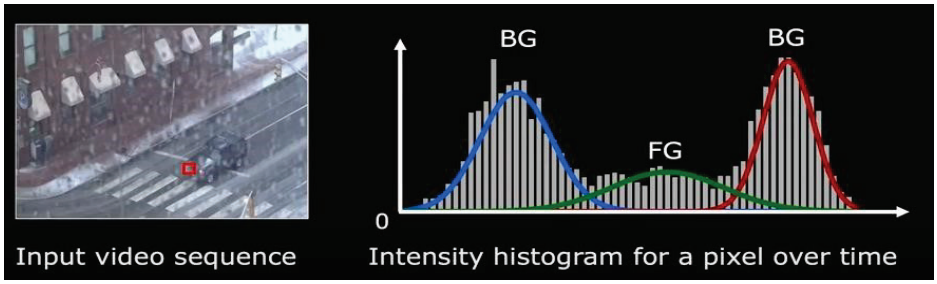
3.4.2 Erstellung der Motion-Heatmaps

Den Kern der Applikation zur Fahrlinienanalyse von Bobs anhand visueller Daten bildet die Erstellung der Motion-Heatmaps. Sie bilden die Grundlage für die Vergleichbarkeit verschiedener Fahrlinien einer Kurve.

Mithilfe der Methodik der Background Subtraction wird eine Vordergrundmaske der sich bewegenden Objekte einer Bildsequenz erstellt. Bild für Bild werden sich bewegende Objekte identifiziert, indem das aktuelle Bild eines gewissen Zeitpunkts mit dem statischen Hintergrundbild des gesamten Videos verglichen wird. Ein Schwellenwert definiert die Stärke der Veränderung eines Pixels, ab dem eine Veränderung als wirkliche Bewegung wahrgenommen werden soll. Das hierfür benötigte Background Model kann anhand verschiedener Algorithmen wie beispielsweise dem *k-Nearest Neighbour* oder dem Gaussian Mixture Model (GMM) generiert werden. Voraussetzung für die Anwendung dieser Technik ist ein statischer Kamerawinkel [OpenCV, „BackgroundSubtractor Class Reference“].

Bei der Verwendung des Gaussian Mixture Models als Background Subtraction Algorithmus wird während des Background Modellings für jeden Pixel und dessen Intensität sowie seiner Farbwerte jeweils ein GMM trainiert. Abbildung 24 verdeutlicht die Funktionsweise des GMM. Anhand der als Histogramm visualisierten Intensitätswerte eines Pixels über die Zeit werden K Gauss-Funktionen $\omega_k \eta_k(x, \mu_k, \sigma_k)$ bestimmt. Die gewichtete Summe der Gauss-Funktionen wird als Gaussian Mixture Model $P(x) \cong \sum_{k=1}^{K} \omega_k \eta_k(x, \mu_k, \sigma_k)$ bezeichnet. Unter der Annahme, dass ein Pixel die meiste Zeit den Wert des Hintergrunds repräsentiert, werden Gauss-Funktionen mit einem großen ω und kleinem σ als Hintergrund klassifiziert (KaewTraKulPong & Bowden, 2002, S. 135-145; Zivkovic, 2004, S. 28-31).

Abbildung 24: GMM – Intensität eines Pixels über die Zeit



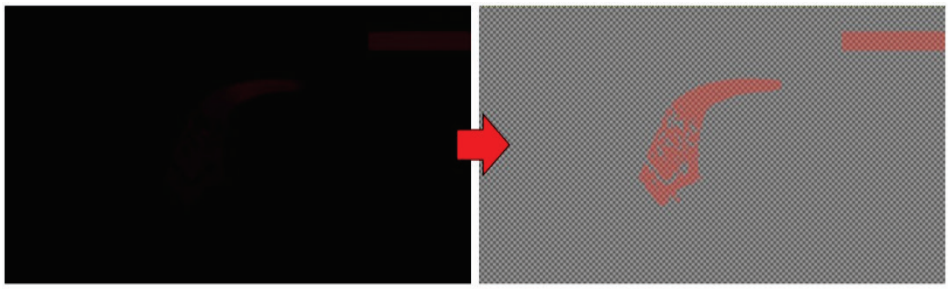
Quelle: Nayar, 2021.

Implementierungen der Background Subtraction Algorithmen, wie dem zuvor angesprochenen Gaussian Mixture Model, finden sich in der „Open Source Computer Vision Library“ OpenCV.

Das Video an sich sowie weitere szenenspezifische Informationen, wie der erste und letzte Frame der Bildsequenz sowie ein Offset, werden der Funktion übergeben. Der Offset reduziert die Bildsequenz am Szenenanfang und -ende um einen definierten Wert. Dies reduziert die Fehleranfälligkeit bei nicht eindeutigen Szenenwechseln. Nachdem der Background Subtraction Algorithmus initialisiert wurde, werden die irrelevanten Frames des Videos übersprungen, bis der Anfang der relevanten Szene erreicht ist. Iterativ wird die Background Subtraction auf jeden Frame der Szene angewendet. Die hierbei generierte Vordergrundmaske jedes einzelnen Frames wird kumuliert und als numpy-ndarray der Form (720, 1280, 3) zurückgegeben.

Am Beispiel der Fahrt von Laura Nolte vom 12. Dezember 2021 durch Kurve 8 der Veltins Eisarena in Winterberg, zeigt Abbildung 25 eine durch die zuvor beschriebene Funktion generierte Motion-Heatmap. In unbehandelter Form erkennt man eine leicht rötliche Schattierung auf schwarzem Hintergrund. Mit Hinblick auf die Vergleichbarkeit verschiedener Kurvenfahrten, durch das Übereinanderlegen mehrerer Fahrten, muss der Hintergrund transparent und die schlecht erkennbare Schattierung verstärkt werden.

Abbildung 25: Motion-Heatmap vor und nach der Bildbearbeitung



Die sehr schwache Schattierung zum Kurveneingang und stärkere Schattierung zum Kurvenende ist begründet in der Geschwindigkeit der Bewegung des Bobs und der Anzahl an Bildern, in denen der Bob sichtbar ist. Die statische Kamera ist oberhalb des Kurveneingangs positioniert, entsprechend ist die Distanz zum Bob am Kurveneingang sehr gering und erhöht sich hingehend zum Kurvenende. Diese Perspektive führt zu verhältnismäßig wenigen Bildern des Bobs am Kurvenbeginn und zu verhältnismäßig vielen Bildern am Kurvenende.

Da sich beide angesprochenen Effekte zur Stärke der Schattierung von Motion-Heatmaps negativ für die Analyse der Kurvenfahrten im Bobsport zeigen, soll die Stärke der Schattierung einheitlich vergeben werden.

Der Code implementiert den angesprochenen zusätzlichen Alpha-Channel und gleicht den Farbwert der Schattierung einheitlich an, um den wie in Abbildung 26 visualisierten gewünschten Effekt zu erreichen. Hierfür wird die als zweidimensionales *numpy-array* repräsentierte Heatmap nach Pixeln sowie Pixelreihen durchlaufen. Ein schwarzer Pixel mit keiner Schattierung bzw. einem RGB-Farbwert von *(0, 0, 0)* wird volle Transparenz bzw. ein BGRA-Wert von *(255, 255, 255, 0)* zugewiesen. Ein schattierter Pixel wird unabhängig seiner aktuellen Farbstärke mit einem einheitlichen BGRA-Wert eingefärbt. Wird kein Farbwert als Parameter in die Funktion übergeben, wird der Standardwert von *(0, 0, 255, 100)* verwendet. Ein Alpha-Wert von null entspricht voller Transparenz, während ein Wert von 255 kompletter Intransparenz bzw. voller Farbsättigung entspricht. Für das spätere Übereinanderlegen verschiedener Motion-Heatmaps mehrerer Fahrten sollte entsprechend ein mittlerer Transparenzwert gewählt werden.

Neben den erzeugten Heatmaps der jeweiligen Fahrten wird ebenfalls ein Bild der Kurve benötigt, um die Fahrlinie in Relation zur Fahrbahn zu setzen. Je nach verwendetem Algorithmus lässt sich direkt das trainierte Background Model als

Hintergrundbild, ohne störende dynamische Objekte, verwenden. Bei der Verwendung des GMM steht diese Funktion nicht zur Verfügung. Entsprechend implementiert der Code eine Methode zur Bestimmung und dauerhaften Speicherung eines geeigneten Hintergrundbilds. Da die automatisiert identifizierten Szenenübergängen nicht immer trennscharf sind, wird der Frame im Mittel also zwischen Szenenanfang und -ende als Hintergrundbild verwendet.

Letztlich erfolgt die Fahrlinienanalyse verschiedener Bobfahrten anhand einer Mehrfachauswahl innerhalb der Benutzeroberfläche. Beim Selektieren oder Deselektieren von Fahrten wird eventbasiert ein neues Bild aus den ausgewählten Motion-Heatmaps und dem Hintergrundbild erzeugt. Die mit einem Alpha-Kanal versehenen Fahrlinien werden nacheinander mit voller Farbsättigung über das Hintergrundbild gelegt.

Das Endergebnis der Bildmanipulation wird nachfolgend der Benutzerin bzw. dem Benutzer in der Benutzeroberfläche angezeigt. Mithilfe weiterer Informationen, wie den Sektorzeiten der jeweiligen Fahrten, die dem Trainer auf anderen Wegen zur Verfügung stehen, können die individuellen Fahrlinien nun zur Generierung neuer Erkenntnisse verwendet werden.

3.5 Datenmanagement

Jede Applikation besitzt auf die ein oder andere Weise eine Datenhaltung bzw. ein Datenmanagement. Grundsätzlich kann das Datenmanagement in zwei Arten unterschieden werden. Eine laufzeitbezogene Datenhaltung im Rahmen einer In-Memory-Speicherung im Arbeitsspeicher oder einer persistierten Datenhaltung, wie beispielsweise durch Datenbanken.

Im konkreten Anwendungsfall der Fahrlinienanalyse von Bobs mithilfe der Python-Bibliothek *PySimpleGui* wurde die Entscheidung hinsichtlich eines minimalistischen Datenmanagements in Form von globalen Variablen implementiert. Die Daten werden zur Laufzeit der Applikation im Arbeitsspeicher verwaltet und zu definierten Ereignissen wie dem Starten oder Beenden der Applikation aus einer lokalen Datei gelesen bzw. in eine lokale Datei geschrieben. Weitere Funktionalitäten, wie die Erstellung der Motion-Heatmaps, weisen deutliche Performancesteigerungen durch das Zwischenspeichern von erzeugten Ergebnissen auf. Der sogenannte Cache der Applikation wird in einem separaten Verzeichnis abgelegt. In Windows-basierenden Applikationen wird der Cache meist im *AppData*-Verzeichnis des aktiven Benutzerkontos abgelegt – dies trifft auch auf das Fahrlinienanalysetool zu.

3.6 Ausführbarkeit als .exe-Datei

Zuvor wurde die Anforderung einer ausführbaren Applikation ohne vorherige Installationen oder Vorkenntnisse einer Programmiersprache definiert. Mithilfe der Bibliothek *pyinstaller* wird eine Python-Applikation zusammen mit all ihren Abhängigkeiten in ein ausführbares Programm gepackt. Diese *.exe*-Datei kann ohne einen zuvor installierten Python Interpreter oder Module ausgeführt werden. Dabei wird das entwickelte Pythonskript eingelesen und alle Module und Bibliotheken identifiziert, die zur Ausführung der Applikation benötigt werden. Alle Dateien der zuvor identifizierten Bibliotheken werden zusammen mit dem Python Interpreter aus einer definierten virtuellen Python Umgebung in die neue ausführbare Datei gepackt (PyInstaller, o. J.).

3.7 Praktische Anwendung

Im Nachfolgenden wird die erstellte Anwendung auf verschiedene Fahrten und deren Videos angewendet. Anschließend wird die Visualisierung durch die Heatmaps genutzt, um Erkenntnisse aus den Fahrlinien zu erhalten.

Die erste Kurve, welche im Rahmen dieses Kapitels analysiert wird, ist die Kurve Nummer acht. Hierbei handelt es sich um eine kürzere Kurve, welche mit einer statischen Kameraposition aufgenommen wird. Die Aufnahmen der gleichen Kamera waren in diesem Beispiel für verschiedene Piloten vom Weltcup 2020/2021 verfügbar. Die nachfolgend betrachteten Fahrten sind hierbei die von Laura Nolte und Francesco Friedrich. Dazu wurden in der Oberfläche die entsprechenden Videos beider Piloten ausgewählt und entsprechend benannt, damit anschließend beim Mapping zu den korrekten Szenen die relevanten Kurven eingetragen werden können. In der aktuellen Ausbaustufe ist es hierbei notwendig, die Videos parallel zu öffnen und die passenden Zeitstempel herauszusuchen. Für die Kurve acht der Fahrt von Laura Nolte ist der Ausschnitt von Frame 2.591 bis zum Frame 2.656 relevant. Beim Video von Francesco Friedrich fährt dieser in den Frames 2.081 bis 2.150 durch Kurve Nummer acht der Veltins Eisarena in Winterberg. Dabei fällt auf, dass Francesco Friedrich die Kurve früher erreicht. Dies lässt zunächst zwei Vermutungen zu: Eine höhere Geschwindigkeit in der Bahn oder einen frühen Start. Nach Analyse der beiden Videos lässt sich die zweite These bestätigen, da Laura Nolte erst nach elf Sekunden in der Aufnahme die Phase des Anschiebens beginnt. Francesco Friedrich startet bereits etwa drei Sekunden nach Start der Aufnahme und somit früher als Nolte. Nachdem die Kurve nun in den Tabellen zu den Videos passend gemappt wurde, lassen sich die Heatmaps

für diese generieren und übereinanderlegen. Bei Analyse der Visualisierung der Fahrlinien in Abbildung 26 wird auf den ersten Blick ein Unterschied deutlich. Die Fahrlinie von Friedrich wird hierbei durch die blaue Heatmap dargestellt und Nolte in einem rötlichen Ton, sodass sich überlappende Bereiche durch einen Mix der beiden Farben in einer violetten Farbe darstellen. Der farbige Streifen oben rechts lässt sich durch die aktuelle Fahrtzeit des Piloten erklären, welche zwischenzeitlich im Video gezeigt wird.

Abbildung 26: Heatmaps Kurve 8 – Nolte und Friedrich

Zu erkennen ist, dass Friedrich höher durch die Kurve fährt. Die Fahrt von Francesco Friedrich fand in der Disziplin des Viererbobs der Männer statt im Kontrast zu einer Zweierbobfahrt im Bereich der Frauen von Laura Nolte.

Als weiteres Beispiel konnte auch die längere Kurve Nummer neun analysiert werden. Diese wurde mit Hilfe einer Weitwinkelkamera aufgenommen. Dadurch, dass die Kameraperspektive nun statisch in den Aufnahmen ist, kann der Algorithmus den Bob erkennen und tracken, da er das einzige sich bewegende Objekt

ist. Für diese Kurve werden nachfolgend die Fahrten der Piloten Zimmer und Zielasko verglichen. Dazu wurden, wie bereits bei den ersten Videos, zunächst die Videoaufnahmen in die Anwendung geladen und entsprechend benannt. Anschließend wurde die Kurve zu der Szene dieser Kurve in der Tabelle über die Anwendung gelegt. Bei der abschließenden Visualisierung der übereinandergelegten Heatmaps lassen sich hier ebenfalls Unterschiede in den Fahrlinien erkennen. Dies lässt sich dadurch begründen, dass der Pilon Zielasko offenbar mit voller Geschwindigkeit die Kurve nimmt und daher eine höhere Fahrlinie fährt. Bei der Pilotin Zimmer erkennt man eine vorsichtigere Fahrt. In der Abbildung 27 ist die Fahrlinie von Zimmer rot dargestellt und die von Zielasko in blau.

Abbildung 27: Heatmaps Kurve 9 – Zimmer und Zielasko

Im Zuge der Analyse weiterer Videos sind verschiedene Heatmaps generiert worden, welche nicht auf den ersten Blick erklärt werden konnten. Eines dieser Beispiele ist bei der Aufnahme von Dosthaler in der Kurve neun mit der Weitwinkelkamera aufgetreten. Bei Generierung der Heatmap aus diesem Video, welche in Abbildung 28 dargestellt ist, sind einige farbige Bereiche unterhalb der Fahrbahn zu erkennen. Dieses lässt sich nach Analyse des Videomaterials durch eine Person, welche durch das Bild läuft, erklären.

Abbildung 28: Heatmap Kurve 9 – Dosthaler

4 Zum Zusammenhang von Fahrlinie und Geschwindigkeit

4.1 Zielsetzung

Ziel dieses Kapitels ist, die Abhängigkeit der Geschwindigkeit von der Fahrlinie zu untersuchen. Die hierbei getroffene Annahme war, dass die Piloten durch die Fahrlinie die Geschwindigkeit beeinflussen können. Zudem wurde überprüft, ob weitere Faktoren wie z. B. der gewählte Messtag, das Gewicht, etc. einen signifikanten Einfluss auf das Ergebnis haben.

4.2 Aufbau der Entwicklungsumgebung

Zur Entwicklung stehen aufgrund der für die Berechnungen benötigten Bibliotheken und Funktionen die Programmiersprachen R und Python zur Auswahl. Insbesondere mathematische und statistische Bibliotheken wie „*statsmodels*“ unter Python und „*Quantmod*“ unter R werden als technische Grundlage für diese Arbeit angesehen. Auf Basis bisheriger Erfahrungen wurde Python als Programmiersprache gewählt. Die hierbei verwendete Version ist Python 3.9.10. Als Entwicklungstool wurde sich für die Anwendung *PyCharm Professional* des Softwareherstellers JetBrains s.r.o entschieden. Die Code-Verwaltung wurde über ein privates Git-Repository gelöst und für den Austausch der Daten wurde die Kollaborationsumgebung in Microsoft Teams genutzt.

4.3 Abgeleitete Vorgaben der Datenaufbereitung

Zur Eingrenzung der Untersuchung dieser Arbeit wurden einige Annahmen und Untersuchungsschwerpunkte definiert, die für die Auf- und Vorbereitung der Daten wichtig sind. Gemäß der hohen Gesamtdatenmenge von insgesamt fast 1.000 Fahrten wurde eine Untersuchung der einzelnen Startvorgänge nicht vorgenommen. Eine Vergleichbarkeit der Fahrweise von Mono-, Zweier- und Viererbob oder zwischen Geschlechtern wurde auf Basis der erkannten wirkenden Kräfte und unterschiedlichen Vorgaben des Reglements nicht angenommen. Zur Untersuchung wurden die Daten daher in identifizierte Gruppen eingeordnet und einzeln betrachtet. Da insbesondere auch eine hohe Abhängigkeit des Reibungskoeffizienten von der Temperatur des Eises herausgestellt wurde, wurden maximal alle Läufe eines Tages unter den gleichen Rahmenbedingungen betrachtet. Die Modellierung wird eine tagesabhängige Abhängigkeit daher in Erwägung ziehen und diese untersuchen.

4.4 Hypothesen und Fragestellungen

Der Datensatz des Eiskanals „VELTINS-EisArena Winterberg“ soll auf die folgenden Hypothesen untersucht werden. Die durch die Literatur aufgestellte Hypothese, dass die Startgeschwindigkeit signifikanten Einfluss auf das Ergebnis hat, soll überprüft werden. Des Weiteren wird davon ausgegangen, dass die Piloten das Ergebnis positiv wie negativ durch die Fahrlinie beeinflussen können. Hierbei ist die Grundannahme zu treffen, dass die optimale Fahrlinie sich zwischen den verschiedenen Sportgeräten unterscheidet. Zudem ist zu klären, ob für die Fahrten ein generelles Optimum zu bestimmen ist oder ob dies, auf Grund einer wechselnden Ausgangslage, für jeden Durchlauf erneut bestimmt werden muss.

4.5 Homogenisierung der Dateistruktur

Die dem Projekt zur Verfügung gestellten Ordner und Dateien befinden sich nicht in einem einheitlichen Format. Für den Import muss ein einheitliches Konzept angewendet werden. Die aufbereitete Struktur beginnt als Basis mit einem Oberordner, welcher sämtliche Daten einer Bahn bzw. einer Lokation enthält. Die Ordner in diesem repräsentieren eine Gruppierung von Datensätzen. Die Atomizität dieser Gruppierungen ist zu maximieren, um in der späteren Analyse eine größere Vielfalt von Subgruppierungen zu erlauben.

Bei der praktischen Umsetzung dieser Vorgaben wurden dazu allen Merkmalen eindeutige Identifikatoren zugewiesen und getrennt durch Unterstriche, im Ordner- und Dateinamen, zusammengefasst. Hierbei wurden alle Merkmale, welche für die gesamte Gruppierung gleich sind, im Ordner zusammengefasst. Nur notwendige Merkmale wurden im Dateinamen behalten. Einige der Quellordner weisen Datensatzgruppierungen mit gemischtem Fahrergeschlecht auf, welche daher in separate Untergruppen sortiert wurden. Das Geschlecht wurde festgestellt mithilfe der Athletenliste des IBSF (vgl. ISBF, 2022a).

4.6 Extraktion der Metadaten und Dateiinhalte

Die Homogenisierung der Dateistruktur resultierte in der in Tabelle 2 dargestellten Struktur. Die Merkmale der Metadaten lassen sich hierbei in zwei Kategorien unterteilen: Merkmale mit statischen oder dynamischen Ausprägungen. Merkmale wie die Anzahl an Athleten treten nur in den Ausprägungen „1“, „2“ oder „4“ auf. Im Gegensatz dazu variieren Fahrer ID je nach Lokation, Datum, Geschlecht, etc. und wurden dynamisch ausgelesen.

Tabelle 2: Ordner- und Metadaten-Struktur

Ebene	Merkmal	Ausprägungen
1. Oberordner	Lokation	Winterberg
2. Ordner	Datum	[Datum]
	Geschlecht	männlich, weiblich
	Schlittentyp	Bobsleigh
	Anzahl	1, 2, 4
	Anlass	Weltmeisterschaft, Training
3. Datei	Fahrer (ID)	[Fahrer-ID]
	Fahrt	1, 2, 3
4. Daten	[Spalten]	[Zeilen]

Die Daten in den Dateien besaßen unterschiedliche Dateikodierungen und Spaltenüberschriften. Die Spaltenreihenfolge ist jedoch bei beiden Varianten gleich. Die Spalten wurden deswegen beim Einlesen einheitlich umbenannt. Die genutzten Bezeichnungen sind in Tabelle 3 zu sehen:

Tabelle 3: Datenspaltenbenennung

Dateikodierung	UTF-8	ANSI	[intern]
Zeit	Time s	Time s	Time
Entfernung	Distance m	Distance m	Distance
Geschwindigkeit	Speed km/h	Speed km/h	Speed
X-Beschleunigung	Acc x m/s2	Acc x m/s2	Acc_X
Y-Beschleunigung	Acc y m/s2	Acc y m/s2	Acc_Y
Z-Beschleunigung	Acc z m/s2	Acc z m/s2	Acc_Z
Rollwinkel	Roll angle deg	Roll angle	Roll
Lichtschranke	Lightbeam	Lightbeam	LB
Lichtschrankenzeit	LB time s	LB time s	LB_Time

4.7 Anreicherung der Daten

Die importierten Daten wurden nun angereichert. Dies beinhaltet die Addition der Metadaten, Datentypkonvertierungen und Berechnungen weiterer für die spätere Analyse sinnvoller Datenfelder.

Zunächst wurden dem Datensatz Lokation, Ordner- und Dateiname hinzugefügt. Dies erlaubt die Extraktion der weiteren Metadaten. Aus diesen Feldern wurde auch eine jede einzelne Fahrt identifizierende ID generiert. Diese erlaubt eine schnelle Gruppierung oder Filterung nach Fahrt. Im Anschluss wurden die Metadatenmerkmale extrahiert. Dazu wurde geprüft, ob im Ordner- oder Dateinamen die eindeutigen Identifikatoren der statischen Merkmale vorliegen. Dynamische Merkmale wurden anhand Ihrer Position oder Formatierung erkannt.

Um gleiche Einheiten zu verwenden, wurden die Werte der Geschwindigkeit von km/h in m/s umgerechnet. Aus den Differenzen dieser und der Zeit wurde die generelle und vektorlose Beschleunigung berechnet. Der absolute Rollwinkel wurde ebenfalls ergänzt. Für den absoluten Rollwinkel wurden mehrere Hilfsmaße berechnet. Die Änderung des absoluten Rollwinkels wurde mithilfe der

Zeitdifferenz auf eine Änderung des absoluten Rollwinkels in °/s skaliert. Diese Werte sollen, unabhängig davon, ob gerade eine Links- oder Rechtskurve durchfahren wird, ein Maß dafür bieten, wie ruhig eine Fahrt verläuft. Zu diesem Zweck wurde dieses Maß auch nochmal mit absoluten Werten berechnet. Die Sektoren wurden jeweils aus den Lichtschranken abgeleitet. Start- und Zielsektor verwenden Platzhalter, da keine weitere Lichtschranke vorhanden ist.

4.8 Datenfilterung und Behandlung von Ausreißern

Bei der Betrachtung der importierten Datensätze fallen einige mit invertierten X-Beschleunigungs-, Y-Beschleunigungs- und Rollwinkelwerten auf. Da diese Datensätze bis auf das inkorrekte Vorzeichen keine Auffälligkeiten besaßen, wurde sich dazu entschlossen, diese mit angepasstem Vorzeichen zu nutzen. Jeder Datensatz, welcher eine negative mittlere X-Beschleunigung vor der Startlinie aufweist, wird mit einer Vorzeichenanpassung aufbereitet.

Die Datenfilterung erfolgte in sechs einzelnen und aufeinanderfolgenden Schritten. In diesen Schritten wurde explorativ sowie mit dem linearen Modell überprüft, ob ein Datensatz für die spätere Auswertung geeignet ist. Der Begriff Datensatz bezieht sich im Kontext der Datenfilterung auf einen einzelnen Lauf und dessen Messergebnisse. Die Filterung, beziehungsweise die Bestimmung der Randwerte, erfolgte dabei zum Teil auf dem vollständigen Datensatz als auch auf einzelnen Datenzeilen. Ist die Voraussetzung auf Grund von fehlerhaften Werten, basierend auf den definierten Vorgaben, oder fehlenden Werten nicht gegeben, wurde der Datensatz verworfen. Die folgende Reihenfolge zur Erläuterung der Filter ist identisch zu der Reihenfolge der Ausführung der Filter, die in Python durchlaufen wird.

Der erste Filter entfernt alle Datensätze mit Non-Werten auf den Feldern „Time“, „Distance“, „Speed“, „Acc_X“, „Acc_Y“, „Acc_Z“ und „Roll“. Enthält das Datenset mindestens einen Non-Wert innerhalb einer der definierten Spalte pro Datensatz, wird dieser entfernt und im Laufe der Auswertung nicht betrachtet.

Als zweites wurden die Sektoren der Bahn in Winterberg innerhalb des Datensatzes geprüft. Hierfür wurde der vollständige Datensatz gegen jeden Sektor des Eiskanals gemappt. Enthält der Datensatz Sektoren, die nicht dem Eiskanal Winterberg zugeordnet werden können, wird dieser Datensatz aus der Auswertung vollständig entfernt. Des Weiteren wird sichergestellt, dass keiner der Sektoren von Winterberg innerhalb des Datensatzes fehlt. Dies spricht für eine falsche Zuordnung oder eine fehlerhafte Messung. Da ein Verschieben der Sektoren und

der damit einhergehenden Rollwinkel etc. die Ergebnisse der Auswertung verfälschen würde, wurden die hiervon betroffenen Datensätze verworfen. Im Zuge der Überprüfung der Sektoren wurde auch sichergestellt, dass die Sektoren innerhalb des Datensatzes in der richtigen Reihenfolge durchfahren wurden. Ist diese Grundannahme für den Datensatz nicht zutreffend, wurde dieser aus der weiteren Auswertung entfernt.

Die Geschwindigkeit im Sektor „S“ dient als weiterer Filter der Daten. Ist diese Geschwindigkeit geringer als 0,25 Meter pro Sekunde, was ca. einem Kilometer in der Stunde entspricht, wird der Datensatz verworfen. Die Startgeschwindigkeit liegt im Durchschnitt aller Läufe bei ca. 10 m/s beziehungsweise 36 km/h.

Durch Sichtung der Daten wurde zudem ermittelt, dass einige der Datensätze unrealistische Rollwinkel beinhalten. Anfangs wurden die Datensätze auf einen maximalen Rollwinkel von 150 Grad geprüft. Alle Datensätze, die diese Grenze überschreiten, wurden, wie in den Schritten zuvor, aus der Auswertung entfernt. Nach explorativer Betrachtung der Daten wurden zudem weitere Fahrten erkannt, die keine realistischen Messungen darstellten. Durch diese weitere Überprüfung der Ergebnisse wurde der maximale Rollwinkel auf 130 Grad reduziert.

Im letzten Schritt wird die Distanz, die während der Fahrt zurückgelegt wurde, überprüft. Die Distanz kann durch die Fahrlinie beeinflusst werden, was für eine Streuung der Distanzwerte innerhalb einer Bobkategorie sorgt. Wird der Eiskanal mit geringem Rollwinkel durchfahren, wird eine im Vergleich zu höherem Rollwinkel kürzere Strecke zurückgelegt. Um diese Streuung zu berücksichtigen, wurde für die Distanz ein Intervall von 1.200 Metern bis 1.800 Metern festgelegt. Dieses Intervall entspricht dem Durchschnitt ± der Standardabweichung multipliziert mit dem Faktor drei.

Durch diese Datenbereinigung wurden 23 Datensätze von insgesamt 1.026 ausgeschlossen. Damit wurden ca. 2 Prozent aller Datensätze als fehlerhaft erkannt und infolgedessen aus dem Datenset entfernt.

4.9 Bestimmung der Kurven

Die Bestimmung der Kurven erfolgt automatisiert mit Python. Zur Bestimmung wurde dabei ausschließlich der absolute Rollwinkel des Durchlaufs verwendet. Die Logik zur Bestimmung der Kurven folgt dem theoretischen Ansatz, dass eine Kurve aus zwei lokalen Minima und einem Maximum besteht. Minimum und Maximum der Daten können durch die Ableitung bestimmt werden. Um fehlerhafte

Messungen oder sonstiges Rauschen innerhalb der Daten bei der Bestimmung von Kurven zu vermeiden, wurde die Annahme getroffen, dass Kurven ein lokales Maximum von 10 Grad oder mehr beim Rollwinkel aufweisen müssen. Alle Werte mit einer geringeren Gradzahl wurden durch die Logik nicht betrachtet. Um die Kurven zu bestimmen, wurden zunächst das lokale Minimum und Maximum bestimmt.

Abbildung 29: Kurvenverlauf am Beispiel der K5

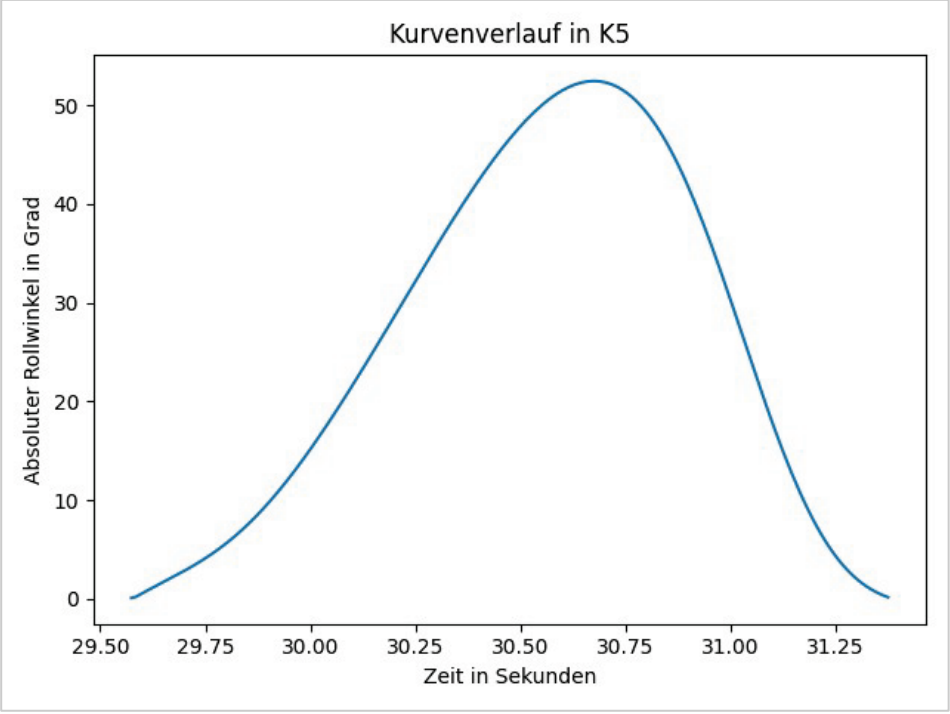


Abbildung 29 zeigt eine beispielhafte Durchfahrt der Kurve 5. Dabei ist zu erkennen, dass die Kurve aus zwei lokalen Minima und einem lokalen Maximum besteht. Mit Hilfe dieser Werte kann im Fall von Kurve 5 der Beginn und das Ende der Kurvenfahrt eindeutig bestimmt werden. Das vorher aufgestellte Modell zur Bestimmung von Kurven würde hier korrekt funktionieren.

Abbildung 30: Kurvenverlauf am Beispiel des Veltins-Kreisel

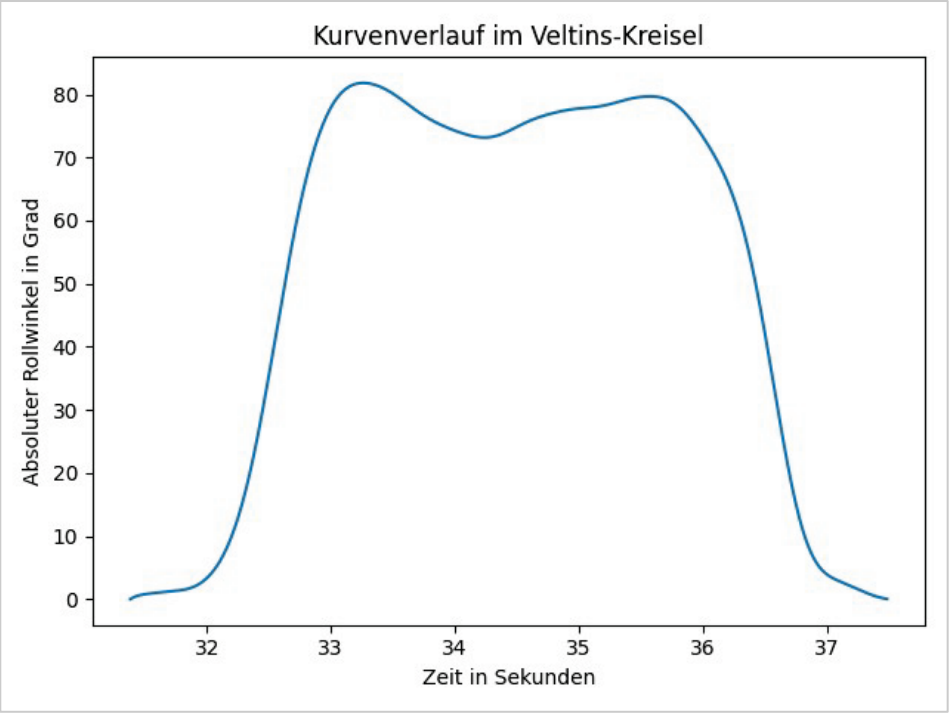


Abbildung 30 zeigt eine beispielhafte Durchfahrt der Kurve „Veltins-Kreisel“. Hier ist zu erkennen, dass die Kurve drei lokale Minima und zwei lokale Maxima aufweist. Dies würde durch die bisher vorgestellte Logik dazu führen, dass die Kurve in zwei unterschiedliche Kurven unterteilt wurde. Um dies zu vermeiden, wurde eine Änderung an der initialen Logik vorgenommen. Als neue ergänzende Regel wurde eingeführt, dass ein lokales Minimum einen geringeren Rollwinkel als 8 Grad haben muss. Ist dies nicht der Fall wurde davon ausgegangen, dass es sich um einen etwas weniger scharfen Teil der gleichen Kurve handelt.

Für die Umsetzung wurde aus der Bibliothek *Scipy* die Funktion „argrelextrema“ genutzt. Die Funktion erlaubt sowohl das Bestimmen von lokalen Minima als auch lokalen Maxima. Zudem kann der Funktion übergeben werden, wie viele Daten links und rechts vom aktuellen Wert überprüft werden sollen. Die Funktion testet jeden der Werte und überprüft je nach Suche, ob sich der Wert doch noch weiter verringert oder erhöht. Auf Grund der hohen Datenanzahl pro Durchlauf wurde der Wert beim Aufruf hier auf 50 gesetzt. Über alle gefundenen Maxima wurde iteriert und nach der oben beschriebenen Theorie der Start und das Ende der

Kurve bestimmt. Alle Zeitstempel innerhalb dieses bestimmten Intervalls wurden anschließend als Kurve im Datensatz markiert. Liegt das nächste lokale Maximum zeitlich innerhalb des letzten Kurvenintervalls, handelt es sich bei der Kurve um einen Fall. Das Maximum wurde daher übersprungen und die Erkennung mit dem nächsten Maximum fortgesetzt.

Durch teilweise sehr geringe Rollwinkel im ersten Sektor, gerade bei den Monobobs oder Zweierbobs, wurden diese teilweise von der Kurvenerkennung übergangen. Um dies zu verhindern, wurde bei zu geringem Rollwinkel der Eintrag durch erweiterte Prüfung zu den lokalen Maxima hinzugefügt.

Abbildung 31: Beispielfahrt – Unterteilung in mittlere Sektorzeit, identifizierte Kurven markiert

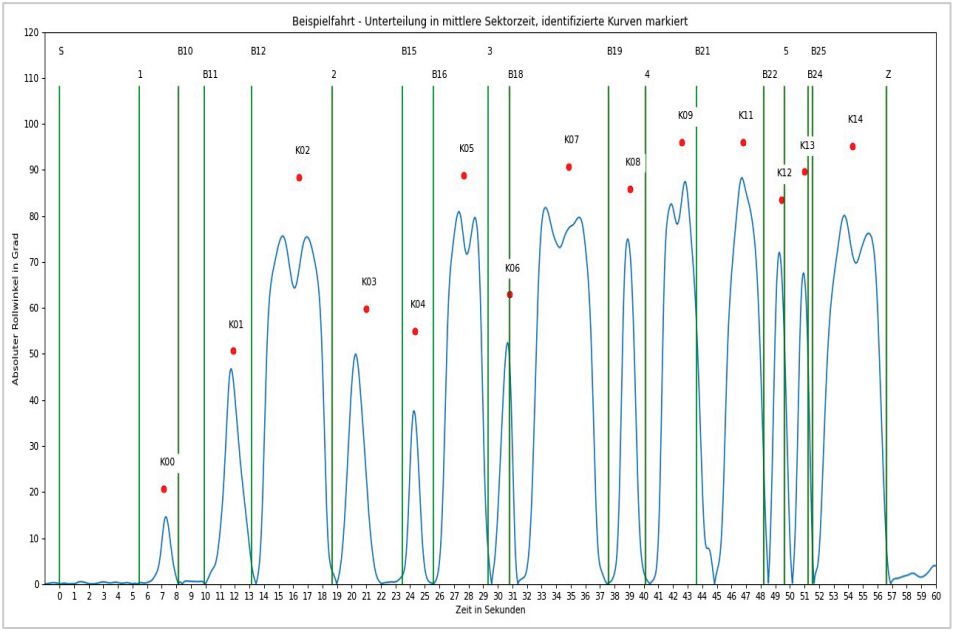


Abbildung 31 veranschaulicht alle erkannten Kurven einer Beispielfahrt und setzt sie ins Verhältnis zum Durchschnitt sowie den Sektoren, in denen die Kurve liegt. Erkennbar ist, dass die hier betrachtete Fahrt einen im Vergleich zum Durchschnitt geringeren Rollwinkel hat. Dies kann beispielsweise an einer geringeren Geschwindigkeit oder an einer leichteren Bobklasse liegen. Alle Kurven der Fahrt wurden jedoch erfolgreich durch die Logik ermittelt.

4.10 Berechnung der aggregierten Modellvariablen

Die Einteilung der Fahrt in Sektoren und Kurven erlaubt nun die Aggregierung der Fahrtvariablen um diese Elemente als Gruppierung. Die aggregierten Werte können dann als Modell-Feature verwendet werden. Auch können diese Aggregierungen genutzt werden, um allgemeingültige Rückschlüsse auf die Fahrt in dem jeweiligen Sektor zu ziehen. Eine Sektordurchfahrt, in welcher eine Korrelation mit negativem Vorzeichen für die Standardabweichung des Rollwinkels gefunden wurde, deutet darauf hin, dass eine ruhigere Durchfahrt zu Zeitreduktionen führt (Minimierung der Fahrtzeit).

Für die Sektoren wurden folgende Aggregationsmaße berechnet: der erste, letzte und mittlere Wert der Geschwindigkeit, der minimal, maximale mittlere Wert der (vektorlosen) Beschleunigung inklusive Standardabweichung, Einfahrts- und Ausfahrtszeit und die bei Einfahrt und Ausfahrt gefahrene Distanz. Die Differenz der Geschwindigkeit, Zeit und Distanz von Einfahrt und Ausfahrt wurden ebenso berechnet.

Für die Kurven wurden folgende Aggregationsmaße berechnet: die erste, letzte und mittlere Geschwindigkeit, minimale, maximale, mittlere Querbeschleunigung sowie Standardabweichung, minimale, maximale, mittlere Vertikalbeschleunigung sowie Standardabweichung, Einfahrts- und Ausfahrtszeit, Einfahrts- und Ausfahrtsdistanz, mittlerer, absoluter Rollwinkel, mittlere Änderungsrate des absoluten Rollwinkels sowie Standardabweichung und mittlere, absolute Änderungsrate des absoluten Rollwinkels sowie Standardabweichung. Die Differenz der Geschwindigkeit, Zeit und Distanz von Einfahrt und Ausfahrt wurden ebenso berechnet.

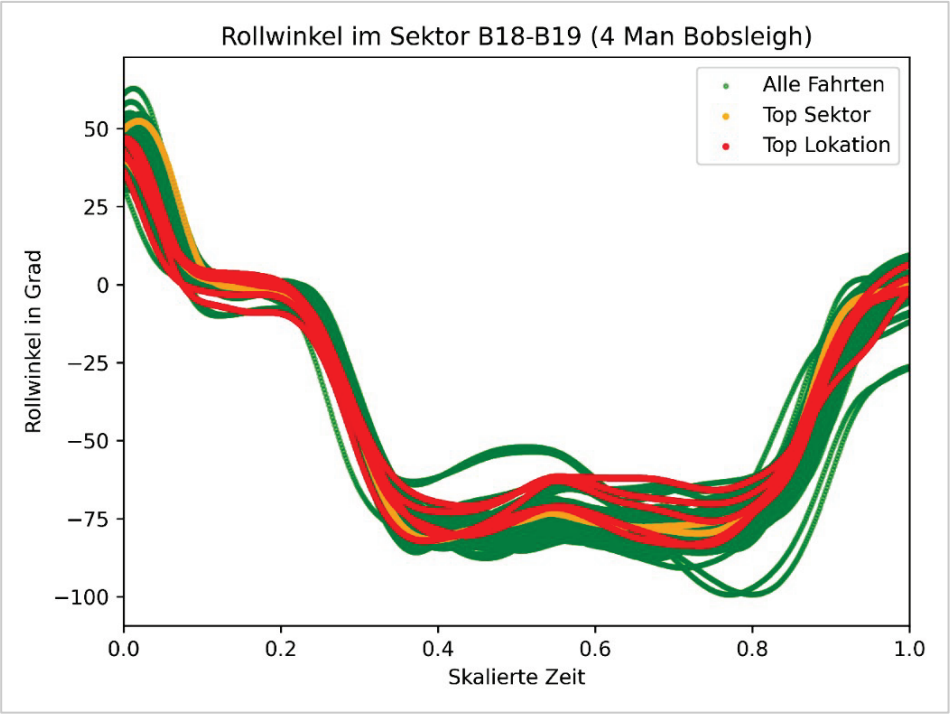
4.11 Modellentwicklung

4.11.1 Datenvisualisierung

Um Fortschritte und Zustände der Daten darstellen zu können, musste eine sinnvolle Visualisierungsform gefunden werden. Im Idealfall wären die Fahrten auf ein virtuelles Modell der Bahn zu projizieren gewesen. Dazu fehlten jedoch ein Modell der Bahn, welche (ohne komplizierte Datenumwandlungen) eine dreidimensionale Darstellung der Fahrten erlaubt hätte. Eine simplere Alternative musste genutzt werden.

Es wurde sich zunächst dazu entschlossen, die Fahrten in die einzelnen Sektoren zu unterteilen. Dabei kann davon ausgegangen werden, dass Sektordurchfahrten sich grundsätzlich ähneln. Operierend mit dieser Annahme, wurden die Fahrten anhand der Zeit, die benötigt wurde, um den Sektor zu durchfahren, normiert. Um die Kurvenform zu visualisieren, wurde als Y-Achsenwert der Rollwinkel genutzt. Eine dieser ersten Visualisierungen ist in Abbildung 32 zu sehen.

Abbildung 32: Initiale Visualisierung der Fahrten (Sektor B18-B19)



Um qualitative Unterschiede zwischen den Fahrten darzustellen, wurden die drei Fahrten mit der besten Zeit im Sektor bzw. mit der besten Fahrtzeit farblich anders kodiert. Auch kann nach einer Subgruppierung gefiltert werden, hier Viererbobfahrten im Sektor B18-B19.

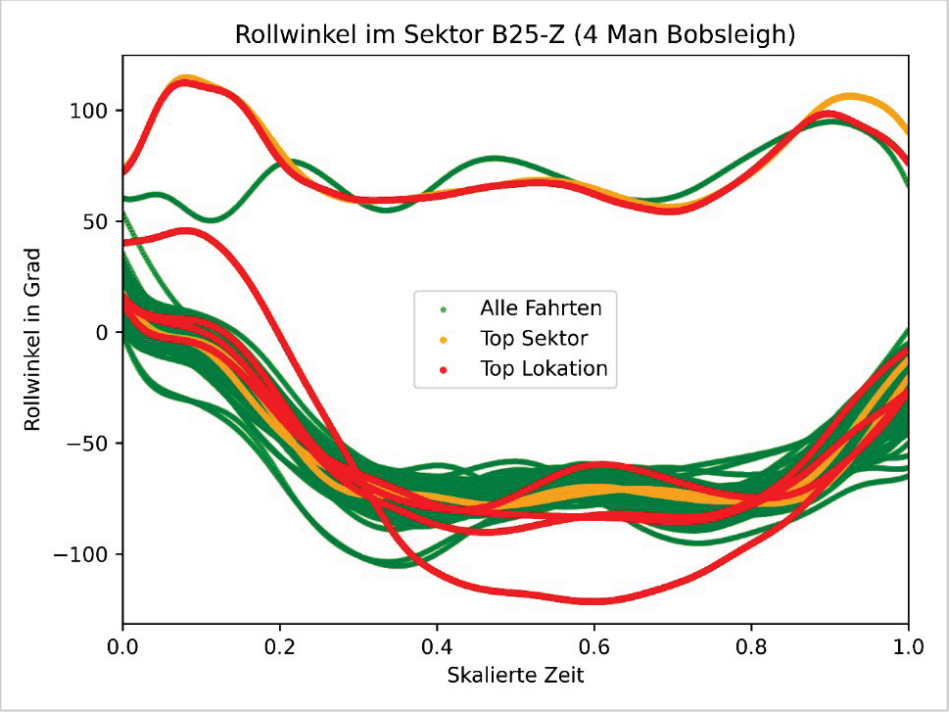
Mit dieser Visualisierung lässt sich ebenfalls prüfen, ob die Annahme, dass Kurvenfahrten sich ähneln, visuell bestätigt wird. Weiter kann der Rollwinkelverlauf auf Plausibilität mit Nutzung des Plans der Bahn geprüft werden. Die Lichtschranke B18 liegt am Ende einer Rechtskurve und leitet in den sich linkswendenden Veltins-Kreisel. Die in Abbildung 33 dargestellten ersten Datenpunkte im

Sektor B18-B19 zeigen zuerst eine Rückkehr zum Nullpunkt des Rollwinkels und folgend eine weitere Umkehrung nach Einfahrt in die Kurve.

