Methodenbaukasten zur Quantifizierung der statistischen Güte und deren Sensitivität von Last- und Verschleißanalysen mit einem Beispiel im Kontext alternativer Antriebskonzepte

Dissertation

zur Erlangung des akademischen Grades

Dr. rer. nat.

an der Fakultät Statistik der Technischen Universität Dortmund

technische universität
dortmund

Thomas Lehmann

Erstgutachter:	Prof. Dr. Claus Weihs
Zweitgutachterin:	Prof. Dr. Christine Müller
Tag der mündlichen Prüfung:	09.10.2020

Dortmund 2020

Meinen Eltern.

Vorwort

An dieser Stelle möchte ich allen Personen danken, die mich auf unterschiedliche Art und Weise dabei unterstützt haben, diese Arbeit fertigzustellen.
Mein besonderer Dank gilt zunächst Herrn Prof. Dr. Weihs für die wissenschaftliche Betreuung dieser Arbeit und die fachliche und persönliche Unterstützung seit vielen Jahren. Frau Prof. Dr. Müller danke ich für die sehr hilfsbereite Betreuung als Zweitgutachterin.
Für die Unterstützung bei der programmierseitigen Umsetzung des im Rahmen dieser Arbeit entwickelten statistischen Methodenbaukastens gilt mein Dank Herrn Dr. Jürgen Schwarz.
Mein Dank gebührt weiterhin der Daimler AG, welche diese Dissertation in Zusammenarbeit mit der Technischen Universität Dortmund ermöglichte. Herrn Dr. Jörg Keller danke ich für die persönliche Betreuung und für die Möglichkeit des Einstiegs in den Konzern 2012. Herrn Dr. Matthias Grabert danke ich ebenso für die persönliche und fachliche Beratung. Herrn Dr. Wolfgang Neher danke ich, dass er mir während meiner Zeit als Doktorand viele Türen geöffnet und mir jegliche Form der persönlichen und fachlichen Weiterentwicklung ermöglicht hat. Für die Abnahme dieser Arbeit aus Sicht der Daimler AG möchte ich besonders Herrn Dr. Tobias Handschuh danken. Vielen Dank an Euch alle, dass Ihr mich bis zum heutigen Tage bereichsübergreifend beratet und unterstützt!
Dank gebührt ebenfalls allen studentischen Kräften, welche ich im Rahmen dieser Dissertation in Form von Studien-, Bachelor-, Master- und Diplomarbeiten betreuen durfte.
Bei meinen Eltern Gerd und Marlis bedanke ich mich für die bedingungslose Unterstützung auf meinem Weg. Euch widme ich diese Arbeit.
Nicht zuletzt möchte ich mich bei meiner Partnerin Veronica bedanken, die auf viel gemeinsame Zeit verzichten musste und mich gerade in herausfordernden Zeiten extrem gestärkt hat.

Einleitung

Die vorliegende Arbeit umfasst die Entwicklung und Beschreibung eines statistischen Methodenbaukastens um Last- und Verschleißanalysen prozessual durchführen zu können. Der methodische Fokus liegt auf der Quantifizierung und Sensitivität der Güte bzw. Unsicherheit auf den einzelnen Analysestufen. Die Arbeit wurde im Rahmen einer Industriepromotion bei der Daimler AG erstellt. Es findet eine Einordnung des Themas in den Kontext alternativer Antriebe statt. In diesem Umfeld ist das Analysebeispiel angesiedelt.

Auf der ersten Analysestufe des Prozesses sollen Gruppen in Belastungsdaten identifiziert werden. Dazu werden verschiedene Clusterverfahren und deren Gütemaße hergeleitet. Auf der zweiten Stufe wird für die Gruppen ihr Verschleißverhalten über die Zeit prognostiziert. Verschiedene lineare und nichtlineare Verfahren inkl. Maße für die Beschreibung der Güte werden definiert. Es findet eine Bewertung der Sensitivität der Güte auf beiden Stufen statt.

Der Methodenbaukasten beinhaltet einen iterativen Prozess, in dem in jeder Iteration sowohl das Clustering als auch die Prognose durchgeführt wird. So kann zum einen in jedem Schritt die Güte des jeweiligen Verfahrens und zum anderen die Sensitivität der Güte bzw. Unsicherheit der Verfahren/Modelle über mehrere Iterationen quantifiziert und bewertet werden.

Abschließend wird der entwickelte Prozess an einem Datenbeispiel aus dem Automobil-Umfeld im Kontext alternativer Antriebe erprobt. Die Qualität und Stabilität von Ergebnissen aus Clustering- bzw. Prognoseverfahren lassen sich mithilfe der prozessual durchgeführten Analyse nun bewerten.

Inhaltsverzeichnis

Abkürzungsverzeichnis

A	Ampere
ABC	Approximate Bootstrap Confidence Interval
Ah	Amperestunde
AIC	Akaike Information Criterion
BC	Bias-Corrected Bootstrap Confidence Interval
BCa	Bias-Corrected Accelerated Bootstrap Confidence Interval
BIC	Bayesian Information Criterion
BLK	Belastungskollektiv
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
EAC	Evidence Accumulation Clustering
EM	Expectation Maximization
HLM	Hierarchical Linear Modeling
HV	Hochvolt
ICC	Intraclass Correlation
km	Kilometer
km/h	Kilometer pro Stunde
KQ	Kleinste Quadrate
kW	Kilowatt
MAD	Median Absolute Deviation
OLS	Ordinary Least Squares
PPM	Parts per Million
s	Sekunde

SoC	State of Charge
SSE	Error Sum of Squares
SSR	Sum of Squared Residuals
SST	Total Sum of Squares
t	Zeit
V	Volt

1 Motivation und Ziel

> Statistics may be defined as *„a body of methods for making wise decisions in the face of uncertainty.“*
>
> W.A. Wallis (1912-1998), US-amerikanischer Wirtschaftswissenschaftler und Statistiker

1.1 Zielsetzung

Elektrische Antriebe gelten zur Zeit als die vielversprechendsten Konzepte für die Zukunft der Automobilindustrie. Nahezu alle führenden Hersteller bringen im Laufe der nächsten Jahre mehrere elektrifizierte Modelle auf den Markt. Auch wenn die Entwicklung von elektrischen Antrieben nicht neu ist (als Geburtsstunde des Elektroautos gilt eine Fahrt von Robert Anderson in Aberdeen im Jahr 1839 (vgl. Zeit Online (2009))), bleibt eine Reihe von Fragestellungen noch ungeklärt, da es bisher wenig Erfahrungen in Seriennutzung gibt. Hier setzt diese Arbeit an.
Sie soll einen Beitrag leisten, das Belastungs- und Verschleißverhalten im Kontext alternativer Antriebe prozessual analysieren und verstehen zu können. Es wird ein statistischer Methodenbaukasten und Analyseprozess entwickelt, der einen Leitfaden darstellt, um die jeweils bestmöglichen Analysen und Prognosen der Belastung und des Verschleißes durchführen zu können, abhängig von den zur Verfügung stehenden Daten. Dazu werden verschiedene statistische Methoden hergeleitet, um zunächst Gruppen in den Daten identifizieren und diese dann bzgl. ihres Verschleißverhaltens prognostizieren zu können.
Ziel dieser Arbeit ist es nicht, eine bestimmte Fahrzeugflotte zu analysieren, sondern die Beschreibung einer prozessual durchzuführenden Analyse.
Der methodische Fokus in dieser Arbeit liegt zum einen auf der Beschreibung und der Sensitivität der Unsicherheit der Analysen und Ergebnisse auf den jeweiligen Analysestufen, zum anderen auf der Entwicklung einer Möglichkeit, die Unsicherheit und dessen Sensitivität über mehrere Analysestufen hinweg quantifizieren zu können. Dieser Prozess wird beschrieben und an einem Beispiel aus dem Umfeld der alternativen Antriebe durchgeführt.

1.2 Aufbau der Arbeit

Die Arbeit ist wie folgt aufgebaut:
In diesem Abschnitt, Kapitel 1, gibt es eine allgemeine Einführung in die Welt der Elektromobilität und die grundlegenden technischen Kenntnisse zu Hochvolt-Batteriesystemen und alternativen Antrieben werden vermittelt. Das Forschungsthema wird hergeleitet und eingeordnet.
In Kapitel 2 werden die Daten beschrieben, anhand derer am Ende der Arbeit der beschriebene Prozess zur Quantifizierung der statistischen Unsicherheit und dessen Sensitivität über die einzelnen Analyseschritte hinweg erprobt wird.
Kapitel 3 leitet den statistischen Methodenbaukasten her und beschreibt diesen als Prozessbild. Alle verwendeten statistischen Methoden werden definiert. Dies beinhaltet verschiedene Clusterverfahren für die Lastanalyse und lineare bzw. nichtlineare Regressionsmodelle zur Verschleißprognose. Für beide Ebenen werden die entsprechenden Maße zur Beschreibung der Güte hergeleitet. Der Gesamt-Prozess beschreibt die Verknüpfung der beiden Methoden, die Quantifizierung sowie die Sensitivität der statistischen Güte.
In Kapitel 4 wird der entwickelte Prozess anhand eines Beispiels aus dem Umfeld der alternativen Antriebe in der Automobilindustrie angewendet.
Der letzte Abschnitt, Kapitel 5, fasst die wichtigsten Erkenntnisse dieser Arbeit zusammen und stellt den Mehrwert, aber auch die Grenzen für zukünftige Analysen dar. Des weiteren gibt es einen Ausblick, wie der entwickelte Analyseprozess in Zukunft verbessert, erweitert oder performanter dargestellt werden kann.

1.3 Umfeld und Einordnung der Arbeit

1.3.1 Elektromobilität

Die individuelle Mobilität wird sich in Zukunft maßgeblich von der aktuellen unterscheiden. Begrenzte Ressourcen und Regularien ergänzend zu globalen Trends wie Sicherheit und Verbrauch führen zu einem Umdenken seitens Regierungen und Automobilherstellern (vgl. u.a. Proff et al. (2014), Kap. 1). Nachweisbare Klimaveränderungen, ausgelöst u.a. durch industriellen Kohlendioxidausstoß haben zu drastischen Senkungsmaßnahmen auch innerhalb der Automobilindustrie geführt. Unter ihrem Dachverband ACEA (Association des Constructeurs Européens d´Automobiles) hatten sich die europäischen Automobilhersteller verpflichtet, den Kohlendioxidausstoß für die gesamte Fahrzeugflotte bis zum Jahre 2008 auf 140 Gramm CO_2 pro km zu senken. Mittlerweile wurde eine Senkung durch die Europäische Kommission auf 95 Gramm CO_2 pro km gesetzlich festgelegt (vgl. Stan (2015), Kap. 1, und Wallentowitz et al. (2010), Kap. 2.1.1 (in Abbildung 1 dargestellt)).

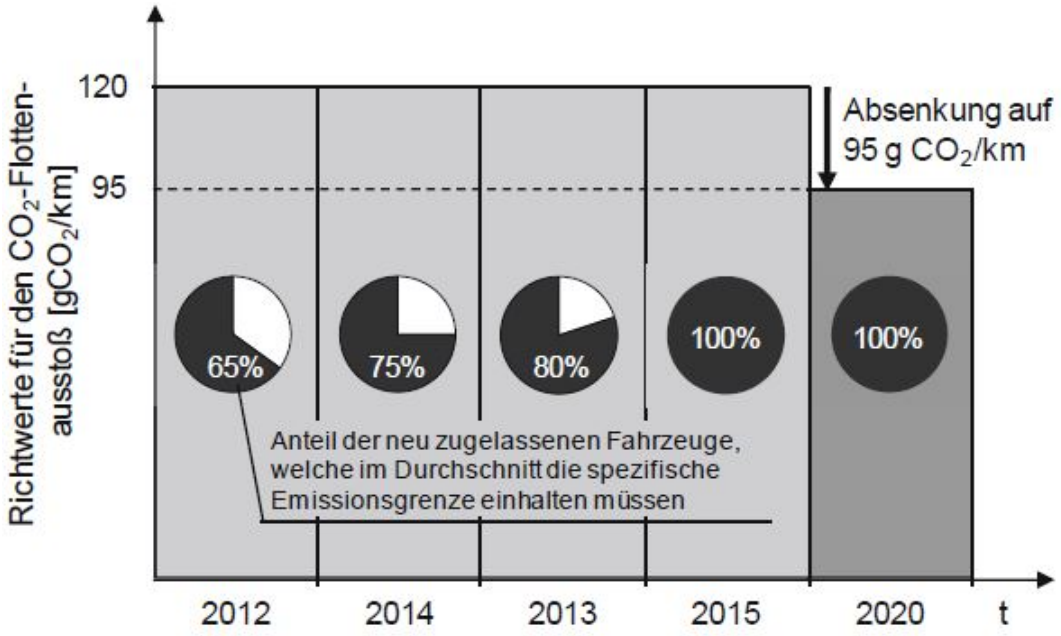

Abbildung 1: Stufenplan von CO_2-Emissionen der Europäischen Union (vgl. Wallentowitz et al. (2010), Kap. 2.1.1).

Das Konzept der Elektromobilität ist dabei ein wichtiger Bestandteil (vgl. Springer (2012)) . Die Bundesregierung unter Bundeskanzlerin Angela Merkel hat es sich zum Ziel gesetzt, bis 2020 eine Million rein elektrisch betriebene Fahrzeuge auf

die deutschen Straßen zu bringen (vgl. FAZ (2013)); im April 2016 lag der aktuelle Stand der Zulassungen bei ca. 25.000 Fahrzeugen (vgl. Zeit Online (2016)). Die Automobilindustrie sieht sich mit der Herausforderung konfrontiert, emissionsarme Antriebskonzepte anzubieten und gleichzeitig die aus aktuellen Konzepten bekannten Bedürfnisse des Kunden hinsichtlich Kosten, Reichweite, Komfort und Fahrgefühl weiterhin zu erfüllen (vgl. Springer (2012)). Ebenso gewinnt der Faktor der Umweltverträglichkeit bei den Kunden zunehmend an Stellenwert. Potenzielle Lösungen müssen insbesondere das Spannungsfeld zwischen Anforderungen des Marktes seitens der Kunden und den technologischen Rahmenbedingungen seitens des Herstellers überbrücken. Angefangen bei der Optimierung der konventionellen Antriebe über verschiedene Hybridsysteme führt der Weg fast zwangsläufig hin zu reinen Elektroantrieben (vgl. Wallentowitz et al. (2010), Kap. 3). Die entscheidenden Treiber für zunehmende Elektrifizierung des Antriebsstrangs sind in Abbildung 2 dargestellt.

Abbildung 2: Treiber für zunehmende Elektrifizierung des Antriebsstrangs (vgl. Freialdenhoven (2009)).

1.3.2 Technische Grundlagen

Fahrzeugtyp

Hybridsysteme stellen, wie beschrieben, einen Zwischenschritt zur Elektrifizierung des Antriebs dar. Ein Hybridfahrzeug (Hybrid, griech.: „von zweierlei Herkunft“) besitzt per Definition mindestens zwei Energiewandler und zwei eingebaute Energiespeichersysteme, um das Fahrzeug anzutreiben. Die Energiewandler sind in dem Fall z.B. Otto-, Diesel- oder Elektromotoren mit den jeweiligen Energiespeichersystemen wie Batterien oder Kraftstofftanks (vgl. Wallentowitz et al. (2010), Kap. 4). Der Gedanke von Hybriden lässt sich so formulieren, dass zwei in ihren Eigenschaften konträre Antriebe so kombiniert werden sollen, dass Nachteile des einen Systems durch Vorteile des anderen Systems kompensiert werden (vgl. Gies (2008)). Die Nachteile der konventionellen Antriebe sind vor allem der schlechte Wirkungsgrad im Teillastbereich und bei Lastwechselzuständen, die Abgasemissionen und die erwähnte Abhängigkeit von fossilen Rohstoffen. Genau dort wiederum zeigen sich elektrische Antriebe von Vorteil, sie weisen über breite Betriebsbereiche einen vom Lastzustand unabhängigen und damit konstanten Wirkungsgrad auf und laufen lokal emissionsfrei. Nach derzeitigem Entwicklungsstand der elektrischen Antriebe ist es noch sinnvoll, diese mit Verbrennungskraftmaschinen zu kombinieren, um z.B. weiterhin Reichweiten zu ermöglichen, die rein elektrisch noch nicht zu erreichen sind (vgl. Wallentowitz et al. (2010), Kap. 5).

Hybridkonzepte lassen sich anhand der Ausprägung der elektrischen Komponente in die vier Klassen Mikrohybrid, Mildhybrid, Vollhybrid und Plug-In Hybrid unterteilen.

Beim Mikrohybrid wird die Lichtmaschine verstärkt und beim Bremsen stärker belastet, wodurch die Batterie geladen wird. Parallel dazu wird meist auch der Anlasser verstärkt, um eine Stopp-Start Technologie anbieten zu können, die im Stand den Motor ausschaltet und ihn bei Einlegen eines Ganges wieder startet.

Bei Mildhybriden wird zusätzlich eine Elektromaschine installiert, die einen Großteil der Bremsenergie zurückgewinnen soll.

Für Fahrzeuge der Kategorie Vollhybrid ist die Installation einer Elektromaschine mit deutlich mehr Leistung erforderlich. Beim Beschleunigen unterstützt der Elektromotor den konventionellen Antrieb. Je nach Größe bzw. Kapazität der HV-

Batterie kann das Fahrzeug wenige Kilometer rein elektrisch fahren.
Stand der Technik bei den Hybridfahrzeugen sind aktuell Plug-In Hybride, nach Definition Vollhybride mit vergrößertem Energiespeicher (Batterie)(vgl. Lienkamp (2012), Kap. 4.1). Der Energiespeicher kann hier extern über einen Stecker (engl.: plug) geladen werden. Plug-In Hybride ermöglichen einen höheren elektrischen Aktionsradius und stellen eine Art Brückentechnologie zwischen Voll-Hybrid und Elektrofahrzeug dar, die die Vorteile beider Systeme kombiniert (vgl. Reif et al. (2012), Kap. 2.4.4.2).
Elektrofahrzeuge vereinen einige Antriebskonzepte mit dem gemeinsamen Merkmal des Elektromotors als einzigem Energiewandler. Darunter fallen als bekannteste Konzepte Brennstoffzellen- und Batterie-Elektrofahrzeuge. Da diese zur Zeit als das dominierende Antriebskonzept der Zukunft gehandelt werden, werden in dieser Arbeit lediglich die Batterie-Elektrofahrzeuge näher beschrieben. Insbesondere durch seine Fähigkeit, einen emissionsfreien Betrieb durch einen universell herstellbaren Energieträger zu bieten, hebt sich der Elektroantrieb von allen anderen Antriebskonzepten ab. Die Antriebstechnik führt zu einer vollständigen Substitution des Verbrennungsmotors, sodass auch ein Umdenken in anderen Fahrzeugbereichen erforderlich ist, die in konventionellen Antrieben auf den Verbrennungsmotor angewiesen sind (z.B. Heizung). Elektrofahrzeuge zeichnen sich durch ihren relativ einfachen Systemaufbau aus, da der Antriebsstrang im Wesentlichen nur noch aus den Komponenten Energiespeicher, Elektromotor(en) und Steuergeräten besteht (vgl. Wallentowitz et al. (2010), Kap. 4).
Die elektrische Energie kann, wie oben beschrieben, in zwei Formen chemisch gespeichert werden: in einem Akkumulator oder in Form von Wasserstoff (vgl. Lienkamp (2012), Kap. 4.2). In dieser Arbeit wird auf die HV-Batterie im speziellen im nächsten Abschnitt näher eingegangen.