4.11.2 Data Preparation und Modelling Synergie

Im Sinne der iterativen Natur der beiden CRISP-DM Phasen Data Preparation und Modelling können die erstellten Visualisierungen genutzt werden, um defekte oder anderweitig beschädigte Datensätze zu erkennen.

Abbildung 33: Sichtung von Ausreißern in Visualisierungen (B25-Z)

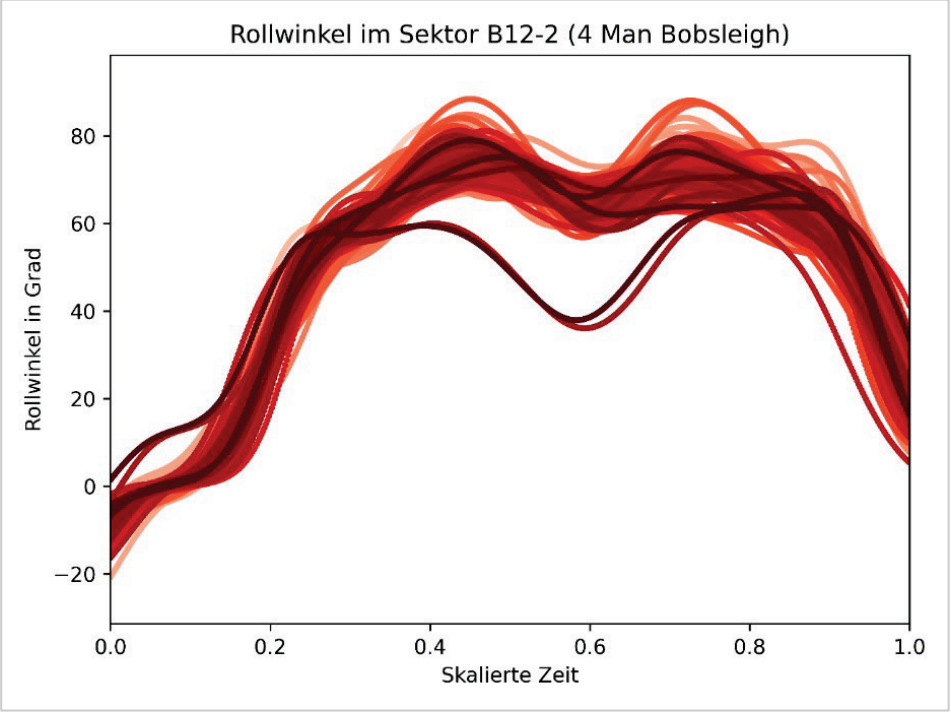


Ein Beispiel ist in Abbildung 33 zu sehen. Bei dem Sektor B25-Z handelt es sich um die Zieleinfahrtskurve, eine hufeisenförmige, lange Linkskurve. Die Lichtschranken befinden sich beide in geraden Streckenpositionen. Der Verlauf der Daten sollte dies entsprechend widerspiegeln. Zu erwarten ist eine Einfahrt um den Nullpunkt, ein Minimum beim Apex der Kurve und eine Annäherung zurück zum Nullpunkt beim Verlassen der Kurve. Die oberen drei Verläufe sind neben

der Abweichung von der deutlich zu erkennenden Gruppierung um den erwarteten Verlauf physikalisch unmöglich. Ähnlich defekte Fahrten wurden aus dem Gesamtdatensatz entfernt.

Bei einer weiteren Art von Ausreißern weichen die Werte in unerwarteter Weise ab. In Abbildung 34 sind im Bereich der skalierten Zeit zwischen 0,3 und 0,8 zwei Verläufe, welche unerklärter Weise niedrig verlaufen. Obwohl unwahrscheinlich, sind diese Verläufe nicht direkt zu entfernen.

Abbildung 34: Sichtung von unwahrscheinlichen Ausreißern (Sektor B12-2)



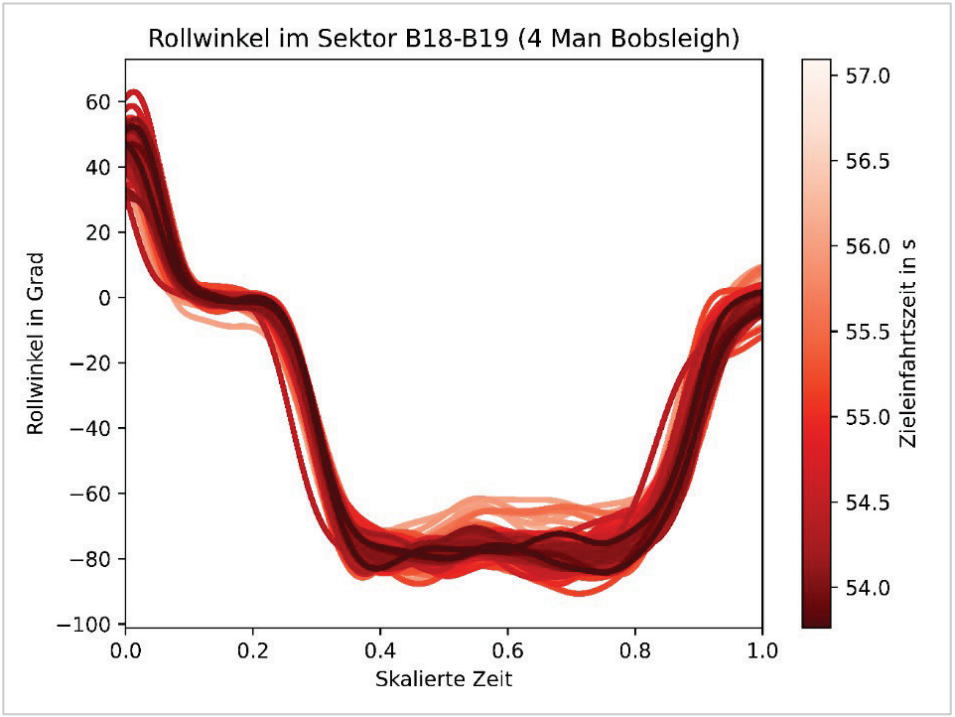
Eine manuelle Untersuchung dieser beiden Verläufe ergab, dass eine Deckelung der Werte für den Rollwinkel im positiven Wertebereich stattfand, jedoch nicht im negativen. An mehreren Stellen fällt der Rollwinkel unter -60°. Jedoch steigt er in beiden Rechtskurven der Bahn nicht über ungefähr 60°. Es wurde von einem Defekt der Messinstrumente ausgegangen und die Fahrten manuell aus dem Gesamtdatensatz entfernt. Unter Nutzung dieser Vorgehensweise wurde eine weitere Gruppe von fehlerhaften Fahrten entfernt, welche über den Streckenverlauf

eine ansteigende Abweichung der Zeit und damit Verschiebung von Werten aufweisen.

4.11.3 Überarbeitete Visualisierung

Wie bereits in Abbildung 35 zu sehen, wurde die Visualisierung angepasst. Um diese Änderungen besser zu verdeutlichen, wurde in Abbildung 36 ein anderer Sektor hervorgehoben:

Abbildung 35: Überarbeitete Visualisierung mit bereinigten Daten (Sektor B18-B19)

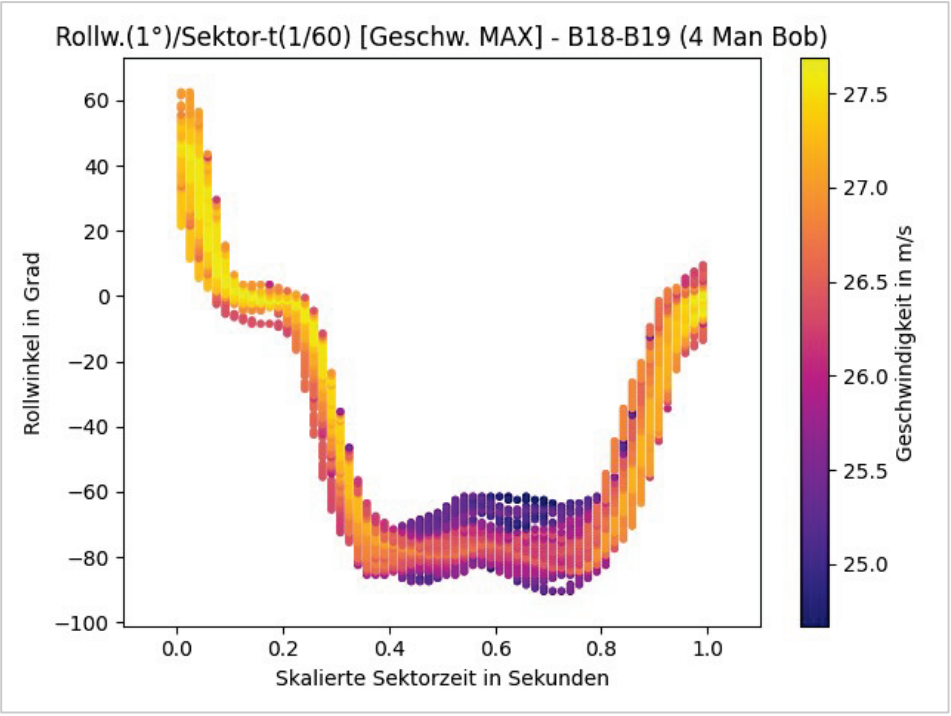


Zum einen sind nach der Entfernung der in anderen Sektoren identifizierten Ausreißer oder abnormalen Fahrten hier keine solche mehr zu erkennen. Alle hier zu sehenden Fahrten sehen plausibel aus. Zum anderen wurde die dreiteilige Sub-Kategorisierung der Fahrten nicht vorgenommen. Da die finale Zielvariable die Idealisierung der Fahrlinie und dadurch indirekt die Verbesserung der Fahrtzeit sein sollen, bestimmt die Zieleinfahrtszeit die Farbe der Kurve. Diese Form der

Visualisierung ist, obwohl sehr informativ, durch die Überlappung der Fahrten nicht für die Visualisierung einer idealen Fahrlinie geeignet.

Unter der weiteren Annahme, dass bei einer höheren Durchfahrtsgeschwindigkeit pro Sektor sich die Zieleinfahrtszeit reduziert, wurde dieses als Farbkodierung benutzt. Des Weiteren wurde, anstatt alle Fahrten einzeln zu betrachten, eine Unterteilung und Aggregierung vorgenommen. Der Sektor wurde in Rollwinkelschritte unterteilt und die skalierte Sektorzeit mit einem Divisor geteilt. Für jeden Abschnitt wurde dann eine Aggregationsfunktion genutzt. Die Visualisierung für den Veltins-Kreisel ist in Abbildung 36 zu sehen:

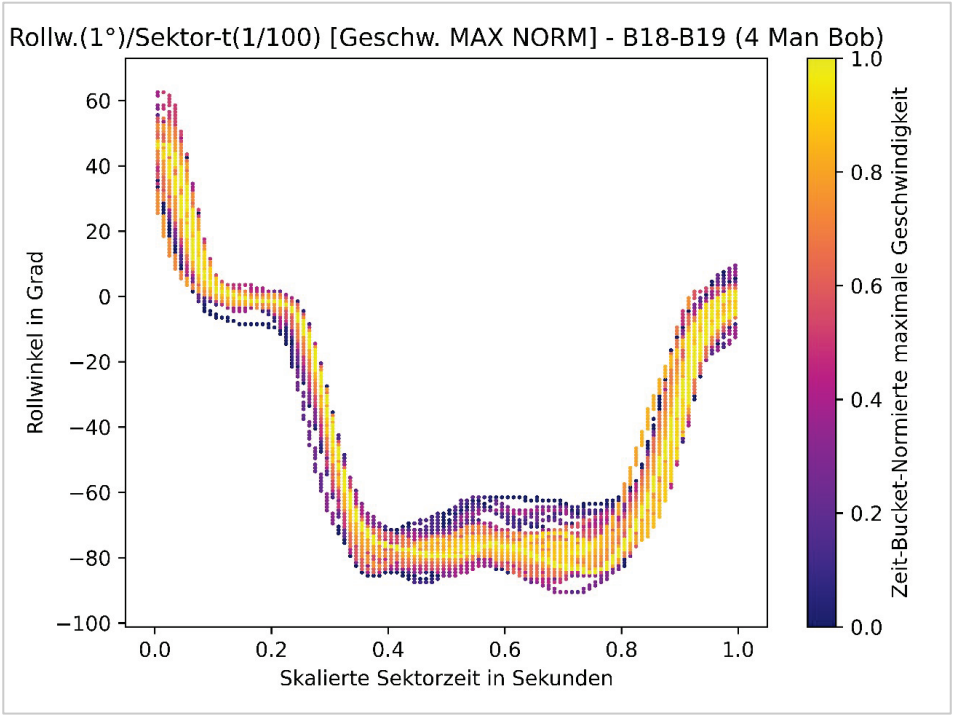
Abbildung 36: Überarbeitete Visualisierung mit Datenunterteilung und Aggregierung (Sektor B18-B19)



Da sich die Skalierungen für Rollwinkel und Sektorzeit anpassen lassen, kann die Sektordurchfahrt unter Nutzung unterschiedlicher Auflösungen betrachtet werden. Auch ist der „ideale“ Fahrkorridor nun deutlicher zu erkennen. Da sich jedoch die Farbkodierung auf den gesamten Sektor und die totale Geschwindigkeit bezieht, ist durch die relativen Geschwindigkeitsunterschiede die „ideale“

Fahrlinie nicht deutlich zu erkennen. Dies soll durch Skalierung jedes einzelnen Zeitabschnittes verbessert werden und ist in Abbildung 37 gezeigt. Der „Idealkorridor“ ist nun ohne weitere Mittel zu erkennen. Insbesondere zwischen den Sektorzeiten 0,4 und 0,8 ist zu erkennen, dass es eine Reduzierung der Geschwindigkeit bei Abweichung von der „idealen“ Fahrlinie stattfindet. Dies ist ebenso eine Validierung der Methodik.

Abbildung 37: Überarbeitete Visualisierung mit zeitabschnittnormierter Datenunterteilung und Aggregierung (Sektor B18-B19)



4.12 Multiple Lineare Regression

4.12.1 Modelldefinition

Zur Durchführung der multiplen linearen Regression wurde die Bibliothek *statsmodels* verwendet. *statsmodels* bietet hierfür mehrere Implementierungen von Regressionstechniken, dazu zählen die Methode der kleinsten Quadrate sowie die verallgemeinerte kleinste Quadrate Methode. Zur Erstellung des Modells

wurde die verallgemeinerte Version der kleinsten Quadrate verwendet. Diese bietet den Vorteil, dass sie im Fall von Heteroskedastizität oder Autokorrelation zuverlässiger funktioniert und dabei bessere Ergebnisse als ein herkömmliches OLS erzielen kann (vgl. Beguería & Pueyo, 2009).

Zur Erstellung und Verifizierung des initialen Modells wurde ausschließlich mit dem Viererbob-Datensatz gearbeitet. Dieser bietet die meisten Durchläufe und lässt sich auf Grund der hohen Anzahl exogener Variablen am besten auswerten. Das Modell wurde mit einem vorausgewählten Subset der möglichen exogenen Variablen erstellt. Diese Vorauswahl wurde durch die Betrachtung von Streudiagrammen getroffen. Alle Diagramme, die optisch eine Korrelation zu der endogenen Variablen aufwiesen, wurden initial für die Modellerstellung übernommen. Der Datensatz wurde in einem Verhältnis von 75:25 in Trainingsdaten und Testdaten unterteilt. Anschließend wurden die Daten mit dem „StandardScaler“ von „Scikit-Learn“ standardisiert. Der Schritt der Standardisierung wurde sowohl für die Trainingsdaten als auch die Testdaten unternommen. Auf Basis der Erkenntnisse von Altman und Royston (2006) wurde die Zielvariable nicht wie beispielsweise bei Zanoletti *et al.* (2006) dichotomisiert, sondern in ihrer metrischen Form beibehalten (vgl. Altman & Royston, 2006; Zanoletti *et al.*, 2006). Zudem ist die Zielvariable, um sie weiterhin interpretieren zu können, ausgenommen von der Standardisierung.

Für eine Analyse der optimalen Fahrweise nach Zhang *et al.* (1995) oder die Erstellung einer Simulation der Bahn nach Rempfler und Glocker (2016) fehlen der Datenbasis die Informationen der Lenkbewegungen und eine mathematische Repräsentation der Bahn am Standort Winterberg (vgl. Zhang *et al.*, 1995; Rempfler & Glocker, 2016). Ohne diese Daten konzentriert sich diese Arbeit auf die Vergleichbarkeit der Fahrten und setzt die relative Performance ins Verhältnis. Dabei wurden die Erkenntnisse referenzierter Arbeiten zur Erarbeitung der Modelle angewandt.

Das hier zur Anwendung kommende lineare Modell verwendet die exogenen Variablen „Date_2020-01-03“, „Date_2020-01-04“, „Date_2021-12-10“ Date_2021-12-11“, „Roll_Abs_Mean_K00“, „Roll_Abs_CR_Abs_Mean_K09“, „Speed_Mean_S-1“ und „Distance_Delta_B11B12“.

Das adjustierte Bestimmtheitsmaß des Modells liegt bei 0,809 und deutet dabei auf einen linearen Zusammenhang zwischen den verwendeten exogenen Variablen und der endogenen Zielvariable hin. Der Wert liegt hierbei immer zwischen Null und Eins, wobei ein Wert von Null bedeutet, dass kein linearer Zusammen-

hang besteht. Im Falle des Werts Eins kann von einem perfekten linearen Zusammenhang gesprochen werden. Das Modell erklärt damit ca. 81 Prozent der Varianz der Zielvariable. Der F-Test bestätigt, bei einem 0,05 α-Niveau, dass eine Korrelation zwischen dem Modell und der Zielvariablen besteht. Damit kann davon ausgegangen werden, dass das lineare Modell statistisch signifikant ist. Um dies weiter zu prüfen, lehnen wir uns an das allgemeine Testvorgehen nach dem Kompendium und der Zusammenfassung qualitativer Methoden und Prüfmethoden von Verbeek an (vgl. Verbeek, 2004).

4.12.2 Prüfung der Modellqualität

Zur Verifizierung der Modellqualität wurde das zuvor erstellte lineare Modell im Folgenden auf Autokorrelation, Heteroskedastizität, Fehlspezifizierung, fehlende Regressoren und Stationarität getestet.

Der Durbin-Watson-Test prüft auf Autokorrelation erster Ordnung und wurde mit Hilfe der „statsmodels“ Methode „durbin_watson“ durchgeführt. Der Test gibt immer ein Ergebnis zwischen Null und Vier zurück. Ein Ergebnis von Zwei spricht dabei für keine Autokorrelation, Werte darunter sprechen für positive Autokorrelation und Werte darüber für negative Autokorrelation (vgl. Durbin & Watson, 1950; Perktold *et al.*, 2021c). Das Ergebnis des Tests liegt bei 2,1876. Damit liegt eine geringe negative Autokorrelation erster Ordnung bei den Residuen vor.

Der Breusch-Godfrey-Test prüft auf Autokorrelation n. Ordnung. Zur Durchführung wurde die mit der Methode *accor_breusch_godfrey* verfügbare Implementierung von *statsmodels* verwendet. Die Nullhypothese des Tests ist, dass keine Autokorrelation bis einschließlich zur n. Ordnung vorliegt (vgl. Godfrey, 1978; Perktold *et al.*, 2021a). Geprüft wurde das Modell auf Autokorrelation bis zur zwölften Ordnung. Die p-Werte für den F-Test und den Lagrange-Multiplier-Test liegen beide weit außerhalb des 0,05 α-Niveaus, dadurch kann die Nullhypothese nicht verworfen werden. Damit widerspricht der Breusch-Godfrey-Test dem Durbin-Watson-Test und sagt aus, dass keine Autokorrelation vorliegt.

Die *statsmodels* Methode *het_breuschpagan* implementiert den Breusch-Pagan-Test basierend auf dem Lagrange-Multiplier-Test für Heteroskedastizität. Der Breusch-Pagan-Test stellt die Nullhypothese auf, dass Homoskedastizität vorliegt. Die Varianz der Störterme ist daher konstant (vgl. Breusch & Pagan, 1980; Perktold *et al.*, 2021d). Die *statsmodels* Methode gibt sowohl für die F-Statistik als auch für den Lagrange-Multiplier-Test einen p-Wert über dem angestrebten

α-Niveau von 0,05 zurück. Dadurch kann die Nullhypothese nicht verworfen werden und es ist von Homoskedastizität auszugehen.

Der Test nach White ist ein zweiter Test für Heteroskedastizität der mit Hilfe der *het_white* Methode von *statsmodels* durchgeführt wurde. Wie der Breusch-Pagan-Test geht der White-Test in der Nullhypothese von Homoskedastizität aus (vgl. White, 1980; Perktold *et al.*, 2021e). Auf Grund des hohen p-Wertes kann die Nullhypothese nicht verworfen werden, damit bestätigt der White-Test die Ergebnisse des Breusch-Pagan-Tests. Auf Grund mangelnder anderer Hinweise ist davon auszugehen, dass das erstellte Modell Homoskedastizität aufweist.

Der linear Regression and Specification Error Test (RESET) nach Ramsey stellt die Nullhypothese auf, dass keine Fehlspezifikation der Linearität oder unbeachtete Regressoren bestehen (vgl. Ramsey, 1969; Perktold *et al.*, 2021g). Der Test wurde mit der *linear_reset* Methode von *statsmodels* zur vierten Potenz durchgeführt, sowohl für den Wald-Test als auch den F-Test. Der Wald-Test stellt die Nullhypothese auf, dass der Koeffizient in Wirklichkeit nicht 0 beträgt (vgl. Wald & Wolfowitz, 1939). Der p-Wert des Wald-Chi-Quadrat-Tests sowie des F-Test liegen über dem α-Niveau von 0,05. Damit kann die Nullhypothese, dass keine Fehlspezifikation oder unbeachtete Regressoren bestehen, nicht verworfen werden.

Der Augmented Dickey-Fuller-Test (ADF) wurde zum Testen der Residuen auf Stationarität verwendet. Dabei prüft der ADF-Test einen zeitbezogenen Datensatz gegen einen autoregressiven gleitenden Mittelwert. Die Nullhypothese des ADF-Tests ist, dass die Daten keine Form der Stationarität aufweisen (vgl. Dickery & Fuller, 1979; Said & Dickey, 1984; Perktold *et al.*, 2021b). *statsmodels* bietet zur Durchführung des Tests eine bereits vollständig implementierte Lösung mit der Funktion *adfuller* an. Mit einer Teststatistik von -5.6757 ist der Wert geringer als der kritische 1 Prozent-, 5 Prozent- und 10 Prozentwert der Teststatistik, dadurch kann die Nullhypothese verworfen werden. Der p-Wert erlaubt uns ebenfalls die Nullhypothese bei einem α-Niveau von 0,05 zu verwerfen, der ADF-Test bestätigt daher eine Stationarität der Residuen.

Der Kwiatkowski-Phillips-Schmidt-Shin-Test (KPSS) prüft auf Trendstationarität. Die Nullhypothese ist hierbei, dass eine Trendstationarität der untersuchten Daten vorliegt. Die Funktion *kpss* von *statsmodels* bietet eine fertige Implementierung des KPSS-Tests (vgl. Kwiatkowski *et al.*, 1992; Perktold *et al.*, 2021f) . Für die Residuen gibt der Test einen p-Wert von 0,1. Basierend auf dem p-Wert und einem α-Niveau von 0,05 kann die Nullhypothese nicht verworfen werden, damit bestätigt der KPSS-Test das Ergebnis des ADF-Tests. Ausgehend vom Ergebnis

des ADF-Tests und KPSS-Test ist zu einem α-Niveau von 0,05 davon auszugehen, dass bei den Residuen Stationarität vorliegt.

4.13 Ergebnis, Ausblick und Fazit

4.13.1 Behandelte Forschungsthemen

Die Arbeit beschäftigte sich mit der Auswirkung der Fahrlinie in Bezug auf die benötigte Zeit bis zur Zieleinfahrt. Dafür wurden die zu untersuchenden Hypothesen mit Hilfe der Literaturrecherche, Visualisierung und einem linearen Modell betrachtet. Dabei konnte sowohl durch die Literatur als auch das lineare Modell bestätigt werden, dass besonders die Startgeschwindigkeit einen signifikanten Einfluss auf das Ergebnis hat. Eine höhere Startgeschwindigkeit korreliert dabei mit der geringeren Gesamtzeit eines Laufs. Damit kann die Nullhypothese, dass die Startgeschwindigkeit keinen signifikanten Einfluss hat, verworfen werden.

Der direkte Einfluss der Fahrlinie auf die Zeit konnte nicht bestimmt werden. Jedoch kann durch das lineare Modell die Aussage getroffen werden, dass die Fahrlinie einen Einfluss auf die Zeit des Laufs ausübt. Die tiefergehende Annahme, dass allgemein ein höherer oder niedriger Rollwinkel von Vorteil ist, kann durch die Visualisierung und das lineare Modell nicht bestätigt werden. Der ideale Rollwinkel ist nach den Untersuchungen von Kurve zu Kurve individuell.

Abbildung 38: Korrelation der Zielzeit mit der durchschnittlichen Sektorgeschwindigkeit

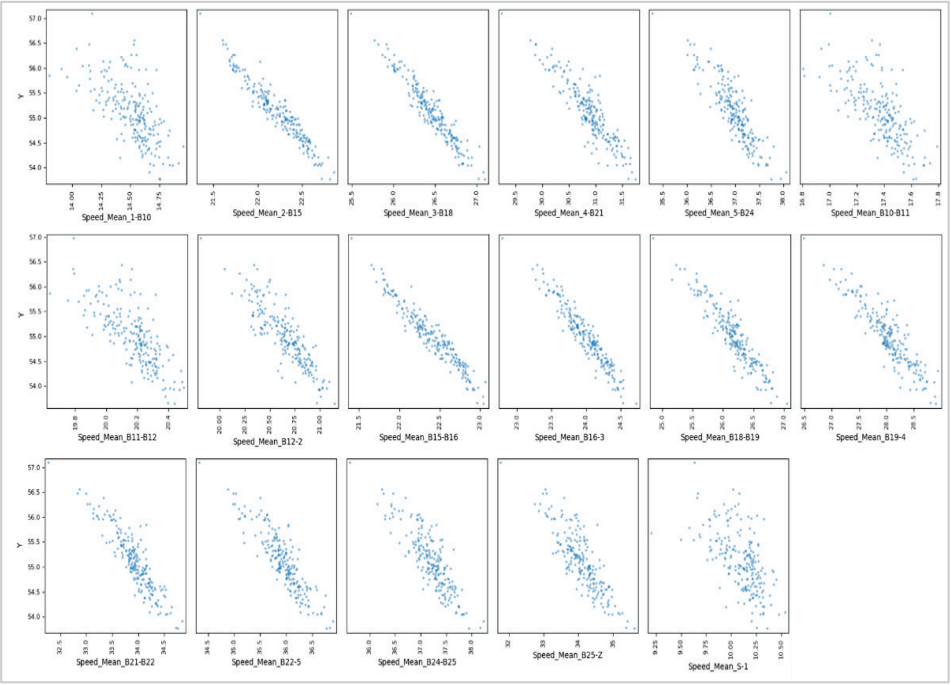


Abbildung 38 zeigt die Korrelation der durchschnittlichen Geschwindigkeit innerhalb eines Sektors zur Zieleinfahrtszeit. Dabei lässt sich erkennen, dass eine Überperformance beziehungsweise eine höhere durchschnittliche Geschwindigkeit mit der Zieleinfahrtszeit stark korreliert. Daher wurde die Hypothese, dass eine Überperformance in allen Sektoren nicht zu einem besseren Ergebnis führt, verworfen.

4.13.2 Diskussion der eingesetzten Methode

Datenanomalien konnten mithilfe der genutzten Visualisierungsmethoden erkannt und entfernt werden. Durch die genutzten Aggregierungen um die Sektoren und Kurven wurden viele Features für ein potenzielles Modell generiert. Mit dem Einsatz der Least Squares Methode konvergierte das Modell auf die relevantesten Features und ignorierte irrelevante. Im finalen Modell blieben nur Features

mit hoher Signifikanz zurück. Eine erneute Betrachtung des Themas sollte andere Methoden in Erwägung ziehen. Die Überparameterisierung des Datenmodells könnte durch eine Auflösung der Gruppierung je Kurve und Sektor unter Abstrahierung der Eigenschaften dieser Gruppen aufgelöst werden.

Ein Problem der eingesetzten Methodik in Kombination mit den Daten ist das häufige Auftreten von Autokorrelation. Aufgrund der vielen nichtlinearen physikalischen Zusammenhänge ist dies zu erwarten. Zur Mitigation dieser Problematik wurde ein hohes Maß an Aggregation vorgenommen und die ermittelten Werte in Relation zueinander gesetzt, zudem wurde im Laufe der Umsetzung auf ein GLS gesetzt. Weiterhin wurde die Performance zwischen OLS und GLS verglichen.

4.13.3 Aussicht und Optimierungsbedarf

Eine mathematische Repräsentation der Bahn, um die Berechnungen von Zhang *et al.* (1995) oder die Simulationen anhand der Vorgehensweise von Rempfler und Glocker (2016) durchführen zu können, würde einen guten Vergleichswert für die getroffene Fahrt bieten können. Ebenso sind alternative Modellierungsformen wie bei Mössner *et al.* (2011) in Betracht zu ziehen (vgl. Zhang *et al.*, 1995; Mössner *et al.*, 2011; Rempfler & Glocker, 2016). Ein Vergleich der Abweichung von dieser optimalen Fahrlinie sollte bei weiteren Untersuchungen aufgenommen werden.

Durch diese Arbeit wurden diverse mathematische Modelle zur Berechnung und Simulation von Reibungskoeffizienten, Lenkverhalten und wirkenden Kräften in der Literatur identifiziert. Die Datenbasis dieser Arbeit lässt jedoch aufgrund fehlender Metadaten zu verwendetem Material, exakten Umweltbedingungen zum Wettkampfzeitpunkt und zur Physis der Sportler, keine Verifikation der Ergebnisse der Literatur zu. Die Erweiterung des Datensatzes durch zusätzliche Messwerte der Umgebungsparameter könnte die Ergebnisqualität weiter erhöhen.

Der Zugriff auf die originalen Messdaten ohne weitere Vorbereitung sollte für weitere Betrachtungen angestrebt werden. Hinzukommend sollten die Messgeräte einheitlich kalibriert und installiert sein. Wie die Datenaufbereitung zeigt, wurden mehrere Fahrten im Rahmen der Möglichkeiten rekalibriert, um für diese Arbeit verwendet werden zu können. Zudem wurden mehrere Fahrten ausgeschlossen, da entweder wichtige Messpunkte nicht erreicht oder verfälschend verschoben sind.

Wie bereits angesprochen, wurde statt OLS die Methode GLS eingesetzt, um die erwartete und erwiesene Autokorrelation und Multikollinearität des Datensatzes zu behandeln. Für weitere Ausarbeitungen sollte der Datensatz um weitere physikalische Größen und Messwerte erweitert werden, um, angelehnt an die Literatur, genauere Modellberechnungen vornehmen zu können.

5 Zum Zusammenhang von Bahnabschnitt und Gesamtlaufzeit

5.1 Zielsetzung

Im Rahmen dieser wissenschaftlichen Arbeit soll der Zusammenhang zwischen den Bahnabschnitten und der Gesamtlaufzeit auf der Bobstrecke in Winterberg analysiert werden. Die folgenden Forschungsfragen sollen dazu mittels verschiedener Methoden beantwortet werden:

1. Lassen sich Bahnabschnitte feststellen, die einen signifikanten Einfluss auf die Gesamtlaufzeit haben?
2. Welche Faktoren pro Bahnabschnitt haben einen signifikanten Einfluss auf die Laufzeit in diesem Abschnitt?
3. Lassen sich optimale Werte für die einzelnen Faktoren pro Bahnabschnitt bestimmen?

Mit den Ergebnissen dieser wissenschaftlichen Arbeit soll eruiert werden, auf welche Faktoren während der Fahrt mit dem Bob ein besonderer Fokus gelegt werden sollte. Darüber hinaus kann ein Ergebnis dieser Arbeit sein, dass gewisse Faktoren keinen signifikanten Einfluss auf die Gesamtlaufzeit des Bobs haben und somit eine Fokussierung und Optimierung dieser Faktoren nicht notwendig ist.

5.2 Theoretische Grundlagen

5.2.1 Gradient Tree Boosting

Gradient Boosting ist eine Technik des maschinellen Lernens, bei der in einem iterativen Prozess additive Regressionsmodelle erstellt werden, die jeweils auf den sogenannten Pseudo-Residuen der vorangegangenen Iteration antrainiert werden (vgl. Friedman, 2002, S. 367).

Beim Gradient Tree Boosting wurde das Konzept des Gradient Boosting mit dem Einsatz von Entscheidungsbäumen kombiniert, um Regressions- oder Klassifikationsaufgaben zu lösen (vgl. Friedman 2001, S. 1189). Die Grundidee hinter Gradient Tree Boosting besteht im iterativen Aufbau von Entscheidungsbäumen, die jeweils auf dem Prognosefehler des zuletzt erstellten Entscheidungsbaums trainiert werden. Durch das iterative Vorgehen wurde dabei mit jedem Baum ein kleiner Schritt in Richtung einer besseren Prognose begangen, bis eine definierte

Anzahl an maximalen Bäumen erreicht wurde oder weitere Bäume den Prognosefehler nicht weiter reduzieren (vgl. Friedman, 2001, S 1195ff).

In einer Regressionsaufgabe bildet der zugrundeliegende Algorithmus zuerst eine initiale Prognose, bei der es sich beispielsweise um den Durchschnitt des zu prognostizierenden Attributs handeln kann. Basierend auf dieser Prognose werden über eine Verlustfunktion die Fehler, die sogenannten Pseudo-Residuen, ermittelt. Nun bildet der Algorithmus iterativ Entscheidungsbäume, welche die gebildeten Pseudo-Residuen der jeweils vorausgegangenen Iteration prognostizieren. Die Pseudo-Residuen werden so schrittweise minimiert, wobei die sogenannte *learning rate* die Ergebnisse der Entscheidungsbäume skaliert und somit die Schrittgröße bestimmt. Die finale Prognose setzt sich anschließend aus der Addition von initialer Prognose und den skalierten Prognosen aller gebildeten Entscheidungsbäumen zusammen. Abbildung 39 veranschaulicht den Gradient Tree Boost Algorithmus (vgl. Friedman, 2002, S 367ff).

Abbildung 39: Gradient Tree Boost Algorithmus

Algorithmus: Gradient_TreeBoost

1 $F_0(x) = argmin_\gamma \sum^{N}_{i=i} \Psi\ (y_i,\gamma)$

2 For $m = 1$ to M do:

3 $\tilde{y}_{im} = -\left[\frac{\partial \Psi(y_{i,F(x_i)})}{\partial}\right]_{F(xi)F(x)=Fm-1(x)}, i = 1, N$

4 $\{R_{lm}\}_1{}^L = L - terminal\ node\ tree(\{\tilde{y}_{im}, x_i\}_1{}^N)$

5 $\gamma_{lm} = argmin_\gamma \sum_{xi \in Rlm} \Psi(y_i, F_{m-1}(x_i) + \gamma)$

6 $F_m(x) = F_{m-1}(x) + \nu * \gamma_{lm} 1(x \in R_{lm})$

7 endFor

Quelle: In Anlehnung an Friedmann (2002), S. 368.

Im ersten Schritt wurde der initiale Prognosewert bestimmt, indem ein einzelner Prognosewert $F_0(x)$ gesucht wird, welcher die Summe aller Verlustfunktionen $\Psi(y_i, \gamma)$ für alle Datensätze N minimiert, wobei γ für den prognostizierten und y_i für den jeweiligen beobachteten Wert steht. Bei Ψ kann es sich dabei um eine beliebige differenzierbare Verlustfunktion handeln, wie zum Beispiel die Kleinste-Quadrate-Verlustfunktion: $\Psi(y, F) = (y - F)^2$ (vgl. Friedman, 2001, S. 370). Im zweiten Schritt startet der iterative Prozess, indem die folgenden Schritte für jeden Entscheidungsbaum m wiederholt werden, bis die definierte Anzahl an Entscheidungsbäumen M erreicht wurde. Der dritte Schritt veranschaulicht die Ermittlung der Pseudo-Residuen $\tilde{y}_{im}$ anhand der Verlustfunktion Ψ für jeden Datensatz i und Entscheidungsbaum m. Im vierten Schritt wurde schließlich der Entscheidungsbaum erstellt, welcher die ermittelten Pseudo-Residuen prognostiziert. Im fünften Schritt wurde nun der Prognosewert γ gesucht, welcher die Summe über die Verlustfunktionen zwischen dem beobachteten Wert y_i und der Summe aus Prognose des vorherigen Entscheidungsbaumes $F_{m-1}(x_i)$ und Prognosewert γ minimiert. Dabei werden jeweils nur Datensätze betrachtet, die in einem Endknoten gebündelt sind ($x_i \in R_{lm}$), sodass für jeden Endknoten des Entscheidungsbaums ein Prognosewert entsteht. Im sechsten Schritt wurde der neue Prognosewert gebildet, indem die letzte Vorhersage mit dem durch die *learning rate* ν skalierten Ergebnis aus Schritt 5 kombiniert wird.

Bei *Extreme Gradient Boosting* (XGBoost) handelt es sich um eine skalierbare Implementation von Gradient Tree Boosting, welche die Approximation beim Erstellen der Entscheidungsbäume, den Umgang mit fehlenden Werten und den effizienten Einsatz von Systemressourcen und Kompression adressiert, um die Analyse von großen Datenmengen zu ermöglichen (vgl. Chen & Guestrin, 2016, S. 785).

5.2.2 Shapley Values

Eine Herausforderung im Bereich des maschinellen Lernens ist, das Ergebnis eines Vorhersagemodells auf die zugrundeliegenden Attribute zurückzuführen. Neben einer guten Vorhersage ist in vielen Bereichen ebenso von Bedeutung, welche Faktoren welchen Einfluss auf die konkrete Vorhersage hatten. Für einige Verfahren, wie beispielsweise die lineare Regression, lässt sich diese Attribuierung über die ermittelten Koeffizienten ableiten. Viele komplexere Modelle, wie zum Beispiel Neuronale Netze, besitzen jedoch keine intrinsische und aussagekräftige Attribuierung (vgl. Sundararajan & Najmi, 2019, S. 1).

Ein weit verbreitetes Verfahren, um eine Attribuierung zu ermöglichen und so die Modellergebnisse erklärbar zu machen, basiert auf den sogenannten *Shapley Values*. Dabei handelt es sich um ein Konzept aus der kooperativen Spieltheorie, bei der ein Ertrag anteilig auf eine Gruppe von Spielern verteilt werden muss. Jeder Spieler soll dabei anteilig nach seinem eigenen Beitrag bedacht werden (vgl. Shapley, 1953, S. 316).

Shapley (1953, S. 309) stellt zur Lösung des Problems drei Axiome auf:

- Symmetrie:
  Zwei Spieler sind austauschbar, wenn sie in jeder beliebigen Kombination der anderen Spieler immer denselben Beitrag zum Ergebnis leisten. Austauschbare Spieler erhalten einen gleichen Ertragsanteil.
- Null-Spieler:
  Ein Spieler ist ein Null-Spieler, wenn sein Beitrag zum Ergebnis in jeder Kombination von anderen Spielern immer Null ist. Null-Spieler erhalten keinen Ertragsanteil.
- Additivität:
  Lässt sich der zugrundeliegende Prozess, um die Erträge zu erzielen, in zwei Prozesse aufteilen, so ist der Ertrag jedes Spielers die Summe seiner Erträge der beiden Prozesse.

Mittels dieser drei Axiome lässt sich die Formel für den anteiligen Ertrag pro Spieler i aufstellen (vgl. Shapley, 1953, S. 311):

Formel 1: Shapley Value

$$\phi_i(N, v) = \frac{1}{N!} \sum_{S \subseteq N \setminus \{i\}} |S|! \, (|N| - |S| - 1)! \, [v(S \cup \{i\}) - v(S)]$$

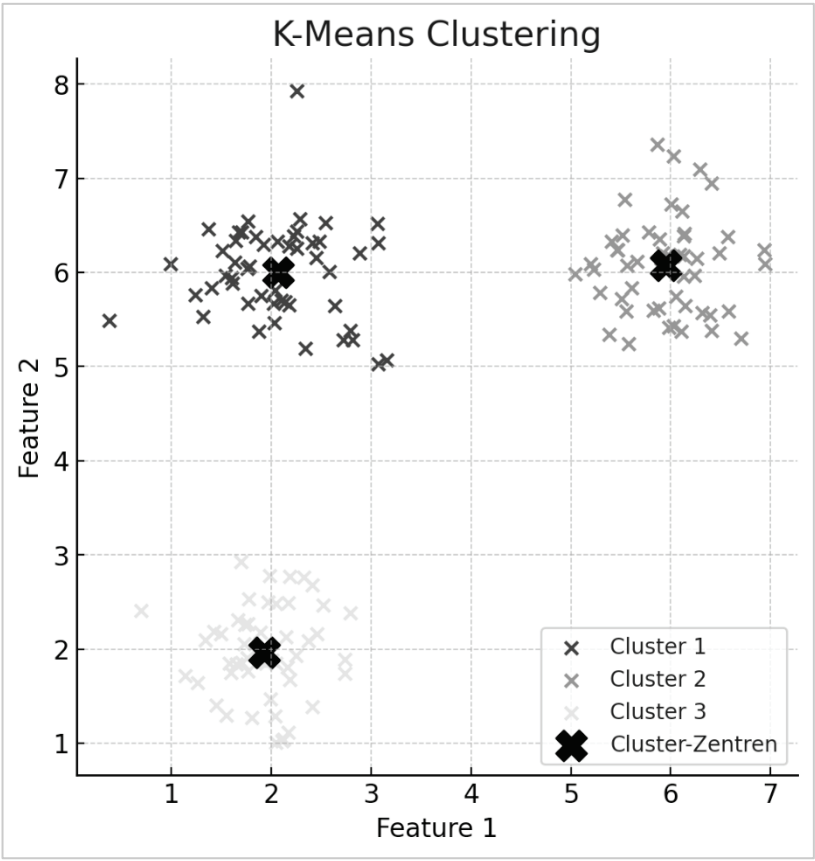
Dabei ergibt ϕi den anteiligen Ertrag pro Spieler bei einem Spiel mit N-Spielern und einem Gesamtertrag v, wobei S die Kombination von anderen Spielern ohne Spieler i darstellt. Übertragen auf maschinelles Lernen lässt sich nun der Gesamtertrag durch die Modellprognose und die Spieler durch die Modellattribute ersetzen (vgl. Sundararajan & Najmi, 2019, S. 1). Somit lassen sich Modellattribute basierend auf ihrem Beitrag zur Modellprognose einordnen und das Modellergebnis erklärbar machen.

5.2.3 K-means Clustering

Der K-means-Clusteralgorithmus ist ein unüberwachtes Lernverfahren, welches Cluster auf der Grundlage eines vordefinierten Optimierungsproblems unter Verwendung des bestimmten Kriteriums erstellt. K-means Clustering ist auch nach vielen Jahren noch beliebt. Die Forschung zu K-means lässt sich bis in die Mitte des letzten Jahrhunderts zurückverfolgen und wurde von zahlreichen Forschern aus verschiedenen Disziplinen durchgeführt, insbesondere von Lloyd in den Jahren 1957 und 1982 (vgl. Lloyd, 1957) und MacQueen im Jahr 1967 (vgl. MacQueen, 1967). Im Laufe der Zeit sind verschiedene Versionen des K-means-Algorithmus entstanden, die sich darauf konzentrierten, den Algorithmus zu verbessern, indem sie einige Vorverarbeitungsschritte durchführten oder die Anzahl der Iterationen reduzierten (vgl. Borlea *et al.*, 2022, S. 63f.).
Kurz gesagt ist K-means ein prototypenbasierter, einfacher Clusteralgorithmus, der versucht, K nicht überlappende Cluster zu finden. Diese Cluster werden durch ihre Schwerpunkte dargestellt. Dabei ist ein Cluster-Schwerpunkt in der Regel der Mittelwert der Punkte in diesem Cluster.

Der Clustering-Prozess von K-means läuft wie folgt ab. Zunächst werden K anfängliche Schwerpunkte ausgewählt, wobei K vom Benutzer festgelegt wurde und die gewünschte Anzahl von Clustern angibt. Jeder Punkt in den Daten wurde dann dem nächstgelegenen Schwerpunkt zugewiesen, und jede Sammlung von Punkten, die einem Schwerpunkt zugewiesen sind, bildet ein Cluster. Der Schwerpunkt jedes Clusters wurde dann auf der Grundlage, der diesem Cluster zugewiesenen Punkte aktualisiert. Dieser Vorgang wurde so lange wiederholt, bis kein Punkt mehr sein jeweiliges Cluster wechselt (Abbildung 40) (vgl. Wu, 2012, S. 7f).

Abbildung 40: K-means weist die Punkte dem nächstgelegenen Schwerpunkt zu



Quelle: EMC Education Services (2015), S. 121.

K-means bietet im Vergleich zu anderen Clustering-Algorithmen einige Vorteile. So ist K-means sehr einfach und robust, äußerst effizient und kann für eine Vielzahl von Datentypen verwendet werden. Daten mit komplizierten Eigenschaften, wie z. B. große Datenmengen oder eine hohe Dimensionalität, erfordern eine Anpassung des klassischen K-means an diese neuen Szenarien, was eine ständige Optimierung des ursprünglichen K-Means zur Folge hat. Die Nachteile von K-means, wie z. B. die schlechte Leistung bei nichtglobalen Clustern und die Empfindlichkeit gegenüber Ausreißern, werden oft von den Vorteilen überlagert und teilweise durch neue Algorithmusvarianten korrigiert (vgl. Wu, 2012, S. 8).

5.3 Analyse des Datensatzes

5.3.1 Datensatz

Der vorliegende Datensatz beinhaltet Aufzeichnungen von Läufen auf der Bahn in Winterberg. Für jeden Lauf liegt eine separate csv-Datei vor. Diese beinhaltet jeweils Messungen an bestimmten Streckenpunkten: Zeit in Sekunden (*Time s*), Distanz in Metern (*Distance m*), Geschwindigkeit in km/h (*Speed km/h*), Beschleunigung der x-Achse (*Acc x m/s2*), Beschleunigung der y-Achse (*Acc y m/s2*), Beschleunigung der z-Achse (*Acc z m/s2*), der Rollwinkel (*Roll angle*), der Messpunkt (*Lightbeam*) und die Zeit bis zum jeweiligen Messpunkt (*LB time s*). Demnach liegen uns für jede Nummer des Laufs jeweils ein separater Datensatz vor. Ordner unterteilen die Aufzeichnungen nach Strecke und Disziplin. Die Strecke ist der Austragungsort und die Disziplin unterscheidet zwischen Frauen und Männern und zum Einsatz kommendem Bob. Damit liegen insgesamt 1.010.432 Messpunkte für 151 Läufe für die weitergehende Analyse vor.

5.3.2 Datenbereinigung

Im Rahmen der vorliegenden Arbeit wurden die Daten für die weitere Analyse aufbereitet. Hierfür wurden zunächst die verfügbaren Dateien mit Hilfe der Pythonbibliothek *os* und *glob* eingelesen und in einem *pandas* Dataframe zusammengeführt. Zusätzlich wurde der Pfadname verwendet, um den Datensatz mit weiteren Informationen anzureichern. Demnach kann aus dem Pfadnamen die Strecke, die Athleten ID, die Nummer des Laufs, die Disziplin, das Jahr, der Monat und der Tag herausgelesen werden. Diese Aufbereitung erfolgt sowohl für die Aufzeichnungen der Trainings als auch der Weltmeisterschaft in Winterberg. Anschließend wurden die Spaltennamen in der Form angepasst, sodass diese im Rahmen der weiteren Programmierung besser verwendet werden können. Hierfür wurden Leerzeichen und Sonderzeichen entfernt. Somit sind alle Dateien in einem Dataframe zusammengeführt und können weiter aufbereitet werden. Die Variable *Lightbeam* wurde zum Datentyp „Float“ umgewandelt, um in der weiteren Analyse Berechnungen durchführen zu können. Zudem wurde der Datensatz gruppiert nach Athletinnen und Athleten, Disziplin und Tag, um den Datensatz auf Basis einzelner Läufe zu betrachten. Hierfür wurde jeweils der erste Datenpunkt jedes Laufs mit einem „PRE“ und der letzte Datenpunkt mit einem „POST“ markiert. Daraufhin wurden zwei weitere Spalten hinzugefügt, die jeweils den vorherigen bzw. nachfolgenden Datenpunkt wiedergeben. Anhand dieser beider Spalten kann der jeweilige Bahnabschnitt ermittelt werden. Anhand der gleichen

Methode wurde die Laufzeit für den Bahnabschnitt bestimmt, womit der erste Teil der Datenaufbereitung abgeschlossen war.
Im zweiten Schritt erfolgte das Feature Engineering, um den Datensatz für die folgenden Modelle vorzubereiten. Dabei wurden die vorhandenen Variablen dahingehend bearbeitet, dass sie pro Streckenabschnitt aggregiert wurden. Für alle nummerischen Variablen wie Geschwindigkeit, Rollwinkel oder Distanz wurde das Minimum, das Maximum, der Durchschnitt, der Median und die Standardabweichung berechnet. Durch die Aggregation wurden die Aufzeichnungen von über einer Million auf ca. 2.800 eingegrenzt. Zuletzt wurden Zeilen ohne Werte aus dem Datensatz entfernt. Weitere Features wurden hinzugefügt, wie die Laufzeit des Streckenabschnitts, die zurückgelegte Distanz im Streckenabschnitt, die Spannweite der Variablen Geschwindigkeit und die Beschleunigung der x-, y- und z-Achse des Bobs. Mit Hilfe des One-Hot-Encoding wurden für die kategoriale Variable *Disziplin* vier Dummys erzeugt (Frauen, Bobsleigh, Männer2, Männer4). Zu diesem Zeitpunkt bestand der Datensatz aus 47 Spalten und im Dataframe befanden sich pro Streckenabschnitt und Lauf ein Datensatz.

Da die einzelnen Streckenabschnitte Variablen für die Bemessung der Zielzeile sind, wurden diese für das XGBoost Modell in einer Zeile zusammengefasst, wodurch aus 47 Spalten nunmehr 570 wurden (Anzahl der Variablen multipliziert mit der Anzahl der Streckenabschnitte).

5.3.3 Deskriptive Analyse

Das Ergebnis der Datenaufbereitung beinhaltet zwei unterschiedliche Datensätze. Ein Datensatz ist auf Ebene des Bahnabschnitts aggregiert und der andere Datensatz bildet jeweils eine gesamte Fahrt ab. Grundlage für die deskriptive Analyse bildet der Datensatz pro Bahnabschnitt. Dieser besteht aus 47 Spalten und 2.582 Zeilen.

Tabelle 4: Deskriptive Analyse – Gesamtlaufzeit in Sekunden

Gesamt-Laufzeit	alle Disziplinen	Women	Men2	Men4	Bobsleigh
Minimum	54,042	56,74	55,02	54,042	55,79
Mittelwert	56,171	57,711	55,733	54,76	56,38
Maximum	59,190	59,19	56,91	55,23	57,22

Die Tabelle beschreibt die Gesamtlaufzeit in Sekunden auf der Strecke in Winterberg. Im Durchschnitt über alle Disziplinen dauert ein Lauf 56,171 Sekunden. Aus der Analyse lässt sich schließen, dass der Viererbob (Men4) die schnellste Disziplin ist, da für diese Disziplin der niedrigste Wert mit 54,043 Sekunden gemessen wurde. Zudem weisen auch die Mittelwerte für das Maximum und den Durchschnitt die niedrigsten Werte auf. Darüber hinaus zeigt sich, dass der Frauenbob (Women) tendenziell am längsten für die Strecke benötigt. Für diese Disziplin wurde der höchste Wert von 59,19 Sekunden gemessen und auch der Mittelwert und das Maximum liegen im Frauenbob durchschnittlich über den Werten der anderen Disziplinen.

Tabelle 5: Deskriptive Analyse – Geschwindigkeit in km/h

Geschwindigkeit	Alle Disziplinen	Women	Men2	Men4	Bobsleigh
Mittelwert	90,78	87,99	92,34	93,93	88,97

Über alle Disziplinen hinweg wurde ein Mittelwert von 90,78 km/h ermittelt. Zudem bestätigt die Analyse, dass der Frauenbob (Women) mit durchschnittlich 87,99 km/h am langsamsten und der Viererbob (Men4) mit 93,93 km/h am schnellsten auf dieser Strecke fährt.

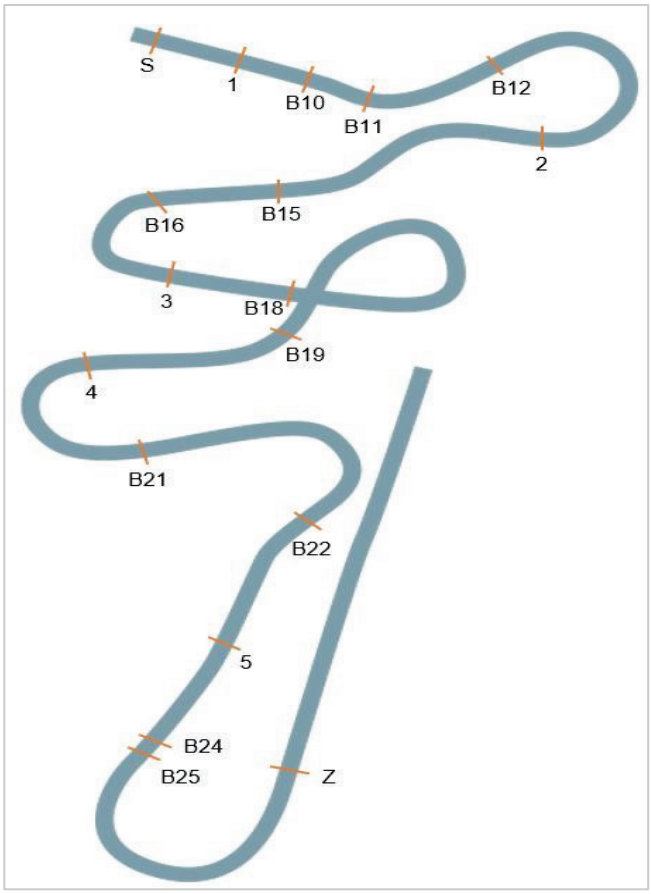
Im nächsten Schritt wurden die einzelnen Streckenabschnitte betrachtet. Hierfür wurde zunächst die Laufzeit pro Streckenabschnitt auf Basis der jeweiligen Start- und Endzeit ermittelt und entsprechend nach der Laufzeit pro Streckenabschnitt über alle Disziplinen hinweg sortiert:

Tabelle 6: Deskriptive Analyse – Durchschnittliche Laufzeit in Sekunden pro Streckenabschnitt

Streckenabschnitt	Laufzeit
B24_B25	0,271
B22_5	1,431
3_B18	1,453
5_B24	1,595
B10_B11	1,719
B15_B16	2,072
B19_4	2,504
1_B10	2,688
B11_B12	3,274
4_B21	3,474
B16_3	3,733
B21_B22	4,527
2_B15	4,793
B25_Z	4,982
S_1	5,371
B12_2	5,502
B18_B19	6,764

Die Tabelle zeigt, dass der Streckenabschnitt B24_B25 im Durchschnitt mit 0,271 Sekunden Laufzeit über alle Disziplinen am schnellsten absolviert wurde, während für den Streckenabschnitt B18_B19 durchschnittlich 6,764 Sekunden benötigt werden und dieser damit den langsamsten Abschnitt darstellt. Demnach haben die langsameren Streckenabschnitte einen deutlich größeren Anteil an der Gesamtlaufzeit, was darauf hindeuten könnte, dass diese Streckenabschnitte zur Optimierung der Gesamtlaufzeit am relevantesten sind.

Abbildung 41: Schaubild der Strecke in Winterberg



Quelle: in Anlehnung an IBSF (2022).

Ein Vergleich der Durchschnittszeit mit den Streckenabschnitten in der Abbildung 41 der Strecke zeigt, dass der Abschnitt B24_B25 der kürzeste Abschnitt ist, wohingegen der Abschnitt B18_B19 einen sehr kurvigen Teil der Bahn darstellt.

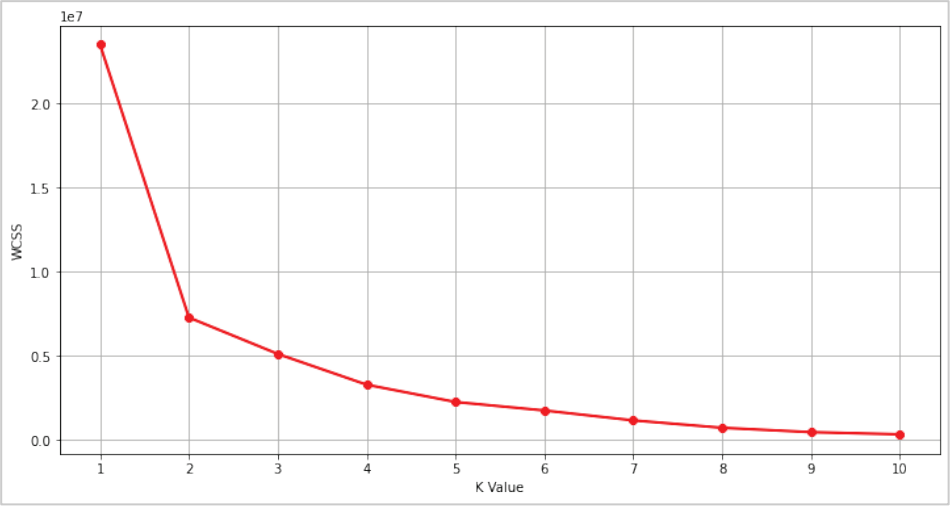
5.4 Modellierung

5.4.1 K-means Clustering

In dieser Arbeit wurde das K-means Clustering eingesetzt, um zwei verschiedene Faktoren zu erkennen. Zum einen soll es mittels des Algorithmus möglich sein,

pro Bahnabschnitt Cluster zu bilden und dadurch Ausreißer in den Daten zu erkennen. Dies hat den Vorteil, dass diese Ausreißer im Nachgang detaillierter analysiert werden können. Andererseits soll mit dem Clustering analysiert werden, ob über die Bahnabschnitte hinweg für eine Fahrt Cluster gebildet werden können, die Erkenntnisse über ähnlich verlaufende Bahnabschnitte liefern und so einen Mehrwert in der Optimierung des Bobsports beinhalten.
Zur Überprüfung auf Ausreißer wurde der Datensatz verwendet, der pro Zeile einen gesamten Lauf abbildet. Dies bedeutet, dass in den Spalten pro Bahnabschnitt die jeweiligen Faktoren zu Geschwindigkeit, Beschleunigung und Zeit enthalten sind. Im K-means Clustering wurde die sogenannte Elbow-Method verwendet, um die optimale Anzahl an Clustern für den Algorithmus zu bestimmen. Bholowalia und Kumar (2014, S. 18) sagen in ihrer Arbeit darüber aus, dass ab einer bestimmten Clustermenge die Vergrößerung dieser Menge kein besseres Ergebnis im K-means Clustering erzielen wird. Wird die Elbow-Method auf dem zugrundeliegenden Datensatz ausgeführt, wurde das folgende Ergebnis erzielt.

Abbildung 42: Elbow Method für Datensatz pro Lauf



Das Ergebnis zeigt, dass die optimale Clustermenge bei vier liegt. Denn ab Clustermenge fünf ist eine starke Veränderung im „Within-Cluster Sum of Square“ (im Folgenden WCSS) Wert nicht mehr festzustellen. WCSS ist die Summe des quadratischen Abstands zwischen jedem Punkt und dem Schwerpunkt eines Clusters. Wird der WCSS-Wert mit dem K-Wert vergleichen, sieht das Diagramm wie ein Ellenbogen aus. Mit zunehmender Anzahl von Clustern nimmt der WCSS-Wert ab (vgl. Bholowalia & Kumar, 2014, S. 18f.). Dies bedeutet, dass das K-

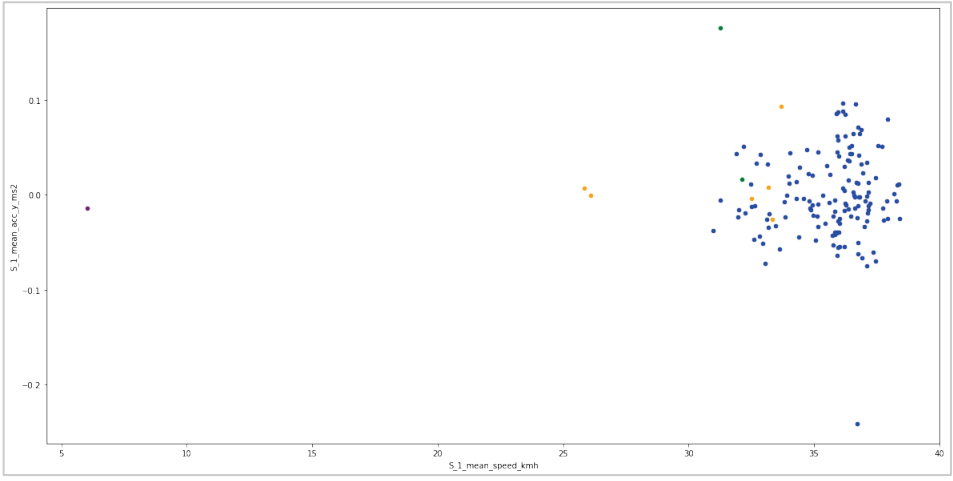
means Clustering für diesen Datensatz mit der Clustermenge vier ausgeführt wird. Bei insgesamt 151 Datensätzen ist die Verteilung auf die vier Cluster wie folgt.

Tabelle 7: Clustergröße pro Cluster für Datensatz pro Lauf

Cluster Nummer	Clustergröße
0	142
1	1
2	6
3	2

An dem Ergebnis ist zu erkennen, dass mittels des Clusterings die Ausreißer gut zu erkennen sind. Der Großteil der Datensätze befindet sich in Cluster 0 und auf die Cluster 1, 2 und 3 verteilen sich die Ausreißer. Wird beispielsweise die durchschnittliche vertikale Beschleunigung mit der durchschnittlichen Geschwindigkeit für den Bahnabschnitt S zu 1 in Winterberg visualisiert, zeigen sich die Ausreißer deutlich.

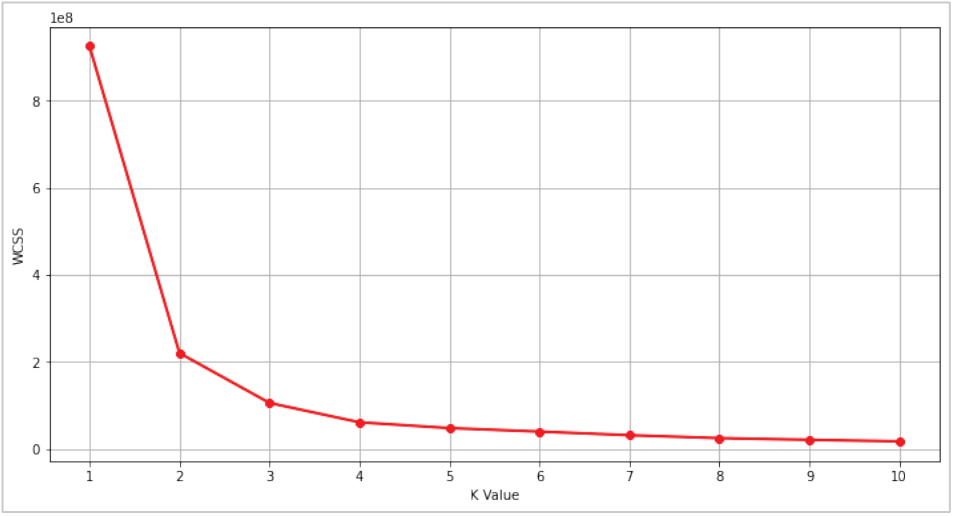
Abbildung 43: Durchschnittliche vertikale Beschleunigung zu durchschnittlicher Geschwindigkeit für den Bahnabschnitt S zu 1



Diese Ausreißer können nun im Detail betrachtet werden. Dabei kann ermittelt werden, ob es sich in diesen Fällen um Messfehler handelt oder ob daraus wertvolle Erkenntnisse gewonnen werden können.

Der zweite Anwendungsfall des Clusterings liegt im Erkennen von Clustern beim Vergleichen der Bahnabschnitte. Dafür wurde der Datensatz verwendet, der die Daten pro Fahrt und Bahnabschnitt aggregiert. Auch hier wurde zunächst die Elbow-Method zur Ermittlung der optimalen Clustermenge genutzt. Dies führte zum folgenden Ergebnis.

Abbildung 44: Elbow Method für Datensatz pro Fahrt und Bahnabschnitt



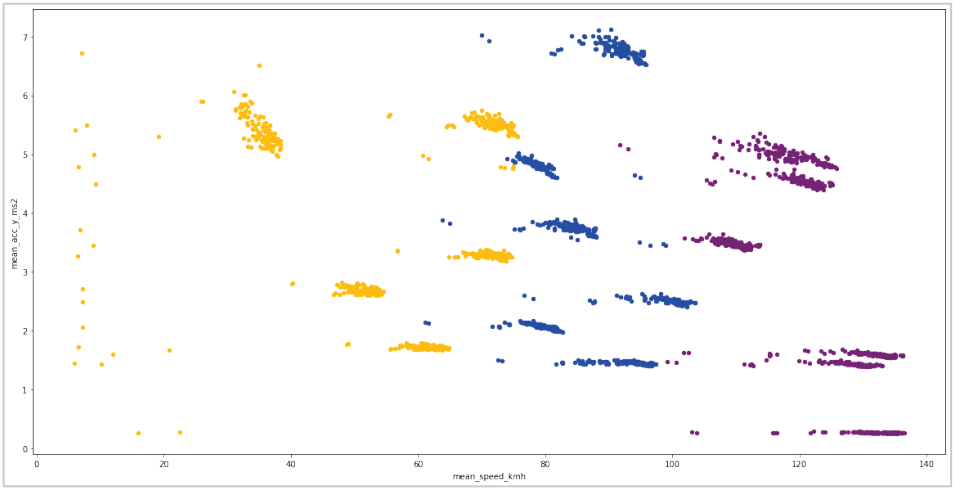
Bei diesem Datensatz zeigte sich, dass bereits ab einer Clustermenge von vier keine große Veränderung im WCSS-Wert zu erkennen ist. Aus diesem Grund wurde für das folgende K-means Clustering die Clustermenge drei gewählt. Die Menge an Datensätzen pro Cluster verteilt sich hier deutlich gleichmäßiger und deutet darauf hin, dass vom Algorithmus nicht nur Ausreißer erkannt werden (Tabelle 8).

Tabelle 8: Clustergröße pro Cluster für Datensatz pro Fahrt und Bahnabschnitt

Cluster Nummer	Clustergröße
0	908
1	893
2	781

Wird hier nun ebenfalls die durchschnittliche vertikale Beschleunigung mit der durchschnittlichen Geschwindigkeit visualisiert, sind die drei gebildeten Cluster deutlich sichtbar.

Abbildung 45: Clustergröße pro Cluster für Datensatz pro Fahrt und Bahnabschnitt



In einer weiterführenden Analyse lässt sich nun untersuchen, welche Bahnabschnitte welchem Cluster zugeordnet wurden und ob dort Ähnlichkeiten zu erkennen sind, die bei der Optimierung der Fahrt hilfreich sind.

5.4.2 XGBoost

Für die Analyse des Zusammenhangs zwischen Bahnabschnitten und dessen Einfluss auf die Gesamtlaufzeit wird im Folgenden ein Prognosemodell mittels

XGBoost erstellt, welches die Gesamtlaufzeit anhand erhobener deskriptiver Statistiken pro Bahnabschnitt vorhersagt. Dazu wurde zuerst Hyperparametertuning betrieben, um die optimalen Modellparameter zu identifizieren. Anschließend wurde der Prognosefehler bestimmt und mittels Shapley Values der Einfluss von verschiedenen Attributen auf die Modellprognose ermittelt.

Die XGBoost Implementation des Gradient Tree Boost Algorithmus besitzt eine Vielzahl von Parametern, welche die Modellprognose in unterschiedlichen Aspekten beeinflussen. Um herauszufinden welche Kombination von Parametern für den zugrundliegenden Datensatz die beste Prognosegüte nach sich zieht, lässt sich in einem iterativen Prozess eine Parametersuche durchführen. Dabei werden alle möglichen Kombinationen eines zuvor definierten Parameterraums im Hinblick auf den Prognosefehler miteinander verglichen, um die optimale Parameterkombination zu ermitteln. Für folgende Parameter wurde ein Hyperparametertuning durchgeführt:

- Die Anzahl an zu erstellenden Entscheidungsbäumen (*n_estimators*).
- Die maximale Tiefe eines Entscheidungsbaums (*max_depth*), also die maximale Anzahl an zu erstellenden Ebenen.

Jede zusätzliche Ebene erhöht die Komplexität des Entscheidungsbaums und führt somit auch zu einem erhöhten Risiko für das sogenannte Overfitting.

Der Wert *colsample_by_tree* ist die anteilige Größe der Teilmenge an Attributen, welche zufällig für die Erstellung jedes Entscheidungsbaums ausgewählt werden können. Der Wert liegt dabei zwischen 0 und 1. Die zufällige Auswahl einer Teilmenge an Attributen pro Entscheidungsbaum führt zu geringerem Overfitting und einer geringeren Korrelation der verschiedenen Entscheidungsbäume untereinander, sodass korrelierte Fehler vermieden werden.

Die *learning rate* ν legt fest, wie die einzelnen Entscheidungsbäume skaliert werden. Je kleiner der Wert von ν, desto kleiner ist die Schrittgröße der Veränderung an der Prognose durch einen einzelnen Entscheidungsbaum.

Der Regularisierungsparameter λ legt fest, welcher Einfluss die einzelnen Observationen abschwächt und somit das Modell unabhängiger vom Trainingsdatensatz macht (*reg_lambda*).

Tabelle 9 veranschaulicht die betrachteten Hyperparameter, deren getesteten Ausprägungen und die jeweils optimale Ausprägung.

Tabelle 9: Hyperparameter

Hyperparameter	Ausprägungen	Optimale Ausprägung
n_estimators	100, 500, 1.000	1.000
max_depth	3, 6, 10, 13	3
colsample_by_tree	0.3, 0.7	0.3
learning_rate	0.01, 0.05, 0.1	0.01
reg_lambda	0, 1, 5, 10, 15	1

Um den Prognosefehler des so spezifizierten Modells zu ermitteln, kommt der RootMean-Square Error (RMSE) zum Einsatz. Die Wurzel der durchschnittlichen quadratischen Abweichung von Modellprognose $\hat{y}$ und Beobachtung y quantifiziert dabei den modellseitigen Prognosefehler (Formel 2).

Formel 2: Root-Mean-Square Error

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}^i - y^i)^2}$$

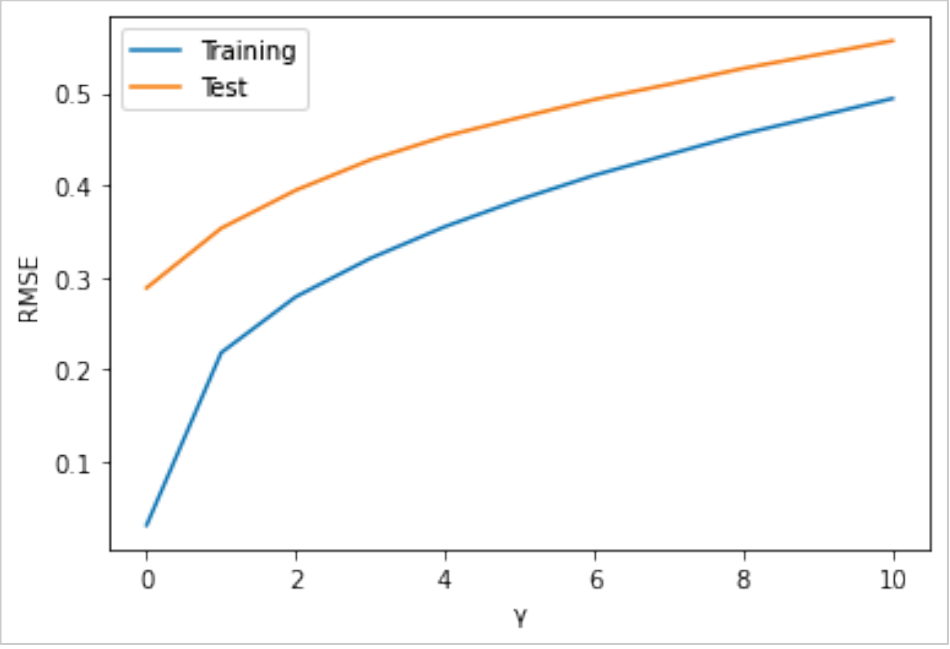
Das Modell erreicht einen Trainings-RMSE von 0,03 und einen Test-RMSE bei ungesehenen Daten von 0,29. Es lässt sich erkennen, dass ein großer Unterschied zwischen dem Trainings- und dem Test-Fehler vorliegt. Dies deutet darauf hin, dass das Model trotz Regularisierung durch $\lambda = 1$ noch zu Overfitting neigt, indem zu viele Informationen aus den Trainingsdaten abgeleitet werden und so das Model nicht gut über den Testdatensatz generalisieren kann, was sich in einem deutlich größeren Test-RMSE ausdrückt.

Der Parameter γ kann genutzt werden, um Entscheidungsbäume zu beschneiden, sogenanntes Pruning. Dabei wird jede Verzweigung des Entscheidungsbaumes entfernt, bei der die entstehenden Knoten zusammen keinen Informationsmehrwert im Bezug zu ihrem Wurzelknoten besitzen. Je höher der Wert für γ gewählt wird, desto größer muss der Informationsmehrwert sein, damit der Ent-

scheidungsbaum nicht beschnitten wird. Dieses Vorgehen reduziert die Komplexität des Entscheidungsbaumes und lässt das Modell so besser auf ungesehenen Daten generalisieren.

Abbildung 46 veranschaulicht den Trainings- und Testfehler im Bezug zu γ-Werten zwischen 0 und 10. Es lässt sich erkennen, dass mit steigenden γ-Werten auch der Prognosefehler steigt, da die Komplexität des Modells reduziert wird. Bei einem Wert von $\gamma = 1$ ist eine signifikante Reduktion der Differenz zwischen Trainings- und Testfehler zu erkennen, sodass diese Konfiguration für das finale Modell genutzt wird, um das Overfitting des Modells zu reduzieren.

Abbildung 46: Einfluss von verschiedenen γ-Werten auf Trainings- und Test-RMSE



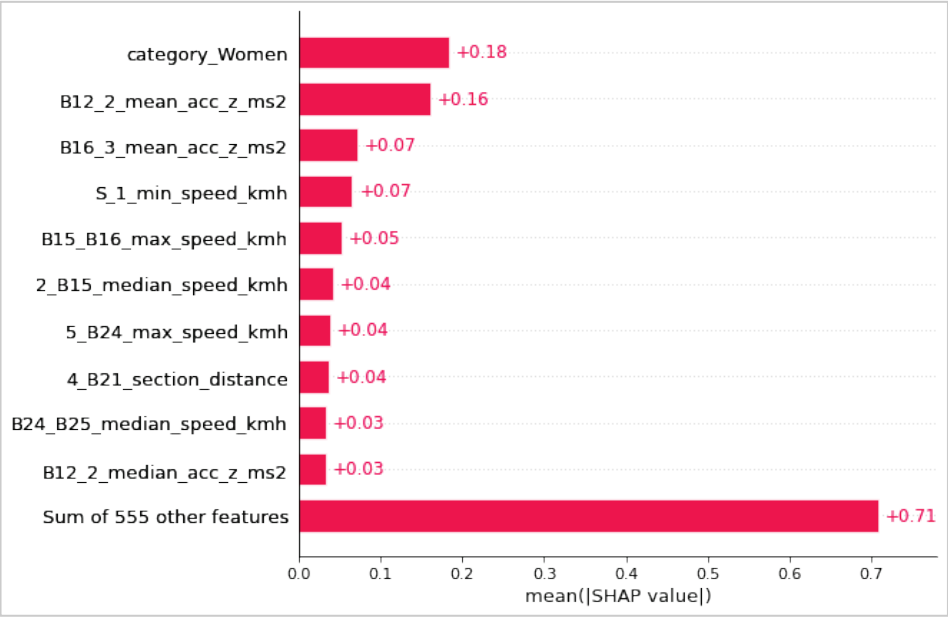
Das finale Modell erreicht somit einen Trainings-RMSE von 0,22 und einen Test-RMSE von 0,35. Dies bedeutet, dass die Modellprognose im Durchschnitt um 0,35 Sekunden vom beobachteten Wert der Gesamtlaufzeit abweicht. Tabelle 10 fasst die Ergebnisse zusammen.

Tabelle 10: Trainings- und Test-RMSE

Modell	Trainings-RMSE	Test-RMSE
XGBoost: n_estimators = 1.000, max_depth = 3, colsample_by_tree = 0.3, learning_rate = 0.01, reg_lambda = 1, gamma = 1	0,22	0,35

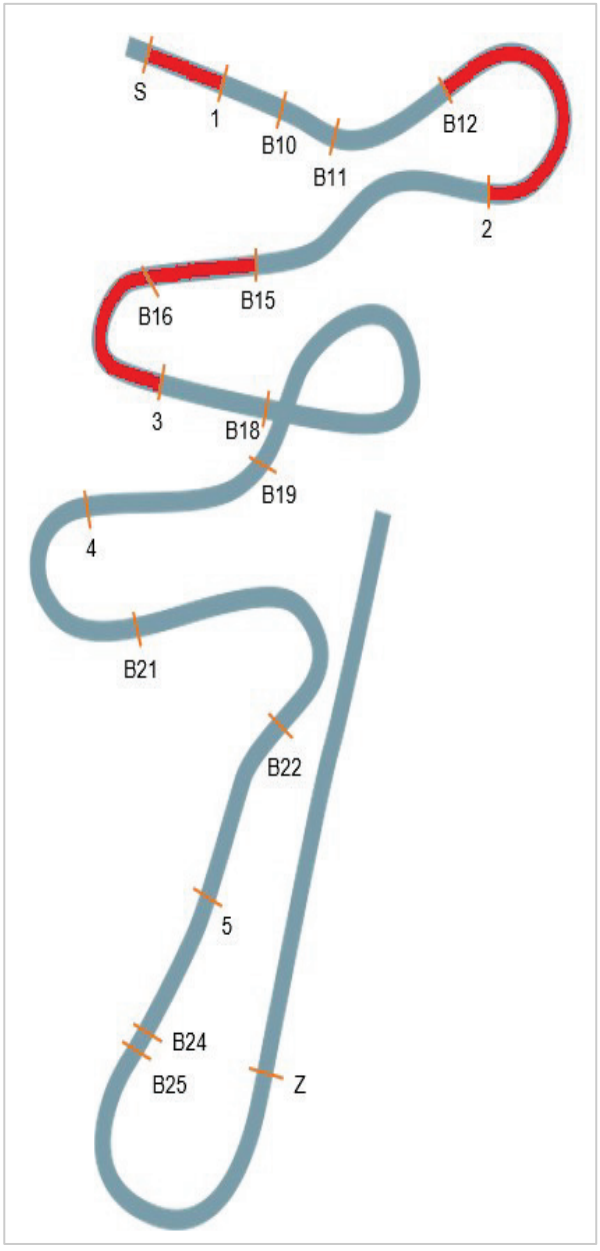
Mithilfe der Shapley Values lässt sich nun der Einfluss der unterschiedlichen Attribute ermitteln und so die Modellentscheidungen erklärbar machen. Abbildung 47 zeigt die zehn Attribute mit dem größten absoluten Einfluss auf die Modellprognose. Es lässt sich erkennen, dass das wichtigste Attribut für das Modell die Kategorie *Women* darstellt. Es ist also davon auszugehen, dass es signifikante Zeitunterschiede zwischen den verschiedenen Kategorien gibt. Von weiterer besonderer Wichtigkeit für das Modell hat sich auch die durchschnittliche z-Achsenbeschleunigung im Abschnitt zwischen den Lichtschranken B12 und B2 erwiesen.

Abbildung 47: Durchschnittlicher absoluter Shapley Value für die wichtigsten 10 Attribute



Es lässt sich feststellen, dass generell das Fahrverhalten in den Abschnitten „B12 bis 2“, „B16 bis 3“, „S bis 1“ und „B15 bis B16“ für das Modell von besonderer Relevanz sind. Abbildung 48 veranschaulicht diese Bereiche auf der Karte.

Abbildung 48: Schematische Darstellung für das Modell wichtiger Streckenabschnitte



Quelle: in Anlehnung an IBSF (2022).

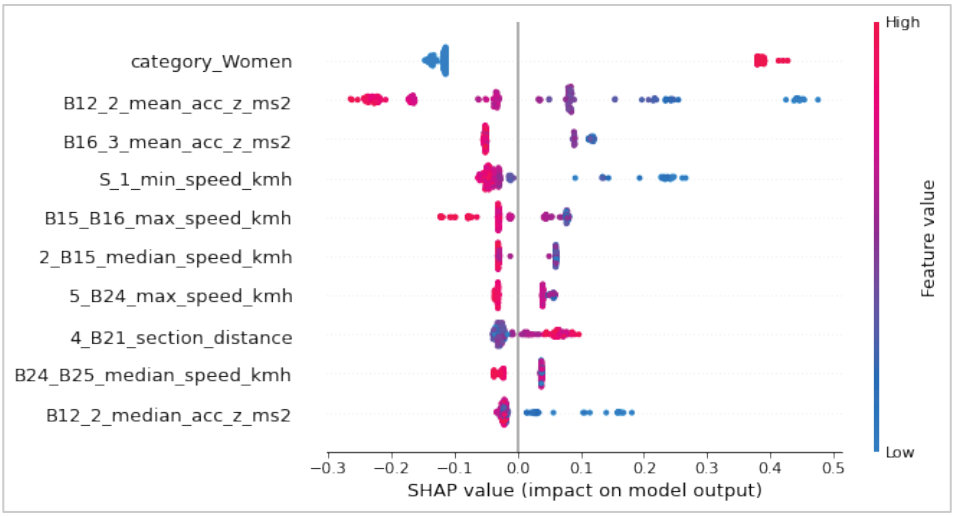
Neben dem durchschnittlichen absoluten Einfluss auf die Modellprognose, lässt sich mittels der Shapley Values auch der Modelleinfluss in Relation zur Attributsausprägung darstellen. Somit lässt sich ein besseres Verständnis gewinnen, welchen Einfluss verschiedene Ausprägungen des Attributs auf die Modellprognose nehmen.

Abbildung 49 setzt die Höhe der Attributsausprägung mit ihrem Einfluss auf die Modellprognose für die modellseitig wichtigsten 10 Attribute in Beziehung. Bei den einzelnen Punkten in der Punktwolke handelt es sich jeweils um die Observationen jedes Boblaufs. Für die binäre Kategorie *Women* lässt sich beispielsweise sehr deutlich erkennen, dass die Zugehörigkeit zur Kategorie mit Shapley Values zwischen 0,3 und 0,5 einen deutlich positiven Einfluss auf die Modellprognose besitzt. Da das Modell die Gesamtlaufzeit prognostiziert, sind positive Shapley Values mit einer längeren Laufzeit zu assoziieren. Für das Modell bedeutet dies, dass die Zugehörigkeit zur Kategorie *Women* die prognostizierte Gesamtlaufzeit um 0,3 bis 0,5 Sekunden verlängert.

Für die durchschnittliche z-Achsenbeschleunigung in Abschnitt „B12 bis 2“ (B12_2_mean_acc_z_ms2) ist ebenfalls deutlich zu erkennen, dass eine besonders hohe durchschnittliche Beschleunigung in diesem Abschnitt die prognostizierte Gesamtlaufzeit verringert.

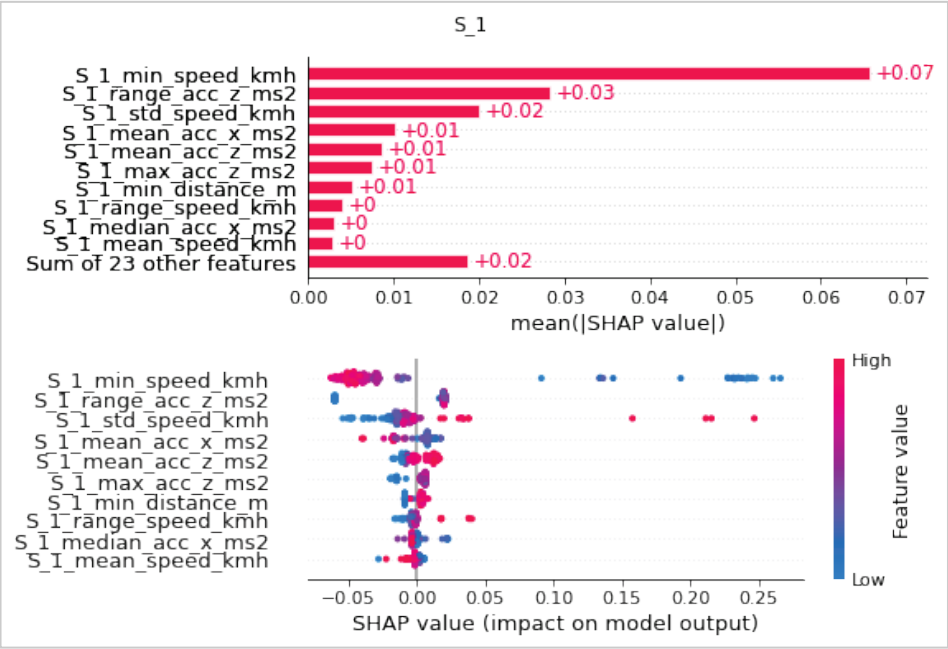
Betrachtet man allerdings beispielsweise die maximale Geschwindigkeit im Abschnitt „5 bis B24“ (5_B24_max_speed_kmh), so stellt man fest, dass tendenziell hohe Maximalgeschwindigkeiten die prognostizierte Gesamtzeit verringern, aber es ebenfalls Läufe gab, bei denen trotz einer hohen Geschwindigkeit kein positiver Einfluss auf die Gesamtlaufzeit ersichtlich ist. Die Maximalgeschwindigkeit in Abschnitt „5 bis B24“ verhält sich also nichtlinear zur Gesamtlaufzeit.

Abbildung 49: Attributausprägungen im Verhältnis zur Modellprognose



Um einen besseren Überblick des Einflusses der verschiedenen Streckenabschnitte zu erhalten, lassen sich beide Analysen auch jeweils pro Abschnitt darstellen. Durch diese Darstellungsform lassen sich die wesentlichen Einflussfaktoren eines einzelnen Streckenabschnittes auf die Modellprognose isolieren. Abbildung 50 zeigt beispielhaft die 10 wichtigsten Attribute für die Modellprognose, welche sich auf Abschnitt „S bis 1“ beziehen.

Abbildung 50: Einfluss und Ausprägung der 10 einflussreichsten Attribute aus Streckenabschnitt „S bis 1“



Die Shapley Values der 16 weiteren Streckenabschnitte sind im Anhang zu finden.

5.5 Schlussbetrachtung

5.5.1 Fazit

Im Rahmen der vorliegenden Arbeit sollten die folgenden Forschungsfragen beantwortet werden:

1. Lassen sich Bahnabschnitte feststellen, die einen signifikanten Einfluss auf die Gesamtlaufzeit haben?
2. Welche Faktoren pro Bahnabschnitt haben einen signifikanten Einfluss auf die Laufzeit in diesem Abschnitt?
3. Lassen sich optimale Werte für die einzelnen Faktoren pro Bahnabschnitt bestimmen?

Auf Basis von Messungen auf der Bobbahn in Winterberg wurden statistische Analysen durchgeführt, um wichtige Erkenntnisse für die Bobsportler und das Trainerteam zu ermitteln und damit wesentliche Faktoren für die Gesamtlaufzeit der Strecke abzuleiten. Hierfür wurden die Rohdaten zunächst aufbereitet und in einem Datensatz zusammengeführt. Die Daten wurden aggregiert und erweitert, sodass eine große Anzahl an deskriptiven Variablen für die Modelle zur Verfügung stand.

Anhand der deskriptiven Analyse zeigte sich, dass die Gesamtlaufzeit für die Disziplinen (Kategorien) unterschiedlich ist. Insbesondere der Frauenbob (Women) liegt bei Geschwindigkeit und Gesamtlaufzeit unter dem Durchschnitt und der Viererbob (Men4) weist sowohl die höchste Geschwindigkeit als auch die kürzeste Gesamtlaufzeit auf.

In Bezug auf die Laufzeit pro Bahnabschnitt lässt sich anhand der deskriptiven Analyse eine Rangfolge ermitteln. Die aus den Rohdaten berechneten Zeiten der einzelnen Streckenabschnitte stimmen mit der Länge und dem Schwierigkeitsgrad der jeweiligen Streckenabschnitte überein.

Darüber hinaus wurden anhand der Methode des Clusterings zum einen Ausreißer identifiziert und zum anderen weitere Erkenntnisse ermittelt, die mit Hilfe von Experten im Bobsport tiefergehend analysiert werden können. Hierfür wird anhand der Elbow-Methode die optimale Clustergröße ermittelt. Dabei wurden zur Ermittlung der Ausreißer vier Cluster identifiziert, während für die Gewinnung weiterer Erkenntnisse hingegen drei Cluster verwendet wurden.

Anhand des XGBoost-Verfahrens konnten zudem wesentliche Erkenntnisse zur Beantwortung der Forschungsfragen ermittelt werden. Das Verfahren ermittelt zehn Attribute, die den größten absoluten Einfluss auf die Modelprognose haben. Hierbei zeigte sich, dass die Fahrabschnitte „B12 bis 2“, „B16 bis 3“, „S bis 1“ und „B15 bis B16“ von besonderer Relevanz für die Gesamtlaufzeit sind (Forschungsfrage 1). Darüber hinaus wurde deutlich, dass weitere Faktoren einen signifikanten Einfluss haben. Dabei hat die Disziplin (Kategorie) *Women* einen deutlichen positiven Einfluss auf die Gesamtlaufzeit, sodass sich die prognostizierte Gesamtlaufzeit sich durch diesen Faktor verlängert. Des Weiteren verringert insbesondere eine hohe durchschnittliche z-Achsenbeschleunigung in Abschnitt „B12 bis 2“ die prognostizierte Gesamtlaufzeit (Forschungsfrage 2).

Zuletzt lassen sich die Analysen anhand des XGBoost-Verfahrens auch für einzelne Streckenabschnitte durchführen. Dabei sind die Ergebnisse für jeden Streckenabschnitt individuell zu betrachten (Forschungsfrage 3).

5.5.2 Grenzen dieser Arbeit

Im Rahmen der Interpretation der Erkenntnisse sind zudem einige Grenzen der Arbeit zu berücksichtigen. Zunächst beschränkt sich die Analyse auf insgesamt 151 Läufe und einen limitierten zeitlichen Betrachtungszeitraum. Zudem fehlen wichtige (qualitative) Informationen in Bezug auf das Bobmodell, externe oder athletenspezifische Einflüsse (z. B. Gewicht). Darüber hinaus sind für die weitergehende Interpretation der Ergebnisse und die Übertragung auf die Praxis eine tiefere Fachexpertise mit Bezug auf Physik (Beschleunigung) und den Bobsport erforderlich.

5.5.3 Ausblick

Es bieten sich einige Möglichkeiten für weitergehende Analysen, die auf den vorliegenden Erkenntnissen aufbauen. Zunächst bietet sich die Arbeit als eine Grundlage an, die gewonnenen Erkenntnisse mit Experten des Bobsports zu diskutieren und somit domänenspezifischer zu interpretieren. Auf dieser Basis lässt sich das Modell ggf. erweitern und es können weitere Analysen durchgeführt werden. Somit können die identifizierten Einflussfaktoren in der Praxis berücksichtigt und überprüft werden. Zudem könnte die Analyse auf weitere Bobbahnen ausgeweitet werden, z. B. im Rahmen eines längeren Betrachtungszeitraums oder durch die Berücksichtigung weiterer Standorte und Disziplinen. Hierzu bietet sich bereits der vorliegende Datensatz an, da neben den Werten für Winterberg bereits weitere Läufe auf anderen Bahnen verfügbar sind. Darüber hinaus können weitere unabhängige Variablen im Modell ergänzt werden. Neben den bereits vorliegenden quantitativen Variablen könnten demnach weitere Daten, wie z. B. das Wetter oder technische Spezifikationen des Bobs berücksichtigt werden.