Die HV-Batterie

Zunächst soll die Bezeichnung Hochvolt-„Batterie“ hergeleitet werden. „Akkumulator“ wäre der korrekte Begriff für einen chemischen Speicher von elektrischer Energie; jedoch wurde bei der Übersetzung des englischen „battery“ auf den Begriff „Batterie“ zurückgegriffen, die eigentlich, im Gegensatz zum Akkumulator, nicht wieder aufladbar ist (vgl. Lienkamp (2012), Kap. 4.3). Im Folgenden wird von der „HV-Batterie“ als Speichermedium gesprochen.
Im Vergleich zur klassischen Blei-Fahrzeug-Starterbatterie, die für nahezu 100 Jahre als einziger Bordnetz-Energiespeicher ihren Platz im Fahrzeug hatte, müssen HV-Batterien für Hybrid- oder Elektrofahrzeuge deutlich leistungsfähiger sein (vgl. Reif et al. (2012), Kap. 3.4.1). Schon vor ca. 100 Jahren sind die damals dominierenden Elektrofahrzeuge an der Leistungsfähigkeit gescheitert; die Batterien waren zu schwer, zu groß und hatten zu geringe Speicherkapazitäten. Jahrzentelange Entwicklung führte dann zu Alternativen wie Nickel-Cadmium-, Nickel-Metall-Hybrid- oder schließlich der Lithium-Ionen-Batterie. Diese stellt sich mittlerweile als Stand der Technik dar (vgl. Sauer, Kowal (2012)). Prinzipiell haben alle elektrochemischen Batterien den gleichen Aufbau. Sie bestehen aus zwei, aus unterschiedlichem Material bestehenden, Elektroden, die durch einen Elektrolyten miteinander verbunden sind. Beide Elektroden haben jeweils ein Potenzial gegenüber dem Elektrolyten, die Differenz der beiden Potenziale wiederum ergibt das Zellpotenzial. Im Betrieb einer Batteriezelle werden Elektronen in die Elektroden ein- oder ausgeleitet (vgl. Sauer, Kowal (2012)). Die Anode als negative Elektrode oxidiert beim Entladen und gibt somit Elektronen ab, die Kathode (positive Elektrode) nimmt Elektronen auf. Ist die Batterie geladen, befinden sich die Elektronen in der Anode, während des Entladens wandern diese über den Elektrolyten in Richtung Kathode (vgl. Brill (2012)). In Abbildung 3 ist dieser Vorgang am Beispiel einer Lithium-Ionen-Zelle mit einer Lithium-Metalloxid-Kathode und einer typisch vewendeten Graphitanode dargestellt. In der Forschung gibt es bzgl. der Materialen vielseitige Ansätze (vgl. z.B. Sedelmeier (2015)).

In der Regel hat eine einzige Batteriezelle zu wenig Kapazität oder zu wenig Spannung, um eine ausreichende Versorgung des Verbrauchers sicherzustellen. Die Ka-

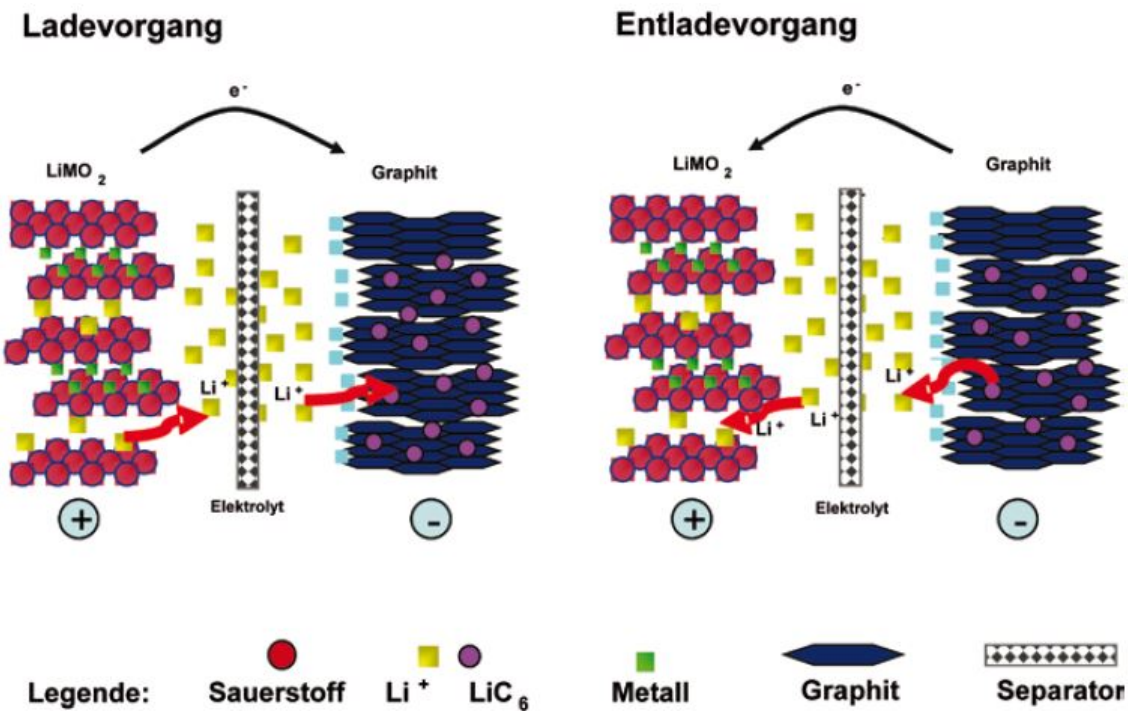

Abbildung 3: Reaktionsmechanismus einer Lithium-Ionen-Zelle (vgl. Lamm et al. (2009)).

pazität beschreibt hier die Ladungsmenge, die eine Batteriezelle aufnehmen kann. Sie wird in Amperestunden Ah angegeben, die Spannung wiederum in Volt V. Die einzelnen Batteriezellen können in der Batterie entweder seriell oder parallel geschaltet werden. Im Falle einer Serienschaltung bleibt die Kapazität unverändert, die Gesamtspannung ergibt sich als Multiplikation der Einzelzellspannungen mit der Anzahl der Zellen. Um die Kapazität zu erhöhen, müssen die Zellen parallel geschaltet werden. In der Praxis kommen oft Kombinationen aus beiden Herangehensweisen zum Einsatz (vgl. Sauer, Kowal (2012)).

Da in dieser Arbeit ein Methodenbaukasten zur Verschleiß- und Ausfallanalyse mit einem abschließenden Analysebeispiel anhand von HV-Batterien vorgestellt wird, wird an dieser Stelle etwas näher auf die dafür relevanten Einflussgrößen (im Speziellen für Lithium-Ionen-Zellen) eingegangen. Verschleiß bzw. Alterung der HV-Batteriezellen nimmt der Kunde z.B. durch Leistungsverlust oder Reichweitenverringerung wahr. Diese begründen sich auf den Kapazitätsverlust der Zellen, der in dieser Arbeit als Messgröße der Alterung betrachtet wird. Jeder Alterungseffekt resultiert aus Materialveränderungen der Zelle. Diese wiederum hängen von den Alterungsfaktoren Zellchemie, Produktionsqualität und den Konditionen ab, unter

denen sie betrieben wird (Stressfaktoren) (vgl. Andre (2014) und Sedelmaier (2015)). Als erster Stressfaktor ist die Betriebstemperatur der HV-Batterie zu nennen. Sowohl sehr niedrige (<0°C) als auch sehr hohe (oberhalb der definierten Spezifikation) Zelltemperaturen sind zu vermeiden (vgl. Ecker, Sauer (2013) und Kreibich (2014)). Zwar erhöht sich bei hohen Temperaturen zunächst die Leistungsfähigkeit der Batteriezellen, da die Materialien reaktionsfreudiger sind, es können dabei allerdings auch Nebeneffekte auftreten, die sich negativ auf die Lebensdauer auswirken. Als optimales Betriebsfenster der Temperatur wird der Bereich 20-40°C angesehen (vgl. Terzimehic (2012)).

Eine weitere Einflussgröße auf die Alterung der Batteriezellen und somit auf die Alterung und den Verschleiß der HV-Batterie ist der Ladezustand (State of Charge). Große SoC-Hübe, d.h. hohe Ladungen/Entladungen und insbesondere Tiefentladungen wirken sich negativ auf die Lebenserwartung der Zellen aus. Ebenfalls sollte es vermieden werden, HV-Batterien bei hohen SoC-Ständen zu lagern. Hoher Energieinhalt geht einher mit gesteigerter Reaktionsfreudigkeit der Materialien, wodurch die Elektroden schneller degradieren (vgl. Terzimehic (2012)).

Fließende Ströme gelten als weiterer Hauptfaktor des Verschleißes. Erwartungsgemäß zeichnen sich sowohl hohe Lade- als auch Entladeströme als negative Einflüsse auf die Lebensdauer der Batterie aus (vgl. Terzimehic (2012) und Kreibich (2014)).

Der zeitliche Verlauf der Alterung, gemessen an dem Kapazitätsverlust der Zellen, hängt stark von den Betriebskonditionen ab (vgl. Sedelmaier (2015)) und kann gesondert nach Alterung während der Lagerung (kalendarische Alterung) und Alterung während des Betriebes (zyklische Alterung) betrachtet werden. Über den zeitlichen Verlauf der Alterung gibt es verschiedene Annahmen und viele akademische Arbeiten sowie veröffentlichte, wissenschaftliche Paper. Da der Fokus dieser Arbeit auf den zur Verfügung stehenden statistischen Methoden zur Verschleiß- und Ausfallprognose liegt, soll auf die Variation des zeitlichen Verlaufs der Alterung hier nicht näher eingegangen werden. Zur Untersuchung und Beschreibung des Alterungsverhaltens von Lithium-Ionen-Zellen siehe u.a. Belt et al. (2003), Broussely et al. (2005), Ecker et al. (2012), Herb (2010), Peterson et al. (2010), Sarre et al. (2004) und Vetter et al. (2005).

2 Datengrundlage

Im Folgenden wird die Charakteristik der Daten für die Last- und Verschleißanalyse dargestellt. Dabei wird auf die Beschaffung, Aufzeichnung und Zielsetzung eingegangen.

2.1 Lastanalyse - Belastungskollektive

Die Lastanalyse beschäftigt sich mit der auf die Komponente einwirkende Last. Die Einflussgrößen werden in diesem Fall als bereits bekannt angenommen und durch Expertenschätzung vorgegeben. Dabei lässt sich die Belastung von HV-Batteriesystemen durch sogenannte, im Fahrzeug aufgezeichnete, Belastungskollektive (BLKs) beschreiben.
Belastungskollektive haben ihren Ursprung in der Betriebsfestigkeit (vgl. Bergmeir et. al (2015)) und dienen in der Automobilbranche zu der Gewinnung von Erkenntnissen über Nutzung, Belastung und Verschleiß der Fahrzeuge und Komponenten. Anhand dieser kann z.B. überprüft werden, ob Bauteile entsprechend der realen Kundennutzung ausgelegt oder über- bzw. unterdimensioniert wurden (vgl. Bauersachs (2015)). Die aufgezeichneten Daten dienen der Verbesserung der Transparenz des Feldgeschehens über die Einsatzarten sowie die sonstigen Betriebsbedingungen. Motor, Batterie und sonstige Komponenten können so optimiert werden, indem aktuelle Aufzeichnungen in die Entwicklung zurückgespiegelt und berücksichtigt werden. Des weiteren ermöglicht die Kenntnis der spezifischen Belastungssituationen bei häufig auftretenden Fehlern oder Schäden eine Ursachenanalyse und Risikoklassifizierung (vgl. Lehmann (2013)).
Belastungskollektive sind stark komprimierte Nutzungs- und Belastungsarten, die in Serien-Steuergeräten aufgezeichnet und weltweit bei der Diagnose des Fahrzeugs in der Werkstatt oder bei Zustimmung des Kunden regelmäßig über die Telediagnose ausgelesen werden. Die Zeiträume zwischen den Werkstattaufenthalten sind in der Regel unterschiedlich, ein Übertrag der Daten über Telediagnose findet regelmäßig (z.B. einmal pro Monat) statt. Die statistische Auswertung der BLKs z.B. für bestimmte Baureihen, Motoren, Märkte oder eben HV-Batterietypen dient dabei der Abbildung des realen Kundenverhaltens. Im präventiven Sinne kann diese Art der

Auswertung für die Auslegung und Erprobung von Komponenten und die Charakterisierung des Fahr- und Nutzungsverhaltens genutzt werden. Automobilhersteller und Zulieferer sind so in der Lage, ihre Produkte im Entwicklungsprozess noch besser auf die Nutzung auszulegen, ohne dabei auf vereinfachte Annahmen zurückgreifen zu müssen. Der Kunde profitiert durch neue Dienstleistungen wie Ferndiagnose oder präventive Reparaturen (vgl. Köttermann et al. (2015)). Betrachtet man eine reaktive Herangehensweise, liegen die Anwendungsfälle z.B. in der Anlauf-und Feldabsicherung bzw. Fehlerursachenanalyse. Die Daten in den Kundenfahrzeugen werden in Form von Häufigkeitsverteilungen über bestimmte Intervalle von Wertebereichen (im Folgenden als Klassen bezeichnet) aufgezeichnet. Belastungskollektive können in unterschiedlichen Formaten vorliegen. Im einfachsten Fall wird nur ein einziger Wert aufgezeichnet und hochgezählt, so z.B. der Kilometerstand. Ein- bzw. zweidimensionale Häufigkeitsverteilungen werden in zuvor festgelegten Klassen aufgezeichnet. Ein Beispiel für eine zweidimensionale Häufigkeitsverteilung ist in diesem Kontext das Belastungskollektiv Batterietemperatur über den Ladezustand (State of Charge = SoC) der Batterie. Die im Bereich der ein- bzw. zweidimensionalen Belastungskollektive angewandte Methodik wird als Momentanwertzählung bezeichnet. Dabei wird in konstanten Zeitabständen jeweils ein Wert ermittelt und in der entsprechenden Klasse des Belastungskollektivs als Ereignis erfasst (vgl. Kreibich (2014)). Es existieren insgesamt fünf verschiedene Zählverfahren für BLKs: Momentanwertzählung über die Zeit, Momentanwertzählung über km, Ereigniszählung, Integral- und Rainflowzählung. Zeitliche Abfolgen und Querbeziehungen zwischen mehreren BLKs können somit nicht erfasst werden. Aus Gründen des Datenschutzes gilt, dass keine Daten erfasst werden dürfen, die eine Rekonstruktion von personenbeziehbaren, singulären Ereignissen (z.B. eine Geschwindigkeitsüberschreitung) zu einem bestimmten Zeitpunkt/an einem bestimmten Ort erlauben. Zur Zeit gibt es ca. 140 verschiedene Belastungskollektive, die in den Steuergeräten aufgezeichnet werden können. Einen beträchtlichen Teil machen dabei die Kollektive aus den Alternativen Antrieben aus.

In Tabelle 1 sieht man die im Steuergerät durchgeführte Klassierung am Beispiel des zweidimensionalen Belastungskollektivs *Leistung über Geschwindigkeit des Fahrzeugs*. In diesem Fall werden sowohl die Leistung des Motors in kW, als auch die Geschwindigkeit in km/h in jeweils vier Klassen unterteilt. Daraus ergeben sich also

16 Felder in der Häufigkeitsmatrix. Die Werte liegen dann für jedes Fahrzeug vor, dessen entsprechendes Steuergerät in der Werkstatt mindestens einmal ausgelesen wurde. Die erfassten Daten beschränken sich auf nur wenige Kilobyte pro Fahrzeug, liefern aber einen enormen Beitrag zur Absicherung der Zuverlässigkeit neuer Produkte (vgl. Köttermann et al. (2015)).

Tabelle 1: Belastungskollektiv „Leistung - Geschwindigkeit“ eines Fahrzeugs.

Geschw. in km/h / Leistung in kW	≤ 0	$(0\ ;\ 50]$	$(50\ ;\ 100]$	> 100
≤ 0				
$[0\ ;\ 50)$				
$[50\ ;\ 100)$				
≥ 100				

2.2 Verschleißanalyse - Alterungsgröße

Die Datenaufzeichnung und -übertragung von Verschleiß- bzw. Alterungsgrößen unterscheidet sich nicht von den in 2.1 beschriebenen Belastungskollektiven. Sie werden onboard in den Steuergeräten der Fahrzeuge aufgezeichnet bzw. berechnet und bei der Diagnose in der Werkstatt oder durch regelmäßige Übertragung per Telediagnose ausgelesen.

Die Verschleißgröße, in dieser Arbeit die Zielgröße der Prognose, kann verschiedene Ausprägungen haben. Der Verschleiß der HV-Batterie wird oftmals über die noch zur Verfügung stehende Kapazität gemessen.

Werden beide Datensätze im selben Steuergerät gespeichert, so ist gewährleistet, dass bei einer Übertragung jeweils ein kompletter Analyse-Datensatz des Fahrzeugs vorliegt. Dieser beinhaltet dann sowohl die hier als Einflussgrößen beschriebenen Daten (Belastungskollektive zur Lastanalyse) als auch die Verschleißgröße als Zielvariable. Beobachtungseinheit ist jeweils ein einzelnes Fahrzeug.

3 Statistische Methoden und Prozess-Modell

In diesem Kapitel wird der Methodenbaukasten hergeleitet und als Prozessbild beschrieben. Es werden zunächst Clusteringverfahren und Regressionsmodelle als die hauptsächlich für den Methodenbaukasten verwendeten Verfahren vorgestellt. Im speziellen wird auf die im Datenbeispiel genutzten Verfahren Evidence Accumulation Clustering und hierarchisch lineare Modelle eingegangen. Methoden zur Beschreibung der Güte und Quantifizierung von deren Sensitivität werden beschrieben. Speziell auf die Verknüpfung der beiden Schritte (Clustering und Modellberechnung) durch gleichzeitiges, mehrfaches Ausführen und die Betrachtung der Verteilungen der Gütekriterien wird im Rahmen des Prozess-Modells eingegangen.

3.1 Clusterverfahren

An dieser Stelle wird das Clustering als Instrument zum Erkennen von Ähnlichkeitsstrukturen in Datensätzen vorgestellt.
In vielen Anwendungsbereichen (z.B. Medizin, Archäologie, Soziologie, Linguistik, Biologie, Wirtschaftswissenschaften) ist es von Interesse, wie ähnlich die Untersuchungsobjekte zueinander sind. Dabei ist oftmals das Ziel, Gruppen in den Daten zu finden, die untereinander möglichst homogen und gegenüber den anderen Gruppen möglichst heterogen sind (vgl. Backhaus et. al (2016), Teil II, Kap. 8.1).

3.1.1 Überblick

Ziel jedes Clusterverfahrens ist es, n Objekte in m Klassen $K_1, ..., K_m$ einzuteilen, um eine Klassifikation $K = \{K_1, \ldots, K_m\}$ zu erhalten. Dabei muss jede Klasse mindestens eines und höchstens alle dieser n Objekte enthalten. Objekte im gleichen Cluster sollen sich möglichst ähnlich sein, Objekte aus verschiedenen Clustern sollen möglichst unähnlich zueinander sein. Bevor dies geschieht, muss der Anwender sich situationsbedingt für einen Klassifikationstyp, ein Bewertungsmaß der Ähnlichkeit und ein Clusteranalyseverfahren entscheiden.
Klassifikationstypen werden in der Literatur in Überdeckungen, Partitionen und Hierarchien unterschieden (Hartung, Elpelt (2007), Kap. VII 1.). Dürfen Klassen sich in einer Klassifikation zwar überschneiden, aber eine Klasse nicht komplette

Teilmenge der anderen sein, so spricht man von einer **Überdeckung**. Für jedes Klassenpaar $K_i, K_j (i \neq j)$ aus einer Überdeckung gilt somit $K_i \cap K_j \notin \{K_i, K_j\}$. Die **Partition** als spezielle Art der Überdeckung verlangt, dass kein Objekt mehr als einer Klasse zugeordnet werden darf. Es gilt für jedes Klassenpaar $K_i, K_j (i \neq j)$: $K_i \cap K_j = \emptyset$. Als „feiner" Klassifikationstyp gelten Hierarchien. Die **Hierarchie** ist eine Folge von Partitionen und besitzt folgende Eigenschaften: Sei K_i eine Klasse aus einer Hierarchie K, so gilt für die Vereinigung aller echter Teilklassen $K_j \in K$: $\bigcup_{K_j \subsetneq K_i} K_j \in \{\emptyset, K_i\}$ und $K_i \cap K_j \in \{K_i, K_j, \emptyset\}$.