6 Zum Zusammenhang von Fahrlinie und Laufzeit

6.1 Zielsetzung

Dieses Kapitel konzentriert sich unter Berücksichtigung der bisherigen Erkenntnisse der Literatur auf die Frage nach der optimalen Fahrlinie. Hierzu werden auf Basis der physikalischen Gesetzmäßigkeiten die Daten aus einer Fallstudie genutzt, um anhand des Rollwinkels und weiteren Geschwindigkeits- sowie Beschleunigungsparametern den Zusammenhang von unterschiedlichen Fahrlinien und realisierten Laufzeiten zu erklären. Hieraus sollen Impulse über Kausalitäten und Laufzeitoptimierungen unter Berücksichtigung diverser Rahmenbedingungen, wie etwa Streckencharakteristik, abgeleitet werden.

Zunächst folgt ein Überblick über die genaue Datenlage, die Vorverarbeitung der Datenbasis in Python, sowie eine Erläuterung der zusätzlich erzeugten Daten für die weitere Analyse. Der so vorbereitete Datensatz dient als Grundlage für die Untersuchung für eine Datenanalyse in Python und eine visualisierungsgetriebene Analyse über PowerBI.

Die hier durchgeführte Analyse wird sich anschließend zunächst aus einer physikalischen Perspektive der Frage nach dem Optimum nähern. In einem zweiten Schritt wird untersucht, welche Fahrweise zu den besten Zeiten geführt hat und wie sich diese Fahrweise von den anderen Fahrstilen unterscheidet. Hierbei werden die Analysen sowohl auf Gesamt- als auch auf Teilstreckenebene durchgeführt. Anschließend folgt ein Erklärungsversuch mit den Faktoren der Fahrphysik, um idealerweise ein Ergebnis zu erhalten, welches auf alle Strecken anwendbar ist und somit streckenunabhängigen Mehrwert schafft.

Zum Abschluss wird auf die Limitationen der Untersuchung eingegangen, welche sich insbesondere in nicht berücksichtigten Faktoren äußern wird, die einen erheblichen Einfluss haben könnten, aber in der Analyse nicht erfasst wurden. Weiterhin wird das Forschungsvorhaben kritisch reflektiert und auf Limitationen dieser Fallstudie hingewiesen. Hierauf basierend erfolgt im Rahmen des Fazits eine endgültige Beurteilung der aufgestellten Forschungsfragen, sowie eine Diskussion der hieraus resultierenden Forschungsansätze für weitere wissenschaftliche Auseinandersetzungen mit der Thematik.

6.2 Ableitung der Untersuchungsthesen

Auf Basis der physikalischen Grundlagen (Kap. 1.2) ergeben sich die zu untersuchenden Thesen für den Fahrstil:

1. Da eine höhere Geschwindigkeit den Reibungskoeffizienten reduziert, ist ein schneller Start fundamental für eine gute Gesamtzeit.
2. Da ein höherer Druck, d.h. eine größere Reibleistungsdichte, zunächst den Reibungskoeffizienten verringert, sollten die Kurven eng gefahren werden. Ein geringer Rollwinkel und eine hohe Vertikalbeschleunigung sind vorteilhaft.

Bei höheren Temperaturen nahe Null und einer hohen Geschwindigkeit ist aufgrund der hohen Vertikalbeschleunigung in scharfen Kurven gegebenenfalls ein höherer Rollwinkel günstig. Da die Umgebungstemperaturen für die Fahrten jedoch nicht vorliegen, wird auf die Untersuchung dieses Aspekts verzichtet.

Die erste These wurde bereits in der Literatur untersucht. So wurden 77 Prozent der Varianz der Gesamtzeiten für Bobfahrten mit der Startzeit erklärt, was die Relevanz dieser Phase unterstreicht (Brüggemann, Morlok und Zatsiorsky, 1997, S. 103). Bei einem kleinen Rollwinkel und einer engen Kurvenfahrt ist zu beachten, dass auch die zu fahrende Strecke kürzer ist.

Beide Thesen werden im nachfolgenden Abschnitt auf Basis der zur Verfügung gestellten Daten für die Veltins-Arena in Winterberg untersucht.

6.3 Data Understanding – Einführung in das Datenmodell

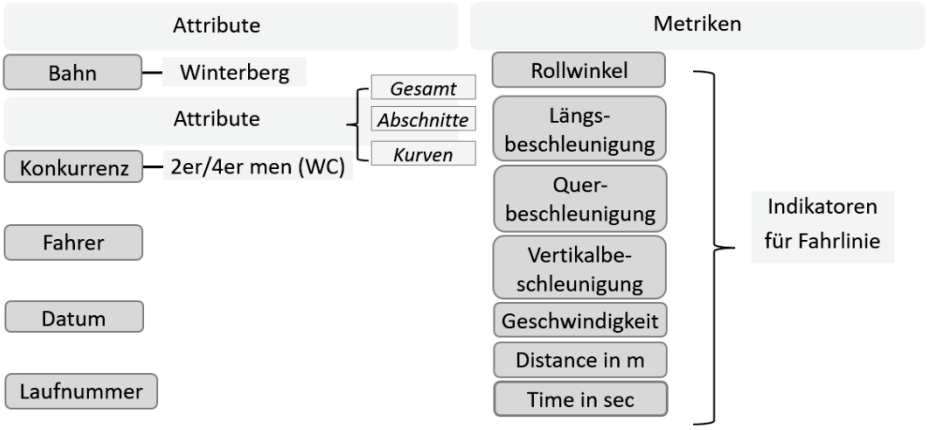
Die Rohdaten der Fallstudie liegen für die betrachteten Bahnen und Konkurrenzen jeweils pro Lauf analog dem nachfolgenden Schema vor.

Tabelle 11: Schema der vorliegenden Rohdaten

Time s	Distance m	Speed km/h	Acc x m/s2	Acc y m/s2	Acc z m/s2	Roll angle deg	Light-beam	LB time s
x	a	e	i	m	q	u		
x+0,01	b	f	j	n	r	v		
x+0,02	c	g	k	o	s	w		
x+0,03	d	h	l	p	t	x	y	z

Demnach werden für jede Zehntelsekunde die gefahrene *Distanz* in Meter, die *Geschwindigkeit* in Kilometer pro Stunde, die Längs- (*Acc x*), Quer- (*Acc y*) sowie Vertikalbeschleunigung (*Acc z*) in Meter pro Quadratsekunde sowie der Rollwinkel in Grad festgehalten. Über die graduelle Aufzeichnung der Geschwindigkeits- und Beschleunigungsparameter sowie des Rollwinkels ist eine Indikation des Fahrstils möglich, indem die Parameter miteinander ins Verhältnis gesetzt werden und hierdurch das Fahrverhalten sowie die Lage des Bobs und somit auch die Fahrlinie des Bobs im Fahrverlauf durchgängig abgebildet werden kann. Eine adäquate Analyse der Fahrlinie und Laufzeit bedarf darüber hinaus jedoch auch einer jeweils aktuellen und präzisen Positionsangabe auf der Strecke, um etwaige Unterschiede anhand von Bahnabschnitten und -kurven zu identifizieren, lokalisieren und hiermit inhaltlich interpretierbar machen zu können. Diese Anforderung ist demnach durch eine Datenanreicherung über Sekundärdaten im Rahmen der Data Preparation umzusetzen. Hierfür ist die Spalte *Lightbeam* von Nutzen: Diese weist die Durchfahrt der jeweiligen Lichtschranken aus, wodurch in Kombination mit der absolvierten Distanz eine genaue Lokalisation des Bobs im Fahrverlauf ermöglicht wird. Zusätzlich erfolgt durch *LB times s* die exakte Zeitmessung beim Passieren der Lichtschranken, sodass neben der Gesamtlaufzeit auch die Laufzeiten innerhalb der einzelnen Streckenabschnitte präzise erfasst werden können. Dies ist im Hinblick auf die zugrundeliegenden Forschungsfrage essenziell, da hiermit der Zusammenhang von Fahrlinie und Laufzeit detailliert für einzelne Abschnitte analysiert werden kann, was die Analysequalität und somit die Aussagekraft der Studie deutlich erhöht. Aus den vorgestellten Rohdaten ergibt sich letztlich eine Datenbasis für das grundlegende Ziel-Datenmodell gemäß Abbildung 51.

Abbildung 51: Datengrundlage der Fallstudie



Demgemäß sind die Rohdaten auf die Attribute Bahn, Konkurrenz, Fahrer sowie Datum und Laufnummer jeweils zu aggregieren, respektive herunterzubrechen, was diverse Auswertungsmöglichkeiten zulässt. Als Nachteil ist hierbei jedoch anzuführen, dass durch den Einbezug unterschiedlicher Bahnen und Konkurrenzen entsprechend auch die Einflussgrößen, wie etwa Streckencharakteristika, stark zunehmen und somit die Komplexität des Forschungsvorhabens sowohl hinsichtlich der Datenaufbereitung als auch -analyse deutlich erhöht. Da diese Forschungsarbeit jedoch als einführende Studie einen fundamentalen Ansatz zur Behandlung der zugrundeliegenden Thematik zum Ziel hat, richtet sich der Forschungsfokus vorliegend primär auf die Viererbob-Weltcup-Wettbewerbe der Herren in Winterberg. Dies ist aus methodischer Perspektive vertretbar, weil für die angesprochenen Konkurrenzen Datenmengen vorliegen, welche aus statistischer Sicht hinreichend groß für die Ableitung valider Schlussfolgerungen sind. Wie bereits diskutiert, sind darüber hinaus zusätzliche Bahnattribute in das Datenmodell zu überführen, um die Analysen auf Kurven- und Streckenabschnittsebene auszuweiten. Hierdurch sind präzisere Aussagen über divergierende Fahrlinien in den jeweiligen Abschnitten sowie ihren spezifischen Auswirkungen auf die Laufzeit möglich. Hierfür erfolgt die Anreicherung der benötigten Sekundärdaten im Rahmen der folgenden Data Preparation.

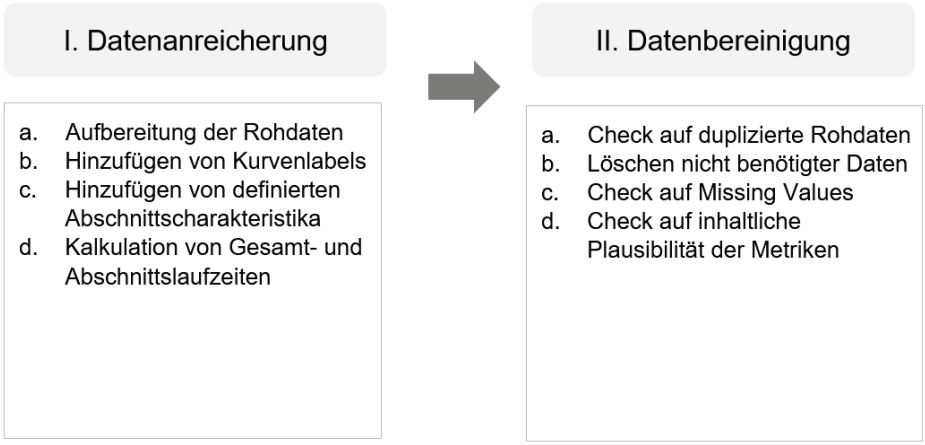
Hinsichtlich der zur Verfügung stehenden Metriken ist anzuführen, dass diese die entsprechend notwendige Datenbasis bereitstellen, um als Indikatoren zur Ableitung von Fahrlinien zu fungieren und somit eine Untersuchung der Kausalität von

Fahrlinien und Laufzeiten ermöglichen. Hierzu sind Extremwerte und statistische Lageparameter der Geschwindigkeits- respektive Beschleunigungsdaten zu erheben, um den gewählten Fahrstil zu konkretisieren und implizit vergleichbar zu machen. Angesprochene Vergleiche sind hierbei nicht nur zwischen einzelnen Piloten anzustellen, sondern auch in zu definierenden Clustern – wie etwa den schnellsten und langsamsten Läufen – sinnvoll, um den Datensatz noch zielgerichteter auf die Fragestellungen analysieren zu können. Diese Forschungsansätze sind auf Basis des vorgestellten Ziel-Datenmodells, bestehend aus den Rohdaten sowie den diskutierten zusätzlichen Datenanreicherungen, möglich. Die entsprechend umzusetzende Aufbereitung des Datenmodells inklusive Bereinigung zur Gewährleistung anforderungsgerechter Datenqualität erfolgt hierzu nachfolgend im Rahmen der Data Preparation.

6.4 Data Preparation – Datenaufbereitung und -bereinigung

Vor dem Einstieg in die detaillierte Analysearbeit ist zunächst das Datenmodell als Fundament jeglicher Analyseprozesse vollumfassend und valide aufzusetzen. Hierfür werden die definierten Anreicherungs- sowie Bereinigungsprozesse gemäß Abbildung 52 mittels Python durchgeführt.

Abbildung 52: Bestandteile der Datenanreicherung und -bereinigung



6.4.1 Aufbereitung der Rohdaten

Im ersten Schritt der Datenanreicherung werden die Rohdaten aller ausgewählten Wettkämpfe, welche in separaten Dateien im comma-separated-values (csv) Format in unterschiedlichen Ordnern liegen, in ein gemeinsames Dataset überführt. Nachfolgende Abbildung stellt einen Auszug dieses Datenkorpus bereit.

Abbildung 53: Aufbereitung der Rohdaten

Time s	Distance m	Speed km/h	Acc x m/s2	Acc y m/s2	Acc z m/s2	Roll angle deg	Lightbeam	LB time s	Fahrer + Fahrt
-4.99400	-0.00000	0.00000	-0.23829	-0.00112	9.74653	-0.25458	nan	nan	Hall_Brad-Run-2.csv 4MEN_2020-01-03
-4.98400	-0.00000	0.00000	-0.23945	-0.00123	9.73200	-0.25455	nan	nan	Hall_Brad-Run-2.csv 4MEN_2020-01-03
-4.97400	-0.00000	0.00000	-0.24091	-0.00133	9.73975	-0.25453	nan	nan	Hall_Brad-Run-2.csv 4MEN_2020-01-03
-4.96400	-0.00000	0.00000	-0.24211	-0.00142	9.74750	-0.25450	nan	nan	Hall_Brad-Run-2.csv 4MEN_2020-01-03
-4.95400	-0.00000	0.00000	-0.24245	-0.00148	9.73294	-0.25449	nan	nan	Hall_Brad-Run-2.csv 4MEN_2020-01-03
-4.94400	-0.00000	0.00000	-0.24322	-0.00154	9.74808	-0.25447	nan	nan	Hall_Brad-Run-2.csv 4MEN_2020-01-03
-4.93400	-0.00000	0.00000	-0.24328	-0.00158	9.74833	-0.25446	nan	nan	Hall_Brad-Run-2.csv 4MEN_2020-01-03
-4.92400	-0.00000	0.00000	-0.24279	-0.00159	9.74112	-0.25445	nan	nan	Hall_Brad-Run-2.csv 4MEN_2020-01-03
-4.91400	-0.00000	0.00000	-0.24230	-0.00159	9.74876	-0.25444	nan	nan	Hall_Brad-Run-2.csv 4MEN_2020-01-03
-4.90400	-0.00000	0.00000	-0.24087	-0.00156	9.73407	-0.25444	nan	nan	Hall_Brad-Run-2.csv 4MEN_2020-01-03
-4.89400	-0.00000	0.00000	-0.23980	-0.00152	9.74907	-0.25444	nan	nan	Hall_Brad-Run-2.csv 4MEN_2020-01-03
-4.88400	-0.00000	0.00000	-0.23780	-0.00146	9.74174	-0.25444	nan	nan	Hall_Brad-Run-2.csv 4MEN_2020-01-03
-4.87400	-0.00000	0.00000	-0.23544	-0.00137	9.73439	-0.25445	nan	nan	Hall_Brad-Run-2.csv 4MEN_2020-01-03
-4.86400	-0.00000	0.00000	-0.23324	-0.00127	9.74929	-0.25447	nan	nan	Hall_Brad-Run-2.csv 4MEN_2020-01-03
-4.85400	-0.00000	0.00000	-0.23033	-0.00115	9.74929	-0.25449	nan	nan	Hall_Brad-Run-2.csv 4MEN_2020-01-03
-4.84400	-0.00000	0.00000	-0.22709	-0.00101	9.74925	-0.25451	nan	nan	Hall_Brad-Run-2.csv 4MEN_2020-01-03

Dieser umfasst die bereits diskutierten Metriken sowie über *Fahrer + Fahrt* einen Primary Key, welcher die entsprechende Datenzeile eindeutig einem bestimmten Lauf eines Piloten zuordnet, auf Basis dessen die zusätzlichen Attribute in das Dataset überführt werden können. Das Resultat ist im folgenden Ausschnitt dargestellt.

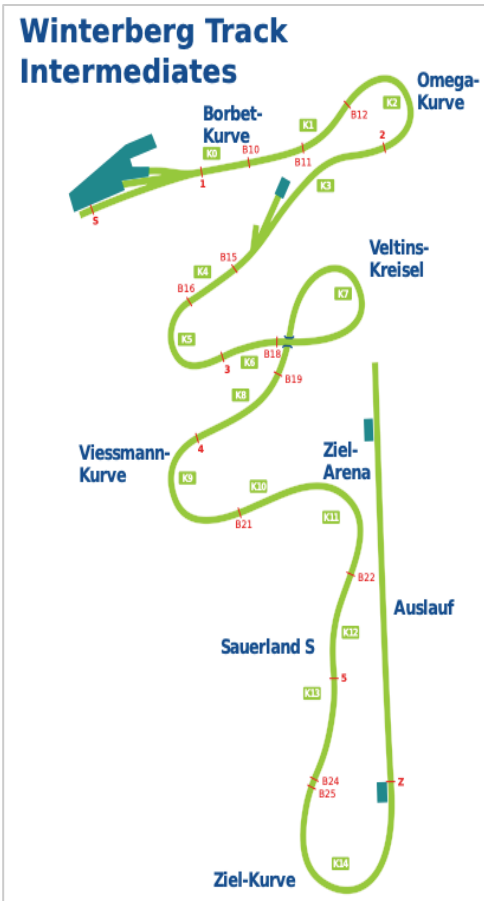
Abbildung 54: Anreicherung der Rohdaten um zusätzliche Attribute

Fahrer + Fahrt	Fahrer	FahrerLauf	Wettkampf	Laufnummer	Konkurrenz	Datum
Hall_Brad-Run-2.csv 4MEN_2020-01-03	Hall_Brad	Hall_Brad-Run-2	4MEN_2020-01-03	2	4MEN	2020-01-03
Hall_Brad-Run-2.csv 4MEN_2020-01-03	Hall_Brad	Hall_Brad-Run-2	4MEN_2020-01-03	2	4MEN	2020-01-03
Hall_Brad-Run-2.csv 4MEN_2020-01-03	Hall_Brad	Hall_Brad-Run-2	4MEN_2020-01-03	2	4MEN	2020-01-03
Hall_Brad-Run-2.csv 4MEN_2020-01-03	Hall_Brad	Hall_Brad-Run-2	4MEN_2020-01-03	2	4MEN	2020-01-03
Hall_Brad-Run-2.csv 4MEN_2020-01-03	Hall_Brad	Hall_Brad-Run-2	4MEN_2020-01-03	2	4MEN	2020-01-03
Hall_Brad-Run-2.csv 4MEN_2020-01-03	Hall_Brad	Hall_Brad-Run-2	4MEN_2020-01-03	2	4MEN	2020-01-03
Hall_Brad-Run-2.csv 4MEN_2020-01-03	Hall_Brad	Hall_Brad-Run-2	4MEN_2020-01-03	2	4MEN	2020-01-03
Hall_Brad-Run-2.csv 4MEN_2020-01-03	Hall_Brad	Hall_Brad-Run-2	4MEN_2020-01-03	2	4MEN	2020-01-03
Hall_Brad-Run-2.csv 4MEN_2020-01-03	Hall_Brad	Hall_Brad-Run-2	4MEN_2020-01-03	2	4MEN	2020-01-03
Hall_Brad-Run-2.csv 4MEN_2020-01-03	Hall_Brad	Hall_Brad-Run-2	4MEN_2020-01-03	2	4MEN	2020-01-03
Hall_Brad-Run-2.csv 4MEN_2020-01-03	Hall_Brad	Hall_Brad-Run-2	4MEN_2020-01-03	2	4MEN	2020-01-03
Hall_Brad-Run-2.csv 4MEN_2020-01-03	Hall_Brad	Hall_Brad-Run-2	4MEN_2020-01-03	2	4MEN	2020-01-03
Hall_Brad-Run-2.csv 4MEN_2020-01-03	Hall_Brad	Hall_Brad-Run-2	4MEN_2020-01-03	2	4MEN	2020-01-03
Hall_Brad-Run-2.csv 4MEN_2020-01-03	Hall_Brad	Hall_Brad-Run-2	4MEN_2020-01-03	2	4MEN	2020-01-03
Hall_Brad-Run-2.csv 4MEN_2020-01-03	Hall_Brad	Hall_Brad-Run-2	4MEN_2020-01-03	2	4MEN	2020-01-03
Hall_Brad-Run-2.csv 4MEN_2020-01-03	Hall_Brad	Hall_Brad-Run-2	4MEN_2020-01-03	2	4MEN	2020-01-03

6.4.2 Hinzufügen von Kurvenlabels

Im nächsten Schritt ist das Datenmodell um Streckencharakteristika manuell zu erweitern, indem die offiziellen Streckenabschnitte in das Dataset eingebettet werden. Hierdurch sind im weiteren Verlauf der Arbeit wesentlich detailliertere Analysen möglich, da die identifizierten Fahrlinien eindeutigen Streckenlokationen zuzuordnen und somit innerhalb von (Fahrer-)Vergleichsgruppen präzise gegenüberzustellen und interpretierbar sind. Die korrekte Zuordnung erfolgt über *Lightbeam*, da dieses Attribut die Durchfahrt einer Lichtschranke entsprechend kennzeichnet und somit als Secondary Key für eine eindeutige Zuordnung genutzt werden kann. Die detaillierte Zuordnungslogik für die Bahn in Winterberg ist der Abbildung 55 zu entnehmen.

Abbildung 55: Definition von Streckenabschnitten für Winterberg



Lightbeam		Streckenabschnitt (Kurve)
von	*bis*	
s	1	Startphase
1	B10	1 K0
B10	B11	1 K0
B11	B12	2 K1
B12	2	3 K2
2	B15	4 K3
B15	B16	5 K4
B16	3	6 K5
3	B18	7 K6
B18	B19	8 K7
B19	4	9 K8
B4	B21	10 K9
B21	B22	11 K10 & 11
B22	5	12 K12
5	B24	13 K13
B24	B25	13 K13
B25	Z	14 K14

6.4.3 Hinzufügen von definierten Abschnittscharakteristika

Die zuvor definierten Streckenabschnitte sind in einem weiteren Schritt um Streckencharakteristika gemäß folgender Kategorielogik zu erweitern.

Tabelle 12: Definition von Abschnittscharakteristika in Winterberg

Streckenabschnitt (Kurve)	Charakteristika
Startphase	Start
1 K0	Start
1 K0	Start
2 K1	Gerade
3 K2	Kurve
4 K3	Gerade
5 K4	Gerade
6 K5	Kurve
7 K6	Gerade
8 K7	Kurve
9 K8	Gerade
10 K9	Kurve
11 K10 & 11	Kurve
12 K12	Gerade
13 K13	Gerade
13 K13	Gerade
14 K14	Kurve

Durch die Klassifikation der Streckenabschnitte ist ein inhaltliches Clustering einzelner Abschnitte realisierbar, was die Untersuchung weiterer Forschungsfragen ermöglicht: Exemplarisch sei hierbei die These angebracht, dass auf geraderen Streckenabschnitten der Reibungskoeffizient tendenziell geringer ausfällt, sofern die vorherige Kurve mit einem höheren Rollwinkel passiert wurde. In diesem Fall könnte bei dem Übergang auf die Gerade eine höhere Beschleunigung realisiert werden, welche durch das *Frictional Heating* den Reibungskoeffizienten reduziert. Durch die Berücksichtigung von Abschnittscharakteristika ist eine Aggregation und somit dezidierte Analyse auf jene definierten Streckensegmente möglich. Einschränkend angemerkt sei hierbei die subjektive Zuordnung der Charakteristika – für streng wissenschaftliche Anforderungen wäre hierbei ein objektivierter Ansatz anzuraten, indem die Charakteristika anhand fundierter Neigungs-

und Krümmungseigenschaften der jeweiligen Streckenabschnitte erfolgt. Zudem gilt zu berücksichtigen, dass die fortan als Geraden definierten Abschnitte realiter auch Krümmungen aufweisen – insbesondere in den Abschnittseingängen sowie -ausgängen – sodass auch in diesen Streckenabschnitten Rollwinkel ungleich null zu beobachten sind.

6.4.4 Kalkulation von Gesamt- und Abschnittslaufzeiten

Eine weitere notwendige Operation im Zuge der Datenaufbereitung liegt in der Kalkulation der Laufzeiten. Aufgrund der vorliegenden Rohdatenstruktur, nach der die Daten jeweils in Zehntelsekunden-Intervallen festgehalten werden – ohne hierbei Teilabschnitte zu berücksichtigen – ist eine Kalkulationslogik auf Gesamt- sowie Abschnittsebene für die spätere Datenanalyse anzuraten. Hierfür werden in separaten Spalten die kumulierten Laufzeiten sowohl auf Gesamt- als auch auf Abschnittsebene festgehalten. Bei der Datenauswertung kann hierdurch unmittelbar auf diese Werte zurückgegriffen werden, sodass der Datensatz in spezifische Cluster, etwa einzelnen Streckenabschnitte unter auszuwählenden Konditionen – beispielsweise definierten Laufzeitintervallen – eingeteilt werden kann, um diese hinsichtlich der gewählten Fahrlinien zu vergleichen. Die Implementierung entsprechender Kalkulationslogik ist anhand des folgenden exemplarischen Ausschnittes ersichtlich.

Abbildung 56: Erfassung von Abschnitts- und Gesamtzeit

Kurvenzeit	Fahrzeit	Fahrer	Streckenabschnitt	Streckenabschnitt_Char
5.42000	18.17000	Dvorak_Dominik	3 K2	Kurve
5.43000	18.18000	Dvorak_Dominik	3 K2	Kurve
5.44000	18.19000	Dvorak_Dominik	3 K2	Kurve
5.45000	18.20000	Dvorak_Dominik	3 K2	Kurve
5.46000	18.21000	Dvorak_Dominik	3 K2	Kurve
0.01000	18.22000	Dvorak_Dominik	4 K3	Gerade
0.02000	18.23000	Dvorak_Dominik	4 K3	Gerade
0.03000	18.24000	Dvorak_Dominik	4 K3	Gerade
0.04000	18.25000	Dvorak_Dominik	4 K3	Gerade
0.05000	18.26000	Dvorak_Dominik	4 K3	Gerade

Diesem ist zu entnehmen, dass bei vorliegendem Lauf die Laufzeit für den Streckenabschnitt 3 K2 bei 5,46 Sekunden liegt, während die Gesamtlaufzeit zu diesem Zeitpunkt 18,21 Sekunden beträgt. Mit diesen Informationen ist der Datensatz auf diversen Ebenen zu analysieren: Neben Slicing- und Clustering-Ansätzen existiert darüber hinaus auf Basis der kumulierten Gesamtlaufzeit zu jedem Zeitpunkt des Laufs die Information über die Gesamtperformance des Piloten im Vergleich zur Konkurrenz. Hierdurch sind zusätzlich Analysen über die relative Laufzeitentwicklung sowie die aktuelle Positionsentwicklung möglich, deren Erkenntnisse wiederum mit gewählten Fahrlinien in Verbindung gebracht werden können.

Nachdem das initiale Datenmodell mit vorgestellter Methodik alle relevanten inhaltlichen Aspekte für die Durchführung anforderungsgerechter Analysen erfüllt, ist im letzten Schritt der Data Preparation die Validität der Daten im Rahmen einer Datenbereinigung zu prüfen.

6.4.5 Prüfung auf duplizierte Rohdaten

Der erste Schritt der Datenbereinigung umfasst die Prüfung auf duplizierte Datensätze. Die Identifikation und Löschung von Duplikaten aus dem Dataset ist von essenzieller Bedeutung für die Qualität der Untersuchung, da hiermit Verzerrungen in den statistischen Auswertungen, wie etwa statistischen Lageparametern, auf denen im Verlauf der Analyse wesentliche inhaltliche Schlussfolgerungen basieren, vermieden werden. Tabelle 13 fasst die Resultate der Prüfung zusammen.

Tabelle 13: Größe des Datenkorpus nach Entfernung von Duplikaten

Betrachtungsebene	Anzahl Zeilen
Datenkorpus nach Datenaufbereitung	2.158.432
davon einmalige Datensätze	1.962.472
davon Duplikate	195.960
Datenkorpus nach Duplikatenlöschung	1.962.472

Demnach konnten ca. 196.000 Duplikate identifiziert und aus dem Datensatz entfernt werden, sodass der bereinigte Datenkorpus ca. 1,97 Millionen Datensätze umfasst.

6.4.6 Löschen nicht benötigter Daten

Neben der Vollständigkeit und Validität stellt Lean Data eine zusätzliche Komponente eines anforderungsgerechten Datenmodells dar. Die Intention hierbei ist, dass nur jene Daten berücksichtigt werden, welche für die Analyse im weiteren Verlauf auch realiter benötigt werden. Entsprechend sind alle weiteren Daten, wie etwa Hilfsdaten, welche zur Datenanreicherung benötigt wurden, aber auch Rohdaten, welche inhaltlich unter Berücksichtigung des Forschungsfokus nicht von analytischem Interesse sind, aus dem Datenkorpus zu entfernen. Hierdurch werden Effizienzgewinne im Storage sowie der Computation erzielt, sodass die Datenverarbeitung und -analyse effizienter, schneller und günstiger durchgeführt werden kann. In der vorliegenden Fallstudie sind hierunter jene Daten zu subsumieren, welche in der Startzone vor dem Start und nach der Zieldurchfahrt im Auslauf erhoben werden. Tabelle 14 fasst die Identifikation und Löschung nicht benötigter Daten für den vorliegenden Datenkorpus zusammen.

Tabelle 14: Größe des Datenkorpus nach Entfernung nicht benötigter Daten

Betrachtungsebene	Anzahl Zeilen
Datenkorpus nach Duplikatenlöschung	1.962.472
davon Daten vor Start	144.372
davon Daten nach Zieldurchfahrt	220.574
Datenkorpus bereinigt	1.597.526

Demgemäß konnten ca. 360.000 Datenzeilen aus dem Datenkorpus entfernt werden, sodass der Datensatz nach Bereinigung ca. 1,6 Millionen Daten umfasst.

6.4.7 Prüfung auf Missing Values

Weiterhin sind die Daten auf inhaltliche Plausibilität zu prüfen: So stellen im Data Analytics-Umfeld neben duplizierten Werten gemeinhin auch Not a Number-Values häufig ein Problem dar. Diese können unter anderem aus fehlerhaften Erhebungen stammen, etwa durch fehlerhafte Sensordaten. Die Prüfung für die vorliegende Fallstudie ergab folgende Resultate:

Tabelle 15: Prüfung des Datenkorpus auf fehlende Werte

Attribute	nan-Values
Time s	0
Distance m	0
Speed km/h	0
Acc x m/s2	0
Acc y m/s2	0
Acc z m/s2	0
Roll angle deg	0
Lightbeam	0
LB time s	1.592.614
Fahrer + Fahrt	0
Lightbeam_new	0
Kurvenzeit	0
Laufzeit	0
Fahrer	0
Streckenabschnitt	0
Streckenabschnitt_Char	0
Fahrerlauf	0
Wettkampf	0
Laufnummer	0
Konkurrenz	0
Datum	0

Es zeigt sich, dass bei den relevanten Attributen und Metriken keine NaN-Values auftreten. Die Ausprägungen in *LB time s* sind erwartungsgemäß, da in dieser Spalte nur dann ein Zeitwert festgehalten wird, sofern in dem entsprechenden Zeitstempel die Durchfahrt einer Lichtschranke registriert wird.

6.4.8 Prüfung auf inhaltliche Plausibilität der Metriken

Im abschließenden Schritt der Datenvalidität werden die Ausprägungen der Metriken anhand von definierten Regeln und Schwellwerten untersucht, um invalide

Beobachtungen zu identifizieren. Dies erfolgt anhand eines zweistufigen Prozesses, bei dem die Ausprägungen der Metriken zunächst auf Duplikate untersucht werden, bevor in einer inhaltlichen Prüfung vordefinierte Schwellwerte als Vergleichsparameter hinzugezogen werden.

Die Prüfung auf duplizierte Beobachtungen ist inhaltlich von der Prüfung in Abschnitt 6.4.5. insofern abzugrenzen, dass der Datenkorpus bei vorheriger Untersuchung auf duplizierte Datensätze, also deckungsgleiche Datenzeilen, geprüft wird, wohingegen die nun durchzuführende Prüfung explizit jede einzelne Metrik separat auf Duplikate untersucht – innerhalb der einzelnen registrierten Fahrten. Der folgende Auszug gibt einen Überblick über die Anzahl der identifizierten Duplikate pro Metrik je Lauf – absteigend sortiert nach jenen Datensätzen mit der höchsten summierten Duplikatanzahl.

Abbildung 57: Identifikation von duplizierten Beobachtungen der Metriken

FahrerLauf	Wettkampf	Dup_Distance_m	Dup_Speed	Dup_Time	Dup_Rollangle	Dup_Acc_x	Dup_Acc_y	Dup_Acc_z	sum dup
Friedrich_Francesco-Run-2	4MEN_2020-01-04	547.00000	655.00000	0.00000	7.00000	12	25.00000	25.00000	1271.00000
Friedli_Simon-Run-2	4MEN_2022-01-09	0.00000	25.00000	0.00000	1.00000	7	4.00000	1.00000	38.00000
Kibermanis_Oskars-Run-1	4MEN_2020-01-04	0.00000	0.00000	0.00000	1.00000	24	3.00000	4.00000	32.00000
Gaitiukevich_Rostislav-Run-1	2MEN_2021-01-09	0.00000	0.00000	0.00000	3.00000	23	2.00000	3.00000	31.00000
Lochner_Johannes-Run-2	2MEN_2022-01-08	0.00000	0.00000	0.00000	4.00000	17	6.00000	3.00000	30.00000
Heinrich_Romain-Run-1	4MEN_2020-01-03	0.00000	0.00000	0.00000	0.00000	22	6.00000	2.00000	30.00000
Andrianov_Maxim-Run-2	4MEN_2021-12-11	0.00000	0.00000	0.00000	3.00000	18	5.00000	3.00000	29.00000
Hall_Brad-Run-1	2MEN_2021-01-09	0.00000	16.00000	0.00000	0.00000	8	4.00000	0.00000	28.00000
Andrianov_Maxim-Run-2	2MEN_2022-01-08	0.00000	0.00000	0.00000	0.00000	17	6.00000	3.00000	26.00000
Vogt_Michael-Run-2	2MEN_2021-01-09	0.00000	0.00000	0.00000	3.00000	14	8.00000	1.00000	26.00000
Friedrich_Francesco-Run-2	4MEN_2022-01-09	0.00000	4.00000	0.00000	2.00000	11	6.00000	3.00000	26.00000
Bascue_Codie-Run-2	4MEN_2021-12-11	0.00000	13.00000	0.00000	0.00000	7	5.00000	1.00000	26.00000
Andrianov_Maxim-Run-1	4MEN_2021-12-12	0.00000	1.00000	0.00000	3.00000	14	7.00000	0.00000	25.00000
Friedrich_Francesco-Run-2	4MEN_2021-01-10	0.00000	0.00000	0.00000	1.00000	12	11.00000	1.00000	25.00000
Heinrich_Romain-Run-2	4MEN_2021-12-12	0.00000	3.00000	0.00000	0.00000	14	7.00000	1.00000	25.00000
Andrianov_Maxim-Run-1	4MEN_2021-12-11	0.00000	0.00000	0.00000	1.00000	18	3.00000	3.00000	25.00000
Friedli_Simon-Run-1	2MEN_2022-01-08	0.00000	10.00000	0.00000	1.00000	6	6.00000	1.00000	24.00000
Friedrich_Francesco-Run-1	2MEN_2022-01-08	0.00000	1.00000	0.00000	0.00000	13	8.00000	2.00000	24.00000
Baumgartner_Patrick-Run-2	2MEN_2021-01-09	0.00000	3.00000	0.00000	1.00000	18	2.00000	0.00000	24.00000
Vogt_Michael-Run-1	4MEN_2020-01-04	0.00000	1.00000	0.00000	2.00000	10	7.00000	3.00000	23.00000
Suk_Youngjin-Run-2	4MEN_2022-01-09	0.00000	0.00000	0.00000	1.00000	13	7.00000	2.00000	23.00000

Auf Grundlage dieser Zusammenfassung ist festzustellen, dass von allen registrierten Läufen einzig der erste Lauf von Francesco Friedrich im Viererbob-Weltcup vom 01.04.2020 infolge einer fehlerhaften Datenerhebung eine relevant hohe Anzahl an Duplikaten enthält. So wurden im Datensatz dieses Laufs über alle Zeit-, Geschwindigkeits- sowie Beschleunigungsparameter hinweg insgesamt 1.271 duplizierte Ausprägungen identifiziert, was bei einer durchschnittlichen Datenmenge je Lauf von ca. 6.700 Zeilen × 7 Metriken einen Anteil von rund 3 Prozent ergibt. Besonders auffällig ist die Menge an Duplikaten bei der Geschwindigkeit sowie der Distanz, die sich auf ca. 8 respektive 10 Prozent aller erhobenen Daten dieser Metriken für diesen Lauf belaufen. Ebenso ist die gleiche Anzahl an Duplikaten – wenn auch auf niedrigerem Niveau – bei der Quer- sowie

Vertikalbeschleunigung auffällig, was dafür spricht, dass die angebrachten Sensoren zwischenzeitlich Fehlfunktionen unterlagen. Ein Ausschnitt aus dem angesprochenen Datensatz unterstreicht die Beobachtungen:

Abbildung 58: Fehlerbehafteter Friedrich-Datensatz

Fahrer + Fahrt	Time s	Distance m	Speed km/h	Acc x m/s2	Acc y m/s2	Acc z m/s2	Roll angle deg
Friedrich_Francesco-Run-1.csv 4MEN_2020-01-04	4.33700	-0.00002	0.00000	-0.90558	-0.11623	9.74942	0.06336
Friedrich_Francesco-Run-1.csv 4MEN_2020-01-04	4.34700	-0.00001	0.00000	-0.91876	-0.11574	9.75646	0.06500
Friedrich_Francesco-Run-1.csv 4MEN_2020-01-04	4.35700	-0.00001	0.00000	-0.92937	-0.11520	9.75606	0.06662
Friedrich_Francesco-Run-1.csv 4MEN_2020-01-04	4.36700	-0.00001	0.00000	-0.93661	-0.11451	9.74081	0.06823
Friedrich_Francesco-Run-1.csv 4MEN_2020-01-04	4.37700	-0.00000	0.00000	-0.94473	-0.11420	9.75532	0.06981
Friedrich_Francesco-Run-1.csv 4MEN_2020-01-04	4.38700	-0.00000	0.00000	-0.94867	-0.11367	9.74754	0.07138
Friedrich_Francesco-Run-1.csv 4MEN_2020-01-04	4.39700	0.00000	0.00000	-0.95057	-0.11317	9.73979	0.07291
Friedrich_Francesco-Run-1.csv 4MEN_2020-01-04	4.40700	0.00000	0.00000	-0.95260	-0.11297	9.75436	0.07443
Friedrich_Francesco-Run-1.csv 4MEN_2020-01-04	4.41700	0.00001	0.00000	-0.95112	-0.11262	9.75408	0.07591
Friedrich_Francesco-Run-1.csv 4MEN_2020-01-04	4.42700	0.00001	0.00000	-0.94689	-0.11223	9.74639	0.07736
Friedrich_Francesco-Run-1.csv 4MEN_2020-01-04	4.43700	0.00001	0.00000	-0.94209	-0.11204	9.75360	0.07879
Friedrich_Francesco-Run-1.csv 4MEN_2020-01-04	4.44700	0.00001	0.00000	-0.93317	-0.11162	9.73852	0.08017
Friedrich_Francesco-Run-1.csv 4MEN_2020-01-04	4.45700	0.00001	0.00000	-0.92517	-0.11158	9.75320	0.08153

Mit steigender Laufzeit reduziert sich die Geschwindigkeit sowie die absolvierte Distanz zwischenzeitlich bis auf 0 km/h bzw. 0 m und darüber hinaus, sodass die ermittelten Daten für diesen Lauf nachfolgend als nicht plausibel angesehen und aus dem Datenkorpus entfernt werden. Auf eine mögliche Datenmodifikation infolge einer detaillierten Auseinandersetzung mit den fehlerhaften Daten wird für die vorliegende Studie aus Komplexitätsgründen verzichtet.

Die folgenden Datensätze, die nach dem diskutierten Datensatz von Francesco Friedrich die meisten Duplikate in den Ausprägungen der Metriken aufweisen, werden aufgrund des geringen Anteils an Duplikaten gemessen an der Gesamtmenge der Datensätze weiterhin im Datenkorpus berücksichtigt. Beispielsweise sei hierbei auf den zweiten Lauf von Simon Friedli verwiesen: Auch hier ist eine (mindestens) stellenweise fehlerhafte Geschwindigkeitsmessung in Anbetracht der 25 Duplikate zu unterstellen, was jedoch in Anbetracht der gesamten Datenmenge nur einen Duplikatanteil von 0,4 Prozent ausmacht. Die Auswirkung der Datenbereinigung auf den Datenkorpus ist der Tabelle 16 zu entnehmen.

Tabelle 16: Größe des Datenkorpus nach Bereinigung um Friedrich-Datensatz

Betrachtungsebene	Anzahl Zeilen
Datenkorpus vor zusätzlicher Bereinigung	1.597.526
davon „Friedrich 4er Bob 2020-01-04“	5.538
Datenkorpus bereinigt	1.591.988

Somit wurde der Datenkorpus durch Entfernung des fehlerhaften Friedrich-Datensatzes um 5.538 reduziert.

Hierbei gilt zu berücksichtigen, dass der vorgestellte Duplikate-Prüfprozess keine wertbasierte Plausibilitätsprüfung, sondern rein die Identifikation und Beurteilung von Duplikaten zum Ziel hat, die bei entsprechenden Beobachtungen in manuellen Prüfprozessen intensiviert werden können. Ein potenzieller Ausschluss der weiteren identifizierten und durchaus als kritisch zu bewertenden Datensätze mit Duplikaten wird fortan implizit im Rahmen des regelbasierten Prüfmechanismus evaluiert, welcher jene Datensätze identifiziert, die von konkreten Schwellen- oder Durchschnittswerten abweichen. Hierzu sei angemerkt, dass eine automatisierte Löschung bei Regelverstoß im zugrundeliegenden Python-Coding integriert ist, welche jedoch nicht angewandt wird. Der Grund hierfür liegt darin, dass eine automatisierte Löschung von Ausreißern ohne manuelle Kontrolle das Risiko birgt, potenziell wertvolle Informationen aus dem Datenmodell zu entfernen, sofern die Daten außerhalb der Norm nicht fehlerhaft sind. Der Hintergrund ist hierbei, dass insbesondere jene Datensätze häufig von erhöhter Aussagekraft und inhaltlichem Mehrwert für die Forschung sind, die außerhalb des Normbereichs liegen, sodass eine Datenlöschung nur nach hinreichender Prüfung erfolgen sollte. Tabelle 17 fasst die definierten Regeln zusammen, nach denen kritische Datensätze identifiziert und für weitere manuelle Prüfungen ausgewählt werden.

Tabelle 17: Regeln zur Plausibilitätsprüfung der Metriken

Metrik	statistischer Parameter	Prüfregel
roll angle deg	max	> 145°
	min	< -145°
	mean*	> 2-fache OR < 0,33-fache des Mittelwerts der Bench**
Speed km/h	max	> 1,05-fache der ∅-Maximalgeschwindigkeit der Bench
	min	< 0 km/h
	mean	> 1,05-fache OR < 0,95-fache des Mittelwerts der Bench
Distance m		
	max	> 1.400 m
	min	< 0 m
Gesamt-laufzeit	max	> 57 s
	min	< 0 s
	mean	> 1,04-fache OR < 0,96-fache des Mittelwerts der Bench
Acc x m/s2	max	> 4° über dem ∅ Acc x m/s2-Maximum der Bench
	min	< -4° über dem ∅ Acc x m/s2-Minimum der Bench
	mean	> + 2° OR <-2° des Mittelwerts der Bench
Acc y m/s2	max	> 6° über dem ∅ Acc y m/s2-Maximum der Bench
	min	< -6° über dem ∅ Acc y m/s2-Minimum der Bench
	mean	> + 1° OR <-1° des Mittelwerts der Bench
Acc z m/s2	max	> 4° über dem ∅ Acc z m/s2-Maximum der Bench
	min	< -4° über dem ∅ Acc z m/s2-Minimum der Bench
	mean	> + 1° OR <-1° des Mittelwerts der Bench
* Mittelwert über gesamte Fahrt hinweg		
** Bench: Vergleichsgruppe ist jeweils die gleiche Konkurrenz		
*** konkurrenzübergreifende Datensätze mit Regelverstoß		

Es sei betont, dass die vorliegenden Prüfregeln subjektiven Definitionen unterliegen, welche auf physikalischen Überlegungen sowie der beobachteten statistischen Verteilungen der Ausprägungen der jeweiligen Metriken basieren. Methodisch wurden für jede Metrik die Maxima, Minima sowie der Mittelwert aller Läufe

– separiert nach Zweier- und Viererbob-Konkurrenz – ermittelt und als Benchmark für die Datensätze der einzelnen Läufe hinsichtlich dieser statistischen Parameter genutzt. Somit ergeben sich – abgesehen für die *Distance m*, da diese implizit in *Speed* bereits mehrheitlich abgebildet wird – für jede Metrik drei Prüfprozesse. Die Resultate dieser separaten Prüfprozesse weisen jeweils die Anzahl identifizierter Läufe aus, welche die entsprechenden Prüfregeln nicht erfüllen konnten. Aufgrund der statistischen Interdependenz zwischen Maximum, Minimum und Mittelwert sind die Einzelresultate je Metrik in der inhaltlichen Prüfung der Datensätze gemeinsam zu interpretieren: Sofern ein Lauf mehrere Prüfregeln verletzt, ist dies ein starker Indikator dafür, dass der Datensatz grundsätzlich fehlerbehaftet sein könnte und es sich nicht um einen einzelnen Ausreißer – der aufgrund das Gesetzes der großen Zahl im Rahmen der Gesamtbetrachtung zu vernachlässigen wäre – handelt. Im Ergebnis werden alle Läufe in einem separaten Datensatz ausgegeben, welche gegen die Prüfregeln verstoßen – zuzüglich einer Angabe über die Art des Verstoßes.

Abbildung 59: Auffällige Datensätze im Rahmen der Plausibilitätsprüfung

Fahrer	FahrerLauf	Regelverstoß	Wettkampf
Won_Yunjong	Won_Yunjong-Run-2	Acc_z_min	4MEN_2022-01-09
Won_Yunjong	Won_Yunjong-Run-2	Acc_x_min	4MEN_2020-01-04
Won_Yunjong	Won_Yunjong-Run-2	Acc_x_min	4MEN_2020-01-03
Won_Yunjong	Won_Yunjong-Run-1	Acc_x_min	4MEN_2020-01-04
Won_Yunjong	Won_Yunjong-Run-1	Acc_x_min	4MEN_2020-01-03
Vogt_Michael	Vogt_Michael-Run-2	Acc_z_min	4MEN_2020-01-03
Vogt_Michael	Vogt_Michael-Run-2	Fahrzeit_max	4MEN_2020-01-03
Vogt_Michael	Vogt_Michael-Run-2	Acc_y_min	4MEN_2020-01-03
Vogt_Michael	Vogt_Michael-Run-2	ra_mean	4MEN_2020-01-03
Vogt_Michael	Vogt_Michael-Run-2	Acc_x_max	4MEN_2020-01-03
Vogt_Michael	Vogt_Michael-Run-2	ra_min	4MEN_2020-01-03
Vogt_Michael	Vogt_Michael-Run-2	Acc_z_mean	4MEN_2020-01-03
Tentea_Mihai Cristian	Tentea_Mihai Cristian-Run-1	Acc_z_min	2MEN_2022-01-08
Sun_Kaizhi	Sun_Kaizhi-Run-2	Acc_x_max	2MEN_2022-01-08
Sun_Kaizhi	Sun_Kaizhi-Run-2	Acc_z_mean	2MEN_2022-01-08

Auf Grundlage dieser Daten ist in der Folge über eine Anreicherung der Daten aus dem Korpus eine umfassende manuelle Prüfung zur abschließenden Beurteilung möglich. Hierzu dienen ebenso die bereits kalkulierten statistischen Parameter der Benchmarks als sinnvolle Referenz, um die Abweichungen vom Normbereich zu interpretieren. Tabelle 18 gibt einen entsprechenden Überblick über alle geprüften Datensätze inklusive der Entscheidung über die weitere Berücksichtigung im Datenmodell.

Tabelle 18: Ergebnisdarstellung der inhaltlichen Plausibilitätsprüfung

Nr.	Fahrer	Wettkampf	Lauf	Anzahl Regelverstöße	Diagnostik	Handhabe
1	Patrick Baumgartner	4MEN_2020-01-04	1	5	Daten zu deutlich außerhalb der Norm (Acc y, Acc z, Rollwinkel)	löschen
2	Francesco Friedrich	4MEN_2021-12-11	1	2	Daten zu deutlich außerhalb der Norm (max und mean Acc x)	behalten (u. V.)
3	Francesco Friedrich	4MEN_2021-12-11	2	2	Daten zu deutlich außerhalb der Norm (max und mean Acc x)	behalten (u. V.)
4	Christoph Hafer	4MEN_2021-12-11	2	1	Abweichung min Acc z im Toleranzbereich	behalten
5	Brad Hall	4MEN_2022-01-09	1	1	Abweichung min Acc x im Toleranzbereich | individueller Fahrstil?	Behalten
6	Brad Hall	4MEN_2021-12-11	1	1	Abweichung min Acc z im Toleranzbereich	behalten
7	Brad Hall	4MEN_2021-12-11	2	1	Abweichung min Acc x im Toleranzbereich | individueller Fahrstil?	Behalten
8	Davis Kaufmanis	4MEN_2021-12-11	1	6	Daten zu deutlich außerhalb der Norm (Acc y, Acc z, Rollwinkel)	löschen
9	Ryo Shinohara	4MEN_2020-01-04	1	8	Daten zu deutlich außerhalb der Norm (Acc y, Acc z, Rollwinkel)	löschen

10	Ryo Shino-hara	4MEN_2020 -01-03	1	1	Abweichung der Gesamtlaufzeit im Toleranzbereich	behalten
11	Youngjin Suk	4MEN_2020 -01-04	1	1	Abweichung mean Acc y im Toleranz-bereich | individuel-ler Fahrstil?	Behalten (u. V.)
12	Youngjin Suk	4MEN_2020 -01-04	2	1	Abweichung mean Acc y im Toleranz-bereich | individuel-ler Fahrstil?	Behalten (u. V.)
13	Youngjin Suk	4MEN_2020 -01-03	1	1	Abweichung mean Acc y im Toleranz-bereich | individuel-ler Fahrstil?	Behalten (u. V.)
14	Youngjin Suk	4MEN_2020 -01-03	2	1	Abweichung mean Acc y im Toleranz-bereich | individuel-ler Fahrstil?	Behalten (u. V.)
15	Kaizhi Sun	4MEN_2022 -01-08	1	2	Abweichung Acc y, z im Toleranzbe-reich	behalten
16	Kaizhi Sun	4MEN_2022 -01-08	2	5	Daten zu deutlich außerhalb der Norm (min & max roll angle)	löschen
17	Mihai Cris-tian Tentea	4MEN_2022 -01-08	1	1	Abweichung min Acc z im Toleranz-bereich	behalten
18	Michael Vogt	4MEN_2020 -01-03	2	7	Daten zu deutlich außerhalb der Norm (min Acc y, min Acc z)	löschen
19	Yunjong Won	4MEN_2022 -01-09	2	1	Abweichung min Acc z im Toleranz-bereich	behalten
20	Yunjong Won	4MEN_2020 -01-04	1	1	Abweichung min Acc x im Toleranz-	Behalten (u. V.)

					bereich | individueller Fahrstil?	
21	Yunjong Won	4MEN_2020-01-04	2	1	Abweichung min Acc x im Toleranzbereich | individueller Fahrstil?	Behalten (u. V.)
22	Yunjong Won	4MEN_2020-01-03	1	1	Abweichung min Acc x im Toleranzbereich | individueller Fahrstil?	Behalten (u. V.)
23	Yunjong Won	4MEN_2020-01-03	2	1	Abweichung min Acc x im Toleranzbereich | individueller Fahrstil?	Behalten (u. V.)

Die Analyse der identifizierten Läufe mit auffälligen Datensätzen bringt neben offensichtlich fehlerhaften Sensordaten auch erste potenzielle Indikationen für individuelle Fahrlinien hervor: So zeigen sich beispielsweise bei allen vier Läufen von Youngjin Suk deutliche Abweichungen bei der mittleren Querbeschleunigung versus Bench, was einerseits eine Folge von falsch angebrachten Sensoren sein könnte, andererseits jedoch auch mit einem spezifischen Fahrstil begründbar sein könnte. Die Abweichungen von der Norm sind in einem tolerierbaren Bereich, sodass die Daten für weitere Analysen – analog zu den Daten von Yunjong Won – beibehalten werden. Einen Sonderfall stellen die Datensätze von Francesco Friedrichs Läufen vom 12.11.2021 dar: Hierbei sind die Daten der Längsbeschleunigung aufgrund der hohen Abweichungen mit an Sicherheit grenzender Wahrscheinlichkeit als fehlerhaft einzuordnen. Da jedoch alle weiteren Metriken unauffällige Ausprägungen aufweisen und Friedrich als Spitzenfahrer trotz unterstellten inkorrekten Querbeschleunigungsdaten wertvolle Informationen über positive Zusammenhänge zwischen Laufzeit und den weiteren Fahrstilindikatoren geben könnte, werden die entsprechenden Datensätze unter Vorbehalt im endgültigen Datenkorpus berücksichtigt. Somit werden insgesamt fünf Datensätze aufgrund inhaltlicher Unplausibilität aus dem Datenkorpus entfernt, sodass dieser nach Abschluss der Data Preparation gemäß nachfolgender Übersicht ca. 1,56 Millionen Datenzeilen umfasst, welche sich insgesamt auf 283 Läufe verteilen.

Tabelle 19: Größe des Datenkorpus nach Bereinigung um invalide Daten

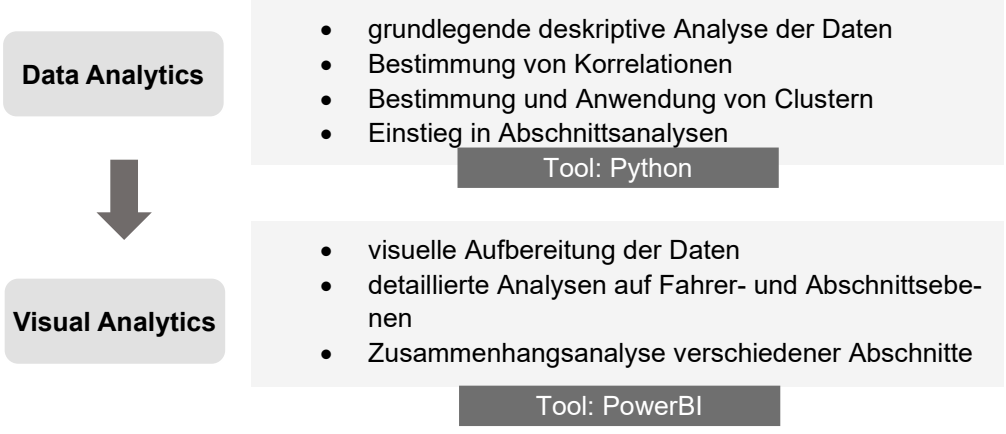
Betrachtungsebene	Anzahl Zeilen
Datenkorpus vor endgültiger Bereinigung	1.591.988
davon unplausible Datensätze	28.267
Datenkorpus bereinigt	1.563.721

Neben der Aufbereitung einer validen Datengrundlage für die folgenden Analysen hat insbesondere die inhaltliche Plausibilitätsprüfung erste wertvolle Erkenntnisse offenbart: So konnten beim impliziten Benchmark-Vergleich deutliche Unterschiede zwischen den Ausprägungen der statistischen Parameter der einzelnen Metriken jeweiliger Fahrer und Läufe identifiziert werden, was nachfolgend als Definitionsgrundlage für verschiedene Fahrlinien zu werten ist. Im Zentrum des folgenden Analyseteils dieser Arbeit steht die Frage, inwiefern diese Unterschiede auch die Variationen der Laufzeiten erklären.

6.5 Modelling

Der Analyseteil dieser Arbeit ist nachfolgend analog der Abbildung 60 in zwei Teilbereiche eingeteilt:

Abbildung 60: Vorgehen im Modelling



Zunächst werden die Daten in Python mit einem zahlenfokussierten Ansatz statistisch ausgewertet. Auf einer einführenden deskriptiven Analyse folgen hierzu Korrelations- sowie Clusteranalysen auf unterschiedlichen Aggregationsebenen, um erste fundamentale Einblicke in potenzielle Zusammenhänge zu gewinnen, diese den Thesen aus der theoretisch-physikalischen Annäherung gegenüberzustellen und hierauf die weiteren Analyseansätze aufzusetzen. Basierend auf den inhaltlichen Erkenntnissen sowie entsprechend aufbereiteten Datensätzen erfolgt über Microsoft PowerBI eine visuelle Analyse der Daten, um durch visuelle Datenexploration zusätzliche Einsichten zu gewinnen. Hierbei sind insbesondere auf Detailebene, etwa über die Auswertung des individuellen Fahrverhaltens einzelner Akteure in einzelnen Streckenabschnitten, wertvolle Informationen zu generieren.

6.5.1 Modelling I – Data Analytics via Python

6.5.1.1 Deskriptive Analyse

Der zugrundeliegende Datenkorpus umfasst insgesamt 283 Läufe, die sich auf 62 unterschiedliche Piloten über acht Wettkämpfe verteilen. Hiervon sind zwei Wettkämpfe dem Zweierbob sowie sechs dem Viererbob-Weltcup zuzuordnen. Da die Auswertungen und Analysen aus Nachvollziehbarkeits- und Simplifizierungsgründen zunächst den datenreicheren Viererbob-Weltcup fokussieren, beziehen sich nachfolgende Ausführungen und Untersuchungen primär auf die Viererbob Konkurrenz. Eine entsprechende Anwendung auf die Zweierbob-Daten geschieht in der praktischen Umsetzung analog. Tabelle 20 gibt hierzu einen Überblick über die Extremwerte sowie Lageparameter der einzelnen Metriken und bietet somit einen sinnvollen Einstieg in die Analyse: Als Quintessenz der deskriptiven Analyse ist festzuhalten, dass trotz der geringen mittleren Zeitunterschiede – die Standardabweichung liegt bei 0,6 Sekunden – die Ausprägungen der Metriken Bandbreiten aufweisen, welche unterschiedliche Fahrlinien vermuten und diese im Rahmen des Analyseprozesses auch als interpretierbar erscheinen lassen. Exemplarisch sei die Spannweite der Maxima des Rollwinkels anzuführen, welche bei circa 20° liegt. Diese ist als Indikator für unterschiedliche Linien bei den Kurvendurchfahrten zu interpretieren und dient somit als erste mögliche Einflussgröße auf die Laufzeit – genau diesen potenziellen Zusammenhang gilt es im Folgenden zu validieren.

Analog sind angesprochene Spannweiten auch bei den Beschleunigungs- sowie Geschwindigkeitsdaten festzustellen. Hierzu sind die jeweiligen Wechselbeziehungen der Parameter und ihre Auswirkungen als Ganzes zu untersuchen, um letztlich Aussagen über mögliche Optimierungspotenziale der Fahrstile treffen zu können. Aus diesem Grund ist ein ganzheitlicher Analyseansatz zu wählen, welcher alle Metriken mit den jeweiligen Interdependenzen berücksichtigt und keine Betrachtung einzelner Metriken, wie etwa der reinen Fokussierung auf den Rollwinkel. Zusätzlich ist zu vermuten, dass identifizierte Intervallausprägungen nicht die Folge gleichmäßiger Fahrlinienunterschiede sind, sondern sich in Anbetracht des vielfältigen Bahnlayouts – bestehend aus engen Kurven, aber auch geraden Passagen – auf bestimmte Streckenabschnitte beschränken. Hierfür sind in den Analysen auch die Charakteristika der einzelnen Streckenabschnitte zu berücksichtigen.

Tabelle 20: Deskriptive Analyse

	2 Men	4 Men
Anzahl Wettkämpfe	2	6
Anzahl Fahrer	27	35
Anzahl Läufe	71	212
Anzahl Datensätze	395.479	1.168.242
Fahrzeit [s]		
max	56,87	57,09
min	54,89	53,76
std	0,35	0,6
Absolvierte Gesamtdistanz [m]		
max	1.378	1.402
min	1.360	1.357
std	3,81	7,42
Top Speed [kmh]		
max	134,0	138,4
min	128,9	128,3
std	0,9	1,4
Rollwinkel [Grad]*		
maxmax	94,2	109,0
minmax	82,3	88,8
mean	33,1	33,5
std	30,6	31,0
Acc x [m/s2]		
maxmax	3,07	6,15 **
minmax	0,94	0,57
maxmin	-1,95	-0,3 **
minmin	-6,61	-8,71
mean	-0,68	-0,64
std	0,94	1,1
Acc y [m/s2]*		
maxmax	8,11	9,1
minmax	3,51	3,76
minmax	-2,77	-3,22
mean	1,12	1,21
std	1,03	1,11
Acc z [m/s2]		
maxmax	44,8	45,8
minmax	39,9	38,6
maxmin	8,6	8,8
minmin	3,4	3,0
mean	16,9	16,7
std	8,6	8,8

* basiert auf Betrag der Ausprägungen
** exkl. Ausreißer F. Friedrich

6.5.1.2 Data Analytics auf Gesamtbahnebene

Die vorliegende Datengrundlage umfasst zahlreiche Parameter, welche den Fahrtverlauf sowohl hinsichtlich der Zeitperformance als auch bezüglich der Lage und Positionierung des Bobs auf der Strecke, woraus die Fahrlinie abgeleitet werden kann, datentechnisch erfassen. Um aus der Vielzahl an Daten konkrete Aussagen über den Zusammenhang von Fahrlinie und benötigter Laufzeit tätigen zu

können, ist es zunächst notwendig, die Ursache-Wirkungs-Beziehungen der einzelnen Metriken zu untersuchen. Ein grundlegendes Verständnis über die Kausalitäten der einzelnen Parameter erlaubt es in der Folge auch, statistische Zusammenhänge inhaltlich validieren zu können. Auf Basis der theoretisch-physikalischen Einführung dieser Arbeit konnten bereits Thesen über potenzielle Zusammenhänge definiert werden, die fortan auf Grundlage der vorliegenden Fallstudie datenbasiert ergänzt werden. Als geeignetes statistisches Instrument zur Identifikation von potenziellen inhaltlichen Kausalitäten aus Daten gilt die Ermittlung von Korrelationen, welche den statistischen Zusammenhang zweier Merkmale beschreiben. Nachfolgende Darstellung bildet die statistischen Beziehungen der vorliegenden Metriken über eine Korrelationsmatrix ab und dient somit als Ansatzpunkt zur Aufdeckung von Interdependenzen.

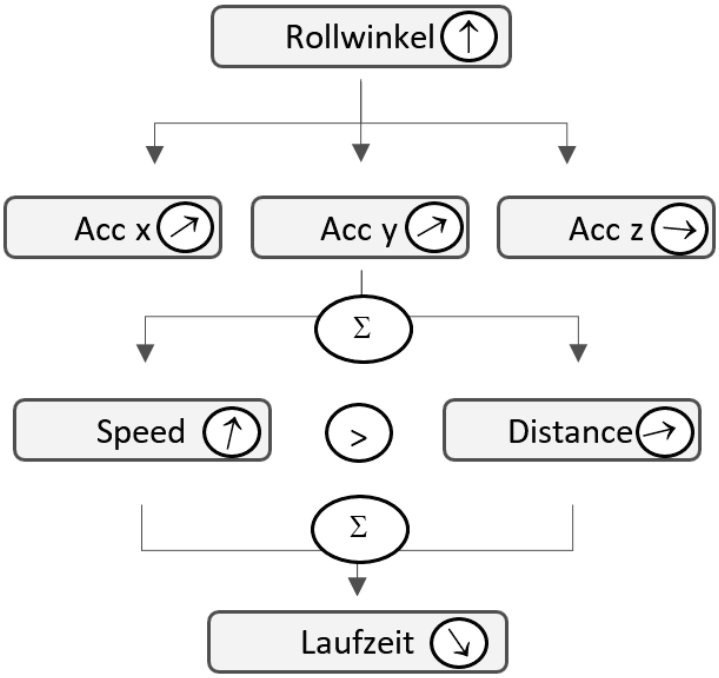
Tabelle 21: Darstellung der linearen Beziehungen der Metriken über eine Korrelationsmatrix

	Rollwinkel	Laufzeit	Speed km/h	Acc x	Acc y	Acc z	Distance m
Rollwinkel	1	-0,344	0,328	0,125	0,129	-0,017	0,037
Laufzeit	-0,344	1	-0,935	-0,204	0,173	-0,117	0,068
Speed km/h	0,328	-0,935	1	0,275	-0,158	0,044	0,215
Acc x	0,125	-0,204	0,275	1	0,07	-0,29	0,155
Acc y	0,129	0,173	-0,158	0,07	1	0,069	0,061
Acc z	-0,017	-0,117	0,044	-0,29	0,069	1	-0,181
Distance m	0,037	0,068	0,215	0,155	0,061	-0,181	1

Hieraus ist zu erkennen, dass die Gesamtlaufzeit sehr stark negativ mit der Geschwindigkeit korreliert: Wie intuitiv zu vermuten ist, reduziert sich die benötigte Laufzeit umso mehr, je höhere Geschwindigkeiten erzielt werden. Zusätzlich weist auch der Rollwinkel mit einem Korrelationskoeffizienten von -0,344 einen aussagekräftigen negativen statistischen Zusammenhang mit der Laufzeit auf. Auffällig ist hierbei auch die positive Korrelation zwischen Rollwinkel, Geschwindigkeit sowie Distanz: Inhaltlich erscheinen diese Interdependenzen plausibel, da durch den höheren Rollwinkel eine tendenziell höhere Längsbeschleunigung realisiert wird. Diese wiederum erhöht tendenziell die Geschwindigkeit, was jedoch in einer Erhöhung der zu absolvierenden Distanz führt, da der Bob in den Kurven weiter nach außen gedrückt wird. Dies erhöht wiederum die zurückzulegende Fahrtstrecke. Der Effekt der höheren Geschwindigkeit scheint gemäß der Daten

den negativen Effekt einer erhöhten Fahrdistanz jedoch zu überkompensieren, sodass die Laufzeit in Summe bei einer tendenziell weiten Fahrlinie mit höheren Rollwinkeln minimal wird. Abbildung 61 schematisiert die abgeleiteten Ursache-Wirkungsbeziehungen aus der Korrelationsanalyse.

Abbildung 61: Schematische Darstellung der identifizierten Kausalitäten



Zur weiteren Validierung dieser abgeleiteten Kausalitäten wird zusätzlich eine Clusteranalyse durchgeführt. Das Ziel hierbei ist es, die Datensätze je Metrik der Höhe ihrer Ausprägung nach in einzelne Cluster einzuteilen und deren Häufigkeitsverteilung mit den Clustern jeweils anderer gruppierten Metriken zu bestimmen. Exemplarisch sei dieser methodische Ansatz anhand der folgenden Häufigkeitstabelle zur Untersuchung des Zusammenhangs der Laufzeit und des Rollwinkels dargestellt:

Abbildung 62: Clusteranalyse über Gesamtlaufzeit vs. mittlere Rollwinkel

	0-20% mean G-Rollwinkel	20-40% mean G-Rollwinkel	40-60% mean G-Rollwinkel	60-80% mean G-Rollwinkel	80-100% mean G-Rollwinkel
0-20% GFahrzeit	0.06977	0.11628	0.20930	0.25581	0.34884
20-40% GFahrzeit	0.15909	0.34091	0.18182	0.13636	0.18182
40-60% GFahrzeit	0.21429	0.21429	0.23810	0.28571	0.04762
60-80% GFahrzeit	0.38636	0.18182	0.13636	0.11364	0.18182
80-100% GFahrzeit	0.30769	0.12821	0.23077	0.17949	0.15385

Hierzu wurden alle Läufe innerhalb der Viererbob-Konkurrenz und separiert nach den einzelnen Weltcups entsprechend der Laufzeiten- und mittleren Rollwinkeln in jeweils zwei getrennte Rankings überführt. Diese Rankings wurden dann jeweils in fünf Quantil-Cluster eingeteilt, welche vorliegend die Zeilen sowie Spalten definieren: Hierbei bildet die *0-20% G-Laufzeit*-Zeile die Top-20 Prozent Läufe – basierend ihres Rankings – ab, wohingegen die *0-20% mean G-Rollwinkel*-Spalte die 20 Prozent der Läufe mit den niedrigsten mittleren Rollwinkeln beinhalten. Durch den Einbezug von Rankings wird ein direkter Vergleich mit der Konkurrenz in die Analyse impliziert, was diese Methode von der reinen Korrelationsanalyse abhebt.

Im Ergebnis ist aus der Häufigkeitsverteilung abzulesen, dass die Gruppe der schnellsten Läufe nur zu 7 Prozent auch dem Cluster der niedrigsten mittleren Rollwinkel angehören, gleichzeitig jedoch 26 Prozent + 35 Prozent = 61 Prozent der 40 Prozent höchsten mittleren Rollwinkel zugehören. Gleichzeitig sind die 20 Prozent langsamsten Läufe durch 31 Prozent der niedrigsten mittleren Rollwinkel gekennzeichnet – und nur durch 15 Prozent der höchsten. Somit ist zu konstatieren, dass die Schlussfolgerung aus der Korrelationsanalyse, nach der höhere Rollwinkel tendenziell mit besseren Laufzeiten korrelieren, auch auf Grundlage der vorliegenden Clusteranalyse zu halten ist. Darüber hinaus gibt sie einen detaillierten Einblick in die Datenlage, indem sie die Analyse von Zusammenhängen in ausgewählten Kohorten ermöglicht. So ist ersichtlich, dass es punktuell auch sehr gute Laufzeiten mit – im Vergleich zur Konkurrenz – niedrigen Rollwinkeln gibt. Hieraus ist wiederum abzuleiten, dass eine tiefergehende Auswertung sowohl auf Fahrerebene als auch für einzelne Streckenabschnitte sinnvoll ist, um anhand der identifizierten Ausnahmen Detailwissen über die notwendigen Voraussetzungen zu erhalten, unter welchen Umständen welches Fahrverhalten vorteilhaft ist – entsprechende Detailanalysen folgen.

Als Fazit der Clusteranalyse bleibt zunächst festzuhalten, dass die aufgedeckten statistischen Zusammenhänge aus der Korrelationsanalyse im Mittel durch diese gestützt werden. Tabelle 22 stellt hierzu die aufgestellten Thesen aus der Korrelationsanalyse den Ergebnissen aus der Clusteranalyse gegenüber. Angemerkt

sei, dass hierbei aus Gründen der Anschaulichkeit lediglich die Top-20 Prozent- sowie Flop-20 Prozent-Quantile dargestellt werden, was vorliegend jedoch als hinreichend angesehen wird.

Tabelle 22: Clusteranalyse über den Zusammenhang der beobachteten Variablen auf Gesamtlaufzeit

Einflussgröße		Laufzeit		Indikation
		Top 20%	Flop 20%	
Rollwinkel	Top 20%	35%	15%	RW ↑⟹ FZ↓
	Flop 20%	7%	31%	
Acc x	Top 20%	28%	28%	X ↑⟹ FZ ↗
	Flop 20%	16%	28%	
Acc y	Top 20%	14%	41%	Y ↑⟹ FZ↑
	Flop 20%	28%	10%	
Acc z	Top 20%	21%	20%	Z ↑⟹ FZ ↘
	Flop 20%	28%	18%	
Speed	Top 20%	56%	0%	SP ↑⟹ FZ↓
	Flop 20%	0%	62%	
Distance	Top 20%	26%	26%	DS ↑⟹ FZ →
	Flop 20%	21%	21%	

Tabelle 23: Clusteranalyse über den Zusammenhang der beobachteten Variablen auf den Rollwinkel

erklärende Variable		erklärte Variable			Indikation
			Top 20%	Flop 20%	
Rollwinkel	Top 20%	Acc x	38%	15%	RW ↑⟹ X ↑
	Flop 20%		19%	20%	
Rollwinkel	Top 20%	Acc y	41%	10%	RW ↑⟹ Y ↑
	Flop 20%		8%	17%	
Rollwinkel	Top 20%	Acc z	44%	23%	RW ↑⟹ Z ↑
	Flop 20%		19%	27%	
Rollwinkel	Top 20%	Distance	28%	10%	RW ↑⟹ DS ↗
	Flop 20%		19%	21%	
Rollwinkel	Top 20%	Speed	49%	13%	RW ↑⟹ SP ↑
	Flop 20%		10%	39%	
	entspricht Korrelationsanalyse				
	weicht tendenziell von Ergebnissen der Korrelationsanalyse ab				

Im Ergebnis wird auf Grundlage des diskutierten Kausalitätsmodells der Rollwinkel, welcher durch die ermittelten Interdependenzen mit den weiteren Geschwindigkeits- sowie Beschleunigungsparameter die Laufzeit mittelbar beeinflusst, folglich als maßgebliche Einflussgröße auf die Fahrlinie betrachtet. Insofern nehmen die folgenden Analysemethoden den Rollwinkel als zentralen Indikator für die Fahrlinienwahl an.

6.5.1.3 Data Analytics auf Abschnittsebene

Nachdem erste grundsätzliche Erkenntnisse hinsichtlich des Zusammenhangs von Fahrlinie und Laufzeit auf aggregierter Ebene gewonnen werden konnten, werden diese im nächsten Schritt durch fokussierte Analysen auf einzelne Streckenabschnitte präzisiert. Hierzu wird zunächst eine Abschnittsanalyse durchgeführt, bei welcher der Einfluss einzelner Teilstreckenabschnitte – sowohl Kurvenabschnitte als auch definierte Streckencharakteristika – auf die Gesamtzeit eruiert wird, um erfolgskritische Streckenabschnitte zu identifizieren. Hierbei ist ein wesentlicher Analyseschwerpunkt auf die Startphase gelegt, um die aufgestellte

These, nach welcher der Startphase eine zentrale Bedeutung für die Performance zukommt, zu überprüfen. Aufbauend auf der Abschnittsanalyse erfolgt eine Untersuchung des Zusammenhangs der einzelnen Metriken auf die wesentlichen Streckenabschnitte, um potenziell zeitoptimierende Fahrlinien in den Schlüsselbereichen der Bahn in Winterberg aufzudecken.

a) Zusammenhang einzelner Abschnitte mit Gesamtzeit

Die Beurteilung des Einflusses der einzelnen Streckenabschnitte auf die Gesamtlaufzeit basiert nachfolgend auf einer separaten Betrachtung der Korrelationskoeffizienten nach Pearson sowie Spearman. Hierbei misst der Korrelationskoeffizient nach Pearson den linearen Zusammenhang zwischen den Abschnittszeiten sowie der Gesamtlaufzeit, wohingegen der Rangkorrelationskoeffizient nach Spearman die Abschnittspositionen den Endpositionen gegenüberstellt und somit bei Untersuchung des Zusammenhangs implizit das Abschneiden im Vergleich zur Konkurrenz berücksichtigt. Die Resultate sind in Tabelle 24 dargestellt.

Tabelle 24: Korrelationsanalyse über den Zusammenhang einzelner Abschnittslaufzeiten vs. Gesamtlaufzeiten

Abschnitt	Korrelationskoeffizient	
	Pearson	Spearman
Startphase	0,625	0,67
K0	0,643	0,705
K1	0,744	0,762
K2	0,933	0,885
K3	0,961	0,825
K4	0,957	0,814
K5	0,965	0,812
K6	0,924	0,571
K7	0,936	0,821
K8	0,914	0,649
K9	0,907	0,613
K10 und K11	0,906	0,674
K12	0,85	0,517
K13	0,839	0,587
K14	0,85	0,708

Generell zeigen sich erwartungsgemäß für jedweden Abschnitt positive Zusammenhänge mit den Endresultaten, wenngleich die Korrelationskoeffizienten nach Pearson durchgängig deutlich höher ausfallen. Dies ist darauf zurückzuführen, dass der Zusammenhang einer kürzeren Abschnittszeit auf die Gesamtzeit größer ist als die Auswirkung einer besseren Abschnittsplatzierung auf das Gesamtergebnis, da bei letzterer durch Berücksichtigung der Konkurrenz eine weitere Einflussgröße hinzukommt.

Da im Bobsport die Performance im Vergleich zur Konkurrenz die Endresultate definiert und somit die relative Verbesserung zum Wettbewerber als primäres Ziel auszugeben ist, wird nachfolgend der Rangkorrelationskoeffizient nach Spearman als zentraler Indikator der Abschnittsanalyse betrachtet. Hierbei ist auffällig, dass insbesondere die Abschnitte von Kurve 2 bis Kurve 7 – mit Ausnahme von Kurve 6 – mit einem Spearman-Koeffizienten > 0,8 sehr starke Korrelationen aufweisen und somit als kritische Erfolgsabschnitte festzuhalten sind. Demgegenüber sind die Passagen kurz vor Ziel weniger entscheidend, was inhaltlich auch plausibel zu begründen ist: Zu berücksichtigen ist hierbei ein kumulativer Effekt der steigenden Geschwindigkeit im Fahrtverlauf, der dazu führt, dass bei einer höheren Ausgangsgeschwindigkeit im Abschnittseingang entsprechend leichter eine bessere Zeit im folgenden Abschnitt zu realisieren ist.

Hieraus folgt demnach die notwendige Einschränkung bei der Interpretation, dass eine komplett separierte Betrachtung einzelner Abschnittszeiten nicht valide ist, da diese – abgesehen vom Start – jeweils auch von der Performance in den vorangegangenen Abschnitten abhängt. Aus diesem Grund sinkt der Zusammenhang zwischen Teil- und Gesamtleistungen im Fahrverlauf tendenziell: Während der kumulative Effekt aus den frühen Abschnitten entsprechend über längere Fahrdauer wirkt und somit die Endplatzierung maßgeblich beeinflusst, ist dieser Effekt kurz vor Ziel sehr limitiert. Zu hinterfragen hierbei ist der geringere Effekt aus der Startphase – eine mögliche Erklärung liegt wahrscheinlich in den sehr geringen Leistungsunterschieden und Zeitabständen unmittelbar zu Beginn des Laufs, sodass die Ausprägung des diskutierten Effekts erst nach den ersten Kurven verstärkt auftritt. Dies wiederum stünde jedoch in Kontrast zur Untersuchung von Brüggemann, Morlok und Zatsiorsky, welche der Startphase eine hohe Bedeutung nachweisen konnten (Brüggemann, Morlok und Zatsiorsky 1997, S. 103). Zur Beurteilung des Einflusses der Startphasen ist demnach eine detailliertere Auswertung der Daten sinnvoll und geschieht nachfolgend über den bereits eingeführten Clusteransatz.

Abbildung 63: Clusteranalyse über den Zusammenhang von Startzeiten vs. Gesamtlaufzeiten

Streckenabschnitt	Abschnittszeit_Quantil	0-20% G-Fahrzeit	20-40% G-Fahrzeit	40-60% G-Fahrzeit	60-80% G-Fahrzeit	80-100% G-Fahrzeit
0 Startphase	0-20% Abschnittszeit	0.59259	0.29630	0.07407	0.00000	0.03704
0 Startphase	20-40% Abschnittszeit	0.52632	0.31579	0.15789	0.00000	0.00000
0 Startphase	40-60% Abschnittszeit	0.47368	0.21053	0.15789	0.10526	0.05263
0 Startphase	60-80% Abschnittszeit	0.11905	0.26190	0.28571	0.23810	0.09524
0 Startphase	80-100% Abschnittszeit	0.02857	0.14286	0.20952	0.30476	0.31429

Hieraus ist ersichtlich, dass sich aus den Läufen mit den 20 Prozent schnellsten Starts knapp 60 Prozent auch in den besten 20 Prozent der Gesamtlaufzeiten wiederfinden, weitere 30 Prozent dieser Klasse rangieren zusätzlich im Bereich der 20-40 Prozent besten Laufzeiten. Demgegenüber sind von den Läufen mit den 20 Prozent langsamsten Starts nur 3 Prozent im Endresultat unter den Top-20 Prozent, wohingegen über 60 Prozent zu den langsamsten 40 Prozent gehören. Aus diesen Daten geht eindeutig hervor, dass ein guter Start eine wesentliche Voraussetzung für eine positive Endplatzierung ist. Gleichwohl zeigt die Clusteranalyse bei der Abschätzung der Effekthöhe auch den sehr hohen Einfluss der folgenden Streckenabschnitte:

Abbildung 64: Clusteranalyse über den Zusammenhang der Streckenabschnittszeiten vs. Top 20 Prozent Gesamtlaufzeiten sortiert nach Effekthöhe

Streckenabschnitt	Abschnittszeit_Quantil	0-20% G-Fahrzeit
3 K2	0-20% Abschnittszeit	0.84375
8 K7	0-20% Abschnittszeit	0.70588
6 K5	0-20% Abschnittszeit	0.68889
2 K1	0-20% Abschnittszeit	0.67647
4 K3	0-20% Abschnittszeit	0.66667
2 K1	20-40% Abschnittszeit	0.63158
14 K14	0-20% Abschnittszeit	0.60976
11 K10 & 11	0-20% Abschnittszeit	0.60000
0 Startphase	0-20% Abschnittszeit	0.59259
1 K0	0-20% Abschnittszeit	0.56250

Demnach weisen insbesondere die frühen Streckenabschnitte nach dem Start – vorliegend K2, K3, K5 und K7 – analog den hohen Korrelationen gemäß Spearman einen sehr starken positiven Zusammenhang zwischen Abschnitts- und Gesamtperformance auf, der auch jenen der Startphase übertrifft. Besonders erfolgskritisch erweist sich hierbei K2, wo festzustellen ist, dass die 20 Prozent schnellsten Läufe in diesem Abschnitt in 84 Prozent der Fälle auch im Gesamtklassement zu den 20 Prozent schnellsten Läufen gehören.

Als Fazit der Abschnittsanalyse ist somit zu konstatieren, dass der Abschnitt von K2-K7 als erfolgskritisch für eine gute Endplatzierung in Winterberg ist. Exakte Quantifizierungen sind hierbei aufgrund der Interdependenzen zwischen den einzelnen Abschnitten nicht möglich, dennoch ist auch die Startphase als fundamentaler Erfolgsfaktor zu bilanzieren: Neben einem hohen Rangkorrelationskoeffizienten von circa 0,7 über die ersten drei Abschnitte – Start bis inklusive K1 – unterstreicht auch die Korrelation zwischen der Startzeit und den folgenden Abschnitten die Wichtigkeit der Startphase.

Tabelle 25: Korrelationsanalyse über den Zusammenhang der Startzeiten vs. folgende Streckenabschnitte

Abschnitt	Korrelationskoeffizient	
	Pearson	Spearman
K0	0,886	0,809
K1	0,848	0,801
K2	0,726	0,691
K3	0,532	0,479
K4	0,5	0,404
K5	0,496	0,438
K6	0,446	0,311
K7	0,375	0,352
K8	0,358	0,209
K9	0,331	0,216
K10 und K11	0,319	0,179
K12	0,29	0,127
K13	0,286	0,15
K14	0,261	0,261

Demnach bleibt festzuhalten, dass der Abschnitt K2, welcher sowohl nach Korrelationsanalyse als auch auf Basis der Clusteranalyse den größten Effekt auf die Gesamtplatzierung hat, mit einem Spearman-Korrelationskoeffizienten von 0,7 maßgeblich von einem guten Start abhängt. Somit sind die Streckenbereiche bis inklusive K7 mit leicht differierenden – und nicht exakt quantifizierbaren – Effekthöhen, wobei hierbei K2, K7, K5 und K3 zu unterstreichen sind, als erfolgskritisch zu definieren. Auf Basis dieser Ergebnisse ist im nächsten Schritt zu prüfen, worin die unterschiedliche Performance in diesen Abschnitten begründet liegt, um somit potenziell zeitoptimierende Fahrlinien in diesen kritischen Bereichen zu identifizieren.

Als zusätzliche Randnotiz ist zu erwähnen, dass bei Clustering der Abschnitte nach Streckencharakteristika folgende Resultate zu beobachten sind:

Abbildung 65: Clusteranalyse über den Zusammenhang von Streckencharakteristika vs. Gesamtlaufzeiten

Streckenabschnitt_Char	Fahrzeit_mean	Anteil_Fahrzeit	r pearson [G-Fahrzeit]	r spearman [G-Fahrzeit]
Gerade	18.81599	0.34146	0.98126	0.96441
Kurve	28.45146	0.51631	0.94695	0.92608
Startphase	7.83830	0.14224	0.68058	0.71181

Hiernach hat die Performance auf geraden Streckenabschnitten im Mittel einen stärkeren positiven Zusammenhang als die besonders kurvigen Abschnitte, obwohl der Anteil der Laufzeit auf den Passagen an der Gesamtlaufzeit wesentlich geringer ist. Eine allgemeine inhaltliche Ableitung auf Grundlage dieser Daten erscheint jedoch nicht angebracht, da die einzelnen Cluster unterschiedliche Lokationen und somit – aufgrund des diskutierten kumulativen Effekts – schwer vergleichbare Interdependenzen zueinander aufweisen. Aus diesem Grund wird neben der Ableitung allgemeingültiger Thesen aus der Fundamentalanalyse in den vorangegangenen Kapiteln eine detaillierte Sicht auf einzelne Streckenabschnitte als aussagekräftiger interpretiert.

b) Zusammenhang einzelner Metriken auf Kurvenzeiten

Nach der Aufstellung der Forschungsthesen für zeitoptimierende Fahrstile sowie der Identifikation erfolgskritischer Streckenabschnitte folgt eine detaillierte Untersuchung der Fahrtparameter in den einzelnen Bahnabschnitten, um konkrete Fahrtoptimierungspotenziale aufzudecken. Hierzu nimmt der Rollwinkel als Indikator für die aktiv zu beeinflussende Fahrlinienwahl eine zentrale Bedeutung ein,

da dieser gemäß den bisherigen Erkenntnissen aus Theorie sowie Datenanalyse durch entsprechende Kausalitäten mittelbar und unmittelbar über die Geschwindigkeits- und Beschleunigungsparameter auf die Laufzeit einwirkt. Als Einstieg in die Analyse zeigt Tabelle 26 den Spearman-Rangkorrelationskoeffizienten zur Messung des Zusammenhangs der einzelnen Streckenabschnittszeiten mit den mittleren Rollwinkeln. Durch Nutzung eines Rangkorrelationskoeffizienten wird ein impliziter Vergleich mit der Konkurrenz erreicht, sodass in der Folge konkrete Aussagen darüber getroffen werden können, wie sich im Mittel die erwartete Platzierung innerhalb des Abschnitts bei einer Änderung des Rollwinkels ändert.

Tabelle 26: Rangkorrelationsanalyse über den Zusammenhang Streckenabschnittsplatzierung vs. mittlere Rollwinkel

Abschnitt	Korrelationskoeffizient
	Spearman
Startphase	0,042
K0	-0,039
K1	-0,195
K2	**0,047**
K3	**-0,118**
K4	0,035
K5	**-0,11**
K6	-0,125
K7	**0,151**
K8	-0,198
K9	-0,070
K10 und K11	0,026
K12	-0,071
K13	-0,095
K14	0,019

Die Resultate bekräftigen die Erkenntnisse aus vorangegangen Analysen, indem im Mittel ein leicht negativer Spearman-Koeffizient festzustellen ist: Dies impliziert, dass ein im Vergleich zur Konkurrenz höherer mittlerer Rollwinkel tendenziell die Endplatzierung verbessert.

Im Rahmen der Streckenabschnittsanalyse haben sich die Abschnitte K2, K3, K5 sowie K7 als besonders erfolgskritisch herausgestellt, sodass diese nachfolgend exemplarisch näher untersucht werden. Hierzu wird analog den vorherigen Analysen auf das Instrument des Clusterings zurückgegriffen. Dabei werden die Läufe basierend auf ihren jeweiligen Abschnittszeiten und mittleren Rollwinkeln in den einzelnen Abschnitten in fünf separate Cluster eingeteilt. Hieraus resultiert nachfolgende Häufigkeitsverteilung, anhand deren Ableitungen über die Rollwinkelausprägungen unter Berücksichtigung der Abschnittszeiten getroffen werden können.

Tabelle 27: Clusteranalyse über den Zusammenhang von Abschnittszeiten vs. mittleren Rollwinkeln erfolgskritischer Streckenabschnitte

Streckenabschnitt		Rollwinkel					
		Spearman	0-20%	20-40%	40-60%	60-80%	80-100%
K2	Top 20%	0,047	**44%**	9%	19%	13%	16%
	20-40%		9%	27%	27%	14%	23%
	40-60%		26%	16%	11%	21%	26%
	60-80%		21%	16%	21%	20%	21%
	80-100%		18%	25%	19%	23%	14%
K3	Top 20%	-0,118	16%	16%	27%	22%	20%
	20-40%		17%	17%	8%	33%	25%
	40-60%		26%	34%	20%	9%	11%
	60-80%		23%	22%	23%	14%	17%
	80-100%		**37%**	7%	15%	22%	19%
K5	Top 20%	-0,11	18%	18%	18%	24%	22%
	20-40%		24%	14%	24%	18%	22%
	40-60%		23%	26%	15%	21%	15%
	60-80%		22%	25%	20%	17%	17%
	80-100%		**35%**	12%	24%	18%	12%
K7	Top 20%	0,151	21%	26%	18%	24%	12%
	20-40%		26%	42%	16%	5%	11%
	40-60%		28%	17%	33%	6%	17%
	60-80%		26%	18%	21%	18%	16%
	80-100%		20%	15%	18%	23%	**23%**

Demnach bleibt festzuhalten, dass die Clusteranalyse die Rangkorrelationen nach Spearman für die ausgewählten Streckenabschnitte bestätigen. Für K2 sowie K7 sind hierbei entgegen der grundlegenden Feststellung bessere Laufzeiten mit tendenziell niedrigeren Rollwinkeln zu erreichen. So sind exemplarisch die 20 Prozent geringsten K2-Abschnittszeiten zu 44 Prozent mit den 20 Prozent minimalen mittleren Rollwinkeln gefahren worden, wohingegen nur 16 Prozent die 20 Prozent höchsten mittleren Rollwinkel aufwiesen. Als zusätzliche Information hält

folgende Übersicht die Extrema sowie die statistischen Lageparameter für ausgewählte Gruppen fest, um die relativen Aussagen mit Zahlenwerten zu hinterlegen und somit konkrete praktische Aussagen tätigen zu können.

Tabelle 28: Deskriptive Analyse der beobachteten Rollwinkel in ausgewählten Streckenabschnitten innerhalb definierter Laufzeitcluster (Top-20 & Flop-20 Prozent)

Streckenabschnitt	Rollwinkel									
	mean		max		min		std		mean Rank_Quantil	
	Top 20% (FZ)	Flop 20% (FZ)	Top 20% (FZ)	Flop 20% (FZ)	Top 20% (FZ)	Flop 20% (FZ)	Top 20% (FZ)	Flop 20% (FZ)	Top 20% (FZ)	Flop 20% (FZ)
K2	51,1	**51,5**	53,6	54,9	47,0	47,1	1,5	1,6	**1,5**	1,9
K3	**16,1**	15,5	31,3	19,9	12,7	12,6	2,9	1,9	2,2	**1,8**
K5	**51,0**	49,7	63,4	53,0	47,3	47,0	3,1	1,5	2,2	**1,6**
K7	48,2	**48,5**	50,8	51,6	43,3	43,2	1,5	1,5	**1,8**	2,2

Demnach ist zu erkennen, dass der mittlere Rollwinkel bei den 20 Prozent schnellsten Läufen für K2 bei 51,1° liegt, wohingegen dieser in der Gruppe der 20 Prozent langsamsten Läufen bei 51,5° liegt. Entsprechend weisen auch die Extrema durchweg höhere Werte auf, was wiederum den Schluss aus der Korrelationsanalyse, wonach für K2 tendenziell niedrigere Rollwinkel eine bessere Laufzeit bewirken, bestätigt. Analog hierzu werden durch die Clusteranalyse auch die weiteren Indizien aus der Korrelationsanalyse für die ausgewählten Streckenabschnitte bestätigt: Neben K2 sind auch in Abschnitt K7 niedrigere mittlere Rollwinkel in den dargestellten Werteintervallen zu präferieren, wohingegen für K3 und K5 jene Fahrlinien erfolgsversprechender sind, die im Mittel eine relativ – im Vergleich zur Konkurrenz – weite Fahrlinie verfolgen. Trotz der wertvollen Erkenntnisse aus den bislang umgesetzten Analytics-Methoden bleibt festzuhalten, dass für eine umfassende Nachvollziehbarkeit spezifischer Fahrlinien eine Darstellung der Daten auf minimaler Ebene benötigt wird – dies erfolgt nachfolgend über Visual Analytics: Hierbei werden alle gesammelten Datenpunkte der jeweiligen Streckenabschnitte visualisiert, wodurch ein transparentes und vollständiges Datenabbild der Fahrlinien erreicht wird. Auf dieser Grundlage ist ein zielgerichtetes Benchmarking für einzelne Streckenabschnitte möglich, was wiederum direkte praktische Implikationen zur umzusetzenden Fahrlinienwahl liefert.