Die dann zu wählenden Distanz- bzw. Ähnlichkeitsmaße beschreiben die Heterogenität zwischen und die Homogenität innerhalb der Klassen. Für beide Bewertungen gibt es verschiedene Maße. An dieser Stelle werden nur die gängigsten aufgeführt, ausführlicher nachzulesen z.B. in Backhaus et. al (2016), Teil II, Kap. 8.2.2.2.2.

Die zu wählenden Heterogenitätsmaße unterscheiden sich je nach Klassifikationstyp (disjunkt oder überschneidend). Grundsätzlich soll ein Maß $v(K_{i_1}, K_{i_2})$, das die Verschiedenheit zweier Klassen K_{i_1} und K_{i_2} beschreibt, nichtnegativ sein und umso kleiner werden, je ähnlicher sich die beiden Klassen K_{i_1} und K_{i_2} sind. Des weiteren wird gefordert, dass gilt: $v(K_i, K_i) = 0$ und $v(K_{i_1}, K_{i_2}) = v(K_{i_2}, K_{i_1})$.

Für disjunkte Klassen K_{i_1} und K_{i_2} lässt sich, ausgehend von einer Distanzmatrix D die Heterogenität z.B. durch **complete linkage:**

$$v(K_{i_1}, K_{i_2}) = \max_{j \in K_{i_1}, k \in K_{i_2}} d(j,k) \tag{1}$$

anhand des unähnlichsten Objektpaares, durch **single linkage:**

$$v(K_{i_1}, K_{i_2}) = \min_{j \in K_{i_1}, k \in K_{i_2}} d(j,k) \tag{2}$$

anhand des ähnlichsten Objektpaares oder anhand der durchschnittlichen Ähnlichkeit der Objekte aus den beiden Klassen **(average linkage)** ermitteln:

$$v(K_{i_1}, K_{i_2}) = \frac{1}{|K_{i_1}| \cdot |K_{i_2}|} \sum_{j \in K_{i_1}} \sum_{k \in K_{i_2}} d(j,k). \tag{3}$$

Bei nicht disjunkten Klassen bleiben bei diesen drei Berechnungen die gemeinsamen Elemente unberücksichtigt.

Als Verfahren zur Ähnlichkeitsbewertung über eine metrische Datenmatrix sei an dieser Stelle das **Ward-Verfahren** (vgl. Ward (1963), Bortz (1993), Kap. 16.3.1, Baier, Gaul (2013)) genannt. Ziel dessen ist die Vereinigung von Objekten, die die Streuung eines Clusters nur möglichst gering erhöhen. Hier werden die Verschieden-

heitsindizes unter Verwendung der Mittelwertvektoren $\overline{a}^{K_{i_1}}$ bzw. $\overline{a}^{K_{i_2}}$ der Objektvektoren der Klassen K_{i_1} und K_{i_2} gebildet. Das Maß ergibt sich dabei als

$$v(K_{i_1}, K_{i_2}) = \frac{|K_{i_1}||K_{i_2}|}{|K_{i_1}| + |K_{i_2}|}(\overline{a}^{K_{i_1}} - \overline{a}^{K_{i_2}})'(\overline{a}^{K_{i_1}} - \overline{a}^{K_{i_2}}). \tag{4}$$

Dieses Verfahren hat sich in einer Studie von Bergs (1981) als empfehlenswert gezeigt, da es gute Partitionen findet und gleichzeitig meistens die richtige Clusteranzahl signalisiert (vgl. Backhaus et al. (2016), Teil II, Kap. 8.2.2.2.2).

Das üblicherweise für quantitative Daten verwendete Maß der Homogenität ist die **L_r-Metrik** (euklidische Metrik für $r = 2$). Ausgehend von einer Datenmatrix Y mit n Objekten und p Merkmalen, berechnet sich der Distanzindex aus

$$d(i,j) = \sqrt[r]{\sum_{k=1}^{p} |y_{ik} - y_{jk}|^r} = ||y_i - y_j||_r, \;\; i,j = 1,...,n. \tag{5}$$

Zum Ende dieses Kapitels wird noch auf die Konstruktionsverfahren eingegangen. Abhängig vom gewählten Klassifikationstyp gibt es zur endgültigen Teilung des Datensatzes eine Reihe von Möglichkeiten.

Für **Überdeckungen** lässt sich entweder ein exhaustives oder iteratives Konstruktionsverfahren wählen. Beide kommen ohne eine vorherige Festlegung einer festen Klassenanzahl aus, stattdessen wird eine Mindesthomogenität der Klassen der Überdeckung bereits mit der Wahl des Homogenitätsmaßes gefordert und eine obere Schranke für die Klassenhomogenitäten festgelegt. Ein **exhaustives** Verfahren (geeignet für das Clustering kleiner Objektmengen) bildet Klassen, zu denen gerade so viele Objekte zugeordnet werden, dass die Homogenitätsschranke nicht verletzt wird. Ein **iteratives** Verfahren, anzuwenden bei einer sehr großen Anzahl an interessierenden Objekten, zeichnet sich durch eine zusätzliche Distanzschranke aus, die verhindert, dass Objekte, die allen anderen Objekten sehr unähnlich sind, die Klassifikation beeinflussen. Dies kann zu einer nichtexhaustiven Überdeckung führen, die vorliegt, wenn nicht alle Objekte klassifiziert werden.

Für den Klassifikationstyp der **Partition** gibt es ebenfalls die Möglichkeiten einer iterativen oder rekursiven Konstruktionsweise. Für ein **iteratives** Verfahren ist zu Beginn eine Festlegung sowohl auf die Anzahl der Klassen als auch auf eine Anfangspartition, die aus zufällig bestimmten, ein-elementigen Klassen besteht, notwendig. Die weitere Zuordnung erfolgt im weiteren dann über die Zentralobjekte der jeweiligen Klassen der Anfangspartition. Abgebrochen wird das Verfahren dann,

wenn ein gewisses Gütekriterium erreicht ist (genauer s. Hartung, Elpelt (2007), Kap. VII 4.1). Das **rekursive** Verfahren kommt ohne eine feste Anzahl von Klassen aus und lässt diese nacheinander bestimmen. Es werden nach und nach die Objekte einer Klasse hinzugefügt, deren Distanz zum Zentralobjekt zum jeweiligen Zeitpunkt minimal ist. Die Bildung einer Klasse ist dann abgeschlossen, wenn eine vorher definierte Homogenitätsschranke erreicht ist. Das gesamte Verfahren gilt als beendet, wenn alle Objekte klassifiziert wurden oder nur noch eine kleine Gruppe von Objekten übrig bleibt.
Als das wichtigste und bekannteste partitionierende Verfahren gilt das **k-means-Verfahren** (vgl. MacQueen (1967), Bortz (1993), Kap. 16.3.2). Dieses funktioniert dabei nach folgendem Schema:

1. Zufällige Anfangspartition mit k beliebigen Clustern
2. Bestimmen der Gruppenzentren (über Gruppenmittelwerte, Zentroide)
3. Bestimmen des Abstands jedes Objekts zu allen anderen Gruppenzentren
4. Neue Zuweisung jedes Objekts zu derjenigen Gruppe, zu deren Zentrum es den geringsten Abstand hat

Schritte 2. bis 4. werden solange wiederholt bis entweder keine Neuzuweisung mehr vorgenommen wird oder eine vorgegebene Iterationsanzahl erreicht wurde.
Bei der Konstruktion einer Hierarchie greift man entweder auf ein divisives oder agglomeratives Verfahren zurück. Beginnt das **divisive** Verfahren mit einer groben Überdeckung und konstruiert dann Schritt für Schritt immer feinere, so beginnt das **agglomerative** Verfahren mit der feinsten Überdeckung und bildet in der Folge immer gröbere.
Eine gute Übersicht zu weiteren möglichen Clusteranalyseverfahren findet sich u.a. in Hastie et. al (2001), Kap. 14.3.

Abweichend von den hier zunächst vorgestellten distanz-basierten Clusteringverfahren existieren auch **dichte-basierte Verfahren**. Im Gegensatz zu den hierarchischen und partitionierenden Verfahren kommt hier die Dichte in einem Cluster als Parameter hinzu. Die Vorstellung des grundsätzlichen Verfahrens orientiert sich

an Ester et. al (1996), Kriegel (2002) und Ester und Sander (2000), Kap. 3.2.6.
Cluster können auch als Gebiete im d-dimensionalen Raum angesehen werden, in denen Objekte dicht beieinander liegen. Diese werden dann getrennt von Gebieten, in denen die Dichte der Gebiete geringer ist. Das Ziel von dichte-basierten Clusterverfahren ist es, diese Gebiete zu identifizieren. Grundidee dabei ist ein gegebener Grenzwert, der von der lokalen Punktdichte bei jedem Objekt innerhalb des Clusters überschritten wird. Die **lokale Punktdichte** eines Objekts o ist dabei gegeben durch die Anzahl der Objekte in einem festgelegten Umkreis um o. Weiter sind die Objekte, die das Cluster charakterisieren, räumlich zusammenhängend. Um die Idee für dichte-basiertes Clustering formal einzuführen, sind zunächst einige Definitionen nötig:
Ein Objekt $o \in O$ heisst **Kernobjekt**, wenn gilt:

$$|N_\epsilon(o)| \geq MinPts, \text{ wobei } N_\epsilon(o) = \{o' \in O | dist(o, o') \leq \epsilon\}. \tag{6}$$

o ist also Kernobjekt, wenn sich in seiner ϵ-Umgebung mindestens $MinPts$ Objekte befinden. ϵ und $MinPts$ sind dabei die Parameter, die einen minimalen Dichtewert spezifizieren. Objekte, die keine Kernobjekte sind, gehören entweder zu einem Cluster (Randobjekte) oder auch nicht (Rauschen).
Ein Objekt $p \in O$ ist **direkt dichte-erreichbar** von $q \in O$ bzgl. ϵ und $MinPts$ in O, wenn gilt:

$$\begin{aligned} &1.\ p \in N_\epsilon(q) \\ &2.\ q \text{ ist ein Kernobjekt in } O \end{aligned} \tag{7}$$

Also sind alle Objekte, die in der ϵ-Umgebung eines Kernobjekts p liegen, direkt dichte-erreichbar. Die Definition der Dichte-Erreichbarkeit ergibt sich daraus wie folgt:
Ein Objekt p ist **dichte-erreichbar** von einem Objekt q bzgl. ϵ und $MinPts$ in der Menge von Objekten O, wenn eine Folge von Objekten $p_1, \ldots, p_n$ in O existiert, sodass $p_1 = q, p_n = p$ ist und p_{i+1} ist direkt dichte-erreichbar von p_i bzgl. ϵ und $MinPts$ in O für $1 \leq i \leq n$. Anschaulich impliziert die Dichte-Erreichbarkeit also die Existenz einer Kette von direkt erreichbaren Objekten zwischen q und p. Diese Ketten von dichte-erreichbaren Objekten sind Teile von Clustern, inklusive der Randpunkte. Die Zusammengehörigkeit solcher Ketten wird formal durch die Dichte-Verbundenheit definiert:

Ein Objekt p gilt als **dichte-verbunden** mit einem Objekt q bzgl. ϵ und $MinPts$ in der Menge O von Objekten, wenn ein $o \in O$ existiert, sodass sowohl p als auch q dichte-erreichbar bzgl. ϵ und $MinPts$ von o sind. Die Dichte-Verbundenheit impliziert also, dass zwei Objekte von einem dritten Objekt dichte-erreichbar sein müssen. Durch die Definitionen der Dichte-Erreichbarkeit und der Dichte-Verbundenheit ist nun die formale Definition von dichte-basierten Clustern und schlussendlich des dichte-basierten Clusterings möglich.

Ein **dichte-basiertes Cluster** C bzgl. ϵ und $MinPts$ in O ist eine nicht-leere Teilmenge von O, für die folgende Bedingungen erfüllt sein müssen:

1. $Maximalität : \forall p, q \in O :$ wenn $p \in C$ und q dichte-erreichbar von p bzgl. ϵ und $MinPts$ ist, dann ist auch $q \in C$.
2. $Verbundenheit : \forall p, q \in C : p$ ist dichte-verbunden mit q bzgl. ϵ und $MinPts$ in O. (8)

Ein Cluster C ist also eine Menge von Objekten, die alle dichte-verbunden miteinander sind. Alle Objekte, die von dem Cluster aus dichte-erreichbar sind, gehören ebenfalls zum Cluster. Als wichtige Eigenschaft für dichte-basierte Cluster, die einfache Algorithmen zur Bestimmung aller dichte-basierten Cluster in der Menge von Objekten O rechtfertigt, gilt folgendes:

Sei C dichte-basiertes Cluster bzgl. ϵ und $MinPts$ in O und sei des weiteren $p \in C$ ein Kernobjekt (s. (Gleichung 6)), dann gilt:

$$C = \{o \in O | o \text{ dichte-erreichbar von } p \text{ bzgl. } \epsilon \text{ und } MinPts\}. \tag{9}$$

Man kann ein Cluster C also dadurch finden, dass man, ausgehend von einem beliebigen Kernobjekt aus diesem Cluster alle dichte-erreichbaren Objekte „aufsammelt“.

Ein **dichte-basiertes Clustering** CL der Menge O bzgl. ϵ und $MinPts$ ist nun eine Menge von oben definierten dichte-basierten Clustern bzgl. ϵ und $MinPts$ in O, $CL = \{C_1, \ldots, C_k\}$, sodass $\forall C$: wenn C ein dichte-basiertes Cluster bzgl. ϵ und $MinPts$ in O ist, dann ist schon $C \in CL$.

Sei weiter $CL = \{C_1, \ldots, C_k\}$ ein dichte-basiertes Clustering der Menge O bzgl.ϵ und $MinPts$, dann heisst die Menge $Noise_{CL}$, definiert als die Menge aller Objekte aus O, die nicht zu einem der dichte-basierten Cluster C_i gehören, das „**Rauschen**“.

Es gilt also:

$$Noise_{CL} = O \setminus (C_1 \cup \cdots \cup C_k). \tag{10}$$

Ein dichte-basiertes Clustering CL ist nach Definition also die Menge aller Cluster bezüglich der (gegebenen) Parameter ϵ und $MinPts$ in der Datenmenge O. Alle Objekte aus O, die zu keinem der Cluster zugeordnet werden, beschreiben die Menge des Rauschens bezüglich DL. Das Verfahren **DBSCAN** ist ein Beispiel für ein dichte-basiertes Verfahren. Da diese im weiteren Verlauf der Arbeit nur eine untergeordnete Rolle spielen, wird auf eine nähere Beschreibung an dieser Stelle verzichtet.

Als weitere Verfahrensgruppe der Clusterverfahren sollen an dieser Stelle kurz die **modell-basierten** Methoden vorgestellt werden.
Im Gegensatz zu modellfreien Methoden der Clusteranalyse werden bei dem hier vorgestellten Ansatz Wahrscheinlichkeitsmodelle unterstellt. Dabei muss zunächst die Annahme getroffen werden, dass die vorliegenden Daten unabhängigen Populationen entstammen, wobei die Anzahl derer unbekannt ist (vgl. Freis (2013)). Notation und Definitionen orientieren sich an Stahl und Sallis (2012).
Sowohl hierarchische als auch nicht-hierarchische Verfahren basieren auf heuristischen Herangehensweisen. Das ist so zu verstehen, dass keine Annahmen über die Struktur der Daten getroffen werden müssen und die Entscheidung über das gewählte Clusteringverfahren meist u.a. auf der Interpretierbarkeit der Ergebnisse basiert. Modellbasiertes Clustering stellt dazu eine Alternative dar, die aufgrund von methodischen und softwaretechnischen Fortschritten zunehmend favorisiert wird. Der Ansatz geht davon aus, dass der Datensatz mehrere verschiedene Untergruppen enthält, die jeweils einer multivariaten Wahrscheinlichkeitsverteilung entstammen. Die Wahrscheinlichkeitsdichten können als die Summe der gewichteten Komponentendichten in einer so genannten **Finite Mixture Dichte** modelliert werden, die für die Population als Ganzes gilt. Das Clustering-Problem wird dann zum Problem der Parameterschätzung für die Mischung aus Wahrscheinlichkeitsverteilungen. Die Parameter werden genutzt, um jeder Beobachtung eine posterior-Wahrscheinlichkeit zuzuordnen, einem bestimmten Cluster anzugehören. Die Probleme der Festlegung der Klassen und Wahl eines geeigneten Clusterverfahrens vereinfachen sich zu einem Modellwahl-Problem, für das zielführende Methoden vorliegen.

In Finite Mixture Modellen wird angenommen, dass die beobachteten Daten einer Mischung von Verteilungen der Form

$$f(x; p, \theta) = \sum_{j=1}^{c} p_j g_j(x; \theta_j) \tag{11}$$

enstammen, wobei x eine p-dimensionale Zufallsvariable, $p' = (p_1, p_2, \ldots, p_{c-1})$ und $\theta' = (\theta'_1, \theta'_2, \ldots, \theta'_c)$ ist. Die p_j sind bekannt als die Mischungsgewichte und die $g_j, j = 1, 2, \ldots, c$ als die Komponentendichten, für die die Dichte g_j durch θ_j parametriert wird. Die Mischungsgewichte sind nichtnegativ und genauso, dass $\sum_{j=1}^{c} p_j = 1$. Die Anzahl der Komponenten, also die postulierte Anzahl der Cluster ist c. Mehr Details zu Mischungsmodellen finden sich u.a. in McLachlan, Peel (2000).

Oftmals wird angenommen, dass die Beobachtungen aus einem Datensatz als Mischung von multivariaten Normalverteilungen beschrieben werden können. Die Verteilungen der Komponenten werden dann durch verschiedene Mittelwertvektoren μ_j und Kovarianzmatrizen $\sum_j$ charakterisiert. Diese Kovarianzmatrizen können sich dann (nicht notwendigerweise) für jede Verteilung unterscheiden. Die Verteilungen der Komponenten haben folgende Wahrscheinlichkeitsdichtefunktion:

$$\phi(x_j; \mu_j; \sum\nolimits_j) = \frac{\exp\{-0.5(x_j - \mu_j)^T \sum_j^{-1} (x_j - \mu_j)\}}{\sqrt{\det(2\pi \sum_j)}} \tag{12}$$

Finite Mixture Modelle können über eine Vielfalt von Wahrscheinlichkeitsfunktionen gebildet werden. Die Parameter der Modelle basieren dann auf Log-Likelihood Schätzungen:

Gegeben seien Beobachtungen $x_1, x_2, \ldots, x_n$ aus einer Mischverteilung wie in Gleichung 11. die **Likelihood** l sieht dann wie folgt aus:

$$l(p, \theta) = \sum_{i=1}^{n} \ln f(x_i; p, \theta) \tag{13}$$

Eine Schätzung der Parameter der Dichte erfolgt dann üblicherweise als Lösung von:

$$\frac{\partial l(\phi)}{\partial(\phi)} = 0, \tag{14}$$

wobei $\phi' = (p', \theta')$. Im Falle von Finite Mixture Modellen ist die Likelihood zu komplex für die Anwendung von gewöhnlichen Methoden zur Maximierung. Die bekannteste und am meisten verbreitetste Methode zur Schätzung des Parametervektors θ_i und der Mischungsgewichte p_j für ein Finite Mixture Modell mit c Komponenten ist dabei der **EM-Algorithmus** (expectation maximization) (vgl. Dempster et al. (1977)). Dieser beschreibt eine iterative Methode zur Findung der Maximum-

Likelihood-Schätzung der Parameter einer zugrundeliegenden Verteilung eines gegebenen Datensatzes. Alternativ finden in der Literatur Bayesianische Methoden (vgl.Bayes (1763)) immer häufiger Verwendung.
Nach Schätzung der Parameter für die Mischverteilung entspricht jede Komponentendichte einem Cluster und die geschätzten posterior-Wahrscheinlichkeiten werden genutzt, um Fälle den Clustern zuzuordnen:

$$\Pr(cluster j|x_i) = \frac{\hat{p}_j g_j(x_i, \hat{\theta})}{f(x_i; \hat{p}, \hat{\theta})},\ j = 1, 2, \ldots, c \tag{15}$$

Ein Fall bzw. eine Beobachtung wird dem Cluster mit der maximalen geschätzten posterior-Wahrscheinlichkeit zugeordnet.