Abschließend ist eine Zusammenfassung der bisherigen Thesen und Erkenntnisse in Tabelle 29 festgehalten.

Tabelle 29: Thesenimplikationen aus Data Analytics

Thesen	Beurteilung auf Basis Data Analytics	Anmerkung
Ein schneller Start ist fundamental für eine gute Gesamtzeit.	✓	bekräftigt durch Korrelations- sowie Clusteranalyse: mittelbar sowie unmittelbarer positiver Einfluss
Ein geringer Rollwinkel und eine hohe Vertikalbeschleunigung sind vorteilhaft.	✗	tendenziell widerlegt durch Rangkorrelations- sowie Clusteranalyse: im Mittel sind im Vergleich höhere Rollwinkel vorteilhaft (aber: abschnittsspezifisch)

6.5.2 Modelling II – Visual Analytics via PowerBI

6.5.2.1 Einführung in Dashboards und Power BI

Das menschliche Gehirn kann Informationen durch graphische Visualisierung besser verarbeiten als durch rein textuelle oder numerische Eingaben. Dies wird als Picture Superiority Effect bezeichnet (Hockley, 2008, S. 1351). Daher wird in dieser Arbeit der Fokus zusätzlich auf die graphische Analyse und Darstellung der Ergebnisse gelegt. Damit befinden wir uns in der Phase des Modelling entsprechend des CRISP-DM. Hierzu wird die Self-Service Software Microsoft Power BI verwendet.

Self-Service Business Intelligence (BI) Lösungen ermöglichen Nutzern einfache Analysen, ohne die statistisch-mathematischen Hintergründe vollumfänglich beherrschen zu müssen. Sie bieten darüber hinaus nutzerfreundliche Interfaces und liefern verschiedene Darstellungsoptionen je nach Anspruchsgruppe. Somit können die Ergebnisse der Datenauswertung jeder Nutzerin bzw. jedem Nutzer unabhängig von deren bzw. dessen Vorwissen zur Verfügung gestellt und leicht verständlich präsentiert werden. Dashboards liefern einen strukturierten Zugang zu großen Datenmengen und bilden diese zusammengefasst ab. Ein vollständiger Power BI Report wird als ein solches Dashboard bezeichnet.

Die Wahl der genutzten Anwendung (Power BI) erfolgte durch die vorhandene Integration in das gängige MS Office Paket, und damit einhergehend die weltweite Abrufbarkeit ohne zusätzliche Lizenzkosten einerseits und die Möglichkeit zur Anbindung jeglicher Datenquellen andererseits. Dies ermöglichte die einfache Veröffentlichung sowie Zugänglichkeit der Dashboards, aber auch der zugrundeliegenden Daten über die Downloadfunktionen. Darüber hinaus können interaktive Dashboards erstellt werden, die dem Nutzer mit simplen Klicks Zugang zu verschiedenen Perspektiven auf die Daten bieten. Diese lassen sich dann zur weiteren Bearbeitung oder Weiterleitung in gängige Formate wie .csv, Excel- und PowerPoint oder Portable Document Format (PDF) exportieren.

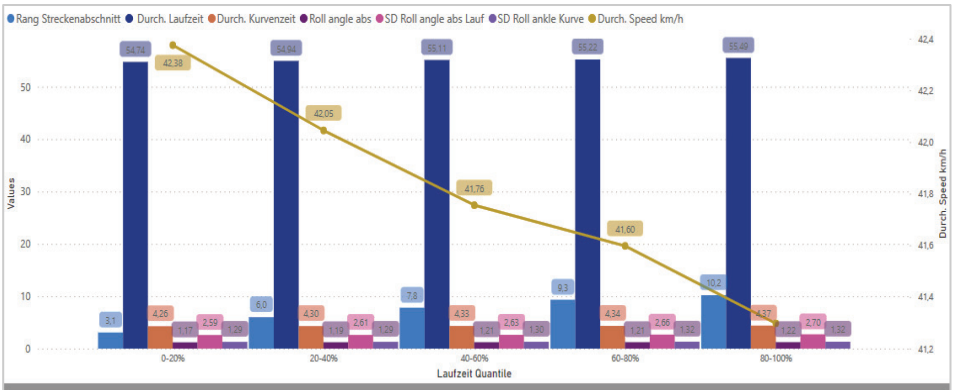
6.5.2.2 Analyse der Gesamtfahrt durch Visualisierung mittels Power BI

Die so gewonnenen Erkenntnisse werden ergänzend zu der im vorigen Abschnitt dargelegten Analyse gewonnen und im Folgenden präsentiert. Dazu basiert das Power BI Modell ebenfalls auf dem bereinigten Datensatz. An dieser Stelle sei zu vermerken, dass die Quantileinteilung der Laufzeiten beibehalten wurde, jedoch keine Rollwinkeleinteilung in Quantile anhand der Häufigkeiten vorgenommen wurde. Daher können sich Durchschnitts- und Standardabweichungswerte aufgrund der leicht veränderten Gesamtheit vom vorherigen Abschnitt leicht unterscheiden. Dies hat jedoch keine Auswirkungen auf die Ergebnisse und den Zusammenhang beider Textabschnitte. Ferner wurde der bestehende Datensatz um zusätzliche Metriken angereichert, um eine umfassende grafische Analyse gewährleisten zu können.

Durch das intuitive Filtern lassen sich zusätzliche Erkenntnisse gewinnen, welche die bereits erwähnten Beobachtungen detaillierter erfassen und zusätzliche Erkenntnisse generieren.

Zunächst werden die Startphase und der Start näher betrachtet. Hier werden die Ergebnisse der ersten Untersuchung bestätigt, wonach das Quantil der besten Laufzeiten bereits beim Start zur besten Laufzeit führen kann und dabei auch die höchste Durchschnittsgeschwindigkeit erzielt. Darüber hinaus zeigt sich im obersten Quantil sowohl der niedrigste durchschnittliche Rollwinkel als auch die niedrigste Standardabweichung des Rollwinkels innerhalb der Phase, was für eine sowohl schnellere, als auch geradere Fahrweise am Start spricht.

Abbildung 66: Ergebnisse Startphase



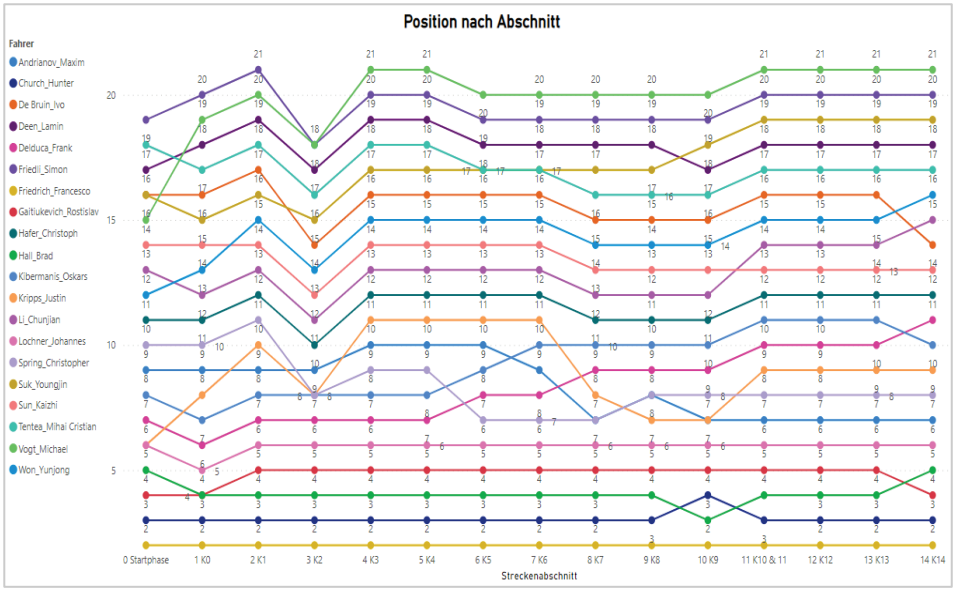
Laufzeit Quantile	Rang Streckenabschnitt	Durch. Laufzeit	Durch. Kurvenzeit	Roll angle abs	SD Roll angle abs Lauf	SD Roll ankle Kurve	Durch. Speed km/h
0-20%	3,05	54,738	4,26	1,17	2,59	1,29	42,38
20-40%	5,99	54,940	4,30	1,19	2,61	1,29	42,05
40-60%	7,78	55,112	4,33	1,21	2,63	1,30	41,76
60-80%	9,30	55,220	4,34	1,21	2,66	1,32	41,60
80-100%	10,20	55,491	4,37	1,22	2,70	1,32	41,30

Abweichungen einzelner Tage ergaben sich am 9. Januar, als das oberste Quantil weiterhin die niedrigste Standardabweichung des Rollwinkels zeigt, jedoch das unterste Quantil die zweitniedrigste. Darüber hinaus fuhren in den Wettbewerben am 3. und 4. Januar 2020 sowie am 12. Dezember 2021 die Fahrer des obersten Quantils im Mittel einen höheren Rollwinkel als die restlichen Fahrer der Konkurrenz, bei der weiterhin höchsten Geschwindigkeit. Zusätzlich ist der durchschnittliche Rang innerhalb der Startphase des obersten Quantils an diesen Tagen höher als an den weiteren Wettkampftagen des Datensatzes, was eine Indikation dafür sein kann, dass manche Fahrer an diesen Tagen trotz eines schwächeren Starts einen der Top-Ränge einfuhren und auch für die höheren Rollwinkel verantwortlich sein könnten. Dies steht im Einklang mit den im vorigen Abschnitt gemachten Beobachtungen zum Start.

Zusammengefasst lässt sich für den Start festhalten, dass ein niedriger Rollwinkel unter Einbezug einer niedrigen Standardabweichung als optimale Fahrweise angesehen werden kann. Zusätzlich ist der Start elementar für den Geschwindigkeitsaufbau in den folgenden Streckenabschnitten und damit für die gesamte Laufzeit. Zur Verdeutlichung dient folgende Darstellung über die Rangentwicklung innerhalb eines ausgewählten Laufes, wonach sich die Endplatzierungen nach Start nicht mehr groß geändert haben. Es gilt bei der Beurteilung der Wich-

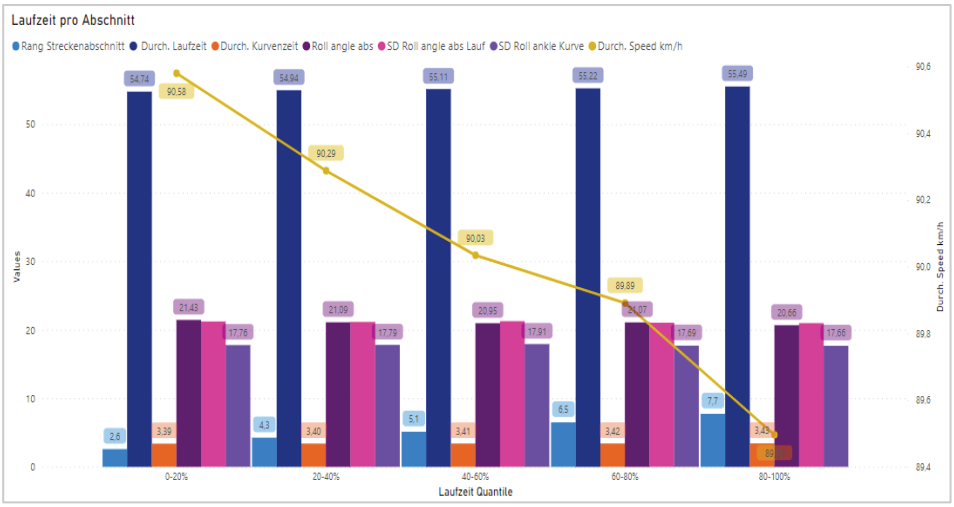
tigkeit der Startphase zusätzlich zu berücksichtigen, dass die Zeitdifferenzen unmittelbar nach dem Start noch sehr gering sind – geringfügige Optimierungen können demnach bereits große Verbesserungen in der Endplatzierung herbeiführen.

Abbildung 67: Positionsentwicklung innerhalb eines Laufes



Bevor die Einzelfahrten detaillierter beleuchtet werden, wird zunächst das Fahrverhalten auf den Geraden und in den Kurven analysiert. Unter Betrachtung aller Geraden als Einheit stellt sich gegenüber der Startphase ein etwas differenzierteres Bild dar. Hier weisen die Fahrten des obersten Quantils den größten Rollwinkel auf, die niedrigste Standardabweichung jedoch die beiden untersten Quantile. Weiterhin zeigen auf den Geraden im Mittel die Quantile eine Geschwindigkeit absteigend vom obersten bis zum untersten Quantil mit einer Durchschnittsgeschwindigkeit von 90,58 km/h im obersten sowie 89,50 km/h im untersten Quantil. Abbildung 68 bietet einen Überblick über die verschiedenen Metriken. Erwähnenswert ist hierbei, dass der in den Diagrammen genannte Begriff „Kurvenzeit“ die Zeit des jeweiligen Streckenabschnitts wiedergeben und somit auch für Geraden gilt.

Abbildung 68: Ergebnisse Startphase

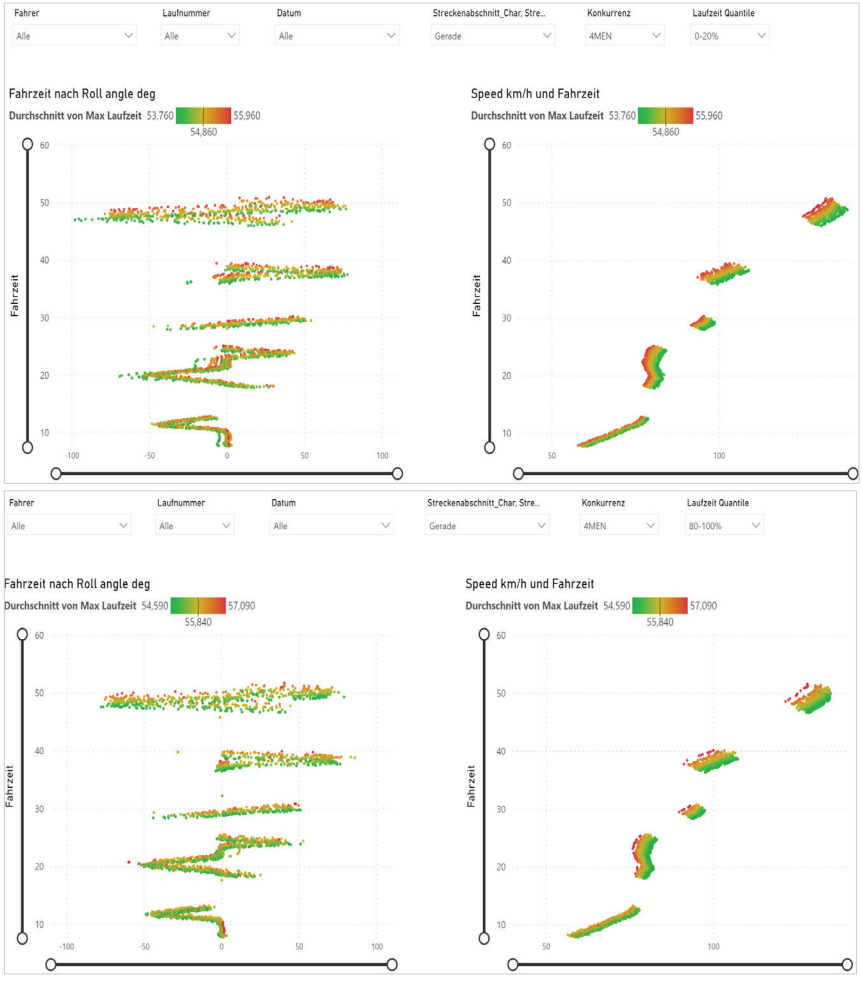


Der Vergleich des obersten mit dem untersten Quantil auf den Geraden bringt die Auffälligkeit hervor, dass die Fahrer des untersten Quantils die Geraden mit positivem Rollwinkel unter einem fast gleichen Rollwinkel fahren wie die Fahrer des ersten Quantils, während diese die Geraden unter negativem Rollwinkel mit einem höheren absoluten Rollwinkel fahren. Hohe Rollwinkel auf geraden Streckenabschnitten können beispielsweise bei dem Übergang von einer Kurve in die Gerade realisiert werden. Es ist zu beachten, dass sich Extremwerte immer auf den gesamten untersuchten Streckenabschnitt beziehen. So beträgt das Maximum des Rollwinkels im obersten Quantil 77,78°, während im dritten Quantil das maximale bei 89,59° und im untersten Quantil bei 85,95° liegt. Hingegen stellt im negativen Bereich mit -97,70° das oberste Quantil den höchsten Betrag. Folglich findet sich auch das durchschnittliche Maximum der absoluten Beträge und der höchste Mittelwert im obersten Quantil, was sich der Tabelle 30 entnehmen lässt. Somit lässt sich als bevorzugte Fahrweise für gerade Streckenabschnitte ein höherer absoluter Rollwinkel unter einer niedrigeren Standardabweichung festhalten.

Tabelle 30: Metriken zum Fahrverhalten der Quantile auf den Geraden

Streckenabschnitt_Char Laufzeit Quantile	Gerade Av Fahrzeit	Av Laufzeit Quantil	Av Kurvenzeit Quantil	Roll angle abs	SD Roll angle abs	SD Roll angle Kurve	Av Max Roll angle Kurve	Minimum von Roll angle deg	Maximum von Roll angle deg	Speed km/h
0-20%	25,79	54,74	3,40	21,43	21,21	17,76	51,19	-97,70	77,78	90,58
20-40%	25,91	54,94	3,41	21,09	21,14	17,79	51,00	-83,56	84,42	90,29
40-60%	26,00	55,11	3,41	20,95	21,26	17,91	51,08	-80,04	89,59	90,03
60-80%	26,06	55,22	3,42	21,07	21,03	17,69	50,58	-81,51	84,99	89,89
80-100%	26,19	55,49	3,44	20,66	20,97	17,66	50,71	-77,54	85,95	89,50
Gesamt	**26,00**	**55,11**	**3,42**	**21,03**	**21,12**	**17,76**	**50,91**	**-97,70**	**89,59**	**90,04**

Abbildung 69: Rollwinkel und Geschwindigkeit im Verhältnis zur Laufzeit



Das obere Diagramm stellt mittels Punktwolken die Datenpunkte der Rollwinkel entlang des Fahrverlaufs dar und visualisiert hiermit implizit die Bandbreite des

gewählten Rollwinkels und damit die Fahrlinie. Diese Linksverschiebung der Grafik stellt einen interessanten Fakt dar, der auf unterschiedliche Fahrstile hindeutet. Dies ist besonders auf der letzten Geraden zu beobachten, welche eine charakteristische leichte Links- und Rechtsverschiebung beinhaltet. Hier sind die Ausprägungen im obersten Quantil stärker, ebenfalls lässt sich innerhalb beider Quantile feststellen, dass die schnellsten Fahrer (grüne Punkte) diese Links-rechts-Kombination unter höherem Rollwinkel fahren als die etwas langsameren Fahrer des Quantils (rote Punkte).

Ob sich diese Beobachtung auf alle Geraden verteilt oder nur durch spezifische Abschnitte bedingt wird, wird nachfolgend untersucht. Als erste Indikation dient die vorangestellte Grafik über die Verhältnisse von Rollwinkel und Geschwindigkeit zur Fahrtzeit des obersten und untersten Quantils.

Dass die Unterschiede auf den Geraden des hinteren Streckenabschnitts größer ausfallen als die der vorderen Abschnitte, deutet auf ein ebenfalls verändertes Fahrverhalten in den Kurven hin. Dies macht eine sich anschließende Untersuchung der Fahrlinien in den Kurven sowie deren Auswirkung auf die Geschwindigkeit der anschließenden Geraden erforderlich. Da eingangs festgestellt wurde, dass die Streckenabschnitte zwar individuell betrachtet, jedoch immer im Kontext der vorangehenden und sich anschließenden Streckenabschnitte analysiert werden müssen, ist diese Anforderung zwingend.

6.5.2.3 Detaillierte Analyse der Streckenabschnitte

Für den ersten Streckenabschnitt K1 lässt sich mit dem höchsten durchschnittlichen Rollwinkel im obersten Quantil ein ähnliches Bild festhalten. Der sich anschließende Streckenabschnitt K2 hat auf Basis der eingangs berechneten Korrelationen einen signifikanten Einfluss auf die Endplatzierung. Der durchschnittliche absolute Rollwinkel besitzt den drittgrößten Wert der Quantile und liegt somit im mittleren Bereich. Die Standardabweichung besitzt den niedrigsten Wert der Konkurrenz. Der Durchschnitt des maximalen Rollwinkels liegt für das oberste Quantil an vierter Stelle der Quantile und ist niedriger als im untersten Quantil, während das absolute Maximum des Rollwinkels im obersten Quantil liegt, was durch folgende Grafiken verdeutlicht wird.

Hier liegt ein Ausreißer in den Fahrten von Alexey Stulnev begründet, welcher auf den Metern 240 bis 280 der Kurve K2 einen großen Rollwinkel fuhr. An dieser Stelle lässt sich sehr gut die Stärke von grafischen Analysen erkennen, da zum einen die Ausreißer direkt erkennbar sind und zum anderen durch das interaktive

Design des Dashboards sich auch direkt der dafür verantwortliche Lauf zeigt. In der zweiten Grafik lässt sich ein sehr individueller Fahrstil beider Piloten an dem spezifischen Tag erkennen.

Abbildung 70: Rollwinkel nach Distanz oberstes und unterstes Laufzeitquantil

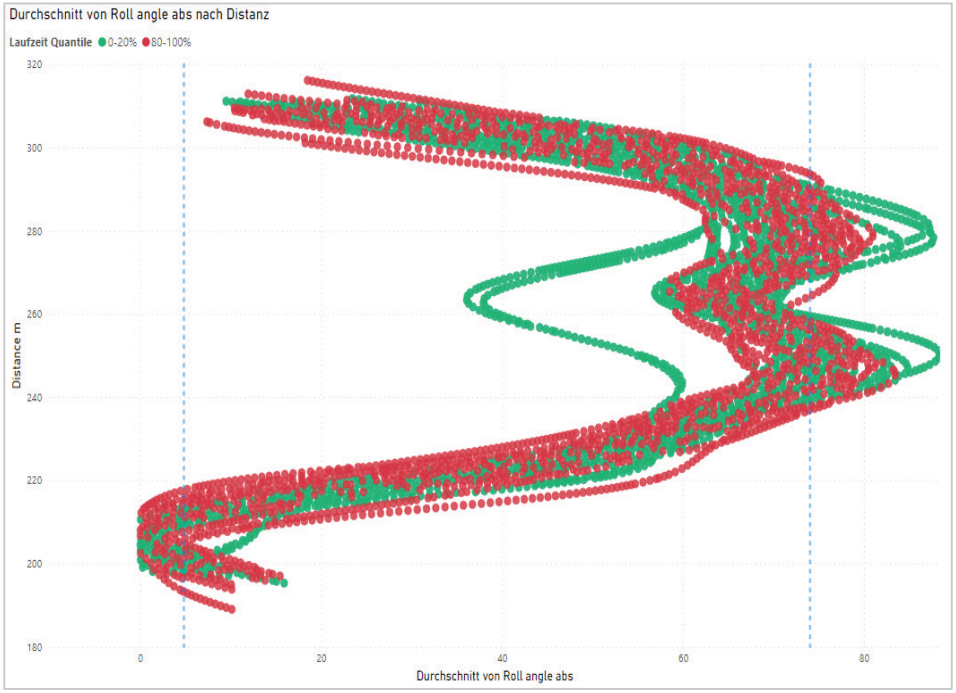
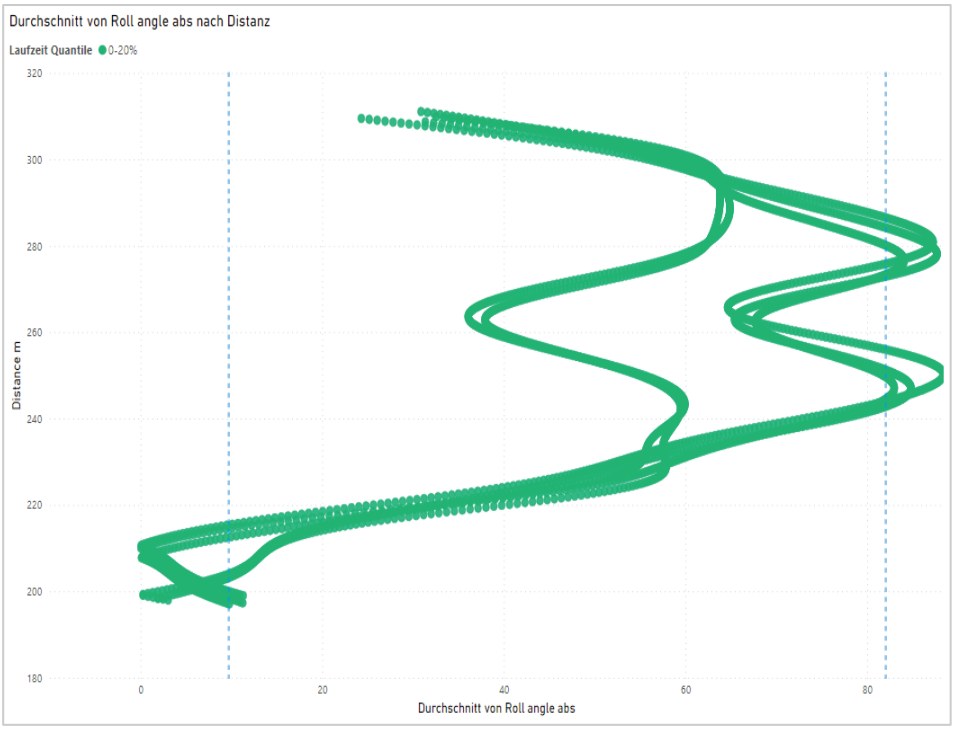


Abbildung 71: Rollwinkel nach Distanz oberstes und unterstes Laufzeitquantil



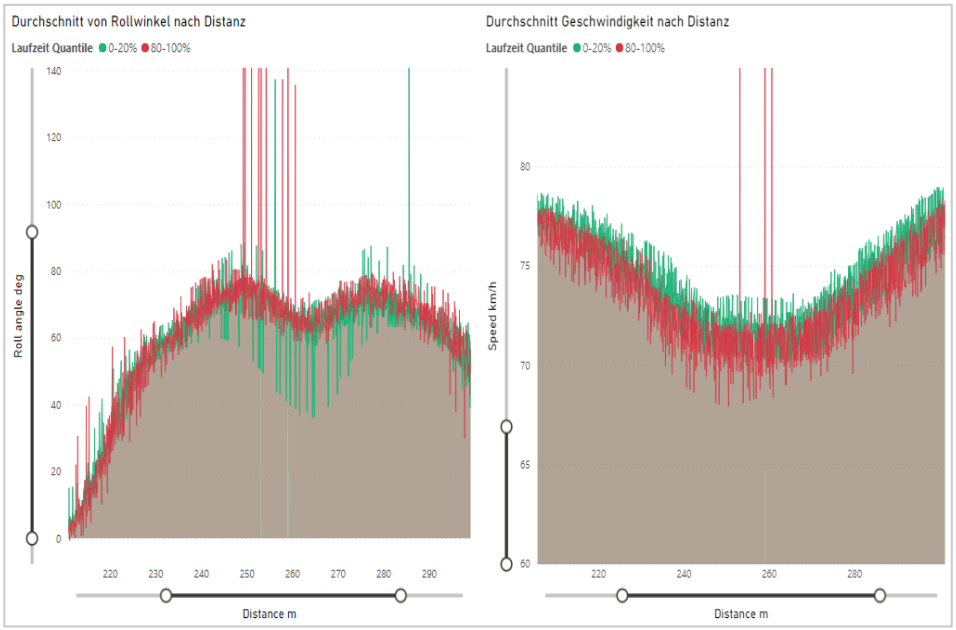
Zum Ausgang von K2 lassen sich folgende Metriken festhalten.

Tabelle 31: Metriken zum Fahrverhalten in den Quantilen in Abschnitt K2

Streckenabschnitt_Char Laufzeit Quantile	Kurve Av Fahrzeit	Av Laufzeit Quantil	Av Kurvenzeit Quantil	Roll angle abs	SD Roll angle abs	SD Roll angle Kurve	Av Max Roll angle Kurve	Minimum von Roll angle deg	Maximum von Roll angle deg	Speed km/h
0-20%	17,99	54,73	5,37	20,06	5,77	24,34	75,42	9,50	34,10	79,19
20-40%	18,10	54,94	5,39	20,65	5,53	24,73	75,94	5,07	32,70	78,93
40-60%	18,17	55,11	5,41	19,36	5,42	24,87	76,24	9,46	34,48	78,67
60-80%	18,22	55,22	5,42	20,14	6,23	24,51	75,11	9,79	34,70	78,58
80-100%	18,31	55,49	5,44	18,93	6,31	25,06	76,33	7,37	40,39	78,30
Gesamt	**18,16**	**55,11**	**5,41**	**19,80**	**5,91**	**24,71**	**75,82**	**5,07**	**40,39**	**78,73**

Der durchschnittliche absolute Rollwinkel sowie das durchschnittliche Maximum des Rollwinkels weisen zum Ausgang der Kurve niedrigere Werte auf. Im Vergleich zur Gesamtkonkurrenz liegen diese jedoch noch über dem letzten Quantil. Folgende Analyse der Einzelabschnitte der Kurve liefern eine Erklärung für diese Verhältnisse.

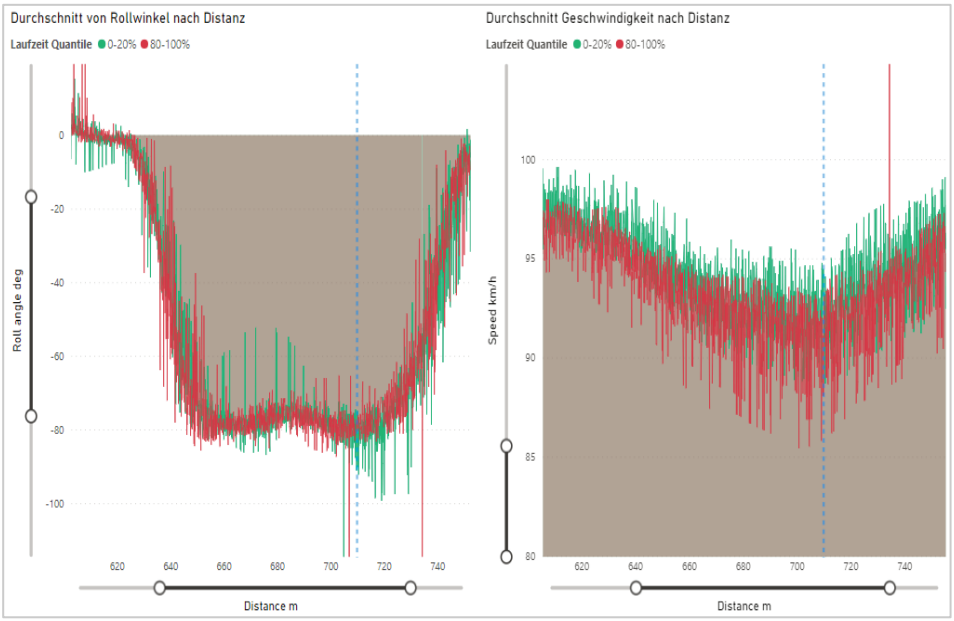
Abbildung 72: Rollwinkel und Geschwindigkeit im Verhältnis zur zurückgelegten Strecke in Abschnitt K2



Die graphische Analyse im kritischen Kurvenbereich um den Scheitelpunkt der Kurve, an dem die Geschwindigkeit abnimmt, ergibt eine stärkere Ausprägung der niedrigeren Rollwinkel des obersten Quantils gegenüber dem unteren. Weiterhin bleibt die Geschwindigkeit des obersten Quantils über der des unteren und dreht im Verlauf der Kurve von Brems- zu Beschleunigungsgeschwindigkeit. Dies deckt sich mit der eingangs festgestellten Beobachtung der größeren Häufigkeiten von niedrigeren Rollwinkeln im obersten Quantil. Diese Beobachtung lässt den Schluss zu, dass die niedrigeren Rollwinkel in den Kurven unter einer Geschwindigkeitsreduktion vorteilhaft sind. Im Beschleunigungsbereich der Kurve steigt der Rollwinkel des obersten Quantils schneller als im untersten.

Um diese These weiter zu untersuchen, wird kurz auf die Kurve K7 vorgegriffen, welche ebenfalls einen signifikanten Einfluss besitzt und im Verlauf um den Scheitelpunkt eine Geschwindigkeitsabnahme aufweist. Die folgende grafische Darstellung kann die vorher getroffene Beobachtung bestätigen. Auch hier sind insbesondere im Bereich der Geschwindigkeitsabnahme die Rollwinkel des obersten Quantils tendenziell geringer als im untersten Quantil.

Abbildung 73: Rollwinkel und Geschwindigkeit im Verhältnis zur zurückgelegten Strecke in Abschnitt K7



Diese Beobachtungen führen zu dem Zwischenfazit, dass in den Abschnitten von Kurven unter Bremswirkung ein niedrigerer Rollwinkel als vorteilhaftere Fahrweise anzusehen ist, während in Beschleunigungsabschnitten ein höherer zu empfehlen ist. Dies liefert einen wichtigen Beitrag für die Untersuchung einzelner Abschnitte und erklärt auch die positive Korrelation der beiden Kurven zur Gesamtlaufzeit, da in diesen die Geschwindigkeit abnimmt, was sich negativ auf die Laufzeit auswirkt.

Da die Stichprobenauswahl aus diesen zwei Kurven zu gering ist, wäre an dieser Stelle eine tiefgreifende statistische Analyse über eine größere Stichprobe sehr interessant.

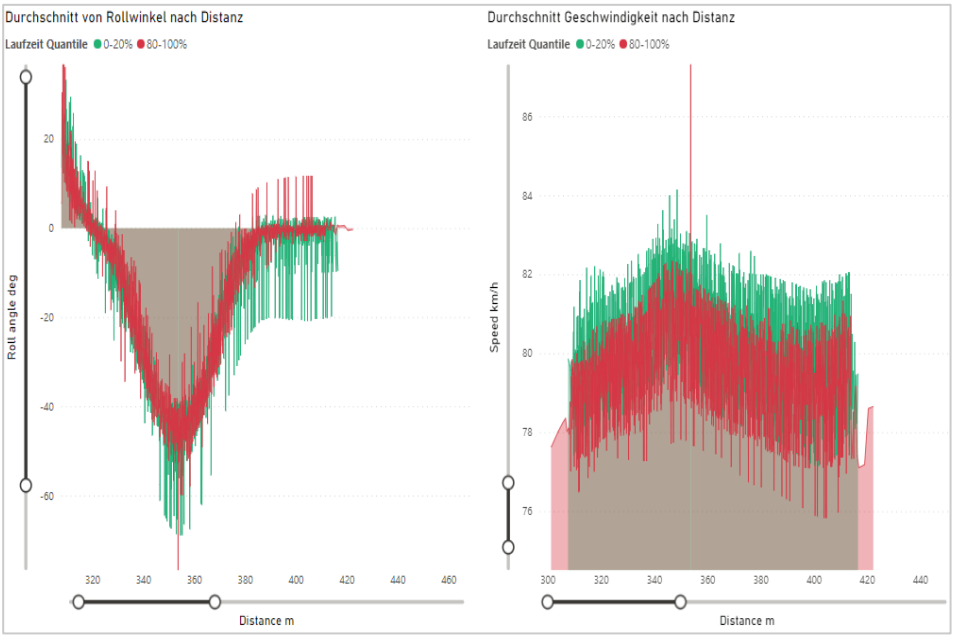
Für die aufeinanderfolgenden Abschnitte K3 und K4 lässt sich die interessante Beobachtung machen: In K3 weist das oberste Quantil mit 16,74° im Mittel einen deutlich höheren Rollwinkel auf und liegt damit über 1° über dem Mittelwert des Benchmarks.

Abbildung 74: Metriken des Abschnitts K3

Streckenabschnitt_Char Laufzeit Quantile	Gerade Av Fahrzeit	Av Laufzeit Quantil	Av Kurvenzeit Quantil	Roll angle abs	SD Roll angle abs	SD Roll angle Kurve	Av Max Roll angle Kurve	Minimum von Roll angle deg	Maximum von Roll angle deg	Speed km/h
0-20%	20,35	54,74	4,70	16,74	16,80	16,32	47,65	-68,88	33,18	80,48
20-40%	20,46	54,94	4,71	15,39	15,83	15,64	44,94	-59,83	31,72	80,23
40-60%	20,53	55,11	4,72	15,29	16,15	16,01	45,55	-54,16	33,82	79,97
60-80%	20,59	55,22	4,73	15,75	15,94	15,70	44,61	-57,20	33,73	79,87
80-100%	20,69	55,49	4,75	15,27	15,94	15,77	45,06	-59,89	39,57	79,53
Gesamt	**20,53**	**55,11**	**4,72**	**15,68**	**16,15**	**15,89**	**45,61**	**-68,88**	**39,57**	**80,00**

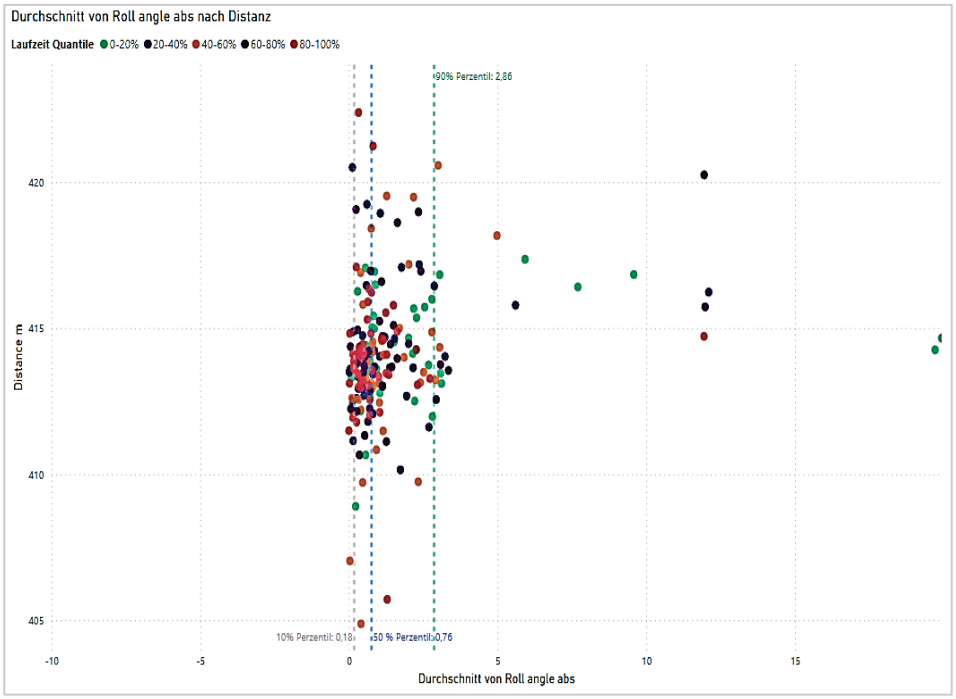
Fließt nun der Fakt mit ein, dass zum Eintritt in Kurve K3 aus der Kurve K2 heraus der mittlere Rollwinkel des obersten Quantils mit 20,06° lediglich der dritthöchste ist und am Ende der Kurve K3 mit 2,59° das Maximum als Doppel des Mittels der weiteren Quantile beträgt, ergibt sich eine erste Feststellung: In Kombination mit dem mittleren Maximum von 47,85° in K3 ergibt sich ein hoher Rollwinkel für die leichte Rechtskurve innerhalb von K3 als indikativ beste Fahrlinie. Folgende Grafik bestätigt diese Indikation und lässt erkennen, dass hier ein höherer Rollwinkel zur Beschleunigung zu wählen ist, was sich mit den vorigen Beobachtungen zu diesem Abschnitt deckt. Dies lässt den Schluss zu, dass die getroffene Beobachtung zum Fahrverhalten in K2 um den Punkt ergänzt werden kann, dass die Wahl eines höheren Rollwinkels zum Ausgang der Kurve und in der anschließenden Geraden eine zu präferierende Fahrlinie darstellt.

Abbildung 75: Rollwinkel und Geschwindigkeit im Verhältnis zur zurückgelegten Strecke in Abschnitt K3



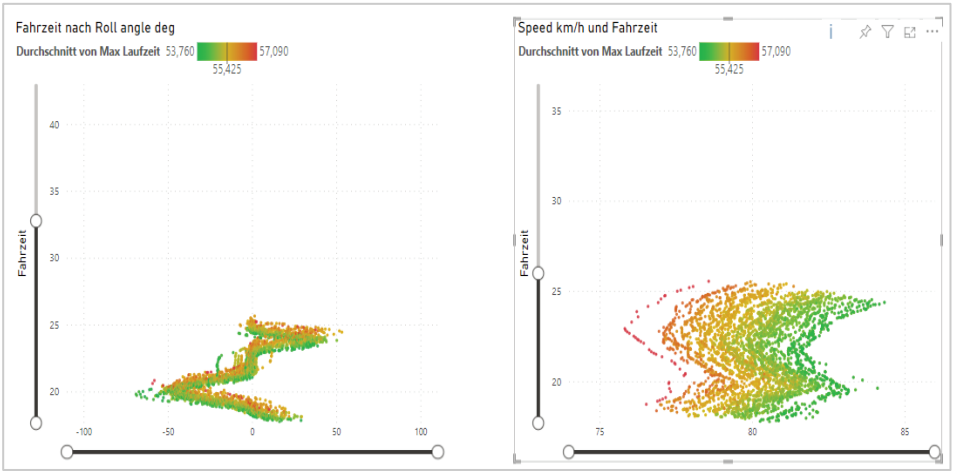
Zum Ausgang aus der Kurve K3 am Messpunkt B15 besteht im Mittel ein Vorsprung von 0,12 Sekunden, ausgehend von 0,06 Sekunden nach dem Start. Nachstehende Punktwolke stellt die Rollwinkelverteilung an der Lichtschranke B15 dar.

Abbildung 76: Rollwinkel in den Quantilen im Verhältnis zur zurückgelegten Strecke am Messpunkt B15



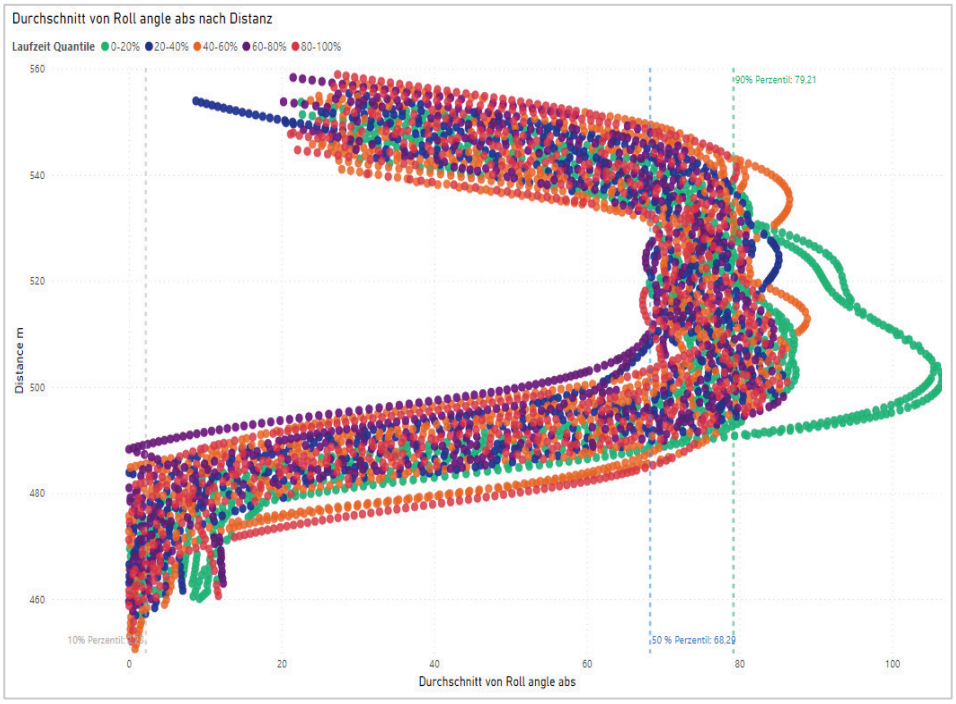
Im Laufe der K4 fahren die Bobs des obersten Quantils den niedrigsten absoluten Rollwinkel und mit 12,65° eine Standardabweichung von einem Grad weniger als im Wettbewerb, um dann zum Ausgang der Kurve K4 mit 2,14° deutlich über dem Mittelwert zu liegen. Dass sich zu diesem Zeitpunkt der mittlere Vorsprung auf 0,13 Sekunden erhöht hat, obwohl an diesem Punkt die inkrementelle Geschwindigkeitszunahme seit dem Start die niedrigste von allen Quantilen ist, hebt die Bedeutung des Starts hervor. Dies deutet einen Fahrstil an, welcher den Geschwindigkeitsvorteil aus dem Start hält und nicht darauf ausgelegt ist, in diesem Streckenabschnitt die größtmögliche Beschleunigung zu erzielen. Dies wird in folgender Grafik verdeutlicht, welche anzeigt, dass die Geschwindigkeiten des obersten Quantils (grün markiert) aufgrund der höheren Eintrittsgeschwindigkeit, verschoben auf der X-Achse, parallel zu den Entwicklungen des untersten Quantils (rot markiert) verlaufen.

Abbildung 77: Rollwinkel und Geschwindigkeit im Verhältnis zur Laufzeit, oberstes und unterstes Quantil in Abschnitt K4



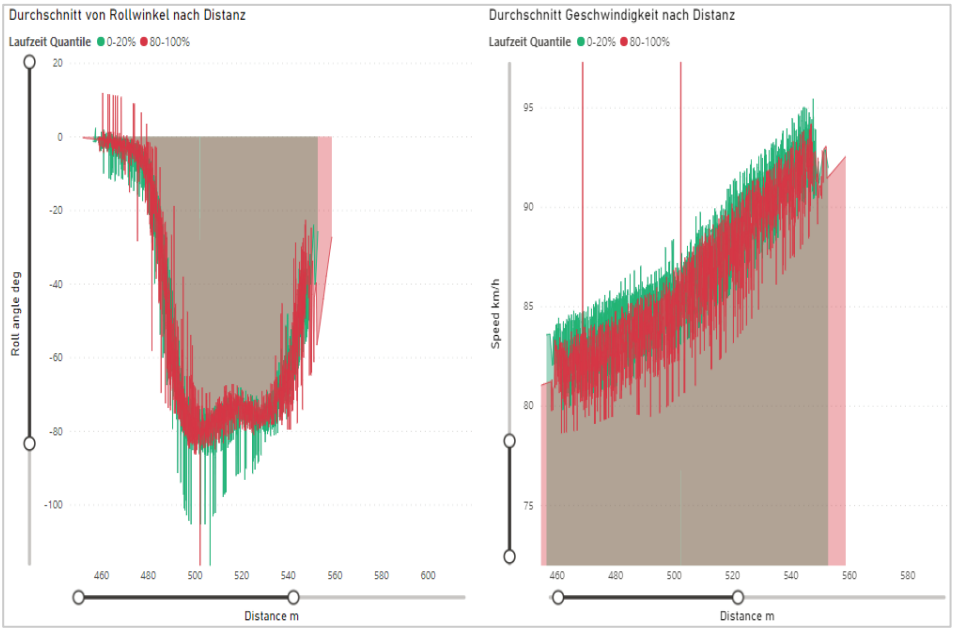
Innerhalb von Kurve K5 bleiben der Abstand der Gesamtzeit und der Geschwindigkeit im Mittel identisch. Der Rollwinkel des obersten Quantils ist im absoluten Mittel der höchste. Auffälligkeiten liefern hier das 40-60 Prozent Quantil mit einem durchschnittlichen Rollwinkel von ca. 2° unter dem Mittel der weiteren Quantile und einem maximalen Rollwinkel von 5° unter dem Mittelwert, sowie das 4. Quantil mit einem Rollwinkel von 8° über dem Mittel der maximalen Werte der Konkurrenz.