3.1.2 Evidence Accumulation Clustering

Als spezielles Clusterverfahren für den Klassifikationstyp der Partition wird an dieser Stelle das Evidence Accumulation Clustering (Fred, Jain (2002)) vorgestellt. Der Vorteil dieser Methode gegenüber den meisten anderen Verfahren ist, dass die Ergebnisse von mehrfachem Clustering zu einer Klassenbildung zusammengeführt werden, statt auf ein einzelnes Ergebnis zu vertrauen. Jedes Cluster-Ergebnis wird dabei als unabhängiger Hinweis auf die wahre Gestalt der Daten angesehen. Die endgültige Klassenbildung erfolgt dann über einen Abstimmungsmechanismus.
In Fred (2001) wird der k-means-Algorithmus als Basis für die Zerlegung des Datensatzes in k Cluster verwendet, um dann über N Clusterings mit zufälligen Startwerten eine Ähnlichkeitsmatrix co_{assoc} zu erhalten, für die $co_{assoc}(i, j)$ die Häufigkeit beschreibt, mit der das Wertepaar (i, j) demselben Cluster über alle N Clusterings zugeordnet wird. Die endgültige Klassifikation geschieht dann durch Anwendung der single linkage Methode auf eben diese Ähnlichkeitsmatrix.
Fred und Jain beschreiben den Parameter k als kritisch, da die Güte der Klassifikation durch dessen Wahl stark beeinflusst wird. Ist k zu klein, wird möglicherweise die Komplexität der Daten nicht richtig wiedergegeben. Wird k zu groß gewählt, kann dies zu einer zu feinen Zerlegung führen. Die Ähnlichkeitswerte zwischen den einzelnen Wertepaaren sinken mit steigendem k und die Verbindungen im Dendrogramm (Beschreibung s. z.B. Hartung, Elpelt (2007), Kap. VII 6.) bilden sich auf höheren Ebenen, wodurch die wirklichen Cluster weniger deutlich definiert werden.

Als „Überlebenszeit des Clusters “ bezeichnet man dann den Abstand zwischen zwei aufeinanderfolgenden Zusammenführungen im Dendrogramm. Dieser Wert wird im Folgenden genutzt, um die endgültige Anzahl der Cluster festzulegen. Die Wahl der Anzahl von k wird über eine Kombination mehrerer k-means-Durchgänge variabel gehalten. Das häufigste Auftreten im Dendrogramm einer Cluster-Konfiguration, abgebildet durch die single linkage Methode über co_{assoc}, entscheidet dann über die finale Partition.

Der Algorithmus des Evidence Accumulation Clustering sieht dabei wie in Algorithm 1 dargestellt aus ((Fred, Jain (2002))).

Algorithm 1

Input:

n d-dimensionale Objekte

k_{min}: minimale Anzahl an Clustern

k_{max}: maximale Anzahl an Clustern

N: Anzahl der Clusterdurchgänge

Output:

Partitionierte Daten

START: Starte mit einer Nullmatrix $n \times n - co_{assoc}$

1: **N-mal folgender Ablauf:**
2: zufällige Wahl des k im Intervall $[k_{min}; k_{max}]$.
3: zufällige Wahl von k Clusterzentren
4: Durchlauf des k-means-Algorithmus mit obigem k und obigen Startwerten und Ausgabe der Partition P
5: Aktualisierung der co_{assoc}-Matrix: Für jedes Paar (i, j) im selben Cluster in der Partition P setze $co_{assoc}(i, j) = co_{assoc}(i, j) + \frac{1}{N}$
6: Erkenne konsistente Cluster in co_{assoc} mit der single linkage Methode: Berechne das single linkage Dendrogramm und identifiziere die finale Klassenbildung als die mit der höchsten Überlebenszeit.

Eine Anwendung des Verfahrens im Kontext der Lastanalyse im automobilen Umfeld findet sich in Lehmann (2013). In der erwähnten Arbeit wurde das Verfahren leicht angepasst, sodass, statt der single linkage Methode zur Bewertung der Ähnlichkeit zweier Klassen, das Ward-Verfahren gewählt wurde (s. 3.1.1). Mit dieser Methodik

wurde eine Fahr- und Lastmustererkennung auf Basis verschiedener Datenwelten, von Erprobungsdaten bis hin zu „echten“ Kundendaten durchgeführt.
Im Laufe der Zeit gab es immer wieder Erweiterungen des Ansatzes von Evidence Accumulation Clustering. In Fred, Jain (2005) wird z.B. vorgeschlagen, die Stabilität der Cluster-Ergebnisse durch mehrfaches Clustering mit demselben Algorithmus, aber unterschiedlichen Parametern oder aber durch Nutzen von verschiedenen Algorithmen zu erhöhen. Die Motivation liegt darin, dass es schwierig sei, sich vor dem Clustering von vorliegenden Daten auf einen speziellen Algorithmus festzulegen.
2010 untersuchten Lourenco et al. (2010) die Performance des Algorithmus bzgl. dessen Rechenbarkeit. 2016 beschäftigte sich Silva (2016) ebenfalls mit der Herausforderung, dass die Methode für größere Datensets an Performance verliert und schlägt drei verschiedene Ansätze vor, um diesem Problem zu begegnen. 2013 fanden Lourenco et al. (2013) eine Möglichkeit, die Wichtigkeit der Partitionen vorab zu differenzieren, statt allen Partitionen grundsätzlich das gleiche „Gewicht“ zuzuteilen.
Song et al. (2016) erweiterten das Verfahren auf hierarchische Verfahren und applizierten dies auch bzgl. Performance auf Big Data Ansätze.

3.1.3 Gütekriterien für Clusterverfahren

An dieser Stelle werden verschiedene Maße zur Bewertung der Güte eines Clusterings vorgestellt, bevor dann auf die Beurteilung der Güte des Bewertungskriteriums eingegangen wird. Die Maße können meist auch schon direkt zur Konstruktion des Clusterings (durch Minimierung dieser Zielgröße) genutzt werden, werden aber an dieser Stelle lediglich für die Bewertung des Clusterings verwendet.

$\mathbf{TD}^2$ und TD

Die Einführung von $\mathbf{TD}^2$ zur Darstellung der Kompaktheit eines Clusterings erfordert zunächst zwei vorhergehende Definitionen. Dies geschieht hier u.a. nach Ester, Sander (2000), Kap. 3.2.1.
Jedes Cluster C sei durch einen sogenannten Centroid μ_C repräsentiert. Dieser stellt

den Mittelwert aller Punkte $p = (x_1^p, \ldots, x_d^p)$ im Cluster C dar:

$$\mu_C = (\bar{x}_1(C), \ldots, \bar{x}_d(C)), \text{ wobei } \bar{x}_j(C) = \frac{1}{n_C} \cdot \sum_{p \in C} x_j^p. \tag{16}$$

n_C sei die Anzahl aller Objekte im Cluster C. Als Maß für die Kompaktheit eines Clusters ergibt sich dann

$$TD^2(C) = \sum_{p \in C} dist(p, \mu_C)^2, \tag{17}$$

das, je näher die Punkte in C beieinander liegen, immer kleiner wird. Anschaulich ist es z.B. die Summe aller quadrierten euklidischen Distanzen zum Centroid.
Aus dieser Definition folgt dann direkt ein Maß für die Kompaktheit des gesamtem Clusterings als die Summe der quadrierten Distanzen jedes Punktes zum Centroid seines Clusters:

$$TD^2 = \sum_{i=1}^{k} TD^2(C_i) \tag{18}$$

Hier gilt: Je kleiner der Wert wird, desto kompakter sind (durchschnittlich) die einzelnen Cluster des Clusterings.
Eine leichte Abwandlung des Maßes TD^2 stellt das Maß **TD** dar. Dabei werden im Cluster statt der Centroide die Medoide als Repräsentanten herangezogen (vgl. Kaufman, Rousseeuw (1990), Kap. 2 und Ester, Sander (2000), Kap. 3.2.1. Im Gegensatz zum Centroid kommt der Medoid $m_C \in C$ auch sicher in der Datenmenge vor. Der Medoid ist anschaulich gesehen das zentralste Objekt des Clusters. Jedes Objekt wird dann demjenigen Medoid zugeordnet, dem es am nächsten ist. Statt auf die quadrierte wird hier üblicherweise auf die einfache Distanz als Kriterium zurückgegriffen. Das Maß für die Kompaktheit eines Clusters ergibt sich dann als:

$$TD(C) = \sum_{p \in C} dist(p, m_C) \tag{19}$$

und das Maß für die Kompaktheit eines Clusterings als:

$$TD = \sum_{i=1}^{k} TD(C_i). \tag{20}$$

Auch hier gilt: Je kleiner TD, desto kompakter sind (durchschnittlich) die einzelnen Cluster des Clusterings.

Silhouettenkoeffizient für distanzbasierte Verfahren

Ein geeignetes Maß, um die Güte von mehreren Clusterergebnissen bezüglich verschiedener Werte von k (entstanden z.B. durch das k-means Verfahren) zu vergleichen, ist der **Silhouetten-Koeffizient** eines Clustering. Bei obigen Verfahren würde TD^2 bzw. TD immer kleiner werden, je größer k wird. In dem Ausbleiben dieses Phänomens liegt hier der entscheidende Vorteil. Zurückführend auf Rousseeuw (1987) definieren Ester, Sander (2000), Kap. 3.2.4 den Silhouetten-Koeffizienten folgendermaßen:
Sei $C_M = \{C_1, \ldots, C_k\}$ die Menge aller Cluster in einer Menge von Objekten O, $C \in C_M$ und $o \in C$. Dann sei der durchschnittliche Abstand des Objekts o zu einem beliebigen Cluster $C_i \in C_M$: $dist(o, C_i) = [\sum_{p \in C_i} dist(o, p)]/|C_i|$. $a(o) = dist(o, C)$ beschreibt dann den durchschnittlichen Abstand des Objekts o zu „seinem“ Cluster C und $b(o) = \min_{C_i \in C_M, C_i \neq C} dist(o, C_i)$ den durchschnittlichen Abstand des Objekts o zum „Nachbarcluster“, also zu dem nächstbesten Cluster.
Mithilfe dieser beiden Abstände kann nun die **Silhouette** $s(o)$ von $o \in C$ definiert werden:

$$s(o) = \begin{cases} 0 & \text{, wenn } |C| = 1, \text{ d.h. } a(o) = 0, \\ \frac{b(o) - a(o)}{\max\{a(o), b(o)\}} & \text{, sonst.} \end{cases} \tag{21}$$

Für die Silhouette als Maß dafür, wie gut die Zuordnung des Objekts o zu seinem Cluster C ist, gilt: $-1 \leq s(o) \leq 1$. Die Werte von $s(o)$ folgen dabei folgender Interpretation:

- $s(o) \approx 0$, d.h. $a(o) \approx b(o)$:
  o liegt ungefähr zwischen seinem eigenen und dem Nachbarcluster.
- $s(o) \approx 1$, d.h. $a(o)$ ist wesentlich kleiner als $b(o)$:
  o ist gut klassifiziert.
- $s(o) \approx -1$, d.h. $b(o)$ ist wesentlich kleiner als $a(o)$:
  o ist schlecht klassifiziert.

Die Zuordnung von o zu seinem Cluster wird also immer besser, desto größer der Wert von $s(o)$ ist.
Der durchschnittliche Wert der Silhouetten $s(o)$ aller Objekte o eines Clusters C

beschreibt somit die Güte des einzelnen Clusters und wird definiert als **Silhouettenweite** von C.

$$s(C) = \left(\sum_{o \in C} s(o) \right) / |C|. \tag{22}$$

Der **Silhouettenkoeffizient** eines gesamtem Clusterings C_M wäre dann die Silhouettenweite der Gesamtmenge O und definiert als

$$s(C_M) = \frac{\sum_{C \in C_M} \sum_{p \in C} s(p)}{|O|} \tag{23}$$

Als Maß für die Güte eines Clusterings, welches unabhängig von der Anzahl k der Cluster ist, ist das Clustering umso besser, je größer der Wert von $s(C_M)$ wird. Kaufman, Rousseeuw (1990), Kap. 2.2 schlagen folgende Interpretation des Koeffizienten vor:

- 0,70< $s(C_M) \leq$ 1,00: starke Struktur
- 0,50< $s(C_M) \leq$ 0,70: brauchbare Struktur
- 0,25< $s(C_M) \leq$ 0,50: schwache Struktur
- $s(C_M) \leq 0,25$: keine Struktur.

Davies-Bouldin Index

Der **Davies-Bouldin Index** beschreibt eine Funktion des Verhältnisses der Streuung im Cluster und der Heterogenität zwischen den Clustern (vgl. Davies, Bouldin (1979).

Der Index wird für jedes Clusterpaar C_i und C_j berechnet. Die Streuung innerhalb der Cluster wird addiert und durch die Separierung zwischen den Clustern geteilt. Über alle N Cluster wird der jeweilige maximale Wert des Quotienten aufaddiert (vgl. Ploehn (2014)).

Die Streuung im i-ten Cluster stellt sich dar als

$$S_i = \frac{1}{|C_i|} \sum_{x \in C_i} dist(x, z_i) \tag{24}$$

und die Distanz zwischen Cluster C_i und C_j, bezeichnet als d_{ij} wird definiert als

$$d_{ij} = dist(z_i, z_j), \tag{25}$$

wobei z_i das Zentrum des i-ten Clusters und z_j das Zentrum des j-ten Clusters repräsentiert. Der Davies-Bouldin Index ergibt sich dann unter der Definition

$$R_i = \max_{i \neq j}\{R_{ij}\} \tag{26}$$

für $R_{ij} = \frac{S_i + S_j}{d_{ij}}$ zu:

$$DB = \frac{1}{N}\sum_{i=1}^{N} R_i. \tag{27}$$

(vgl. Davies, Bouldin (1979). Je kleiner dieser Index wird, desto besser das Clustering.

Dunn´s Index

Seien S und T zwei nichtleere Teilmengen von $\mathbb{R}^n$. Dann wird der Durchmesser Δ von S definiert als:

$$\Delta(S) = \max_{x,y \in S}\{dist(x,y)\}, \tag{28}$$

die Distanz δ zwischen S und T sei

$$\delta(S,T) = \min_{x \in S, y \in T}\{dist(x,y)\} \tag{29}$$

(mit $dist(x,y)$ als Distanz zwischen den Datenpunkten x und y) und der zu definierende **Dunn´s Index** (Dunn (1973) wird eingeführt als

$$\nu_D = \min_{1 \leq i \leq K}\left\{\min_{1 \leq j \leq K, j \neq i}\left\{\frac{\delta(C_i, C_j)}{\max_{1 \leq k \leq K}\{\Delta(C_k)\}}\right\}\right\} \tag{30}$$

Hier spricht ein höherer Wert von ν_D für eine bessere Einteilung durch das gewählte Clusterverfahren.

Calinski-Harabasz Index

Der **Calinski-Harabasz Index** (Calinski, Harabasz (1974)) nimmt eine Gegenüberstellung der beiden Streuungen innerhalb und zwischen den Clustern vor und kann geschrieben werden als:

$$CH(C) = \frac{n-K}{K-1}\frac{Sep(C)}{Comp(C)} \tag{31}$$

(vgl. Ploehn (2014)).

Das Kompaktheitsmaß $Comp(C)$ ist die Intra-Cluster-Varianz und wird über die Quadratsumme der Distanz zwischen den Objekten und deren Clusterzentren berechnet. Das Separierungsmaß $Sep(C)$ hingegen beschreibt die Inter-Cluster-Varianz

und definiert sich über die Quadratsumme der Distanz zwischen den Clusterzentren und dem Mittelpunkt der Daten. Gewichtet wird der Abstand der Clusterzentren mit der Anzahl der Objekte in einem Cluster. Der Index beinhaltet einen Normalisierungsfaktor $\frac{N-K}{K-1}$, der verhindert, dass der Quotient monoton mit der Clusteranzahl steigt (vgl. Ploehn (2014) und Vendramin et al. (2010)) und wird für jede Cluster-Lösung mit K Gruppen berechnet. Bei einem optimalen Clustering wird dieser Wert maximal (vgl. Ploehn (2014)).

Weitere Maße zur Beurteilung der Güte eines distanzbasierten Clusterings werden u.a. in Xie, Beni (1991), Pal, Bezdek (1995) und Maulik, Bandyopadhyay (2002) vorgestellt.

3.1.4 Unsicherheit der Güte von Clusterverfahren

Wird ein Clustering mehrmals nacheinander durchgeführt (z.B. mit verschiedenen Start-Partitionen, verschiedenen Verfahren oder auf verschiedenen Stichproben), ist es möglich, in jeder Iteration das in 3.1.3 definierte Gütemaß zu berechnen und danach über die Iterationen hinweg die Sensitivität zu bestimmen. Im weiteren Verlauf dieser Arbeit geschieht das mehrfache Ausführen des Clusterings im Rahmen des Evidence Accumulation Clustering (vgl. 3.1.2). Entscheidet man sich für ein „klassisches“ Verfahren, so könnte ein mehrfaches Berechnen, z.B. durch eine vorgelagerte Aufteilung in Trainings- und Testdaten, erfolgen. Wiederholt man diesen Schritt mehrmals, ergeben sich auch dann mehrere Berechnungen des Gütemaßes. Die Bewertung der Sensitivität kann z.B. über die Betrachtung der empirischen Verteilung, d.h. Berechnung der Minima, Maxima, Standardabweichung, Varianz und Quantile (hier: 5%, 50% und 95%) durchgeführt werden.
Auf eine Beschreibung dieser Kenngrößen wird an dieser Stelle verzichtet (vgl. dazu z.B. Vogel (2005), Kap. 3.3).

3.1.5 Zusammenfassung

In diesem Teil-Kapitel wurden die grundsätzlichen Verfahren und Begriffe der Clusteranalyse vorgestellt. Partitionierende Methoden zeichnen sich dadurch aus, dass die Anzahl der Cluster vorgegeben wird. In mehreren Schritten werden dann die Objekte dem jeweils nächsten Cluster zugeordnet, das durch seinen Mittelwert (je

nach Definition) repräsentiert wird. Bei gegebenem k muss bezüglich des Cluster-Kriteriums die optimale Aufteilung des Datensatzes gefunden werden. Hierarchische Methoden funktionieren per schrittweiser Aggregation oder Aufteilung der Objekte in Gruppen. Dafür muss vorab ein Maß für die Clusteringgüte definiert werden, das das Abbruchkriterium darstellt. Im weiteren wurde das Evidence Accumulation Clustering als Algorithmus vorgestellt, der mehrfaches Clustering zu einer Klassenbildung zusammenführt. Innerhalb des Algorithmus kann das Cluster-Verfahren wiederum gewählt werden. Um die Qualität des Clusterings zu beurteilen, wurden verschiedene Gütemaße vorgestellt und die Betrachtung der Sensitivität der Güte anhand der empirischen Verteilung der Gütemaße beschrieben.

3.2 Lineare Modelle

In diesem Teil-Kapitel werden allgemeine lineare Modelle von der Modellformulierung bis zur Prüfung der Modellprämissen hergeleitet. Hinter linearen Modellen steht in der Statistik die Annahme eines linearen Zusammenhangs zwischen Einfluss- und Zielvariablen.

3.2.1 Allgemeine lineare Regression

Motivation

Das Ziel der Regressionsanalyse ist es, die Abhängigkeit einer Variablen Y von anderen Variablen X_j darzustellen. Y ist dabei der Regressand, X_j die Regressoren (Meyna, Pauli (2010), Kap. 16.7). Notation und Beschreibung der linearen Regressionsanalyse orientieren sich hier, wenn nicht anders gekennzeichnet, an Backhaus et al. (2016), Teil II, Kap. 1.2.
Die Regressionsanalyse ist sehr flexibel und wird häufig eingesetzt, um Zusammenhänge quantitativ zu beschreiben und zu erklären, sowie Werte der abhängigen Variablen zu schätzen bzw. zu prognostizieren. Im einfachsten Fall lässt sich die Kausalbeziehung zwischen der abhängigen Variablen Y und der unabhängigen Variablen X als deterministisches Modell

$$Y = f(X) \tag{32}$$

beschreiben. Wird Y durch mehrere Größen beeinflusst, so erweitert sich Gleichung 32 zu:

$$Y = f(X_1, X_2, \ldots, X_j, \ldots, X_J). \tag{33}$$

Probleme in Form von Gleichung 32 können mit einfacher Regressionsanalyse und Probleme in Form von Gleichung 33 mit multipler Regressionsanalyse gelöst werden. Die Regressionsanalyse lässt sich immer dann anwenden, wenn sowohl die abhängige als auch die unabhängigen Variablen quantitativ vorliegen. Ist das nicht der Fall, müssen ggf. qualitative Variablen erst in binäre Variablen umgewandelt werden, um dann wie quantitative Variablen betrachtet werden zu können.
Die Vorgehensweise bei der linearen Regressionsanalyse lässt sich in die Schritte Formulierung (Modell), Schätzung (Regressionsfunktion) und Prüfung (Regressionsfunktion und -koeffizienten, Modellprämissen) unterteilen.

Modellformulierung

Die Modellbildung beinhaltet meistens den Zielkonflikt zwischen Komplexität und Einfachheit. Zum einen soll es den Zusammenhang zwischen der abhängigen und den unabhängigen Variablen möglichst einfach, zum anderen möglichst genau darstellen. Die Modellformulierung der Regressionsfunktion beinhaltet bereits die Annahme des Zusammenhangs zwischen der abhängigen und den unabhängigen Variablen. Das Ziel der Regressionsanalyse ist es, diese lineare Regressionsfunktion zu schätzen. Die lineare Regressionsfunktion sieht im einfachsten Fall wie folgt aus:

$$\hat{Y} = b_0 + b_1 X, \tag{34}$$

wobei $\hat{Y}$ die Schätzung der abhängigen Variable Y, b_0 ein konstantes Glied, b_1 der Regressionskoeffizient und X die unabhängige Variable ist. Der Regressionskoeffizient b_1 ist geometrisch gesehen die Steigung der Geraden und dient als Maß der Stärke der Wirkung von X auf Y. b_0 ist der Schnittpunkt der Regressionsgeraden mit der Y-Achse (also der y-Wert, der zu erwarten ist, wenn der x-Wert Null ist). Die Werte von b_0 und b_1 sind zu diesem Zeitpunkt noch unbekannt und müssen auf Basis der beobachteten Daten (Stichprobe) geschätzt werden. Dies passiert im nächsten Schritt.

Schätzung der Regressionsfunktion

Ziel der Schätzung einer Regressionsfunktion ist es immer, die Abweichungen der Wertepaare von abhängigen und unabhängigen Variablen zu der Regressionsgeraden zu minimieren. Die Differenzen zwischen den beobachteten und durch die Regressionsgeraden geschätzten Werte werden als Residuen bezeichnet und meist durch e wie error symbolisiert:

$$e_k = y_k - \hat{y_k}, k = 1, \ldots, K \tag{35}$$

y_k ist dabei der beobachtete Wert der abhängigen Variablen Y für x_k, $\hat{y_k}$ der ermittelte Schätzwert von Y für x_k und K die Anzahl aller Beobachtungen.
Durch Hinzunahme der Residualgröße zur systematischen Komponente $\hat{Y}$ ergibt sich dann

$$Y = b_0 + b_1 X + e. \tag{36}$$

Für eine einzelne Beobachtung gilt: $y_k = b_0 + b_1 x_k + e_k$ $(k = 1, \ldots, K)$. Um durch die Parameter b_0 und b_1 eine gute Anpassung der Funktion an die Daten zu erlangen, gilt es, die Residualgrößen zu minimieren. Dies ist nichts anderes als ein klassisches Optimierungsproblem. Um zu verhindern, dass positive Abweichungen von negativen „ausgeglichen“ werden, kann man nun entweder die Absolutwerte der Residuen oder die quadrierten Residuen minimieren. Üblich ist aufgrund der einfacheren Handhabung die zweite Variante:

$$\sum_{k=1}^{K} e_k^2 \rightarrow min! \tag{37}$$

Das ist das Kleinst-Quadrate-Kriterium (KQ-Kriterium). Setzt man nun Gleichung 36 in Gleichung 37 ein, erhält man die Zielfunktion für das Optimierungsproblem (Methode der kleinsten Quadrate):

$$\sum_{k=1}^{K} e_k^2 = \sum_{k=1}^{K} [y_k - (b_0 + b_1 x_k)]^2 \rightarrow min! \tag{38}$$

Durch partielle Differentiation von Gleichung 38 nach b_0 und b_1 ergeben sich

$$b_0 = \bar{y} - b_1 \bar{x} \text{ und} \tag{39}$$

$$b_1 = \frac{K(\sum x_k y_k) - (\sum x_k)(\sum y_k)}{K(\sum x_k^2) - (\sum x_k)^2} \tag{40}$$

mit b_0 als konstantem Glied und b_1 als Regressionskoeffizienten.

Wird mehr als eine unabhängige Variable in das Modell aufgenommen (vgl. Glei-

chung 33) modifiziert sich Gleichung 34 zu:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_j X_j + \cdots + b_J X_J \tag{41}$$

und Gleichung 38 zu:

$$\sum_{k=1}^{K} e_k^2 = \sum_{k=1}^{K} [y_k - (b_0 + b_1 x_{1k} + b_2 x_{2k} + \cdots + b_j x_{jk} + \cdots + b_j x_{Jk})]^2 \rightarrow min!. \tag{42}$$

e_k sind dabei die Werte der Residualgröße, y_k die Werte der abhängigen Variablen, b_0 das konstante Glied, b_j die Regressionskoeffizienten, x_{jk} die Werte der unabhängigen Variablen, J die Zahl der unabhängigen Variablen und K die Anzahl der Beobachtungen.

Prüfung der Regressionsfunktion und der Regressionskoeffizienten

Nach der Schätzung der Regressionsfunktion geht es nun darum, dessen Güte zu prüfen. Dies lässt sich auf zwei Ebenen tun: globale Prüfung der Regressionsfunktion und Prüfung der einzelnen Regressionskoeffizienten. Für die Ebene der globalen Prüfung wird an dieser Stelle auf das nachfolgende Kapitel in dieser Arbeit 3.2.2 verwiesen.
Ergibt die globale Prüfung der Regressionsfunktion, dass nicht alle Regressionskoeffizienten β_j Null sind, können die Regressionskoeffizienten einzeln überprüft werden. Dies wird hier nicht genauer beschrieben, siehe dazu Backhaus et al. (2016), Teil II, Kap. 1.2.4.

Prüfung der Modellprämissen

Die Annahmen des linearen Regressionsmodells sehen wie folgt aus:

1. Das Modell ist richtig spezifiziert, d.h.:
   (a) Es ist linear in den Parametern β_0 und β_j
   (b) Es enthält alle erklärenden Variablen
   (c) Die Anzahl der zu schätzenden Parameter ist kleiner als die Anzahl der Beobachtungen
2. Die Störgrößen haben den Erwartungswert Null

3. Es besteht keine Korrelation zwischen erklärenden Variablen und der Störgröße
4. Die Störgrößen haben eine konstante Varianz (Homoskedastizität)
5. Die Störgrößen sind nicht korreliert (keine Autokorrelation)
6. Zwischen den erklärenden Variablen besteht keine lineare Abhängigkeit (keine perfekte Multikolliniearität)
7. Die Störgrößen sind normalverteilt

Die Anwendbarkeit des linearen Regressionsmodells scheint durch die vielen Annahmen sehr eingeschränkt. Da es aber recht unempfindlich auf kleinere Verletzungen reagiert, bildet es ein sehr flexibles und vielseitig anwendbares Analyseverfahren (vgl. Backhaus et al. (2016), Teil II, Kap. 1.2.5.7).

3.2.2 Gütekriterien für lineare Regressionsmodelle

Ein optimiertes lineares Regressionsmodell entsteht durch die Minimierung der Summe der quadrierten Residuen SSR (sum of squared residuals) $SSR = \sum_{i=1}^{n} e_i^2$ mit n als Zahl der Beobachtungen und e_i als Werte der Residualgröße ($i = 1, 2, \ldots, n$) Man kann diesen Wert als Maß für die Güte der Anpassung bewerten; je kleiner SSR, desto besser die Anpassung. Isoliert gesehen ist diese Größe als Gütemaß allerdings nicht geeignet. Es ist nötig, SSR mit anderen Größen in Beziehung zu setzen. Dazu wird hier kurz das Prinzip der **Streuungszerlegung** definiert.
Sei SST (total sum of squares) $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ die Gesamtstreuung, SSE (explained sum of squares) $SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ die erklärte Streuung und $SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, dann gilt:

$$SST = SSE + SSR \tag{43}$$

und für das **Bestimmtheitsmaß** R^2:

$$R^2 = \frac{SSE}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{erklärte Streuung}}{\text{Gesamtstreuung}}. \tag{44}$$

R^2 ist eine normierte Größe mit Wertebereich zwischen Null und Eins. Je höher der Anteil der erklärten Streuung an der Gesamtstreuung ist, desto größer wird der Wert. Ist die gesamte Streuung erklärt, ist $R^2 = 1$, im anderen Extremfall entsprechend $R^2 = 0$. Als alleiniges Kriterium zur Beurteilung der Güte eines Modells ist

allerdings auch dieses Maß nicht ausreichend, da weder die Anzahl n der Beobachtungen noch die Komplexität des Modells berücksichtigt wird. Ein größeres Modell mit mehr Variablen wird immer eine bessere Anpassung als ein weniger komplexes Modell liefern, aber nicht unbedingt bessere Schätzwerte. Mit steigender Anzahl der unabhängigen Variablen sinkt die Zahl der Freiheitsgrade für die Schätzung. Jeder Parameter, der geschätzt werden soll, verbraucht einen Freiheitsgrad. Daher wird hier das **korrigierte Bestimmtheitsmaß** definiert. Das einfache Bestimmtheitsmaß wird um eine Korrekturgröße verringert, die umso größer ist, je größer die Zahl der Regressoren und je kleiner die Zahl der Freiheitsgrade ist. Somit kann das korrigierte Bestimmtheitsmaß durch Aufnahme weiterer Regressoren durchaus abnehmen. Mit n = Zahl der Beobachtungswerte, j = Zahl der Regressoren und $n - j - 1$ = Zahl der Freiheitsgrade ergibt sich dieses zu:

$$R^2_{korr} = R^2 - \frac{j \cdot (1 - R^2)}{n - j - 1} \tag{45}$$

Das Maß ist immer kleiner oder maximal gleich dem einfachen Bestimmtheitsmaß und kann auch negativ werden. Die Überlegung ist, dass übergroße Modellkomplexität durch Korrektur bestraft werden soll (vgl.Backhaus et al. (2016), Teil II, Kap. 1.2.3).

3.2.3 Unsicherheit der Güte von linearen Regressionsmodellen

An dieser Stelle soll anschließend zu der Güte eines Modells deren Sensitivität betrachtet und beschrieben werden. In Soper (2016) findet sich ein Kalkulator, um Konfidenzintervalle für das Bestimmtheitsmaß R^2 zu berechnen.
Dazu ist es zunächst nötig, die Varianz und somit den Standardfehler von R^2 zu berechnen. Nach Wishart (1931) lässt sich der Erwartungswert von R^2 mit k als Anzahl der Prädiktoren im Modell, $p = k - 1$ und n als Gesamtstichprobenumfang darstellen als:

$$\begin{aligned} E(R^2) &= 1 - \frac{n-p}{n-1}(1 - R^2)F\{1, 1, \frac{1}{2}(n+1), R^2\} \\ &= R^2 + \frac{p-1}{n-1}(1 - R^2) - \frac{2(n-p)}{n^2-1}R^2(1 - R^2) \end{aligned} \tag{46}$$

F greift dabei auf die Gaußsche hypergeometrische Funktion zurück. In ähnlicher

Weise lässt sich die Varianz von R^2 dann darstellen als:

$$\begin{aligned} Var(R^2) &= \frac{(n-p)(n-p+2)}{n^2-1}(1-R^2)^2 F\{2,2,\frac{1}{2}(n+3),R^2) - \{E(R^2)-1\}^2 \\ &= ... \\ &= \frac{4R^2(1-R^2)^2(n-p)^2}{(n^2-1)(n+3)} \end{aligned} \tag{47}$$

Der Standardfehler für das R^2 wird nach Olkin, Finn (1995) durch

$$SE_{R^2} \approx \left(\frac{4R^2(1-R^2)^2(n-k-1)^2}{(n^2-1)(3+n)}\right)^{\frac{1}{2}} \tag{48}$$

approximiert. Darauf basierend definierten sie und später ergänzend Cohen et al. (2003), Kap. 3.6.2 ein Konfidenzintervall für R^2:

$$[R^2 - z_{\alpha/2}SE_{R^2}; R^2 + z_{\alpha/2}SE_{R^2}] \tag{49}$$

mit α als gewünschtem prozentualen Konfidenzintervall und $z_{\alpha/2}$ als das $100 \cdot \alpha/2$ obere Percentil der Standardnormalverteilung (Notation vgl. Tan (2012)).

Tan (2012) stellt in seiner Arbeit weitere Möglichkeiten, Konfidenzintervalle für R^2 zu finden, zusammen. So lassen sich die verschiedenen Ansätze in fünf Kategorien unterteilen: „Wald-Methode“, „Fisher´s R^2-zu-z-Transformation“, „Exakte Methode“, „Approximierung basierend auf der Dichte von R^2“ und „Bootstrapping-Methode“; für genaue Definitionen s. Tan (2012) und weiterführende Literatur.

Die „Wald-Methode“ wurde bereits definiert. Tan (2012) sieht den Hauptnachteil dieser Methode darin, dass sie Werte außerhalb von [0,1] ausgeben kann.

Konfindenzintervalle für R^2 können ebenso durch die Nutzung Fisher´s z-Transformation

$$z = 1/2log[(1+R)/(1-R)] \tag{50}$$

gebildet werden (Olkin, Finn (1995)). Daraus lassen sich dann untere und obere Vertrauensgrenzen l_z und u_z für z ableiten:

$$log\left(\frac{1+\sqrt{R^2}}{1-\sqrt{R^2}}\right) \pm z_{\alpha/2}\sqrt{4/n} \tag{51}$$

$4/n$ beschreibt hier die Varianz.

Da z eine monoton steigende Funktion von R^2 ist und für den Fall, dass sowohl untere als auch obere Vertrauensgrenze nichtnegativ sind, erhält man für R^2 folgendes

Konfidenzintervall:

$$\left(\frac{exp(l_z)-1}{exp(l_z)+1}\right)^2, \left(\frac{exp(u_z)-1}{exp(u_z)+1}\right)^2 \tag{52}$$

Aufgrund einiger Schwächen in der Performance (u.a. bei R^2 nahe 0) wird diese Methode in (Tan (2012)) als schwächste der hier vorgestellten beurteilt.

Die „Exakte Methode“ basiert auf der geometrischen Interpretation multipler Korrelation. Fisher (1928) gab die exakte Dichtefunktion von R^2 an, die später von Lee (1972) in folgende Form angepasst wurde:

$$\begin{aligned} f_{R^2}(x) &= B(p/2, n_1/2)^{-1}(1-R^2)^{(n-1)/2}(x)^{p/2-1}(1-x)^{n_1/2-1} \\ &= F\left((n-1)/2, (n-1)/2; p/2; R^2x\right) \end{aligned} \tag{53}$$

B bzw. F beschreiben hier die Beta- bzw. Gaußsche hypergeometrische Funktion, $n_1 = n - p - 1$, n ist die Stichprobengröße.

Sowohl die Beta- als auch die Gaußsche hypergeometrische Funktion sind komplizierte Integranden bzw. Reihen. Weiterhin gibt es keine Formel, die die exakte kumulative Verteilungsfunktion von R^2 berechnen kann. Deswegen erfordert die Berechnung von exakten Vertrauensgrenzen Iterationen der Berechnung der Percentile des Integrals der Dichte $f_{R^2}(x)$. Auf Basis dieser Verteilung lassen sich mit dem Kalkulator Steiger (2017) Vertrauensgrenzen für R^2 berechnen (Lee (1972), Tan (2012)).

Aufgrund der Komplexität der exakten Dichte von R^2 entwickelten viele Forscher verschiedene Methoden, um Konfidenzintervalle über die approximierte Dichtefunktion von R^2 zu bestimmen. Genannt seien hier u.a. Approximationen von Khatri (1966) und Lee (1971). Beschreibung der Approximationen und Herleitung von Konfindenzintervallen finden sich in Tan (2012).

Auch Bootstrapping-Methoden können genutzt werden, um Konfidenzintervalle für R^2 zu finden. Die am meisten verbreiteten Techniken sind hier die Nutzung von Percentilen, bias-corrected (BC), bias-corrected accelerated Bootstrap (BCa), bootstrapt und approximate bootstrap Konfidenzintervalle (ABC) (vgl. Tan (2012)).

3.2.4 Hierarchisch lineare Modelle

Allgemeine Motivation und Einleitung

Der Nachteil von OLS-Regressionsmodellen und Ansätzen, die dem generalisierten linearen Modell unterliegen wie logistische und Poisson-Regression, liegt darin, dass sie die Unabhängigkeit aller Beobachtungen untereinander voraussetzen (vgl. Cohen et al. (2003), Kap. 14). In dieser Arbeit wird ein Methodenkonzept vorgestellt, das vorrangig mit geclusterten Daten umgehen muss. Durch Nutzung eines der in 3.1 vorgestellten Clusteringverfahren werden Ähnlichkeiten zwischen den Beobachtungen gesucht und aufgedeckt. Die Annahme der Unabhängigkeit stellt nun also OLS-Ansätze für geclusterte Daten vor ein Problem. Cohen et al. (2003), Kap. 14.1 stellen drei verschiedene Möglichkeiten dar, diesem Problem zu begegnen, um trotzdem OLS-Regression auf geclusterte Daten anwenden zu können:

1. Das vorgenommene Clustering wird ignoriert. Jede Beobachtung wird so analysiert als gäbe es keine Struktur in Form einer Gruppierung in den Daten (disaggregierte Analyse)

2. Die Daten werden auf Gruppenebene (Cluster) aggregiert. Es werden Mittelwerte für jeden Prädiktor und für die Zielvariable für jede Gruppe gebildet. Die Gruppen sind dann die neuen Individuen der Analyse (aggregierte Analyse).