Abbildung 78: Rollwinkel im Verhältnis zur zurückgelegten Strecke in Abschnitt K5. mittels Power BI



Da Kurve K5 eine Beschleunigungskurve mit zunehmender Geschwindigkeit ist, lassen die bisher getroffenen Beobachtungen die Vermutung zu, dass im Verlauf der Kurve und um den Scheitelpunkt im Besonderen, ein höherer Rollwinkel vorteilhaft sei. Dies kann durch nachfolgende Grafik erneut belegt werden.

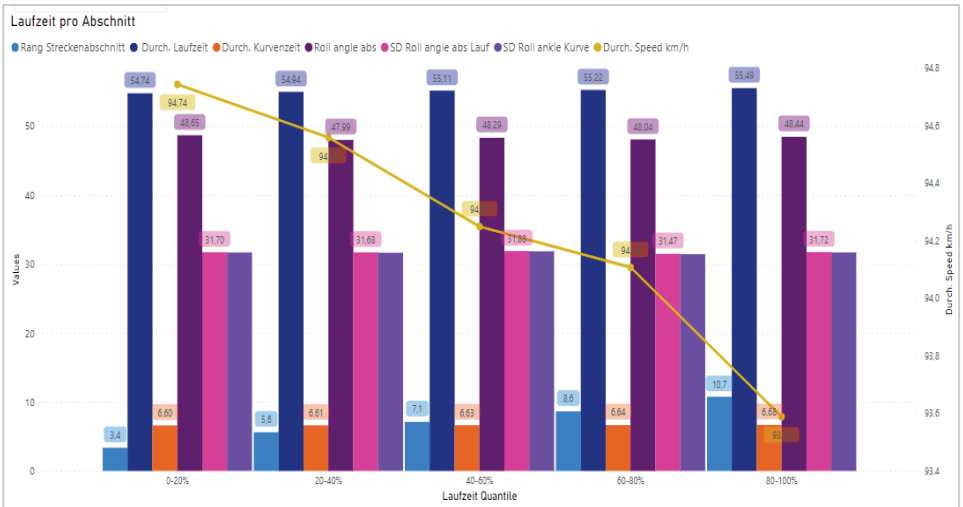
Abbildung 79: Rollwinkel und Geschwindigkeit im Verhältnis zur zurückgelegten Strecke in Abschnitt K5



Im Verlauf der Kurve K6 verläuft der Rollwinkel ins Negative und zur Einfahrt in den Veltins-Kreisel ins Positive mit einem mittleren Rollwinkel über alle Quantile von 44,23°. Der Rollwinkel des ersten Quantils liegt an dieser Stelle unter dem Durchschnitt, jedoch abermals mit der niedrigsten Standardabweichung. Der mittlere Vorsprung erhöht sich um eine hundertstel Sekunde und auch die Differenz der Geschwindigkeit erhöht sich in diesem Abschnitt um 2 km/h.

Der sich anschließende Veltins-Kreisel stellt einen der schwierigsten und aus Analysesicht interessantesten Streckenabschnitte dar. Diese Kurve stellt den ersten Streckenabschnitt mit einem Geschwindigkeitsverlust im Mittel aller Quantile dar. Ebenfalls lässt sich festhalten, dass hier das letzte Quantil stärker abfällt mit einem Zeitverlust von 0,04 Sekunden auf das vorletzte Quantil und einem Geschwindigkeitsverlust von 0,42 km/h zur Einfahrt in die Kurve. Beides sind in diesem Abschnitt die negativen Höchstwerte. Ob Einzelfahrten als Ausreißer dieses Ergebnis beeinflussen, wird im weiteren Verlauf dieser Arbeit erläutert.

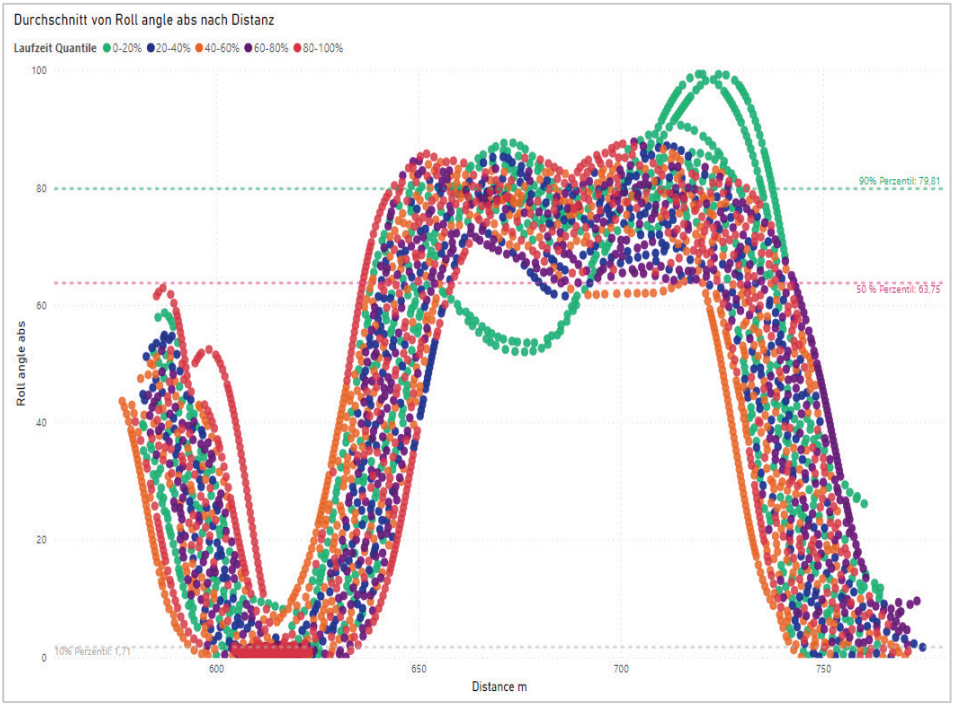
Abbildung 80: Metriken des Abschnitts K7



Streckenabschnitt_Char Laufzeit Quantile	Kurve Av Fahrzeit	Av Laufzeit Quantil	Av Kurvenzeit Quantil	Roll angle abs	SD Roll angle abs	SD Roll angle Kurve	Av Max Roll angle Kurve	Minimum von Roll angle deg	Maximum von Roll angle deg	Speed km/h
0-20%	33,08	54,74	6,61	48,65	31,70	31,66	83,06	-99,35	58,66	94,74
20-40%	33,22	54,94	6,63	47,99	31,68	31,63	81,49	-87,14	55,09	94,56
40-60%	33,34	55,11	6,65	48,29	31,88	31,84	82,11	-86,58	52,74	94,25
60-80%	33,41	55,22	6,66	48,04	31,47	31,41	81,23	-87,79	53,75	94,11
80-100%	33,57	55,49	6,69	48,44	31,72	31,67	82,08	-87,49	62,94	93,59
Gesamt	**33,33**	**55,11**	**6,65**	**48,29**	**31,69**	**31,64**	**82,01**	**-99,35**	**62,94**	**94,23**

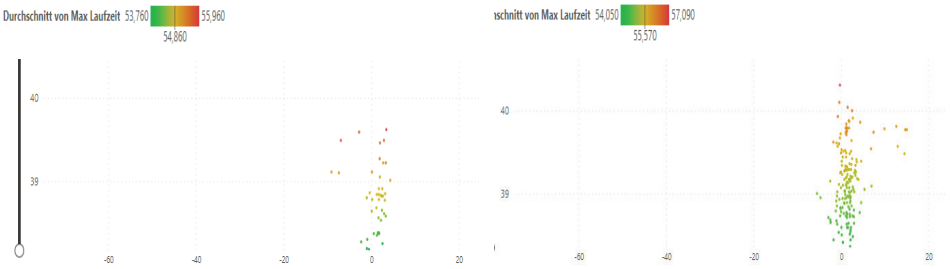
Weiterhin ist zu konstatieren, dass das oberste Quantil einen etwas stärkeren Geschwindigkeitsverlust als die nächste Konkurrenz zulässt und einen weitaus größeren Rollwinkel fährt. An dieser Stelle sei Francesco Friedrich hervorgehoben, für den in der eingangs erwähnten Fahrt am 11.12.2021 ein extremer Rollwinkel an dieser Stelle registriert wird. Auch bei Ausklammerung dieses Ausreißers bleibt der Trend bestehen. Auf die Besonderheiten der Fahrlinie von Francesco Friedrich wird im Anschluss gesondert eingegangen. Unter Betrachtung der folgenden Rollwinkelanalyse mit der zurückgelegten Distanz auf der X-Achse und dem Rollwinkel auf der Y-Achse lässt sich konstatieren, dass die Fahrer des obersten Quantils den Rollwinkel innerhalb des Veltins-Kreisels im Mittel halten und eine geringere Streuung aufweisen als die anderen Quantile. Dies spricht an dieser Stelle für die Vorteilhaftigkeit einer flüssigeren, geradlinigeren Fahrlinie.

Abbildung 81: Rollwinkel und Geschwindigkeit im Verhältnis zur zurückgelegten Strecke in Abschnitt K7



Innerhalb der sich an den Veltins-Kreisel anschließenden Kurve K8 zeigt sich, dass die Fahrlinie des ersten Quantils innerhalb der K7 erfolgreich war, da dieses Quantil mit 9,82 km/h nun den höchsten Geschwindigkeitszuwachs aufzeigt. Darüber hinaus hat sich der Vorsprung im Mittel auf 0,16 Sekunden erhöht. Als Ausgleich des hohen Rollwinkels der vorangegangenen Kurve stellt die Fahrlinie des obersten Quantils auf dieser Geraden einen niedrigen Rollwinkel im Vergleich dar, wieder in Kombination mit der niedrigsten Standardabweichung. Die Ausfahrt dieser Kurve nehmen die Fahrer des obersten Quantils unter einem deutlich negativeren Rollwinkel als die restlichen. Dieser Unterschied lässt sich als Linksverschiebung auf der X-Achse in folgender Darstellung erkennen.

Abbildung 82: Rollwinkel im Verhältnis zur zurückgelegten Laufzeit, oberstes und unterstes Quantil, in Abschnitt K8



Die folgende Viessmann-Kurve K9 wird vom ersten Quantil abermals unter einem deutlich höheren absoluten Rollwinkel gefahren als von der Konkurrenz, konträr zu vorigen Kurven mit der höchsten Standardabweichung des Rollwinkels. Auffällig ist ebenfalls, dass auch in dieser Kurve der positive Rollwinkel im Maximum deutlich geringer ausfällt als bei der Konkurrenz, während der negative Rollwinkel deutlich ausgeprägter ist. Die Piloten fahren die Kurve folglich weiter und steuern sehr wenig gegen. Da dies im Vergleich zum Mittel aller weiteren Quantile festzustellen ist, kann davon ausgegangen werden, dass dies nicht durch Ausreißer bedingt ist. Diese Fahrlinie zeigt sich am Messpunkt ausgangs der Kurve, da dort der Rollwinkel, ebenso wie die Standardabweichung, niedriger ist als bei der Konkurrenz. Die zusätzlich gewonnene Geschwindigkeit nimmt im obersten Quantil mit 11,41 km/h am stärksten zu.

Die Fahrlinie in der sich anschließenden Kombination aus kurzer Kurve K10 und Kurve K11 kennzeichnet sich durch die geringste Standardabweichung des Rollwinkels im obersten Quantil, bedingt durch den niedrigeren Rollwinkel bei der Einfahrt in diesen Streckenabschnitt. Am Messpunkt ausgangs der Kurve K11 werden für das oberste Quantil der niedrigste maximale Rollwinkel gemessen in Kombination mit der niedrigsten Standardabweichung. Dies könnte ein Indiz sein, dass bei dieser hohen Geschwindigkeit die weniger steile Kurve enger genommen wird. Zu diesem Zeitpunkt beträgt der Vorsprung im Mittel 0,18 Sekunden und für die Geschwindigkeit 0,35 km/h.

Die zwei folgenden Kurven K12 und K13 werden zum weiteren Geschwindigkeitsaufbau genutzt, bevor es dann in die finale Kurve K14 vor dem Zieleinlauf geht. In diesem Abschnitt mit der höchsten Durchschnittsgeschwindigkeit – Streckenabschnitt K13 – fahren die Piloten im obersten Quantil im Mittel 134,04 km/h bei einem absoluten Rollwinkel von 36,59° und einer Standardabweichung des

Rollwinkels von 21,79°, was den drittniedrigsten Wert des Rollwinkels und zweitniedrigsten Wert der Standardabweichung bedeutet, während die vorangegangene Kurve K12 noch die höchsten Rollwinkel und Standardabweichungen aufweist. Dies lässt darauf schließen, dass die Fahrer des obersten Quantils ihre Fahrlinie so anpassen, dass sie als Vorbereitung der anschließenden Kurve den Rollwinkel reduzieren. Die letzte Kurve K14 zum Ziel hin fährt das oberste Quantil mit einer mittleren Ausprägung des maximalen absoluten Rollwinkels im Vergleich zur Benchmark. Dies erscheint sinnvoll, um zum Schluss Bandenkontakt zu vermeiden und die Position zu erhalten, da die letzte Kurve nur einen geringen Einfluss auf das Endergebnis besitzt. Folgende Metriken lassen sich für die Zieleinfahrt festhalten.

Tabelle 32: Metriken zum Fahrverhalten in den Quantilen in Abschnitt K14

Streckenabschnitt_Char Laufzeit Quantile	Kurve Av Fahrzeit	Av Laufzeit Quantil	Av Kurvenzeit Quantil	Roll angle abs	SD Roll angle abs	SD Roll angle Kurve	Av Max Roll angle Kurve	Minimum von Roll angle deg	Maximum von Roll angle deg	Speed km/h
0-20%	54,73	54,73	4,88	22,25	10,65	24,67	80,28	-44,97	2,13	108,30
20-40%	54,94	54,94	4,90	24,49	10,39	24,90	79,97	-45,82	-7,37	107,79
40-60%	55,11	55,11	4,92	29,90	11,66	25,09	81,22	-64,45	-12,04	107,10
60-80%	55,22	55,22	4,93	28,96	9,15	24,71	79,59	-45,72	-11,71	106,78
80-100%	55,49	55,49	4,95	26,28	8,46	24,79	80,21	-49,18	-11,46	106,02
Gesamt	**55,10**	**55,11**	**4,92**	**26,34**	**10,47**	**24,83**	**80,25**	**-64,45**	**2,13**	**107,17**

Um im Folgenden einzelne Fahrverläufe zu analysieren, wird eine zufällige Stichprobe aus den obersten und untersten Quantilen der jeweiligen Läufe gezogen, jeweils 2 Fahrer pro Quantil. Dies ergibt folgende Auswahl:

Tabelle 33: Stichproben einzelner Athleten

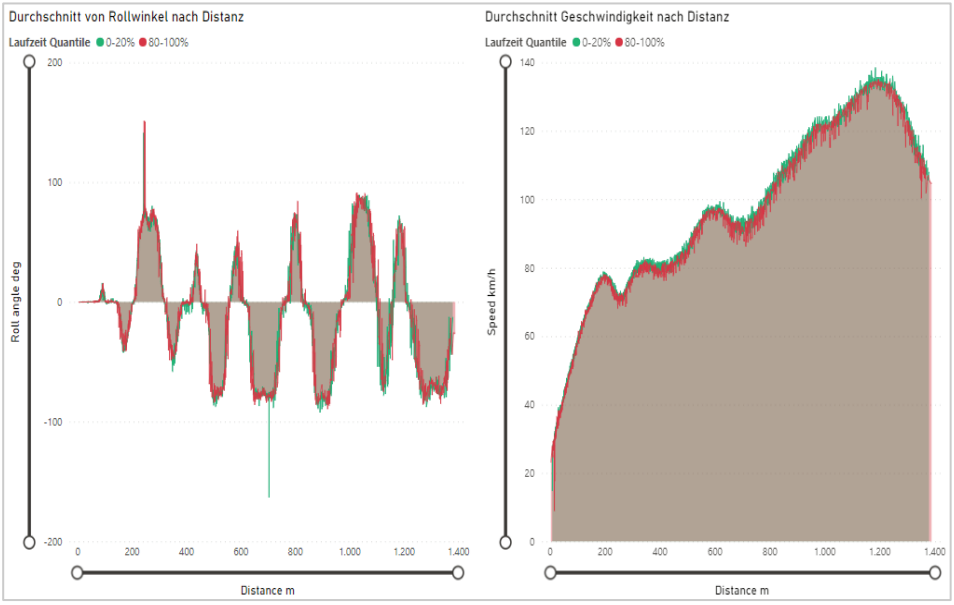
Datum	Lauf	Fahrer	Quantil	Rang	Zeit, in s
03.01.2020	1	Walther_Nico	0-20%	1	55,50
03.01.2020	1	Kibermanis_Oskars	0-20%	2	55,82
03.01.2020	1	Baumgartner_Patrick	80-100%	15	56,55
03.01.2020	1	Shinohara_Ryo	80-100%	16	57,09
03.01.2020	2	Kibermanis_Oskars	0-20%	1	55,11
03.01.2020	2	Stulnev_Alexey	0-20%	2	55,33
03.01.2020	2	Rohner_Timo	80-100%	9	55,67
03.01.2020	2	Tentea_Mihai Cristian	80-100%	11	55,69
04.01.2020	1	Walther_Nico	0-20%	1	55,58
04.01.2020	1	Kibermanis_Oskars	0-20%	3	55,74
04.01.2020	1	Berzins_Ralfs	80-100%	13	56,12
04.01.2020	1	De Bruin_Ivo	80-100%	15	56,47

04.01.2020	2	Kibermanis_Oskars	0-20%	1	55,16
04.01.2020	2	Friedrich_Francesco	0-20%	3	55,18
04.01.2020	2	Won_Yunjong	80-100%	11	55,52
04.01.2020	2	Suk_Youngjin	80-100%	13	55,60
10.01.2021	1	Friedrich_Francesco	0-20%	1	54,04
10.01.2021	1	Maier_Benjamin	0-20%	3	54,38
10.01.2021	1	Friedli_Simon	80-100%	13	55,01
10.01.2021	1	Vogt_Michael	80-100%	14	55,02
10.01.2021	2	Friedrich_Francesco	0-20%	1	54,09
10.01.2021	2	Kripps_Justin	0-20%	2	54,42
10.01.2021	2	Bascue_Codie	80-100%	13	55,01
10.01.2021	2	Friedli_Simon	80-100%	14	55,01
11.12.2021	1	Lochner_Johannes	0-20%	1	53,76
11.12.2021	1	Maier_Benjamin	0-20%	4	53,91
11.12.2021	1	Deen_Lamin	80-100%	22	54,67
11.12.2021	1	Kripps_Justin	0-20%	2	54,19
11.12.2021	2	Hall_Brad	0-20%	3	54,20
11.12.2021	2	Treichl_Markus	80-100%	14	54,60
11.12.2021	2	Church_Hunter	80-100%	16	54,66
11.12.2021	2	Lochner_Johannes	0-20%	2	54,90
12.12.2021	1	Hafer_Christoph	0-20%	5	55,05
12.12.2021	1	Bascue_Codie	80-100%	18	55,39
12.12.2021	1	Shinohara_Ryo	80-100%	22	56,12
12.12.2021	1	Friedrich_Francesco	0-20%	1	54,64
12.12.2021	2	Gaitiukevich_Rostislav	0-20%	2	54,76
12.12.2021	2	Kibermanis_Oskars	80-100%	15	55,10
12.12.2021	2	Tentea_Mihai Cristian	80-100%	18	55,32
12.12.2021	2	Friedrich_Francesco	0-20%	1	54,40
09.01.2022	1	Hall_Brad	0-20%	3	54,67
09.01.2022	1	Friedli_Simon	80-100%	17	55,54
09.01.2022	1	Tentea_Mihai Cristian	80-100%	19	55,82
09.01.2022	1	Friedrich_Francesco	0-20%	1	54,67
09.01.2022	2	Gaitiukevich_Rostislav	0-20%	3	54,83
09.01.2022	2	Suk_Youngjin	80-100%	18	55,67
09.01.2022	2	Vogt_Michael	80-100%	20	55,98
09.01.2022	2	Walther_Nico	0-20%	1	55,50

Anhand dieser Stichprobe wird grafisch dargestellt, ob die zuvor getroffenen Annahmen sich auch auf diese Auswahl übertragen lassen und sich spezielle Fahrmuster ergeben. Dazu wird der Fahrverlauf der beiden Quantile der Stichprobe

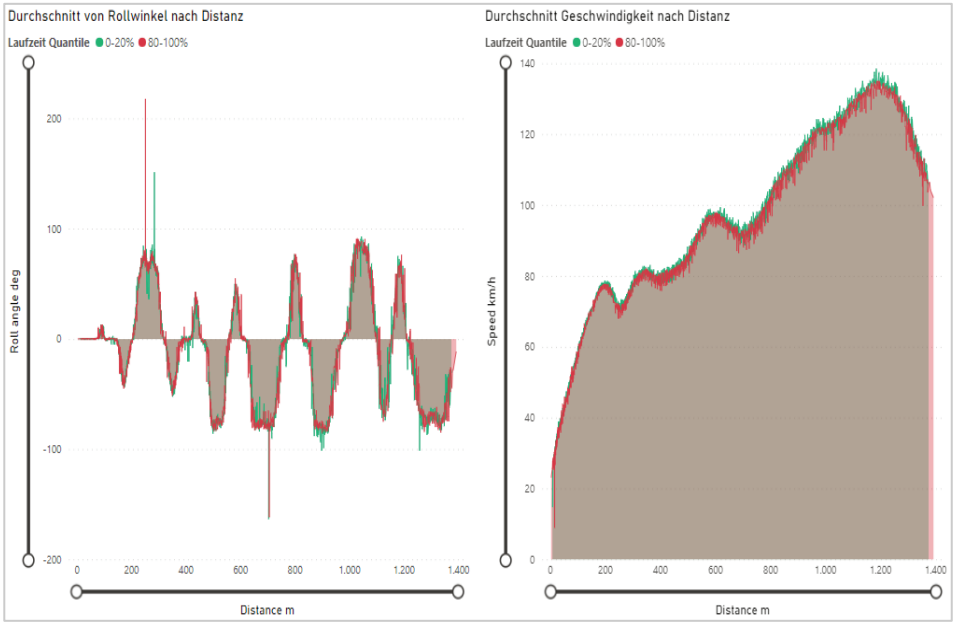
visualisiert und in folgendem Diagramm zusammen mit dem Geschwindigkeitsverlauf abgebildet.

Abbildung 83: Rollwinkel und Geschwindigkeit ausgewählter Fahrer im Verhältnis zur zurückgelegten Strecke des Laufes



Auffällig ist, dass die Fahrten der Bobs des obersten Quantils auf den ersten Blick an den meisten Stellen einen höheren Rollwinkel aufweisen, was sich an den grünen Rändern erkennen lässt. Ebenfalls liegen die durchschnittlichen Geschwindigkeiten ab dem Start durchweg oberhalb der Geschwindigkeiten des untersten Quantils. Um diesen Eindruck zu bestätigen, wird nachfolgend die gleiche Grafik für die Quantile inklusive aller Fahrer gezeigt. Bis auf wenige Ausreißer zeigt sich ein nahezu deckungsgleiches Bild.

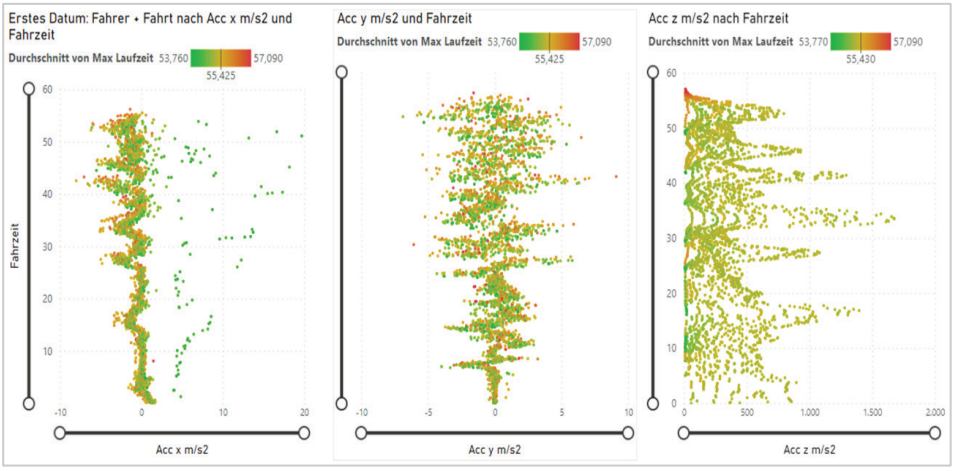
Abbildung 84: Rollwinkel und Geschwindigkeit aller Fahrer im Verhältnis zur zurückgelegten Strecke des Laufes



Die Einzelanalysen der Streckenabschnitte decken sich ebenfalls mit den Beobachtungen der Gesamtstichprobe der Fallstudie und die dazugehörigen grafischen Erkenntnisse können im Power BI Dashboard nachverfolgt werden.

Bei der Analyse der Beschleunigungsdaten fällt auf, dass die eingangs erwähnte Fahrt von Francesco Friedrich einen Ausreißer darstellt. Unter Ausklammerung dieser lässt sich feststellen, dass die schnellsten Fahrten der Quantile eine geringe Ausprägung der horizontalen Beschleunigung aufweisen. Erst in den letzten Geraden wird die Streuung größer und die horizontale Beschleunigung nimmt mit steigender Geschwindigkeit zu. Besonders auf der letzten Geraden wirkt die Horizontalbeschleunigung im obersten Quantil stärker als im untersten.

Abbildung 85: Achsenbeschleunigungen im Verhältnis zur zurückgelegten Zeit des Laufes



Die detaillierte Betrachtung der Beschleunigungswerte deckt sich mit den bisher gefundenen Erkenntnissen bezüglich des Rollwinkels und kann mittels grafischer Analyse innerhalb des Dashboards vertieft werden. Um den Rahmen der Analysen nicht zu sprengen, wird nun reduziert auf die Beschleunigungen eingegangen, welche aufgrund des Zusammenhangs zwischen den Achsen-Beschleunigungen und dem Rollwinkel zulässig ist. Unterstellte Kausalitäten zwischen den Beschleunigungsparametern und dem Rollwinkel werden durch Visual Analytics bekräftigt, da x und y tendenziell positiv vom Rollwinkel abhängen und z keine konkrete Tendenz aufweist. Aus diesem Grund ist für die Ableitung von inhaltlichen Schlüssen hinsichtlich der Fahrlinie der Rollwinkel als zentraler Indikator maßgeblich und ausreichend, sodass auf eine detaillierte Auswertung der Beschleunigungsparameter an dieser Stelle verzichtet und auf das Power BI-Dashboard verwiesen wird. Nichtsdestotrotz bietet sich eine tiefergehende Analyse für weitere Studien an. Abschließend für dieses Kapitel folgt eine Zusammenfassung der wichtigsten Erkenntnisse, welche in die sich anschließende Evaluation einfließen.

Tabelle 34: Zusammenfassung der Erkenntnisse aus der Analyse mittels Visualisierung

Thesen	Beurteilung auf Basis Data Analytics	Anmerkung
Ein schneller Start ist fundamental für eine gute Gesamtzeit.	✓	- Zusätzlich verstärkt durch Darstellung der Positionsentwicklung nach dem Start - Erweitert um die Bedeutung des Starts für die Geschwindigkeit der sich anschließenden Abschnitte
Ein geringer Rollwinkel und eine hohe Vertikalbeschleunigung sind vorteilhaft.	✗	- Aufteilung der These auf Beschaffenheit der Abschnitte - Widerlegung für Geraden und Beschleunigungsabschnitte von Kurven - Valide für bremsende Kurvenabschnitte

6.6 Evaluation

Anknüpfend an die Erkenntnisse aus Data und Visual Analytics wird im folgenden Abschnitt der Kreis des CRISP-DM geschlossen, indem eine Evaluation der Ergebnisse und eine Bewertung der getroffenen Beobachtungen erfolgt. Diese werden mit den eingangs aufgestellten Forschungshypothesen und dem theoretischen Hintergrund verknüpft und eine Grundlage des Fazits erarbeitet.

6.6.1 Analyse der Ergebnisse

Im Rahmen der initialen Analyse mittels Clustering in Python wurde die erste These „Ein schneller Start ist fundamental für eine gute Gesamtzeit“ untersucht und konnte nicht abgelehnt werden. Die Untersuchung mittels Korrelationskoeffizienten bietet einen guten ersten Schritt zur Einordnung. Zur Validierung der Erkenntnis waren jedoch tiefgreifendere statistische Verfahren, wie beispielsweise eine Regression, erforderlich. Da diese Arbeit ihren Fokus ebenfalls auf die Visualisierung der Ergebnisse und eine grafische Analyse legt, wurden die

eingehenden Beobachtungen durch eine detailliertere Analyse des Fahrverhaltens und damit zusammenhängend dem Erfolg validiert. Dazu wurde eine grafische Visualisierung der einzelnen Streckenabschnitte vorgenommen, welche die individuellen Fahrlinien erfasst und mit der Benchmark vergleicht.

Diese Analyse stützt die Nicht-Falsifizierbarkeit der ersten These. Es konnte aufgezeigt werden, dass ein schneller Start verbunden mit einer hohen Geschwindigkeit signifikanten Einfluss auf das Endergebnis hat. Dies zeigt sich durch die niedrigste Abschnittszeit in Kombination mit der höchsten Geschwindigkeit im obersten Quantil der Konkurrenz nach dem Start. Erweitert werden konnten diese Feststellungen durch das Fahrverhalten in den folgenden Streckenabschnitten, in welchen sich insbesondere die Fahrer des besten Quantils die gewonnene Geschwindigkeit durch den Start zu Nutze machen.

In Bezug auf die zweite These konnte in der initialen Analyse festgestellt werden, dass sie nicht bestätigt werden kann. Vielmehr lassen sich umgekehrte Effekte feststellen. Diese können jedoch nicht allgemeingültig über einen gesamten Lauf gehalten werden, sondern sind abschnittsspezifisch zu betrachten. Hierzu brachte der grafische Teil der Analyse hervor, dass eine Unterteilung der Strecke und entsprechende Fahrweisen für eine optimale Fahrweise unabdingbar ist. Zum einen konnte aufgezeigt werden, dass die Fahrweisen sich aus vorigen Streckenabschnitten bedingen und sich auf die Nachfolgenden auswirken. Daher muss die Fahrweise in einzelnen Abschnitten in Zusammenhang mit den umliegenden betrachtet werden. Als Ergebnis aus dieser Analyse lässt sich festhalten, dass der Rollwinkel je nach Abschnittsbeschaffenheit sowohl in niedriger als auch hoher Ausprägung optimal sein kann. Dies deckt sich mit der physikalischen Beobachtung, dass ein höherer Druck (= höhere Reibleistungsdichte) zunächst den Reibungskoeffizienten verringert. Daher sind Abschnitte, welche sich negativ auf die Beschleunigung auswirken (= der Reibungskoeffizient tendenziell ansteigt), in einem niedrigeren Rollwinkel zu fahren, um den Bremseffekt nicht unnötig zu verstärken.

Um dann nachfolgend wieder Geschwindigkeit aufzubauen, ist ein hoher Rollwinkel in den Kurven vorteilhaft, was sich mit der initialen Beobachtung deckt, dass auf geraden Streckenabschnitten der Reibungskoeffizient tendenziell geringer ausfällt, sofern die vorherige Kurve mit einem höheren Rollwinkel passiert wurde. In diesem Fall könnte bei dem Übergang in die Gerade eine höhere Beschleunigung realisiert werden, welche durch das *Frictional Heating* den Reibungskoeffizienten reduziert. Ferner werden Fliehkräfte durch einen höheren Rollwinkel in Form von Höhe gespeichert, welche sich beim Übergang in die Gerade wieder

abbaut und dadurch eine höhere Beschleunigung realisiert wird. Somit sollte der Rollwinkel in beschleunigenden Kurven tendenziell erhöht werden.

6.6.2 Limitationen

Die Analyse der zugrundeliegenden Forschungsthesen basiert im Wesentlichen auf der Datengrundlage einer Fallstudie und unterliegt demzufolge diversen Limitationen. Eine fundierte wissenschaftliche Abhandlung erfordert aus Transparenz- sowie Nachvollziehbarkeitsgründen entsprechend eine kritische Reflexion der gewählten Verfahrensschritte mitsamt den unterstellten Annahmen. Hierzu sind die wesentlichen Restriktionen dieser Arbeit in Tabelle 35 zusammengefasst.

Tabelle 35: Limitationen der Arbeit

Thema	Limitation	Bedeutung
Datengrundlage	• limitiert auf Winterberg • keine Berücksichtigung weiterführender Streckencharakteristika (Neigungen etc.) • limitiert auf Viererbob Herren Konkurrenz • reine Weltcup-Daten • kein Einbezug der Streckenerfahrung der Fahrer • keine weiterführenden physikalischen Daten	hoch
Methodik	• Datenpool über mehrere Wettkampftage • begrenze Auswahl statistischer Parameter zur Fahrlinienindikation • manuell festgelegtes Clustering • einfaches statistisches Modell gewählt	mittel

externe Einflüsse	• keine Berücksichtigung von Umwelteinflüssen • keine Berücksichtigung von laufindividuellen Ereignissen (Fahrfehler, Fahrergewicht) • keine Berücksichtigung von Materialeinflüssen • Reihenfolge, in der die Fahrten abliefen, wird ausgeblendet, bspw. inwiefern vorige Fahrten die Bahnbeschaffenheit beeinflussen	sehr hoch

Grundsätzlich sind die Limitationen hierbei in drei Kategorien einzuteilen. Hinsichtlich der Datengrundlage ist festzuhalten, dass die Beschränkungen auf die Weltcups der Viererbob-Herrenkonkurrenz in Winterberg dazu führen, dass die vorliegenden Ableitungen und Implikationen rein strecken- und konkurrenzspezifisch anzusehen sind. Zusätzlich fehlen zur Validierung der physikalischen Annahmen und Thesen weiterführende Daten, sodass die unvollständige Datengrundlage als bedeutungsvolle Limitation dieser Arbeit einzuschätzen ist.

Ferner unterliegen auch die gewählten Verfahrensschritte im Analyseteil diversen Vereinfachungen – exemplarisch angeführt sei hierzu die subjektive Definition der Bereinigungslogiken im Rahmen der Data Preparation sowie der vereinfachten Bestimmung von Clustern. Das Risiko dieser Einschränkungen für die Ergebnisqualität wird hierbei als mittel eingestuft.

Die dritte Kategorie umfasst sämtliche externe Einflüsse, für die keine Berücksichtigung im Datenmodell vorliegt, obwohl diese als Einflussfaktoren auf die untersuchten Thesen zu vermuten sind. Hierbei seien exemplarisch Temperaturangaben über das Eis oder Materialqualität der einzelnen Fahrer anzuführen – beides jeweils Faktoren, die erwartungsgemäß einen wesentlichen Einfluss auf die spezifische Reibung von Eis und Kufe haben. Ohne adäquate Berücksichtigung dieser Faktoren ist eine Verzerrung in den Ableitungen nicht auszuschließen, sodass aufgeführte fehlende Einflüsse als potenziell sehr bedeutungsvoll einzuschätzen sind.

6.7 Fazit und Ausblick

Das Ziel dieser Arbeit war es, den Zusammenhang von gewählten Fahrlinien auf die realisierten Laufzeiten im Bob-Sport zu untersuchen, um Eigenschaften von laufzeitoptimierenden Fahrlinien zu identifizieren. Methodisch wurde hierfür ein zweistufiger Analyseprozess angewandt: Zunächst erfolgte die Ableitung von Forschungsfragen aus der Literatur, indem auch die physikalischen Gesetzmäßigkeiten, welche beim Bobsport wirken, eruiert und berücksichtigt wurden. Hieraus sind die beiden folgenden Kernthesen definiert worden, welche als zentrale Untersuchungsgegenstände den Fokus der Analysen ausmachen.

I. Eine kurze Startzeit ist fundamental für eine gute Gesamtzeit.

Hintergrund: Je höher die Geschwindigkeit, desto geringer der Reibungskoeffizient und somit der Reibungsverlust. Hieraus folgt einem guten Start ein sich verstärkender, kumulativer Positiveffekt aus einer hohen Ausgangsgeschwindigkeit über die gesamte Fahrt.

II. Ein geringer Rollwinkel und eine hohe Vertikalbeschleunigung sind vorteilhaft.

Hintergrund: Infolge eines geringen Rollwinkels in Kurven wird durch den erhöhten Druck auf das Eis der Reibungskoeffizient reduziert, was zu geringeren Reibungsverlusten führt. Erhöhter Druck resultiert in erhöhter Reibleistungsdichte.

Angeführte Thesen wurden im zweiten Schritt anhand der Datenauswertung von offiziellen Wettkämpfen im Rahmen des Bob-Weltcups in Winterberg untersucht und bewertet. Eine Untersuchung der Ursache-Wirkungszusammenhänge der vorliegenden Fahrparameter ergab, dass der Rollwinkel sowohl physikalisch als auch statistisch die maßgebliche Einflussgröße auf die Laufzeit darstellt. Da die weiteren Parameter mittelbar wie unmittelbar vom Rollwinkel abhängen, wird dieser als zentrale Metrik zur Fahrliniendefinition und als Instrument einer potenziellen Fahroptimierung definiert, sodass die Analysen sowie inhaltlichen Ableitungen primär den Rollwinkel fokussieren.

Basierend auf durchgeführten Rangkorrelationsanalysen konnte hierbei ein positiver Zusammenhang zwischen der Performance in der Startphase sowie dem endgültigen Gesamtresultat identifizieren werden, sodass die erste These auf Basis der Datenanalyse nicht falsifiziert werden kann. Hierbei gilt zu beachten, dass eine eindeutige Bestätigung dieser These nicht möglich ist, da der beobachtete Zusammenhang auch auf andere Faktoren, wie etwa der Fahrerqualitäten, zurückzuführen sein könnte. Durch den impliziten Benchmark-Vergleich ist die

These jedoch dahingehend zu stützen, dass kurze Startzeiten überproportional eine bessere Endplatzierung begünstigen. Zur Untersuchung der zweiten These wurden vorliegend Korrelations- sowie Clusteranalysen durchgeführt, welche das Ergebnis hatten, dass in der Gesamtbetrachtung ein tendenziell höherer Rollwinkel zu besseren Endplatzierungen führt. Aus diesem Grund kann die aufgestellte These, wonach ein geringerer Rollwinkel als vorteilhaft anzusehen ist, nicht beibehalten werden. Hierbei ist zu berücksichtigen, dass bei Betrachtung von einzelnen Streckenabschnitten keine übereinstimmenden Ableitungen über die zeitoptimalen Rollwinkel zu treffen sind. Die Ergebnisse lassen hierzu vielmehr den Schluss zu, dass die Fahrweise sehr spezifisch auf das jeweilige Bahnlayout anzupassen ist. Der Hintergrund liegt darin, dass Kurvenabschnitte, in denen der Bob eher gebremst wird, tendenziell mit einem niedrigeren Rollwinkel zu durchfahren sind, um den Reibungsverlust zu minimieren, wohingegen in beschleunigenden Abschnitten – etwa unmittelbar vor geraden Abschnitten – der Rollwinkel zu erhöhen ist, um einen Teil der Vertikalbeschleunigung in Höhe sowie einem höheren Rollwinkel zu übersetzen und damit eine höhere Beschleunigung auf der folgenden Gerade zu erreichen. Demzufolge hängt die Fahrlinie unmittelbar vom Bahnlayout ab und kann somit auch innerhalb eines Kurvenabschnitts Fahrstiladaptionen verlangen, sodass eine vereinfachte, allgemeingültige These wie die definierte Ausgangsthese in der Praxis nicht haltbar erscheint. Auf Basis einer umfangreicheren und heterogeneren Datengrundlage erscheint hierzu eine Kollektion mehrerer bedingter Thesen – in Form von Fallunterscheidungen je nach Bahn- und Abschnittslayout – erforderlich zu sein, um die Voraussetzungen einer zeitoptimalen Fahrlinie adäquat wiederzugeben.

Somit lässt sich als Fazit und weiterführenden Ausblick konstatieren, dass auf Basis der vorliegenden Arbeit keine allgemeingültigen Schlüsse hinsichtlich zeitoptimierender Fahrlinien im Bobsport zu ziehen sind. Gleichwohl kann gezeigt werden, dass durch entsprechende Analyseverfahren strecken- sowie streckenabschnittsspezifische laufzeitoptimierende Fahrlinieneigenschaften zu identifizieren sind. Vorliegend ergeben sich hieraus bereits äußerst wertvolle praktische Implikationen zur Performanceverbesserung. Gleichzeitig dienen die Erkenntnisse bereits als erste fundierte Ausgangsthesen über potenziell allgemeingültige, jedoch sehr bedingte laufzeitoptimierende Fahrlinienwahlen. Eine Ausweitung dieser Erkenntnisse auf allgemeingültige Feststellungen bedarf zusätzlicher Datenerfassung hinsichtlich des Bahnzustands sowie der Kontaktqualität von Eis und Kufen, um die zugrundeliegenden und unterstellten physikalischen Gesetzmäßigkeiten überprüfen zu können. Zusätzlich ist eine Ausweitung des For-

schungsumfangs um weitere Bahnen sowie Konkurrenzen notwendig, um wissenschaftlich fundierte, allgemeingültige Aussagen treffen zu können. Hierbei ist der potenzielle Mehrwert einer tiefergehenden wissenschaftlichen Auseinandersetzung als gegeben einzuschätzen, da die Ergebnisse der vorliegenden Datenanalyse zeigen, dass bereits kleine Anpassungen bei den gewählten Fahrlinien signifikante Auswirkungen auf die Performance haben können.

Literatur

Altman, D.G.; Royston, P (2006): The cost of dichotomising continuous variables, in: British Medical Journal, 2006, 332(7549), 1080.

Beguería, S.; Pueyo, Y.: A comparison of simultaneous autoregressive and generalized least squares models for dealing with spatial autocorrelation, in: Global Ecology and Biogeography, 2009, 18(3), 273–279

Bholowalia, P. & Kumar, A. (2014). EBK-Means: A Clustering Technique based on Elbow Method and KMeans in WSN. International Journal of Computer Applications, 105(9), 17-24.

Breusch, T. S.; Pagan, A. R.: The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics, in: The Review of Economic Studies, 1980, 47(1), 239

Brüggemann, G.-P., Morlock, M. & Zatsiorsky, V. M. (1997). Analysis of the Bobsled and Men's Luge Events at the XVII Olympic Winter Games in Lillehammer. Journal of Applied Biomechanics, 13(1), 98–108, https://doi.org/10.1123/jab.13.1.98 .

Buchkremer, R. et al. (2019a). "The Application of Artificial Intelligence Technologies as a Substitute for Reading and to Support and Enhance the Authoring of Scientific Review Articles," IEEE Access, 7, 65263–65276, doi: 10.1109/ACCESS.2019.2917719.

Buchkremer, R., Demund, A., Ebener, S., Gampfer, F., Jagering, D., Jurgens, A., Klenke, S., Krimpmann, D., Schmank, J., Spiekermann, M., Wahlers, M., & Wiepke, M. (2019). The Application of Artificial Intelligence Technologies as a Substitute for Reading and to Support and Enhance the Authoring of Scientific Review Articles. IEEE Access, 7(c), 65263–65276. https://doi.org/10.1109/ACCESS.2019.2917719.

Castellano, B. (2014): „PYSCENEDETECT", https://bcastell.com/projects/PySceneDetect/ [Zugriff am 23 02 2022].

Chen, T., & Guestrin, C. (2016). XGBoost. A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

Colyer, S. L., Stokes, K. A., Bilzon, J. L. J., Holdcroft, D., & Salo, A. I. T. (2018). The effect of altering loading distance on skeleton start performance: Is higher pre-load velocity always beneficial? Journal of Sports Sciences, 36(17), 1930–1936. https://doi.org/10.1080/02640414.2018.1426352.

Cotsaces, C., Nikolaidis, N., Pitas, I. (2006): „Video shot detection and condensed representation. a review," IEEE Signal Process. Mag. (IEEE Signal Processing Magazine), 28-37.

Cui, Y., Qiu, C., Cai, Y., & Gao, X. (2017): „Scene detection of news video using CNN features," in 2017 10th International Congress 10/14/2017 - 10/16/2017, 2017.

Dabnichki, P. (2015). "Bobsleigh performance characteristics for winning design", Procedia Engineering, 112, 436-442.

Data Story: "Gephi - Clustering layout by modularity", https:// parklize.blogspot.com/2014/12/gephi-clustering-layout-by-modularity.html. [Zugriff am 22.02.2022].

Daudpota, S. M., Muhammad, A., Baber, J. (2019): „Video genre identification using clustering-based shot detection algorithm," SIViP (Signal, Image and Video Processing), 1413-1420.

Dickey, D. A.; Fuller, W. A.: Distribution of the Estimators for Autoregressive Time Series with a Unit Root, in: Journal of the American Statistical Association, 1979, 74(366a), 427–431

Dumm, M., Hainzlmaier, C., Boerboom, S. & Wintermantel, E. (2006). "The Effect of Pressure on Friction of Steel and Ice and Implementation to Bobsleigh Runners", erschienen in: Moritz E.F., Haake S. (eds) "The Engineering of Sport 6", Springer, New York, NY, https://doi.org/10.1007/978-0-387-45951-6_19 .

Durbin, J; Watson, G S: Testing for Serial Correlation in Least Squares Regression: I, in: Biometrika, 1950, 37(3/4), 409.

EMC Education Services (2015). Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, in: John Wiley & Sons, Indianapolis.

Fraunhofer Institut, „Nutzerzentrierung", : https://www.ipa.fraunhofer.de/de/aktuelle-forschung/laborautomatisierung-und-bioproduktionstechnik/digital_lab_services/nutzerzentrierung.html [Zugriff am 15.04.2025].

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine, in: Annals of Statistics, 29(5), 1189-1232.

Friedman, J. (2002). Stochastic gradient boosting, in: Computational Statistics & Data Analysis, 38(4), 367-378.