3. Die Regression wird auf der Ebene der einzelnen, ursprünglichen Beobachtungen durchgeführt. Daraus ergeben sich nochmals zwei unterschiedliche Ansätze zur Modellbildung:

    - Es gibt ein lineares Regressionsmodell für jedes Cluster. Unter der Annahme, dass die Steigung der Geraden über alle Gruppen konstant ist, unterscheiden sich die Modelle nur in ihrem Intercept.
    - Es gibt ein lineares Regressionsmodell für jedes Cluster. Die Modelle unterscheiden sich sowohl in ihrem Intercept als auch in der jeweiligen Steigung.

Die disaggregierte und aggregierte Analyse (s. 1. und 2.) schätzen unterschiedliche Regressionskoeffizienten. Während der disaggregierte Ansatz den Gesamt- Regressionskoeffizienten schätzt, schätzt der aggregierte Ansatz den Regressionskoeffizienten

zwischen den einzelnen Gruppen, da jede Gruppe nun als eigene Beobachtung betrachtet wird (vgl. Cohen et al. (2003), Kap. 14.2). In Regressionsanalysen über hierarchisch lineare Modelle bleiben die Rohdaten die Grundlage der Analyse (vgl. Walter und Rack (2009)).

Hierarchisch lineare Modelle sind der Disziplin der **Mehrebenenanalysen** zuzuordnen. Diese sind im Prinzip nichts anderes als Regressionsmodelle, können aber hierarchische Strukturen in den Daten mitberücksichtigen. Sie finden häufig Anwendung in den Sozialwissenschaften oder der Psychologie, in denen hierarchische Datenstrukturen sehr oft gegeben und zu untersuchen sind (z.B. verschiedene Gruppen). Die Untersuchung von Wechselwirkungen zwischen einzelnen Variablen auf den verschiedenen Ebenen und der abhängigen Zielvariablen ist mit konventionellen Methoden nicht möglich und kann zu verzerrten Schätzungen führen (vgl. Rusch (2011)).

Raudenbush und Bryk (2002), Kap. 1) nennen „Multilevel Modelle“ (Soziologie), „Mixed Effects Modelle“ (Biometrie), „Random Coefficient Regression Modelle“ (Ökonometrie) oder „Kovarianz Modelle“ (Statistik) als mögliche weitere, von ihrer bevorzugten Bezeichnung der hierarchisch linearen Modelle abweichenden, Bezeichnungen.

Alle genannten Ansätze verfolgen dieselben drei Ziele (vgl. Langer (2009), Kap. 4):

1. Untersuchung der Varianz von Beziehungen auf der untersten Ebene in hierarchischen Datenstrukturen

2. Erklärung dieser Varianz der Beziehungen aus den Kontextmerkmalen der nächsthöheren Aggregationsebene und Bestimmung der Cross-Level-Interaktionen

3. Möglichst unverzerrte Schätzung der Regressionskoeffizienten und deren Standardfehler unter ausdrücklicher Berücksichtigung der hierarchischen Datenstruktur

Bei einer Mehrebenenanalyse gibt es mindestens zwei Ebenen (vgl. Baltes-Götz (2013), Cohen et al. (2003), Kap. 14.4.3, Raudenbush und Bryk (2002)), die Individualebene und die Aggregatebene.

Beschreibung und Notation im Folgenden orientieren sich, wenn nicht anders gekennzeichnet, an Walter und Rack (2009). Das grundsätzliche Konzept von hierarchisch linearen Modellen ist es, den Einfluss von Variablen der Ebene 1 zunächst durch separate (für jede Gruppe (Cluster)) Regressionsanalysen zu untersuchen. Bei Vorliegen von signifikanter Varianz zwischen den einzelnen Regressionskonstanten β_{0j} und Regressionskoeffizienten β_{1j}, fließen diese im nächsten Schritt als abhängige Variablen in eine Regressionsgleichung der Ebene 2 ein. Die Gleichungen der verschiedenen Ebenen werden ineinander überführt.

Notation und Definition

Auf Ebene 1 sei Y_{ij} die abhängige Variable, X_{ij} ein Prädiktor, β_{0j} eine Regressionskonstante, β_{1j} ein Regressionskoeffizient, r_{ij} das Residuum und σ^2 die Residuenvarianz (Individuum i, Gruppe j). Auf Ebene 2 sei G_j ein Prädiktor, γ_{00} die Regressionskonstante, γ_{10} ein Regressionskoeffizient, u_{0j} das Residuum bei Schätzung von β_{0j}, u_{1j} das Residuum bei Schätzung von β_{1j} und T die Matrix der Varianz-Kovarianz-Komponenten τ_{00}, τ_{01}, τ_{10} und τ_{11}. Die Annahme, dass Varianzunterschiede in den Regressionskonstanten (Intercept) bzw. in den Regressionskoeffizienten (Steigung) zwischen den Gruppen vorliegen, ist die Motivation für die Anwendung der hier beschriebenen Modelle. In Abbildung 4 zeigt Hofmann (1997) vier Möglichkeiten von Beziehungen zwischen den Gruppen auf (A: kein Varianzunterschied, B: unterschiedliche Intercepts, C: unterschiedliche Steigungen, D: unterschiedliche Intercepts und Steigungen).

Walter und Rack (2009) skizzieren fünf Submodelle der hierarchisch linearen Modelle:

1. Einfaktorielle Varianzanalyse mit Zufallseffekten
2. Regression mit Vorhersage der Mittelwerte
3. Regression mit Zufallskoeffizienten
4. Einfaktorielle Kovarianzanalyse mit Zufallseffekten
5. Regression mit Vorhersage der Mittelwerte und Regressionskoeffizienten

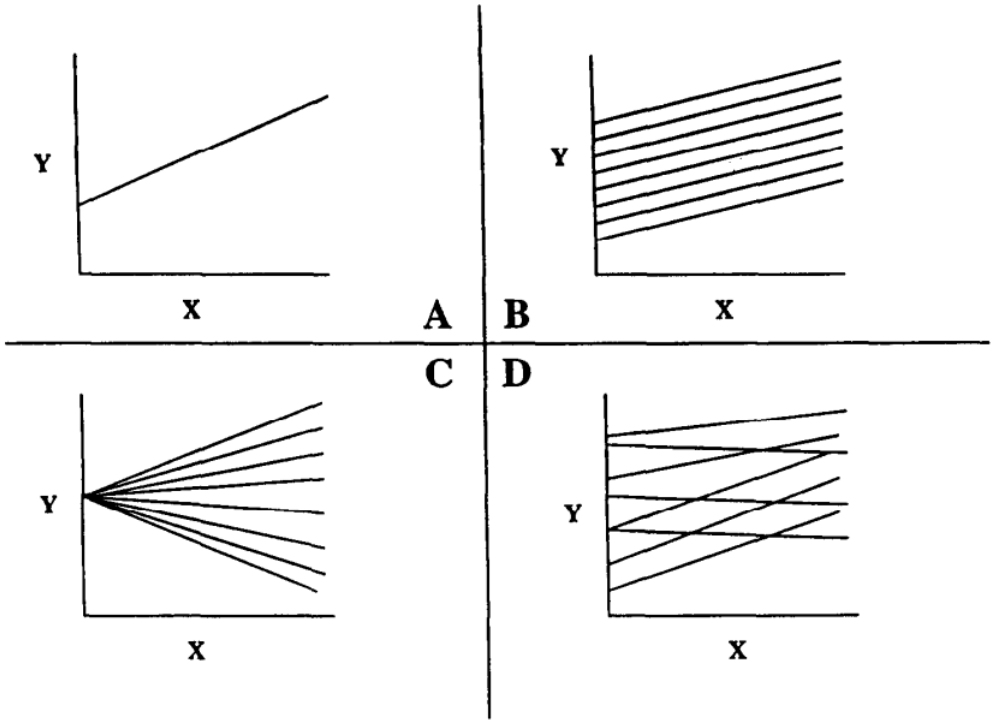

Abbildung 4: Mögliche Beziehungen von Regressionskoeffizienten und -steigungen (vgl. Hofmann (1997) und Walter und Rack (2009))

Durch das erste Submodell lässt sich die Variation der abhängigen Variablen Y_{ij} zwischen den Gruppen untersuchen und die Frage beantworten, ob die Zugehörigkeit zu einer Gruppe zusätzliche Varianz erklärt. Außer der Gruppenzugehörigkeit wird keine weitere Variable aufgenommen. Die Gleichungen der beiden Ebenen lauten dann wie folgt:

$$\text{Ebene 1: } Y_{ij} = \beta_{0j} + r_{ij};\ \text{Ebene 2: } \beta_{0j} = \gamma_{00} + u_{0j} \tag{54}$$

β_{0j} ist der Intercept, der den erwarteten Wert von Y_{ij} darstellt, wenn der Prädiktor und das Residuum r_{ij} den Wert 0 annehmen. Auf Ebene 2 sind entsprechend γ_{00} und u_{0j} Intercept bzw. Residuum. Die Gesamtvarianz der abhängigen Variablen ergibt sich durch die Varianz innerhalb einer Gruppe und die Varianz zwischen den Gruppen. Über die Intraklassen-Korrelation ICC lässt sich das Verhältnis dieser beiden Größen berechnen (vgl. Kreft und de Leeuw (1998), Kap. 4.2.2):

$$ICC = \frac{VAR(u_{0j})}{VAR(u_{0j}) + VAR(r_{ij})} = \frac{\hat{\tau}_{00}}{\hat{\tau}_{00} + \hat{\sigma}^2} \tag{55}$$

Ein geringer ICC-Wert ist gleichbedeutend mit einem unbedeutenden Varianzanteil zwischen den Gruppen.

Das zweite Submodell bezieht auf Ebene 2 (Aggregatebene) unabhängige Aggregat-

variablen G_j mit ein. So soll überprüft werden, ob sich Unterschiede in den Mittelwerten der einzelnen Gruppen durch diese Variablen erklären lassen. Die Modellgleichungen in den Ebenen lauten dann:

$$\text{Ebene 1: } Y_{ij} = \beta_{0j} + r_{ij};\ \text{Ebene 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}G_j + u_{0j} \tag{56}$$

Im dritten Submodell können die Steigungen zwischen den Gruppen variieren, d.h. nicht nur die Regressionskonstanten (Intercepts) sondern auch die Regressionskoeffizienten sind zufällig bzw. variabel (vgl. Szenario D in Abbildung 4). Daraus folgen die Gleichungen

$$\text{Ebene 1: } Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij};\ \text{Ebene 2: } \beta_{0j} = \gamma_{00} + u_{0j} \text{ und } \beta_{1j} = \gamma_{10} + u_{1j} \tag{57}$$

Die unterschiedlichen Steigungen (in Ebene 2 beschrieben) werden dabei als Differenzwerte von der mittleren Steigung beschrieben. In dem Modell kommen außer der Regressionskonstanten und Regressionskoeffizienten keine Prädiktoren mehr hinzu. Es lässt sich aber anhand der Varianzanteile ermitteln, für welche Koeffizienten die Annahme fixer Effekte (geringer Varianzanteil) und für welche Koeffizienten die Annahme zufälliger Effekte (hoher Varianzanteil) zulässig ist.

Das vierte Modell der einfaktoriellen Kovarianzanalyse mit Zufallseffekten als Spezialfall des dritten Submodells findet in Walter und Rack (2009) nur kurze Erwähnung. Mögliche Effekte der Individualmerkmale X_{ij} auf Ebene 1 können durch eine Adjustierung der Gruppeneffekte durch Individualeffekte kontrolliert werden. Die Autoren verweisen dabei auf Ditton (1998), Kap. 2.2.3.

Im Modell der Regression mit Vorhersage der Mittelwerte und Regressionskoeffizienten (Submodell 5) lassen sich jetzt (im Gegensatz zu Submodell 3) die Unterschiede in den Regressionskoeffizienten erklären. Dazu werden geeignete Aggregatmerkmale als Prädiktor G_j eingebunden. Auf Ebene 1 wird Y_{ij} durch eine oder mehrere Individualvariablen X_{ij} vorhergesagt. Die Ebene 1-Gleichung bleibt zu Submodell 3 identisch:

$$\text{Ebene 1: } Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij} \tag{58}$$

Auf Ebene 2 werden die Regressionskoeffizienten β_{1j} jetzt zu abhängigen Variablen, deren Varianz zwischen den Gruppen durch ein Aggregatmerkmal G_j erklärt werden

soll. So ergeben sich für Ebene 2 folgende Gleichungen:

$$\text{Ebene 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}G_j + u_{0j} \text{ und } \beta_{1j} = \gamma_{10} + \gamma_{11}G_j + u_{1j} \tag{59}$$

Das Submodell 5 wird als allgemeines hierarchisch lineares Modell bezeichnet, durch Nullsetzen einzelner Parameter können daraus wiederum die Submodelle 1-4 abgeleitet werden.

Parameterschätzung

An dieser Stelle werden verschiedene Möglichkeiten zur Parameterschätzung für hierarchisch lineare Modelle erwähnt. Für eine detailliertere Beschreibung der Schätzmethoden s. z.B. Raudenbush und Bryk (2002), Kap. 3 oder daran angelehnt Rusch (2011), an dessen Arbeit sich dieses Kapitel größtenteils orientiert.
In einem Modell auf zwei Ebenen gilt es, die fixen Effekte, die zufälligen Koeffizienten und die Varianz-Kovarianz-Komponenten zu schätzen. Fixe Effekte sind Parameterschätzwerte ohne Variation zwischen den Gruppen (hier: γ_{00}, γ_{10} etc.). Können die Parameterschätzwerte zwischen den Gruppen variieren, spricht man von zufälligen Koeffizienten (hier β_{0j}, β_{1j} etc.). Die Varianz-Kovarianz-Komponenten beinhalten sowohl die Varianz der Residuen der ersten und zweiten Ebene (hier: σ^2 bzw. τ_{00}, τ_{11} etc.) als auch die Kovarianz der Residuen auf Ebene 2 (hier: τ_{10}, τ_{01} etc.) (vgl. Walter und Rack (2009)).
Die fixen Effekte auf Ebene 1 und Ebene 2 können über eine Punktschätzung durch lineare Modelle ermittelt werden. Die Schätzung der zufälligen Koeffizienten (Zufallsgrößen) kann z.B. durch

- Maximum Likelihood-Schätzung
- Empirische Bayes-Schätzung
- Vollständige Bayes-Schätzung

erfolgen.
Die Schätzung der Varianz-Kovarianz-Komponenten findet üblicherweise über Maximum Likelihood-Schätzungen statt. Dabei gibt es wiederum zwei verschiedene Möglichkeiten:

- Vollständige Maximum Likelihood-Methode
- Restringierte Maximum Likelihood-Methode

(vgl. Rusch (2011))

3.2.5 Gütekriterien für hierarchisch lineare Modelle

Das grundsätzliche Ziel der Regression, möglichst viel der empirisch vorhandenen Varianz der abhängigen Variablen zu erklären, bleibt auch für hierarchisch lineare Modelle bestehen. Hier wird diese allerdings in zwei Komponenten zerlegt: Die Varianz innerhalb des Clusters (Ebene 1) und die Varianz zwischen den Clustern (Ebene 2). Die Bewertung eines hierarchisch linearen Modells durch die Quantifizierung der durch die abhängigen Variablen erklärten Varianz stellt sich etwas komplizierter dar, da, wie oben beschrieben, mehrere Varianzkomponenten vorliegen.
Es ist also notwendig, für jede dieser Varianzkomponenten ein R^2 zu berechnen. An dieser Stelle werden zwei Möglichkeiten für diese Berechnung dargestellt. Die Beschreibung, Herleitung und Notation der ersten Möglichkeit orientiert sich, wenn nicht anders gekennzeichnet, an Snijders und Bosker (1994).
Snijders und Bosker (1994) sehen das Prinzip der proportionalen Verringerung des Vorhersagefehlers als das ansprechendste an, um erklärte (modellierte) Varianz in Multilevel-Modellen zu messen. Dies lässt sich folgendermaßen beschreiben: Gegeben sei eine Population von Werten (X_i, Y_i) mit einer bekannten Wahrscheinlichkeitsfunktion, β sei der Wert des Vektors v, für den der erwartete quadratische Fehler $E(Y_i - X_i v)^2$ minimal wird. Ist X_i unbekannt, so ist $E(Y)$ der beste Schätzer für Y_i mit mittlerem quadratischen Fehler $var(Y_i)$. Wenn X_i gegeben ist, ist der lineare Schätzer von Y_i mit minimalem quadratischen Fehler der Regressionswert $X_i\beta$, dessen mittlerer quadratischer Fehler $E(Y_i - X_i\beta)^2$ beträgt. Die proportionale Verringerung des mittleren quadratischen Fehlers der Prädiktion ist dann definiert als

$$\frac{var(Y_i) - var(Y_i - X_i\beta)}{var(Y_i)} = 1 - \frac{var(Y_i - X_i\beta)}{var(Y_i)}. \tag{60}$$

Dies ist lediglich eine andere Definition des Bestimmtheitsmaßes R^2, vgl. 3.2.2.
Dasselbe Prinzip wird nun angewendet, um die erklärte (modellierte) Varianz in Mehrebenenmodellen zu definieren. Zunächst stellt sich die Frage, ob im Zwei-

Ebenen-Fall ein individueller Wert Y_{ij} auf Ebene 1 oder ein aggregierter Wert $\bar{Y}_{.j}$ auf Ebene 2 vorhergesagt wird. Als Basis dient ein Modell auf zwei Ebenen mit einem zufälligen Intercept und einigen Prädiktor-Variablen mit fixen Effekten, aber ohne weitere zufällige Effekte

$$Y_{ij} = X_{ij}\beta + U_{0j} + E_{ij}, \tag{61}$$

wobei die zufälligen Residuen U_{0j} und E_{ij} unkorreliert (jeweils mit Erwartungswert Null und Varianz τ^2 bzw. σ^2) und die X_i zufällig und unkorreliert mit den Variablen U und E sind. Temporär sei nun angenommen, der Vektor β der Regressionskoeffizienten sei bekannt. Für die erklärte (modellierte) Varianz auf Ebene 1 gelte es, für eine zufällig gewählte Einheit i auf Ebene 1 und einer zufällig gewählten Einheit j auf Ebene 2 den Wert Y_{ij} vorherzusagen. Wenn die Werte der Prädiktoren X_{ij} unbekannt sind, ist der Erwartungswert von Y_{ij} der beste Schätzer: $\mu\beta$ mit $\mu = EX_{ij}$ und mittlerem quadratischen Fehler $var(Y_{ij})$. Wenn die Werte der Prädiktoren X_{ij} gegeben sind, ist der beste lineare Schätzer von Y_{ij} mit minimalem quadratischen Fehler der Regressionswert $X_{ij}\beta$, dessen mittlerer quadratischer Fehler $var(Y_{ij} - X_{ij}\beta) = \sigma^2 + \tau^2$ beträgt. Dann lässt sich der erklärte Anteil der Varianz auf Ebene 1 als die proportionale Verringerung des mittleren quadratischen Vorhersagefehlers definieren:

$$R_1^2 = 1 - \frac{var(Y_{ij} - X_{ij}\beta)}{var(Y_{ij})} \tag{62}$$

Im Falle von unbalancierten Daten ist die Varianz der Stichprobe nicht unbedingt ein guter Schätzer für $var(Y_{ij})$. Snijders und Bosker (1994) empfehlen stattdessen auf $\hat{\sigma_0^2} + \hat{\tau_0^2}$ zurückzugreifen, wobei $\hat{\sigma_0^2}$ und $\hat{\tau_0^2}$ definiert sind als Schätzer für das Zwei-Ebenen-Modell mit einem zufälligen Intercept aber ohne Prädiktoren („Null-Modell“ (Kreft und de Leeuw (1998), Kap. 3.4)):

$$Y_{ij} = \beta_0 + U_{0j} + E_{ij}. \tag{63}$$

Da die Varianz der Stichprobe und $\hat{\sigma_0^2} + \hat{\tau_0^2}$ zwei Schätzer für denselben Parameter sind, sollten diese sich nicht signifikant unterscheiden. Um nun das R^2 zu schätzen, zieht man $\hat{\sigma_0^2} + \hat{\tau_0^2}$ sowohl für Gleichung 63 als auch für Gleichung 61 heran und zieht den Quotienten der beiden Werte von 1 ab. Anschaulich ist das R^2 dann nichts anderes als die proportionale Verringerung des $\hat{\sigma_0^2} + \hat{\tau_0^2}$ (also der Varianz) durch Hinzunahme der X Variablen in das Modell.

Analog wird nun das R^2 für die Ebene 2 als proportionale Verringerung des mittleren

quadratischen Fehlers für die Vorhersage von $\bar{Y}_{.j}$ für eine zufällig gewählte Einheit j auf Ebene 2 definiert.

Wenn die Werte der X_{ij} für das Set von den i Einheiten aus Ebene 1 für die Ebene 2 Einheit j unbekannt sind, ist der Erwartungswert von $\bar{Y}_{.j}$ der beste Schätzer: $\mu\beta$ mit $\mu = EX_{ij}$ und mittlerem quadratischen Fehler $var(\bar{Y}_{.j})$. Wenn die Werte der X_{ij} für alle i in der Gruppe j gegeben sind, ist der beste lineare Schätzer von $\bar{Y}_{.j}$ der Regressionswert $\bar{X}_{.j}\beta$, dessen mittlerer quadratischer Fehler $var(\bar{Y}_{.j} - \bar{X}_{.j}\beta) = \frac{\sigma^2}{n_j} + \tau^2$ beträgt.

Dann lässt sich der erklärte (modellierte) Anteil der Varianz auf Ebene 2 als die proportionale Verringerung des mittleren quadratischen Vorhersagefehlers für $\bar{Y}_{.j}$ definieren:

$$R_2^2 = 1 - \frac{var(\bar{Y}_j - \bar{X}_{.j}\beta)}{var(\bar{Y}_{.j})} \tag{64}$$

Gleichung 64 ist ähnlich der Definition des klassischen R^2 für eine aggregierte Regressionsanalyse. Anschaulich ist das R^2 dann wieder nichts anderes als die proportionale Verringerung der Varianz ($\frac{\hat{\sigma^2}}{n} + \hat{\tau^2}$).

2013 wurden durch Nakagawa und Schielzeth (2013) neue Möglichkeiten, ein R^2 im Rahmen von hierarchisch linearen Modellen zu berechnen, aufgezeigt. So führen sie das **marginale** und **konditionale** R^2 ein. Die folgende Notation und Beschreibung richtet sich, wenn nicht anders gekennzeichnet, an Nakagawa und Schielzeth (2013). Sei vereinfacht ein Modell mit zwei zufälligen Effekten angenommen, um die beiden Maße einzuführen:

$$y_{ijk} = \beta_0 + \sum_{h=1}^{p} \beta_h x_{hijk} + \gamma_k + \alpha_{jk} + \epsilon_{ijk} \tag{65}$$

y_{ijk} sei die i-te Ausprägung des j-ten Objekts in der k-ten Gruppe (Cluster), x_{hijk} der i-te Wert des j-ten Objekts in der k-ten Gruppe für den h-ten Prädiktor, γ_k der Gruppen-spezifische Effekt mit Varianz σ_γ^2, α_{jk} der Objekt-spezifische Effekt mit Varianz σ_α^2 und ϵ_{ijk} das Residuum mit Varianz σ_ϵ^2. So kann ein **marginales** R^2 für hierarchisch lineare Modelle definiert werden als:

$$R_m^2 = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_\gamma^2 + \sigma_\alpha^2 + \sigma_\epsilon^2} \tag{66}$$

σ_f^2 sei hier die Varianz berechnet aus den fixen Effekten des Modells

$$\sigma_f^2 = var(\sum_{h=1}^{p} \beta_h x_{hijk}) \tag{67}$$

Das **marginale** R^2 aus Gleichung 66 ist das Verhältnis der Varianz erklärt durch die fixen Effekte.

Das **konditionale** R^2 ist das Verhältnis der Varianz erklärt durch sowohl die fixen als auch die zufälligen Effekte (vgl. dazu u.a. Nakagawa et al. (2017)):

$$R_c^2 = \frac{\sigma_f^2 + \sigma_\alpha^2}{\sigma_f^2 + \sigma_\gamma^2 + \sigma_\alpha^2 + \sigma_\epsilon^2} \tag{68}$$

3.2.6 Unsicherheit der Güte von hierarchisch linearen Modellen

Wird ein hierarchisch lineares Modell mehrfach berechnet (z.B. auf verschiedenen Stichproben aus der Grundgesamtheit), ist es auch möglich, sich die empirische Verteilung der in 3.2.5 definierten Gütemaße über mehrere Iterationen anzuschauen. Auch wenn das definierte Maß nicht als klassisches Qualitätskriterium für das Modell gedacht ist (Nakagawa und Schielzeth (2013)), so eignet es sich trotzdem zum Vergleich der erklärten Varianz in den verschiedenen Modellen.

Analog zu 3.1.4 geschieht die Bewertung der Sensitivität der Güte über die empirische Verteilung über die Iterationen und Berechnung von Minima, Maxima, Standardabweichung, Varianz und Quantile (hier: 5%, 50% und 95%). Zur Definition dieser Kenngrößen s. z.B. Vogel (2005), Kap. 3.3.

3.2.7 Zusammenfassung

In diesem Kapitel wurde das Aufstellen von Linearen Modellen durch einfache und multiple Regressionsanalyse grundätzlich hergeleitet. Dieses Verfahren ist immer anwendbar, sobald sowohl die abhängigen als auch die unabhängigen Variablen quantitativ vorliegen und ein linearer Zusammenhang angenommen werden kann.

Im weiteren wurde das Bestimmtheitsmaß als Gütekriterium für lineare Modelle eingeführt, das den Anteil der erklärten Streuung an der Gesamststreuung misst. Die Sensitivität des Gütekriteriums über z.B. mehrere Iterationen hinweg kann zum einen klassisch über die empirische Verteilung des Maßes oder durch ein von Olkin, Finn (1995) beschriebenes Konfidenzintervall bewertet werden. Somit kann zum einen die Güte eines gewählten Modells bewertet werden als auch dessen Sensitivität gegenüber z.B. unterschiedlichen, bereits im Prozessschritt des Clusterings gewählten Startbedingungen quantifiziert werden.

Als spezielles lineares Regressionsmodell wurde in diesem Kapitel das hierarchisch

lineare Modell hergeleitet. Es findet später im Kontext dieser Arbeit Anwendung, da es im Gegensatz zu den meisten anderen linearen Modellen mit hierarchischen Strukturen in den Daten umgehen kann.

3.3 Nichtlineare Modelle

Wie in 3.2 beschrieben, ist Modellbildung immer eine Gratwanderung zwischen Simplifizierung und Verkomplizierung. Für viele Fragestellungen ist es ausreichend, nichtlineare Zusammenhänge durch lineare Modelle zu approximieren, da die nichtlineare Regression mit einigen Schwierigkeiten verbunden ist und hohen Rechenaufwand nach sich zieht. Dies bedingt sich darin, dass sich das zu lösende, mathematische Problem nicht mehr analytisch lösen lässt; die Schätzwerte müssen mithilfe iterativer Algorithmen berechnet werden (vgl. Backhaus et al. (2016), Teil III, Kap. 10.1). Des weiteren müssen Startwerte für die zu schätzenden Parameter bereits vom Anwender festgelegt werden. Überdies gibt es keine Garantie, dass iterative Algorithmen ein globales Optimum finden (vgl. Backhaus et al. (2015), Kap. 1.1).

3.3.1 Nichtlineare Regressionsmodelle

Die Vorgehensweise bei der nichtlinearen Regressionsanalyse lässt sich in folgende Schritte unterteilen, an denen sich auch die weitere Gliederung dieses Kapitels orientiert:

1. Formulierung von Modellen
2. Finden von Startwerten
3. Schätzung der Modelle
4. Prüfung der Modelle

Notation und Beschreibung der nichtlinearen Regressionsanalyse orientieren sich hier, wenn nicht anders gekennzeichnet, an Backhaus et al. (2015), Teil II, Kap. 1.2).

Formulierung von Modellen

Es existieren nahezu unendlich viele Möglichkeiten, nichtlineare Modelle zu formulieren, meist beschränkt man sich jedoch auf Standardmodelle (z.B. Exponential- oder Potenzmodell).

Finden von Startwerten

Wie oben beschrieben, ist es bei nichtlinearen Modellen von Nöten, zweckmäßige Startwerte der iterativen Algorithmen zu finden und festzulegen. Dies erfordert neben dem mathematischen Verständnis auch Erfahrung bzw. Vorwissen, was das vorliegende Problem angeht. Von der Wahl der Startwerte hängt oft ab, ob und wenn ja, wie schnell der Algorithmus das globale Optimum findet.
Um geeignete Startwerte zu finden wird oftmals das nichtlineare Modell solange vereinfacht, bis es sich durch lineare Regression lösen lässt. Die erhaltenen Schätzwerte werden dann zu den Startparametern umgeformt.

Schätzung der Modelle

Um die Modellparameter der nichtlinearen Regression zu schätzen, kann analog zur linearen Regression die Methode der kleinsten Quadrate (s. Gleichung 38) verwendet werden. Ist das zu schätzende Regressionsmodell intrinsisch nichtlinear in den Parametern, so müssen die Schätzwerte numerisch ermittelt werden, indem man Gleichung 38 mithilfe iterativer Algorithmen minimiert. Dazu gibt es in der Literatur eine Vielzahl von Optimierungsverfahren; die Betrachtung dieser ist nicht Bestandteil dieser Arbeit. Das Problem des iterativen Algorithmus ist die Ungewissheit, ob und wenn ja, wann man das Optimum (in diesem Fall das Minimum) gefunden hat. Daher wird ein Konvergenzkriterium definiert. Dies liefert jedoch keine Garantie, ein globales Minimum hervorzubringen. Ob und wie schnell ein globales Minimum gefunden wird, hängt also von dem iterativen Algorithmus, der Rechengenauigkeit, der Größe des Konvergenzkriteriums, den Startwerten und ggf. weiteren Parametern ab.

Prüfung der Modelle

Die Prüfung der Modelle umfasst bei den nichtlinearen Modellen neben der statistischen auch eine sachlogische Ebene. Diese ist sehr individuell und wird hier nicht näher beschrieben.
Die statistische Prüfung der globalen Güte erfolgt, ebenso wie bei der linearen Regression, durch das Bestimmtheitsmaß R^2. Dadurch lassen sich auch verschiedene getestete Modelle miteinander vergleichen.
Durch die Störgröße u sind automatisch auch die geschätzten Modellparameter fehlerhaft. Die Schätzwerte der Modellparameter sind Werte von Zufallsvariablen, deren Standardabweichung als Standardfehler der Parameter bezeichnet wird. Dieser ist sowohl proportional abhängig von der Standardabweichung von u als auch proportional abhängig von der Streuung der unabhängigen Variablen. Mit der Störgröße u ist auch deren Standardabweichung nicht beobachtbar und somit durch die Standardabweichung der Residuen, also dem Standardfehler des Modells zu ersetzen.
Die Standardfehler der Parameter lassen sich hier nur asymptotisch berechnen.
Durch Konfidenzintervalle, die mithilfe der Standardfehler gebildet werden, lässt sich eine genauere Aussage über die Genauigkeit der Parameterschätzung treffen. Diese sagen aus, in welchem Bereich die unbekannten, wahren Modellparameter vermutlich liegen. Das Konfidenzintervall um β sieht folgendermaßen aus:

$$b - t_{\alpha/2} \cdot s_b \leq \beta \leq b + t_{\alpha/2} \cdot s_b. \tag{69}$$

β ist der unbekannte Modellparameter, b dessen Schätzwert, $t_{\alpha/2}$ der Wert der Student-Verteilung für die Irrtumswahrscheinlichkeit $\alpha/2$ und s_b der Standardfehler von b. Dieses Konfidenzintervall gilt ebenfalls nur asymptotisch.
Alternativ lässt sich mithilfe von Bootstrapping-Verfahren der Standardfehler berechnen. Dabei werden wiederholt Stichproben aus der Datenmenge gezogen und dadurch Verteilungen von Schätzwerten des Modellparameters ermittelt. Über die Standardabweichung der Schätzwerte kann der Standardfehler des Parameters geschätzt werden.

3.3.2 Gütekriterien für nichtlineare Regressionsmodelle

Zur Prüfung der globalen Güte von nichtlinearen Regressionsmodellen kann ebenfalls das in 3.2.2 definierte Bestimmtheitsmaß R^2 herangezogen werden.

3.3.3 Unsicherheit der Güte von nichtlinearen Regressionsmodellen

Analog zu 3.2.3 lassen sich auch für nichtlineare Regressionsmodelle Konfidenzintervalle für das Bestimmtheitsmaß R^2 definiert nach Soper (2016) ermitteln.

3.3.4 Zusammenfassung

In diesem Kapitel wurden die nichtlinearen von den linearen Modellen abgegrenzt. Analog zu den linearen Modellen wurden die grundsätzlichen Schritte von der Modellformulierung bis zur Prüfung des Modells beschrieben. Die Beschreibung sowohl der Güte als auch der Sensitivität der Güte ist analog zu den linearen Modellen und wird daher in diesem Unterkapitel nur erwähnt.

3.4 Prozess-Modell

Nachdem im bisherigen Verlauf von Kapitel 3 die einzelnen statistischen Methoden beschrieben wurden, werden diese nun zu einem Gesamt-Analyseprozess zusammengeführt. Dieser soll als Methodenbaukasten für Analysen von Belastungs- und Verschleißverhalten geeignet sein und wird entsprechend in Kapitel 4 an einem Beispiel erprobt. Der Methodenbaukasten ist in Abbildung 5 dargestellt und wird nun Schritt für Schritt beschrieben.

Die Datenvorverarbeitung in einer statistischen Analyse stellt in vielen Fällen einen Großteil des Aufwandes dar (bis zu 70%) (vgl. Müller und Lenz (2013), Kap. 3.1.1). Auch in dem später beschriebenen Anwendungsbeispiel des Methodenbaukastens (s. 4.2.1) spielt die Vorverarbeitung eine wichtige Rolle. Da die aber nicht im methodischen Fokus dieser Arbeit steht, wird auf eine nähere Beschreibung dieses Arbeitsschrittes an dieser Stelle verzichtet. Die notwendigen Rechenoperationen und Normierungen werden direkt in 4.2.1 und 4.2.2 beschrieben.

Übergreifend sei zunächst festgelegt, dass im Rahmen der Analyse zunächst ein

Clustering der Daten und schließlich ein Prognosemodell erstellt wird. Dafür werden im Vorfeld sowohl die zu clusternden Daten als auch die Zielgröße der Prognose definiert. Als methodischer Rahmen sei das Evidence Accumulation Clustering hier gegeben. Das in 3.1.2 beschriebene Verfahren umfasst einen Durchlauf über mehrere Iterationen (vgl. in Algorithmus 1 "N-mal"). In jeder Iteration wird sowohl das Clustering als auch die Prognosemodelle inklusive der jeweiligen Gütemaße berechnet. Dadurch soll im weiteren Verlauf des Analyseprozesses die Sensitivität der Zwischenergebnisse beurteilt werden können.

In **Schritt 1** wird zunächst das Clusterverfahren gewählt, d.h. die Entscheidung gefällt, ob es sich um ein distanzbasiertes, dichtebasiertes oder modellbasiertes Verfahren handeln soll (vgl. 3.1.1). So kann im nächsten Schritt **(2)** die Entscheidung für den passenden Algorithmus getroffen werden. Als Beispiel-Algorithmen für die drei genannten Verfahren seien hier k-means (distanzbasiert), DBSCAN (dichtebasiert) und Finite Mixture Modelle (modellbasiert) genannt (vgl. 3.1.1). Das zu wählende Clusterverfahren ist stark abhängig von den vorliegenden Daten. Für Querschnittsdaten, die hier im Fokus des Clustering liegen (s. 2.1), wird in der weiteren Analyse das k-means-Verfahren als partitionierendes, distanzbasiertes weiter betrachtet und bzgl. der Güte beleuchtet. Stellt es im Gegensatz zum gewählten Analysebeispiel keine Grundvoraussetzung dar, dass die Objekte (z.B. Fahrzeuge) nur genau einem Cluster zugeordnet werden sollen, wäre es ebenso möglich, sich für den Klassifikationstyp der Überdeckung zu entscheiden. Ist es gewünscht, eine Mindestanforderung an Homogenität innerhalb der Cluster vorab zu definieren, eignet sich der Klassifikationstyp der Hierarchie. Dies ist sinnvoll, wenn keine Kenntnis über die Anzahl der Cluster vorliegt bzw. keine Vorgabe gemacht werden soll.

Auch dichtebasierte Verfahren kommen ohne vorherige Definition der Anzahl der Cluster aus. Geht man auf Basis der vorliegenden Daten davon aus, dass die resultierenden Cluster keine konvexen Formen aufweisen, so sind dichtebasierte Verfahren von Vorteil. Der Aufwand und die Abhängigkeit von den jeweiligen Eingabeparametern bei dichtebasierten Verfahren wird als größter Nachteil gesehen (vgl. Busch (2005)).

Ist davon auszugehen, dass den einzelnen Clustern unterschiedliche Verteilungen zugrunde liegen, eignen sich modellbasierte Verfahren. Auch hier ist keine vorherige Annahme über die Anzahl der Cluster vorgesehen. Da der Ansatz auf Verteilungs-

annahmen basiert, weist er auch eine hohe Fehleranfälligkeit auf, wenn eben diese nicht korrekt sind (vgl. Wunder (2014)). Interessant sind modellbasierte Verfahren für Anwendungen, in denen statt einer einzigen, definitiven Zuordnung von Objekten zu einem Cluster, Wahrscheinlichkeiten von Zugehörigkeiten zu einem Cluster gefragt sind (vgl. Jäckle (2017), Kap. 3.6). In diesem Fall hat man die Möglichkeit, die Granularität der Partitionierung nachträglich selbst zu wählen, ohne vorher eine feste Clusteranzahl vorzugeben.

Die Wahl für ein geeignetes Gütekriterium für das Clustering (vgl. 3.1.3) kann dann in **Schritt 3** des Prozesses erfolgen. So gibt es in allen drei Verfahren die Möglichkeit, einen Silhouettenkoeffizienten zu berechnen. Für distanzbasierte Verfahren ist die Auswahl an dieser Stelle relativ groß. So seien hier beispielhaft noch Dunn´s Index oder der Calinski-Harabasz Index als mögliche Gütemaße genannt. Grundsätzlich wäre es auch möglich, in jeder Iteration mehrere Gütekriterien zu berechnen. Dies hätte den Vorteil, später auch eine Aussage über die Stabilität der jeweiligen Größen treffen zu können. Da in dieser Arbeit das Evidence Accumulation Clustering den methodischen Rahmen bildet, liegt hier der Silhoettenkoeffizient im Fokus. Dieser ist geeignet, Clusterergebnisse zu vergleichen, die durch unterschiedliche Anzahlen k von Clustern zustande kommen (vgl. 3.1.2). Andernfalls wird das für die Fragestellung passende Maß ausgewählt (z.B. unterschiedliche Wahl von Repräsentanten der Cluster (vgl. 3.1.3).