Godfrey, L. G.: Testing for Higher Order Serial Correlation in Regression Equations when the Regressors Include Lagged Dependent Variables, in: JSTOR: Econometrica, 1978, 46(6), 1303

Hainzlmaier, C. (2005). A new tribologically optimized bobsleigh runner (Doctoral dissertation, Technische Universität München).

Handschin, T.: „Thomas Handschin“, https://www.thomashandschin.ch/bobsport/technik/index.php#. [Zugriff am 20 Februar 2022].

Hockley, W.E. (2008). “The picture superiority effect in associative recognition”, Memory & Cognition, 36, 1351 - 1359 https://doi.org/10.3758/MC.36.7.1351.

Huang, Y.-C., Liao, I.-N., Chen, C.-H., Ik, T.-U., Peng, W.-C. (2019): „TrackNet: A Deep Learning Network for Tracking High-speed and Tiny Objects in Sports Applications,“ 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 1-8.

IBSF (2022). VELTINS EisArena Winterberg, unter: https://www.ibsf.org/en/tracks/track/1/Winterberg, [abgerufen am 28.02.2022].

International Bobsleigh and Skeleton Federation: IBSF | Athletes, https://www.ibsf.org/en/athletes (2022a) [Zugriff am 15. Februar 2022].

International Bobsleigh and Skeleton Federation: IBSF | Tracks, https://www.ibsf.org/en/tracks (2022b) [Zugriff am 24. Februar 2022].

International Bobsleigh and Skeleton Federation (o. J.): IBSF | Bobsleigh History, https://www.ibsf.org/en/our-sports/bobsleigh-history [Zugriff am 24. Februar 2022].

International Bobsleigh and Skeleton Federation: IBSF | Bobsleigh Info Graphics, https://www.ibsf.org/en/our-sports/bobsleigh-info-graphics (2015a) [Zugriff am 24. Februar 2022].

International Bobsleigh and Skeleton Federation: IBSF | Tracks | Winterberg, https://www.ibsf.org/en/tracks/track/1/Winterberg (2015b) [Zugriff am 24. Februar 2022].

International Bobsleigh and Skeleton Federation: IBSF-Bahnreglement, erhalten von < https://www.ibsf.org/fileadmin/user_upload/Resources/Sports/Rules_Quotas/IBSF_Bahnreglement_am2019D.pdf> (2019) [Zugriff am 26. Februar 2025].

International Bobsleigh and Skeleton Federation: Internationales Bob-Reglement, erhalten von https://www.ibsf.org/images/federation/Rules_and_Regulations/2021_Internation ales_Reglement_BOB.pdf (2021a) [Zugriff am 26. Februar 2022].

International Bobsleigh and Skeleton Federation: Internationales Frauen-Monobob Reglement, erhalten von https://www.ibsf.org/images/documents/downloads/Rules/2021_2022/IBSF_Inter nationales_Frauen_Monobob_Reglement_2021.pdf (2021b) [Zugriff am 26. Februar 2022].

Irbe, M., Gross, K. A., Viba, J., & Cerpinska, M. (2018). Analysis of acceleration and numerical modeling of skeleton sled motion. Engineering for Rural Development, 17, 1401–1406. https://doi.org/10.22616/ERDev2018.17.N179 .

Irbe, M.; Gross, K.A., Viba, J.; Cerpinska, M. (2021): Unveiling ice friction and aerodynamic drag at the initial stage of sliding on ice: Faster sliding in winter sports, in: Tribology International, 2021, 160, https://doi.org/10.1016/j.triboint.2021.106967.

Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. (2014). “ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software,” PloS one, 9, 6, e98679, https://doi.org/10.1371/journal.pone.0098679.

KaewTraKulPong, P. & Bowden, R. (2002). An improved adaptive background mixture model for real-time tracking with shadow detection. In P. Remagnino, G. A., Jones, N., Paragios, C. S. Regazzoni (Hrsg.), Video-Based Surveillance Systems: Computer Vision and Distributed Processing (135-144). Springer.

Kanth, R., Skön, J.-P., Lehtomäki, K.(2018): „Image Analysis and Development of Graphical,“ Journal of Image and Graphics, 109-116.

Kietzig, A. M., Hatzikiriakos, S. G., & Englezos, P., 2010b, Physics of ice friction, Journal of Applied Physics, 107 (8), 081101, https://doi.org/10.1063/1.3340792 .

Kietzig, A.-M., Hatzikiriakos, S.G.; Englezos, P. (2010): Ice friction: the effect of thermal conductivity, in: Journal of Glaciology, 2010a, 56(197), 473–479.

Kim, K.-Y., Um, G.-M. (2017): „Usage scenario and user interface for ice hockey game analysis," 2017 International Conference on Information and Communication Technology Convergence (ICTC), 1080-1082.

Koshkina, M., Pidaparthy, H., Elder, J. H. (2021): „Contrastive Learning for Sports Video: Unsupervised Player Classification," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 4523-4531.

Kwiatkowski, D.; Phillips, P.C.B.; Schmidt, P.; Shin, Y. (1992): Testing the null hypothesis of stationarity against the alternative of a unit root, in: North-Holland: Journal of Econometrics, 1992, 54(1–3), 159–178.

Liebermann, D. G. , Katz, L. , Hughes, M. D., Bartlett, R. M., McClements, J., Franks, I.M. (2002): „Advances in the application of information technology," Journal of Sport Sciences, 755-769, 2002.

Lloyd, S. (1957). Least squares quantization in PCM. Technical Report RR-5497, Bell Lab, September 1957.

Lopes, A.D.; Alouche, S.R. (2016): Two-Man Bobsled Push Start Analysis, in: Polish Academy of Science, Committee of Physical Culture: Journal of Human Kinetics, 2016, 50(1), 63–70.

lunaHD GmbH, „Video Überwachung Veltins Eisarena Winterberg," Youtube, 9 April 2018, https://www.youtube.com/watch?v=syv3961b6L0. [Zugriff am 16 Februar 2022].

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, in: L. M. Le Cam & J. Neyman (Hrsg.): Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1, California: University of California Press, 281-297.

Mössner, M.; Hasler, M.; Schindelwig, K., Kaps, P.; Nachbauer, W. (2011): An approximate simulation model for initial luge track design, in: Journal of Biomechanics, 2011, 44(5), 892–896

Nayar, S. K. (2021): „First Principles of Computer Vision - Gaussian Mixture Model“, https://fpcv.cs.columbia.edu/ . [Zugriff am 11.04.2025].

OpenCV, „BackgroundSubtractor Class Reference“, https://docs.opencv.org/3.4/d7/df6/classcv_1_1BackgroundSubtractor.html. [Zugriff am 11.04.2025].

OpenCV, „How to Use Background Subtraction Methods“ [Online]. Available: https://docs.opencv.org/3.4/d1/dc5/tutorial_background_subtraction.html. [Zugriff am 11.04.2025].

Perktold, J.; Seabold, S.; Taylor, J. (2021a): Statsmodels-developers: statsmodels - acorr_breusch_godfrey, https://www.statsmodels.org/devel/generated/statsmodels.stats.diagnostic.acorr_b reusch_godfrey.html [Zugriff am 20. Dezember 2021].

Perktold, J.; Seabold, S.; Taylor, J. (2021b): Statsmodels-developers: statsmodels - adfuller, https://www.statsmodels.org/dev/generated/statsmodels.tsa.stattools.adfuller.html [Zugriff am 25. Dezember 2021].

Perktold, J.; Seabold, S.; Taylor, J. (2021c): Statsmodels-developers: statsmodels - durbin_watson, https://www.statsmodels.org/dev/generated/statsmodels.stats.stattools.durbin_wat son.html [Zugriff am 20. Dezember 2021].

Perktold, J.; Seabold, S.; Taylor, J. (2021d): Statsmodels-developers: statsmodels - het_breuschpagan, https://www.statsmodels.org/dev/generated/statsmodels.stats.diagnostic.het_breus chpagan.html [Zugriff am 20. Dezember 2021].

Perktold, J.; Seabold, S.; Taylor, J. (2021e): Statsmodels-developers: statsmodels - het_white, https://www.statsmodels.org/dev/generated/statsmodels.stats.diagnostic.het_white .html [Zugriff am 20. Dezember 2021].

Perktold, J.; Seabold, S.; Taylor, J. (2021f): Statsmodels-developers: statsmodels - kpss, https://www.statsmodels.org/stable/generated/statsmodels.tsa.stattools.kpss.html [Zugriff am 25. Dezember 2021].

Perktold, J.; Seabold, S.; Taylor, J. (2021g): Statsmodels-developers: statsmodels - linear_reset, https://www.statsmodels.org/dev/generated/statsmodels.stats.diagnostic.linear_res et.html [Zugriff am 20. Dezember 2021].

Poirier, L., Lozowski, E. P., & Thompson, R. I., 2011, "Ice hardness in winter sports", Cold Regions Science and Technology, 67(3), 129-134, https://doi.org/10.1016/j.coldregions.2011.02.005 .

Poirier, L., Lozowski, E. P., Maw, S., Stefanyshyn, D. J., & Thompson, R. I. (2011). Experimental Analysis of Ice Friction in the Sport of Bobsleigh, Sports Engineering, 14(2-4). https://doi.org/10.1007/s12283-011-0077-0

Poirier, L., Lozowski, E.P., Maw, S., Stefanyshyn, D.J, & Thompson, R.I., (2013): “Experimental analysis of ice friction and aerodynamic dragduring a World Cup 2-men bobsleigh competition", Journal of Sports Sciences, http://mc.manuscriptcentral.com/rjsp.

Poirier, L.; Lozowski, E.P.; Thompson, R.I. (2011): Ice hardness in winter sports, in: Elsevier B.V.: Cold Regions Science and Technology, 2011, 67(3), 129-134.

PyInstaller, „PyInstaller - Developer Documentation,“ [Online]. Available: https://pyinstaller.readthedocs.io/en/stable/.

Python GUIs for Humans, [Online], https://pysimplegui.readthedocs.io/en/latest/. [Zugriff am 23.02.2022].

Qiu, L. & Yuan, J. (2014). “Construction of Semantic Associative Network Based on Topic-Maps,” in Construction of Semantic Associative Network Based on Topic-Maps, 325–328.

Ramsey, J. B.: Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis, in: Wiley: Journal of the Royal Statistical Society: Series B (Methodological), 1969, 31(2), 350–371

Rempfler, G.S.; Glocker, C. (2016): A bobsleigh simulator software, in: Multibody System Dynamics, 2016, 36(3), 257–278 Springer Netherlands

Röder, M., Both, A. & Hinneburg, A. (2015). “Exploring the Space of Topic Coherence Measures,” in Exploring the Space of Topic Coherence Measures, New York, NY, USA, 399–408.

Said, S.E.; Dickey, D.A.(1984): Testing for unit roots in autoregressive-moving average models of unknown order, in: Biometrika, 1984, 71(3), 599–607

Saltz, J. S. (2021). “CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps,” in CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps, 2337–2344.

Sannazzaro, R. (2020): „Build a Motion Heatmap Video Using OpenCV With Python", https://medium.com/data-science/build-a-motion-heatmap-video-using-opencv-with-python-fd806e8a2340. [Zugriff am 11.04.2025].

Scene Detection Algorithms [Online]. Available: https://pyscene-detect.readthedocs.io/en/latest/reference/detection-methods. [Zugriff am 23.02.2022].

Scherge, M. (2021). Acceleration analysis in bobsledding – more questions than answers. Gliding Snowstorm Publishing, 1, 8-12.

Scherge, M., Böttcher, R., Richter, M., & Gurgel, U., 2013, "High-speed ice friction experiments under lab conditions: on the influence of speed and normal force", International Scholarly Research Notices, 2013, http://dx.doi.org/10.5402/2013/703202

Schleinitz, J.v.; Wörle, L.; Graf, M.; Schröder, A. (2022): Modeling ice friction for vehicle dynamics of a bobsled with application in driver evaluation and driving simulation, in: Tribology International, 2022, 165 (January 2022), 107344.

Schlipsing, M., Salmen, J. & Igel, C. (2013). Echtzeit-Videoanalyse im Fußball: Ein Live-System zum Spieler-Tracking. KI – Künstliche Intelligenz, 27(1), 235–240.

Seidenschwarz, P., Jonsson, A., Plüss, M., Rumo, M., Probst, L., Schuldt, H. (2020): „The SportSense User Interface for Holistic Tactical Performance Analysis in Football," Proceedings of the 25th International Conference on Intelligent User Interfaces Companion, 45-46.

Shapley, L. S. (1953). A value of n-person games, in: H.W. Kuhn & A.W. Tucker (Hrsg.): Contributions to the Theory of Games, 2, Princeton 1953, 307317.

Sivamani, R. K., Goodman, J., Gitis, N. V., & Maibach, H. I., 2003, "Coefficient of friction: tribological studies in man – an overview", Skin Research and Technology, 9(3), 227-234, https://doi.org/10.1034/j.1600-0846.2003.02366.x

Stein, M., Janetzko, H., Lamprecht, A., Breitkreuz, T., Zimmermann, P., Goldlücke, B., Schreck, T., Andrienko, G., Grossniklaus, M., Keim, D. A. (2018): „Bring It to the Pitch: Combining Video and Movement Data to Enhance Team Sport Analysis," IEEE Transactions on Visualization and Computer Graphics , 13-22.

Sundararajan, M. & Najmi, A. (2019). The many Shapley values for model explanation, in: arXiv:1908.08474 [cs.AI], 1-9.

Tora, M. R., Chen, J., Little, J. J. (2017): „Classification of Puck Possession Events in Ice Hockey," 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 147-154.

Ubbens, H., Dwight, R., Sciacchitano, A. & Timmer, N. (2016). Some Results on Bobsleigh Aerodynamics, in: Procedia Engineering, 147, 92-97.

Ulmer, S., 2009, "Die unspektakulären Gleiter", Dossier Citius, https://www.eth-life.ethz.ch/archive_articles/090226_Dossier_Citius_StahlaufEis_su/index.html, [Zugriff am 14.02.2022].

Verbeek, M. (2004). A guide to modern econometrics (2. Ed.). John Wiley & Sons, Ltd.

Wald, A; Wolfowitz, J.: Confidence Limits for Continuous Distribution Functions, in: The Annals of Mathematical Statistics, 1939, 10(2), 105–118

White, Halbert: A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, in: Econometrica, 1980, 48(4), 817

Wirth, R.; Hipp, J. (2000): CRISP-DM: towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 29-39, in: Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 2000, (24959), 29–39

Zanoletti, C.; La Torre, A.; Merati, G.; Rampinini, E.; Impellizzeri, F.M. (2006): Relationship Between Push Phase and Final Race Time in Skeleton Performance, in: The Journal of Strength and Conditioning Research, 2006, 20(3), 579

Zhang, Y. L.; Hubbard, M.; Huffman, R. K. (1995): Optimum control of bobsled steering, in: Springer: Journal of Optimization Theory and Applications, 1995, 85(1), 1–19

Zivkovic, Z. (2004): „Improved adaptive gaussian mixture model for background subtraction," IEEE, 28-31, 2004.

7 Anhang: Shapley Values einzelner Streckenabschnitte.

Abbildung 86: Abschnitt „S bis 1“

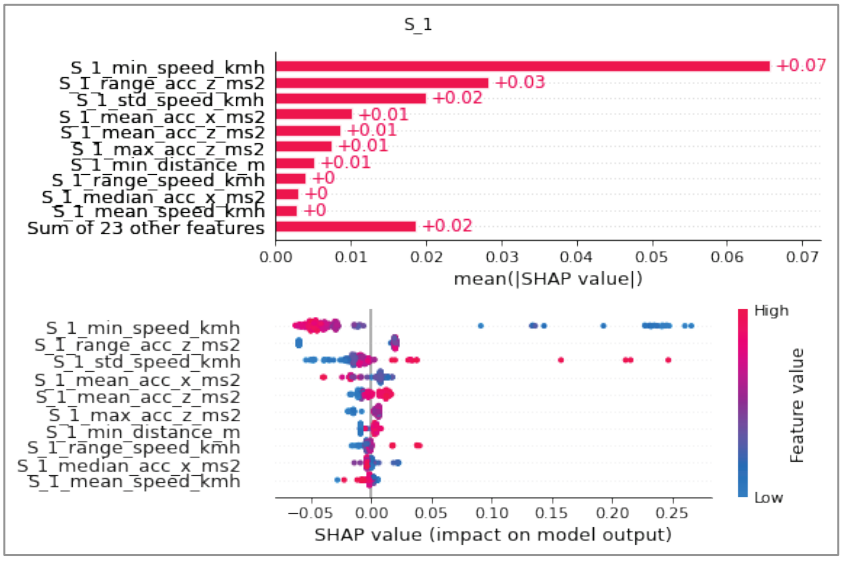


Abbildung 87: Abschnitt „1 bis B10“

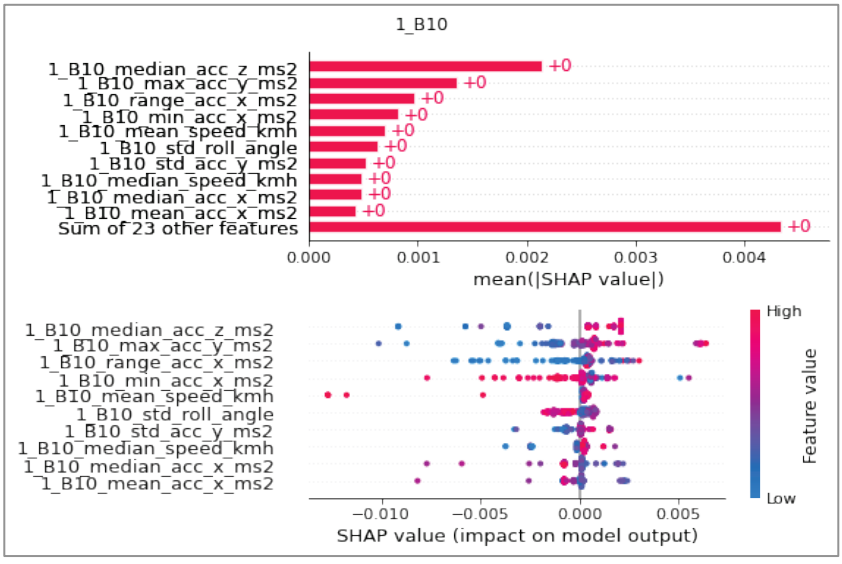


Abbildung 88: Abschnitt „B11 bis B12“

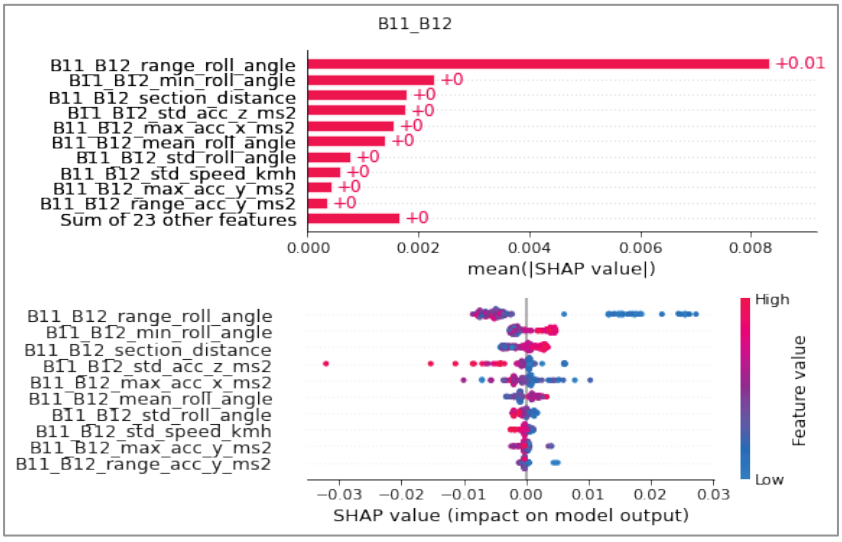


Abbildung 89: Abschnitt „B12 bis 2“

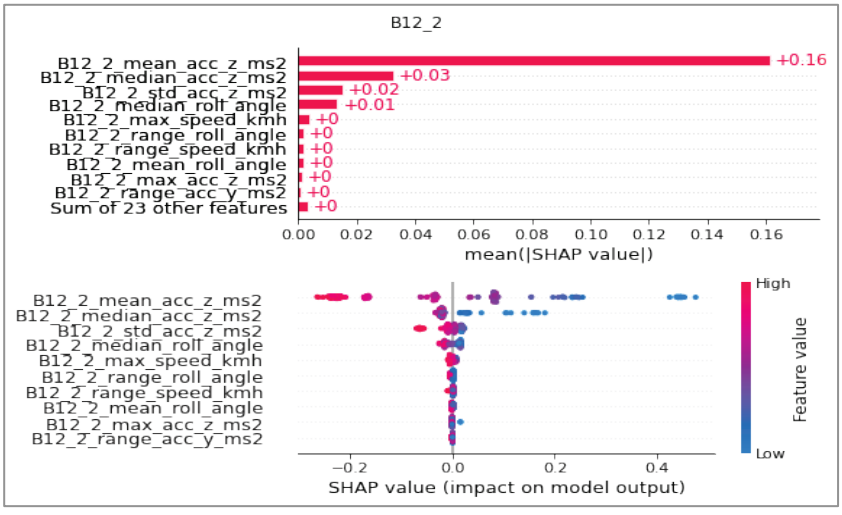


Abbildung 90: Abschnitt „2 bis B15“

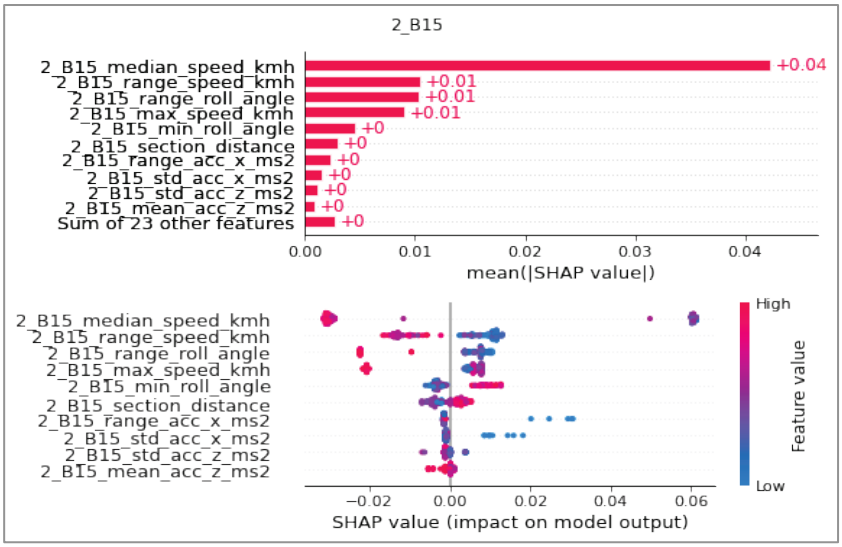


Abbildung 91: Abschnitt „B15 bis B16“

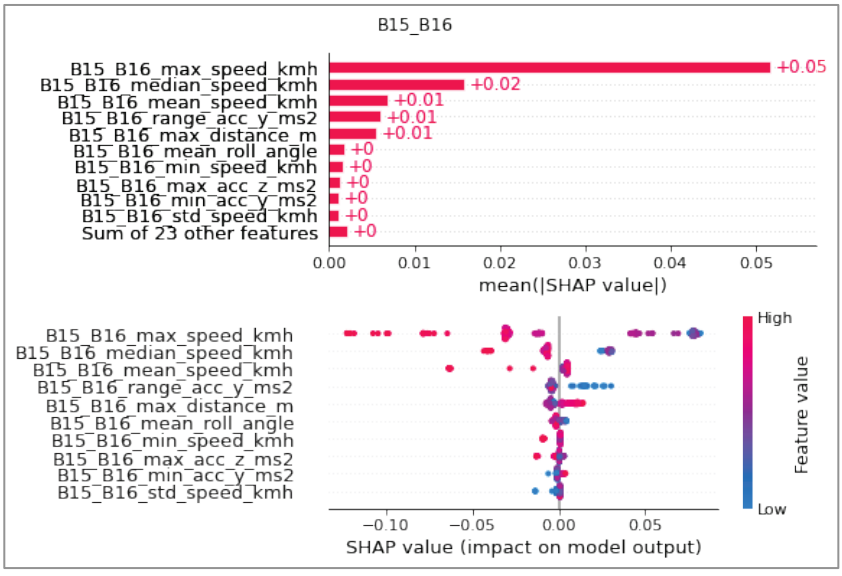


Abbildung 92: Abschnitt „B16 bis 3“

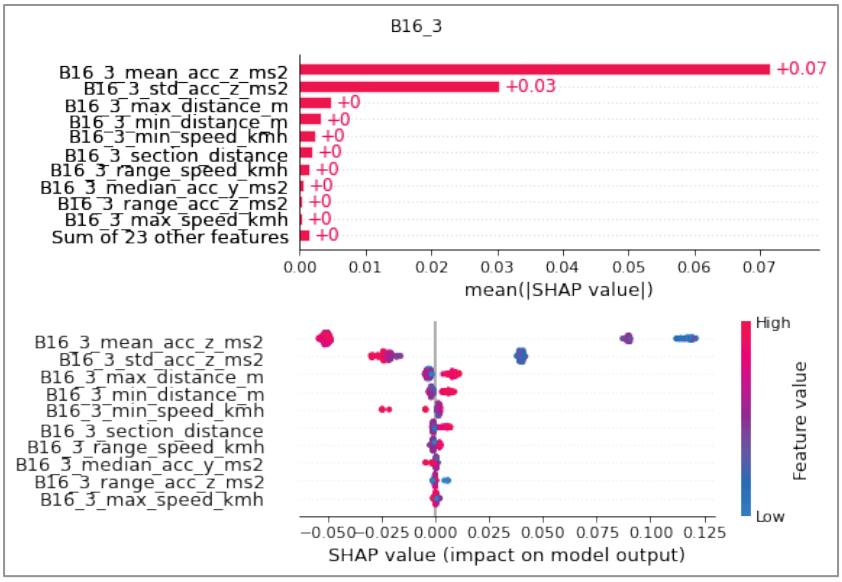


Abbildung 93: Abschnitt „3 bis B18“

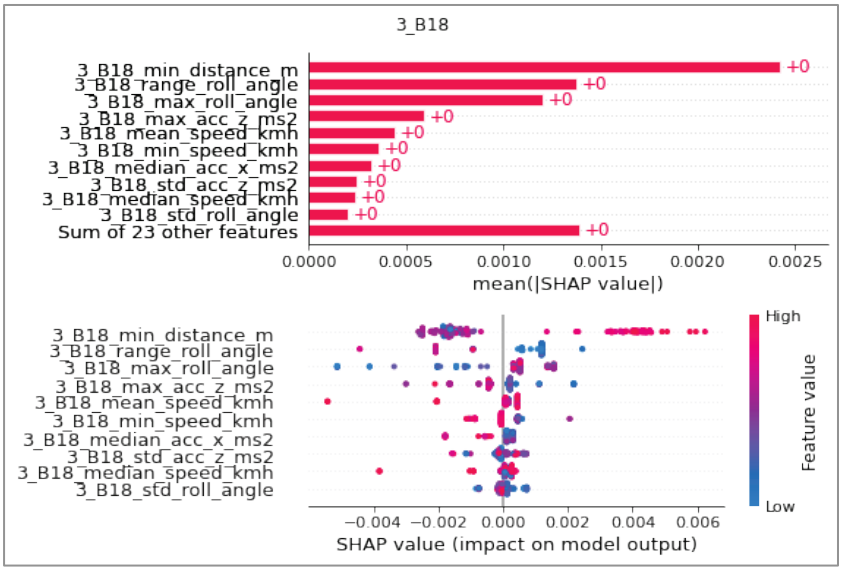


Abbildung 94: Abschnitt „B18 bis B19“

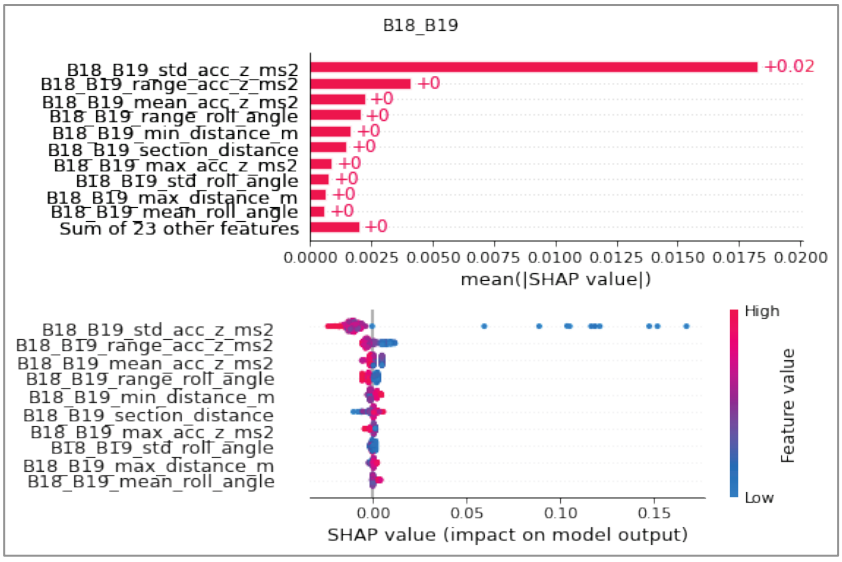


Abbildung 95: Abschnitt „B19 bis 4“

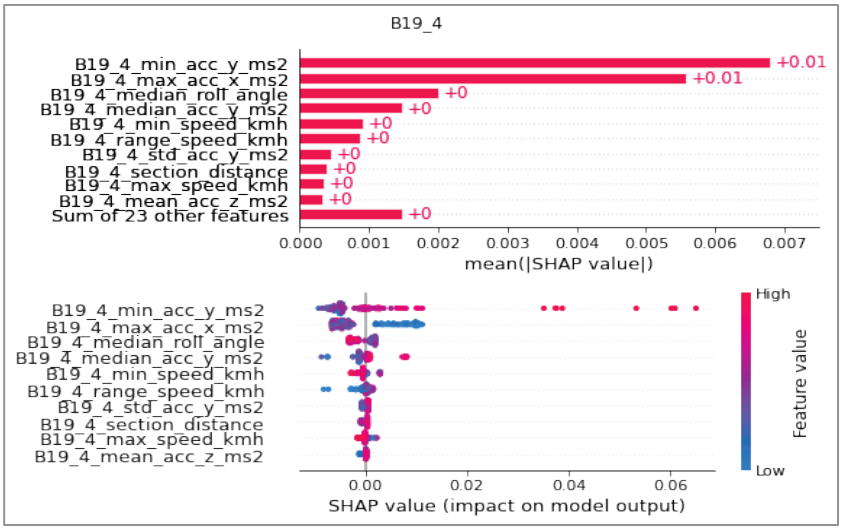


Abbildung 96: Abschnitt „4 bis B21“

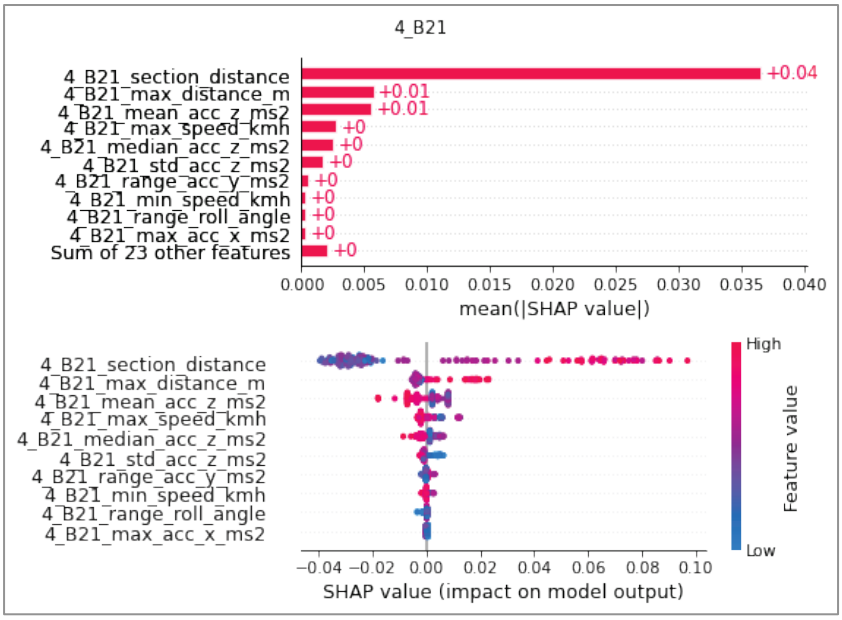


Abbildung 97: Abschnitt „B21 bis B22“

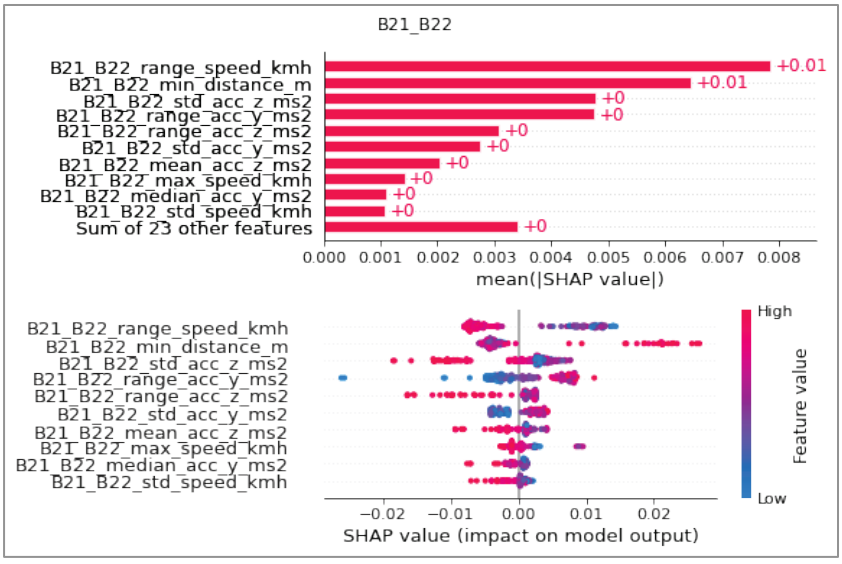


Abbildung 98: Abschnitt „B22 bis 5“

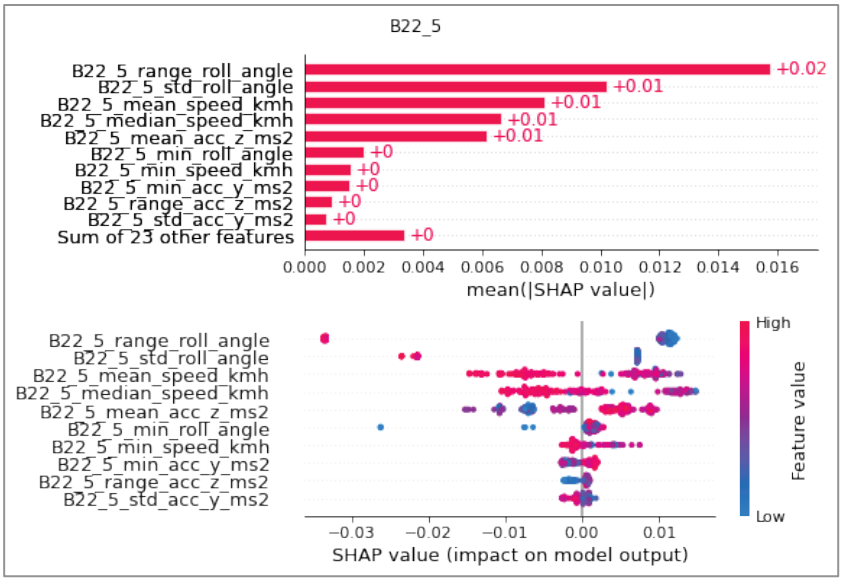


Abbildung 99: Abschnitt „5 bis B24“

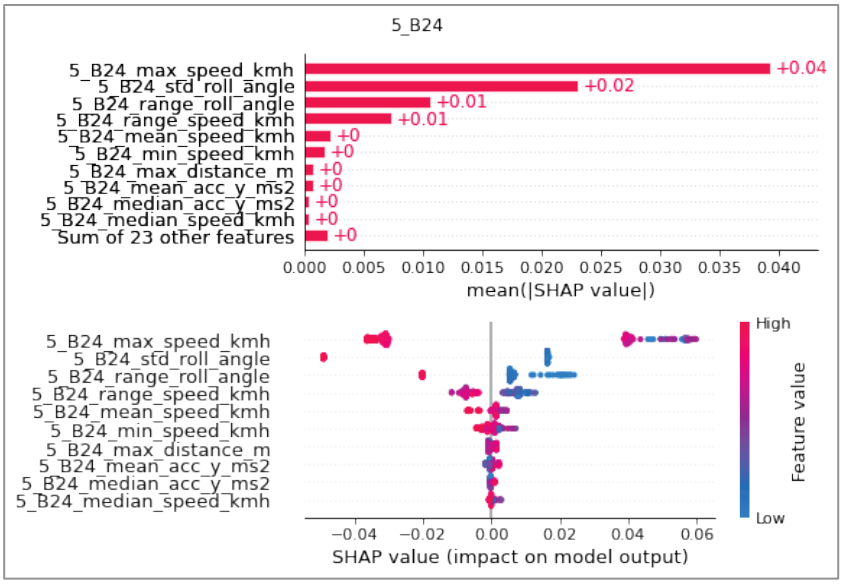


Abbildung 100: Abschnitt „B24 bis B25“

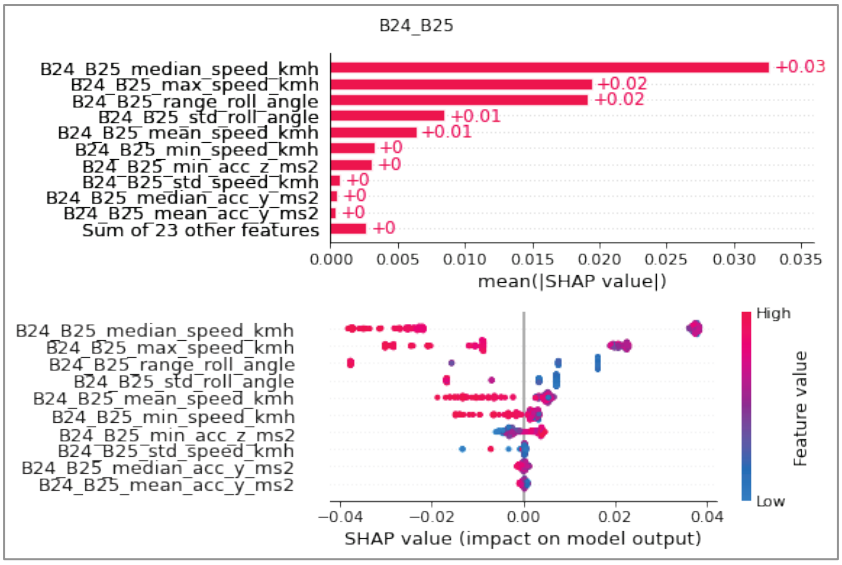
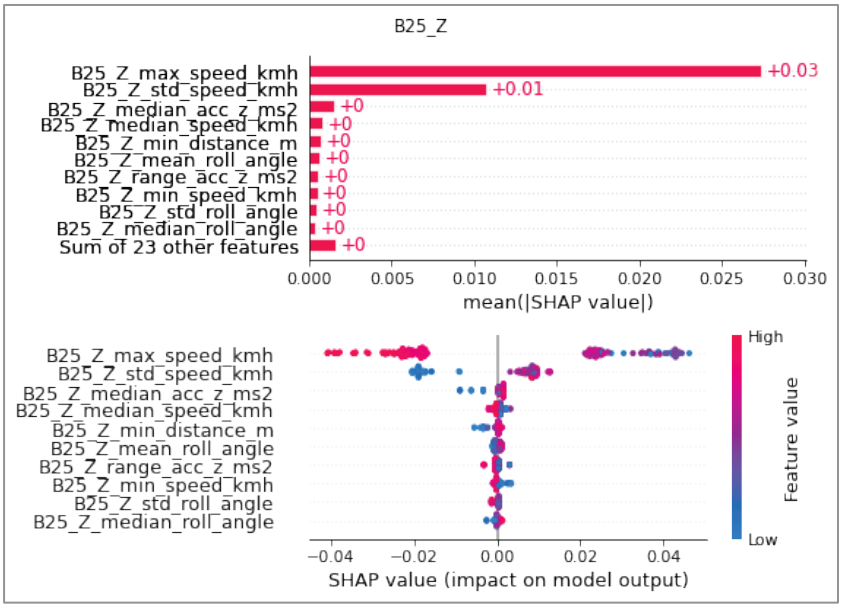


Abbildung 101: Abschnitt „B25 bis Z“



Folgende Bände sind bisher in dieser Reihe erschienen:

Band 1 (2021)
Bähren, T. / Braka, D. / Burchard, P. / Cyron, S. / Demary, M. / Dragieva, M. / Eis, L. / Farid, A. T. / Gomes, D. / Hacker, M. / Kaiser, J. / Krüger, R. / Luu, S. / Maasjosthusmann, R. / Marks, A. / Pachocki, C. / Pongratz, M. / Schade, J. C. / Urban, P. / Walter, A. / Winter, V. / Yesilyurt, E. / Buchkremer, R.
Der Einsatz von grünem Tee und anderen Polyphenolen in der Medizin – eine Big-Data-Analyse der medizinischen Fachliteratur
ISBN (Print) 978-3-89275-119-9 – ISBN (eBook) 978-3-89275-120-5
ISSN (Print) 2699-562X – ISSN (eBook) 2699-5638

Band 2 (2023)
Hamacher, K. / Blach, M. / Kozlik, J. / Muster, F. / Nöllenburg, P.-P. / Ohletz, J.-H. / Franken, G. / Hernes, D. / Hinterding, M. / Höveler, P. / Huppertz, M. / Leppkes, N. / Lopez Rodriguez, A. / Maucy, K. / Petrov, A. / Schäfer, D. / Schneider, R. / Spiegel, B. / Stecker, R. / Steinmann, P. / Tembrink, C. / Titze-Wolter, P. / Vishnyakova, L. / Zimmermann, J. / Buchkremer, R.
Analyse sensorischer E-Commerce-Elemente mittels Big-Data-Methoden und Künstlicher Intelligenz – Automatisierung sensorischer Bewertungen von E-Commerce- und Social-Media-Plattformen auf Basis des Online Sensory Marketing Index
ISSN (Print) 2699-562X – ISSN (eBook) 2699-5638
ISBN (Print) 978-3-89275-320-9 – ISBN (eBook) 978-3-89275-321-6

Band 3 (2025)
Hedfeld, P.
Implicit Decision Voting Made by Humans as Normative and Implementable Rules with the Help of Language Models
ISSN (Print) 2699-562X – ISSN (eBook) 2699-5638
ISBN (Print) 978-3-89275-394-0 – ISBN (eBook) 978-3-89275-395-7

Forschungsstark und praxisnah

FOM. Die Hochschule. Für Berufstätige.

FOM Hochschulzentrum Düsseldorf

Rund 45.000 Studierende, mehr als 20 Forschungseinrichtungen und 500 Veröffentlichungen im Jahr – damit zählt die FOM zu den größten und forschungsstärksten Hochschulen Europas. Initiiert durch die Stiftung für internationale Bildung und Wissenschaft folgt sie einem klaren Bildungsauftrag: Berufstätige und Abiturienten durch qualitativ hochwertige und bezahlbare Studiengänge akademisch zu qualifizieren. Als gemeinnützige Hochschule ist die FOM nicht gewinnorientiert, sondern reinvestiert sämtliche Gewinne – unter anderem in die Lehre und Forschung.

Die FOM ist staatlich anerkannt und bietet mehr als 60 praxisorientierte Bachelor- und Master- Studiengänge an. Studiert wird im Campus-Studium+ mit Vorlesungen im Hörsaal und virtuellen Anteilen oder komplett ortsunabhängig im Digitalen Live-Studium.

Lehrende und Studierende forschen an der FOM in einem großen Forschungsbereich aus hochschuleigenen Instituten und KompetenzCentren. Dort werden anwendungsorientierte Lösungen für betriebliche und gesellschaftliche Problemstellungen generiert. Aktuelle Forschungsergebnisse fließen unmittelbar in die Lehre ein und kommen so den Unternehmen und der Wirtschaft insgesamt zugute.

Zudem fördert die FOM grenzüberschreitende Projekte und Partnerschaften im europäischen und internationalen Forschungsraum. Durch Publikationen, über Fachtagungen, wissenschaftliche Konferenzen und Vortragsaktivitäten wird der Transfer der Forschungs- und Entwicklungsergebnisse in Wissenschaft und Wirtschaft sichergestellt.

Alle Institute und KompetenzCentren unter
fom.de/forschung

FOM Hochschule

ISBN (Print) 978-3-89275-396-4 ISSN (Print) 2699-562X
ISBN (eBook) 978-3-89275-397-1 ISSN (eBook) 2699-5638

Institut für IT-Management & Digitalisierung
der FOM University of Applied Sciences

FOM Hochschule

Mit rund 45.000 Studierenden ist die FOM eine der größten Hochschulen Europas und führt seit 1993 Studiengänge für Berufstätige durch, die einen staatlich und international anerkannten Hochschulabschluss (Bachelor/Master) erlangen wollen.

Die FOM ist der anwendungsorientierten Forschung verpflichtet und verfolgt das Ziel, adaptionsfähige Lösungen für betriebliche bzw. wirtschaftsnahe oder gesellschaftliche Problemstellungen zu generieren. Dabei spielt die Verzahnung von Forschung und Lehre eine große Rolle: Kongruent zu den Masterprogrammen sind Institute und KompetenzCentren gegründet worden. Sie geben der Hochschule ein fachliches Profil und eröffnen sowohl Wissenschaftlerinnen und Wissenschaftlern als auch engagierten Studierenden die Gelegenheit, sich aktiv in den Forschungsdiskurs einzubringen.

Weitere Informationen finden Sie unter **fom.de**

ifid

Das ifid Institut für IT-Management & Digitalisierung bündelt Kompetenzen in den Forschungsbereichen Künstliche Intelligenz (KI), Systemwissenschaften, IT-Management und digitale Transformation.

Die Aufgaben des Instituts umfassen Forschung und Entwicklung, Wissenstransfer und Innovationsförderung an der Schnittstelle von Wissenschaft und Praxis. Auch der Transfer von Forschungserkenntnissen in die Lehre spielt eine große Rolle.
Um diese Aufgaben zu erfüllen, setzt die Forschergruppe auf den Einsatz modernster Big Data-Architekturen und KI-Analysesysteme. Es bestehen Kooperationen mit den großen Technologie-Unternehmen und Instituten der Branche.
Die Wissenschaftlerinnen und Wissenschaftler beschäftigen sich insbesondere mit folgenden Feldern:

- Künstliche Intelligenz / Machine Learning / Data Science / Big Data
- Natural Language Processing (NLP)
- Enterprise Architekturen (insbesondere Big Data)
- Einsatz von Blockchain-Technologien
- Digitalisierung von Prozessen
- Integration der Forschung in die Lehre

Weitere Informationen finden Sie unter **fom-ifid.de**