Um nun die Sensitivität der in Schritt 3 berechneten Güte des Clusterings beurteilen zu können, wird in **Schritt 4** die empirische Verteilung des Gütemaßes (aus den verschiedenen Iterationen) betrachtet. Dabei werden alle Clusterergebnisse mit einbezogen. Dies kann z.B. Ergebnisse aus verschiedenen Anzahlen von Clustern umfassen, wenn im Algorithmus keine feste Anzahl, sondern eine minimale und maximale Anzahl von Clustern vorgegeben wird. Lage- bzw. Streuungsmaße des Gütekriteriums werden abgeleitet (vgl. 3.1.4).

Der Ablauf hinsichtlich des Prognosemodells in **Schritt 5** bis **Schritt 7** läuft ähnlich zu Schritt 2 bis Schritt 4. Zunächst fällt die Wahl auf ein geeignetes Modell. Da hier im speziellen eine Kombination aus Clustering und Prognose gefordert wird, ist in diesem Anwendungsfall die Wahl auf ein hierarchisch lineares Modell naheliegend (vgl. 3.2.4). Wenn die vorliegenden Daten eine Approximation über lineare Modelle nicht erlauben, kann auf nichtlineare Modelle (vgl. 3.3) zurückgegriffen werden. Als

Beispiel für ein sehr häufig verwendetes und flexibles nichtlineares Modell sei hier das Potenz-Modell genannt (vgl. Backhaus et. al (2016), Teil II, Kap. 2.3.1.3).
In Zeitreihenverfahren bezieht die Analyse die Anordnung, in der die Beobachtungen gewonnen werden, mit ein (vgl. dazu z.B. Schlittgen, Streitberg (2001)). Zeitreihenverfahren können also im Kontext des Analysebeispiels interessant sein, wenn z.B. eine Verschleißgröße über die Zeit aufgezeichnet wird (z.B. einmal pro Monat) und ausgewertet werden kann. Hosoya et al. (2014) zeigen eine Möglichkeit der Anwendung von hierarchisch linearen Modellen auf Längsschnittdaten auf. Eine weitere Möglichkeit wäre die Wahl eines für den Anwendungsfall geeignetes Alterungs- bzw. Verschleißmodells.
Nach der Wahl des geeigneten Modells (**Schritt 5**) gilt es wiederum, die Güte des berechneten Modells zu quantifizieren (**Schritt 6**). Dies wäre z.B. in linearen Verfahren klassisch das Bestimmtheitsmaß R^2 (vgl. 3.2.2). Für hierarchisch lineare Modelle existiert dazu eine Abwandlung (vgl. 3.2.5). Bei Zeitreihenmodellen hingegen weicht man auf andere Maße aus (z.B. für lineare Trendmodelle: Median der absoluten Abweichungen (MAD), vgl. dazu z.B. Schlittgen, Streitberg (2001), Kap. 1.4.4).
In **Schritt 7**, dem letzten Schritt des Analyseprozesses, wird nun die Güte des Modells hinsichtlich ihrer Sensitivität untersucht. Dies geschieht analog zum Clustering anhand der empirischen Verteilung über alle Iterationen hinweg und Bestimmung der entsprechenden Lage- und Streuungsmaße (vgl. 3.2.6).
Die in Schritt 6 und 7 berechneten Güte- und Sensitivitätskriterien werden also von allen vorher verwendeten Algorithmen und Modellen beeinflusst. Sie beschreiben die Gesamt-Güte/Sensitivität des Analysepfades (s. vertikale Beschreibung in Abbildung 5).
Alle hier beschriebenen Schritte finden in jeder Iteration der Berechnung statt, d.h. mit jeder Iteration des Clusterings durch das Evidence Accumulation Clustering findet sofort die Berechnung der Clusteringsgüte und die Berechnung des Prognosemodells mit dessen Güte statt. Die Entscheidung für die endgültige Partition und somit auch die Anzahl der Cluster wird mithilfe des Evidence Accumulation Clusterings gefällt. Für diese Partition wird dann das hierarchisch lineare Modell berechnet. So führt der Analyseprozess zu folgenden inhaltlichen und methodischen Ergebnissen.

Inhaltliche Ergebnisse:

1. Die vorliegenden Daten werden auf Basis der gegebenen Einflussfaktoren geclustert.

2. Die abhängige Variable (Zielgröße) wird durch die Einflussfaktoren (unabhängige Variablen) für jedes Cluster über ein Prognosemodell dargestellt.

Methodische Ergebnisse:

1. In jedem Schritt der Analyse kann die Güte des jeweiligen Verfahrens bewertet werden.

2. In jedem Schritt der Analyse kann die Sensitivität der Güte des jeweiligen Verfahrens bewertet werden.

3. Durch die gleichzeitige Durchführung aller Verfahren in jeder Iteration werden Güte und Sensitivität der Güte über beide Analyse-Stufen (Clustering und Prognosemodell) hinweg betrachtet.

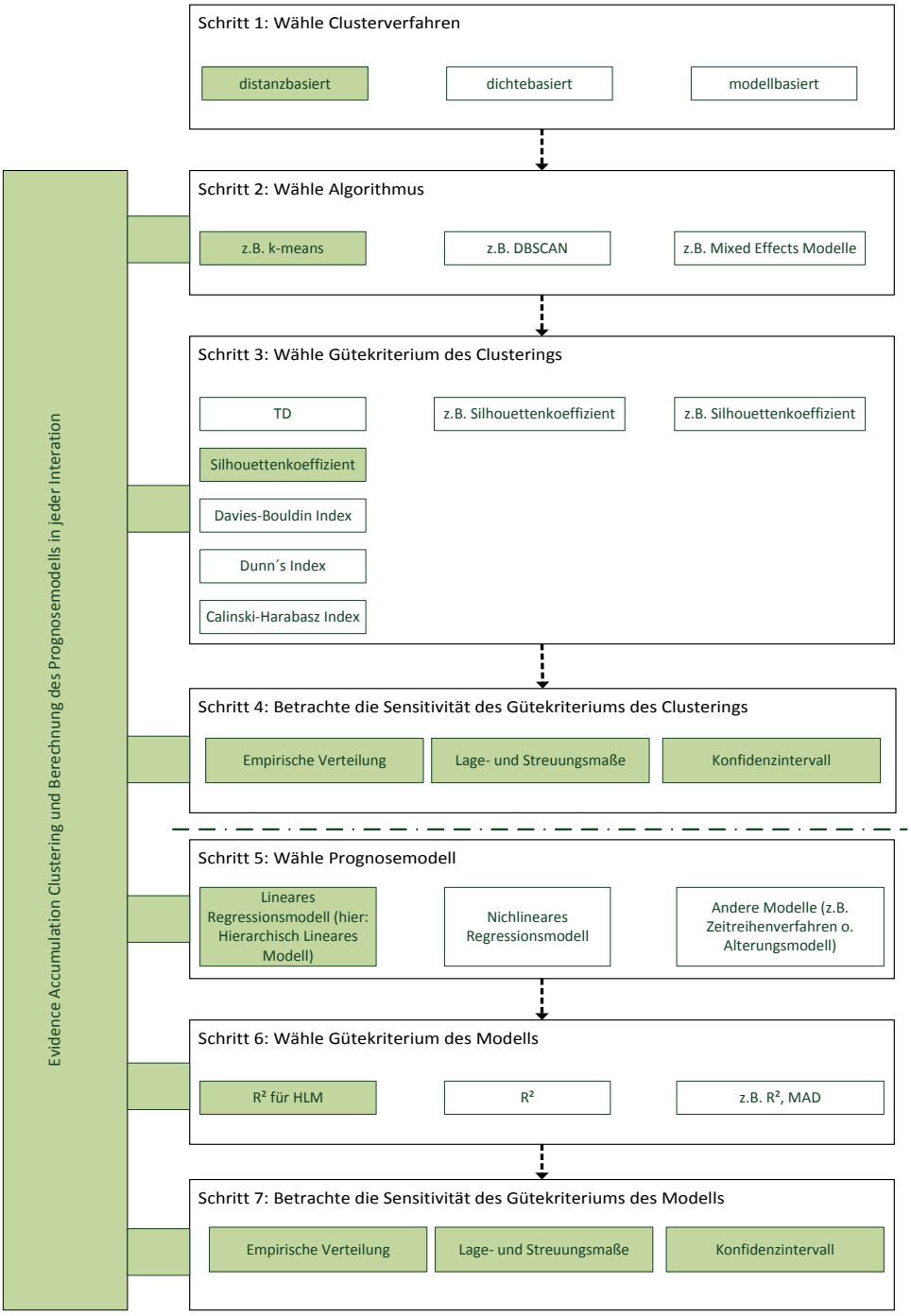

Abbildung 5: Prozessbild/Methodenbaukasten des statistischen Absicherungskonzeptes, Quelle: eigene Darstellung

4 Simulation und Analyse

In diesem Kapitel wird der beschriebene Methodenbaukasten an einem Beispiel im Kontext elektrischer Antriebskonzepte erprobt. Es werden Schritt für Schritt alle Stufen des Baukastens beschrieben und deren Ergebnisse dargestellt.

4.1 Beschreibung des Datensatzes

4.1.1 Allgemein

Für das Analysebeispiel wird eine Fahrzeug- bzw. Batterieflotte aus der Welt der Plug-In Hybride (vgl. 1.3.2) herangezogen, die sich aktuell im Feld befindet. Die Stichprobe umfasst ca. 5.000 Fahrzeuge/HV-Batterien, von denen jeweils ein kompletter Analyse-Datensatz zur Verfügung steht. Dieser umfasst sowohl die Belastungskollektive (vgl. 2.1) als auch die entsprechende Verschleiß- bzw. Alterungsgröße (vgl. 2.2). Es gehen nur Fahrzeuge in die Stichprobe ein, die bereits eine gewisse Zeit und eine gewisse Strecke gefahren sind.
Für die Analyse werden drei zweidimensionale BLKs und eine eindimensionale Alterungsgröße betrachtet. Die BLKs als Nutzungsdaten werden in Folge für das Clustering, die Alterungsgröße als Zielgröße der Regression verwendet.
Als BLKs werden herangezogen: *1. Temperatur HV-Batterie/SoC HV-Batterie*, *2. Lade-Entladezyklen HV-Batterie* und *3. Strom der HV-Batterie/Zeit der Pulse*. Diese Größen wurden im Vorfeld als drei Stellvertreter für das Nutzungs-bzw. Belastungsverhalten auf die HV-Batterie ausgewählt und sind in diesem Analysebeispiel als gegeben zu bewerten. Als Verschleißgröße wird die Kapazität der HV-Batterie ausgewählt.
Die des dargestellten Beispiels in Tabelle 1 in 2.1 entsprechende Matrix des BLKs *1: Temperatur HV-Batterie/SoC HV-Batterie* wird in Tabelle 2 dargestellt. Analog finden sich die Matrizen von BLK *2: Lade-Entladezyklen HV-Batterie* und BLK *3: Strom der HV-Batterie/Zeit der Pulse* in Tabelle 3 und Tabelle 4. An dieser Stelle werden die Matrizen nur gekürzt dargestellt, eine gesamtheitliche Betrachtung findet im Rahmen der deskriptiven Analyse in 4.1.2 (Abbildung 8, Abbildung 11, Abbildung 12) statt.
Das BLK *Temperatur HV-Batterie/SoC HV-Batterie* umfasst ingesamt 64 Einträge

in der Matrix. Die Temperatur der Batterie wird in Grad Celsius ($^{\circ}C$) aufgetragen, der SoC (Ladezustand, vgl. 1.3.2) in Prozent (%).

Die *Lade-Entladezyklen HV-Batterie* werden in der Rainflow-Zählung, einem in der Betriebsfestigkeit etablierten Zählverfahren, aufgezeichnet (vgl. Köhler et al. (2012), Kap. 2.4.4). Dabei wird die Start- und die Zielklasse eines Lastwechsels betrachtet und ein Eintrag in der jeweiligen Zelle der Matrix vorgenommen. Es gibt unterschiedliche Möglichkeiten, die Matrix zu befüllen; diese sind in Abbildung 6 dargestellt.

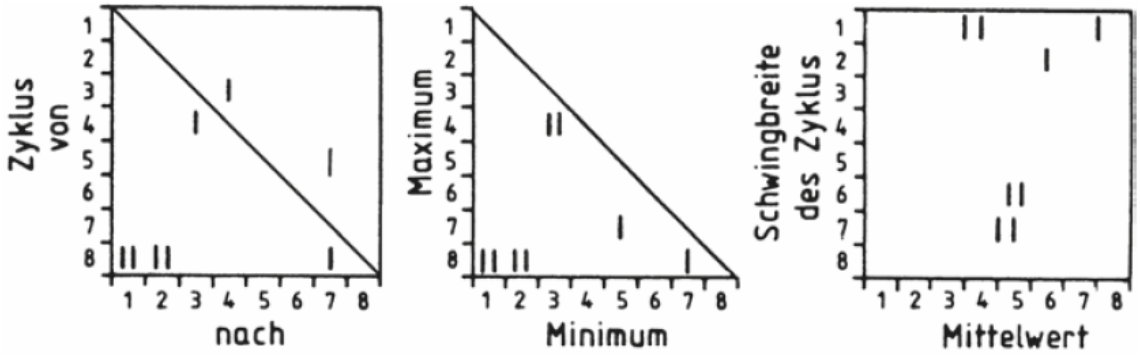

Abbildung 6: Drei verschiedene Möglichkeiten der Matrizendarstellungen (vgl. Köhler et al. (2012), Kap. 2.4.4.4)

Die erste Option beinhaltet die Aufzeichnung der Vollmatrix, d.h. Minimum und Maximum des Zyklus werden aufgezeichnet. In der zweiten Variante werden Daten in der Halbmatrix aufgezeichnet, die Information, in welche Richtung der Zyklus durchlaufen wurde, geht verloren. Die dritte Möglichkeit greift wieder auf eine Vollmatrix zurück, die Mittelwerte der Zyklen und deren Schwingbreite werden aufgetragen (hier z.B.: 8 $\rightarrow$ 1: Mittelwert 4,5 und Schwingbreite 7) (vgl. Köhler et al. (2012), Kap. 2.4.3 und Kap. 2.4.4.4). Die Aufzeichnung im Steuergerät des BLKs (Tabelle 3) orientiert sich an der zweiten Variante. Das bedeutet hier konkret, dass nur die halbe Matrix mit Werten befüllt ist. Die beiden Lastwechsel, symmetrisch gegenüberliegend an der Halbierenden, werden addiert, der Teil unter bzw. oberhalb der Halbierenden bleibt leer. Von den 256 möglichen Klassen-Kombinationen sind 120 befüllt ((256-16 (Halbierende))/2)). Sowohl die Start- als auch die Zielklasse (SoC) der Lade- und Entladezyklen wird in % aufgetragen.

Das BLK *Strom der HV-Batterie/Zeit der Pulse* wird über jeweils 6 Klassen in beiden Größen aufgezeichnet und stellt sich somit über 36 kombinierte Klassen dar. Der Strom wird in Ampere (A), die Pulszeit in Sekunden (s) gemessen.

Die Alterungsgröße *Kapazität* wird in Amperestunden (Ah) über die Zeit (t) aufgetragen.

In der deskriptiven Analyse werden die ca. 5.000 Fahrzeuge der untersuchten Flotte gesamtheitlich betrachtet, d.h. sowohl die Heatmaps als auch die Histogramme bilden die durchschnittliche, relative Belastung der Fahrzeug- bzw. Batterieflotte ab. Die Analyse der Kapazitäten der HV-Batterien erfolgt fahrzeugspezifisch, d.h. jede Batterie geht mit ihrer aktuellen Kapazität in die Analyse ein.

Tabelle 2: Belastungskollektiv 1:„Temperatur - SoC HV-Batterie“

Temp. in °C / SoC in %	≤ -10	(-10 ; 0]	...	(40 ; 45]	> 45
[0 ; 15]					
(15 ; 30]					
...					
(80 ; 90]					
> 90					

Tabelle 3: Belastungskollektiv 2:„Lade-Entladezyklen HV-Batterie“

SoC in % / SoC in %	[0 ; 15]	(15 ; 20]	...	(80 ; 85]	(85 ; 100]
[0 ; 15]					
(15 ; 20]					
...					
(80 ; 85]					
(85 ; 100]					

4.1.2 Deskriptive Analyse

In diesem Unterkapitel werden die untersuchten Größen zunächst deskriptiv analysiert. Die zweidimensionalen BLKs werden bzgl. ihrer Häufigkeitsverteilung über eine Heatmap (zur Orientierung über die Farbgebung dient Abbildung 7) und z.T.

Tabelle 4: Belastungskollektiv 3:„Strom der HV-Batterie/Zeit der Pulse“

Strom in A \ Pulszeit in s	[0 ; 1]	(1 ; 5]	...	(15 ; 35]	> 35
≤ -200					
(-200 ; -80]					
...					
(15 ; 80]					
> 80					

eindimensional über Histogramme beschrieben; die eindimensionale Verschleißgröße wird über die Zeit dargestellt. Aus Gründen der Geheimhaltung wird in den Heatmaps auf konkrete, prozentuale Zahlen und im Punktediagramm auf die quantitative Beschreibung der x-Achse verzichtet.

Abbildung 7: Orientierung für die Farbgebung in der Heatmap

BLK 1: „Temperatur/SoC HV-Batterie“

Im BLK 1 wird im Steuergerät der HV-Batterie aufgezeichnet, bei welchen Ladezuständen in Kombination mit Batterietemperaturen die Komponente betrieben wird. Dabei kann die Batterietemperatur sowohl von der Nutzung als auch von äußeren Effekten (z.B. Jahreszeit, Klimazone) beeinflusst werden. Das Kollektiv wird sowohl in Betriebs- als auch in Ruhezeiten aufgezeichnet.

In der Heatmap in Abbildung 8 ist zu sehen, dass die Eckbereiche [0 ; 15] % und >90 % SoC sowie ≤ -10 °C und >45 °C nur sehr schwach bzw. gar nicht belegt sind. Am häufigsten wird die Batterie in den Bereichen (0 ; 15] °C / (30 ; 40] % SoC, (15 ; 25] °C / (30 ; 40] % SoC und (25 ; 35] °C / (30 ; 40] % SoC betrieben. Alle SoC-Stände (0-100 %)werden in diesem Temperaturfenster zumindest mit einem erwähnenswerten Anteil angefahren.

Dieses Bild wird auch durch die Randverteilungen, dargestellt über die Histogramme in Abbildung 9 und Abbildung 10, bestätigt. Durchschnittlich über die betrachtete

Temp. in °C / SoC in %	≤ -10	(-10 ; 0]	(0 ; 15]	(15 ; 25]	(25 ; 35]	(35 ; 40]	(40 ; 45]	>45
[0 ; 15]								
(15 ; 30]								
(30 ; 40]								
(40 ; 60]								
(60 ; 70]								
(70 ; 80]								
(80 ; 90]								
>90								

Abbildung 8: Heatmap BLK 1: „Temperatur/SoC HV-Batterie“

Stichprobe befindet sich die Batterie zu einem sehr großen Anteil (ca. 95 %) in einem Ladezustand zwischen 15 % und 60 %, davon allein fast zu 90 % in der Klasse (30 ; 40] %. Die Betriebstemperatur der Batterie liegt fast ausschließlich (ca. 98 %) zwischen 0 °C und 40 °C.

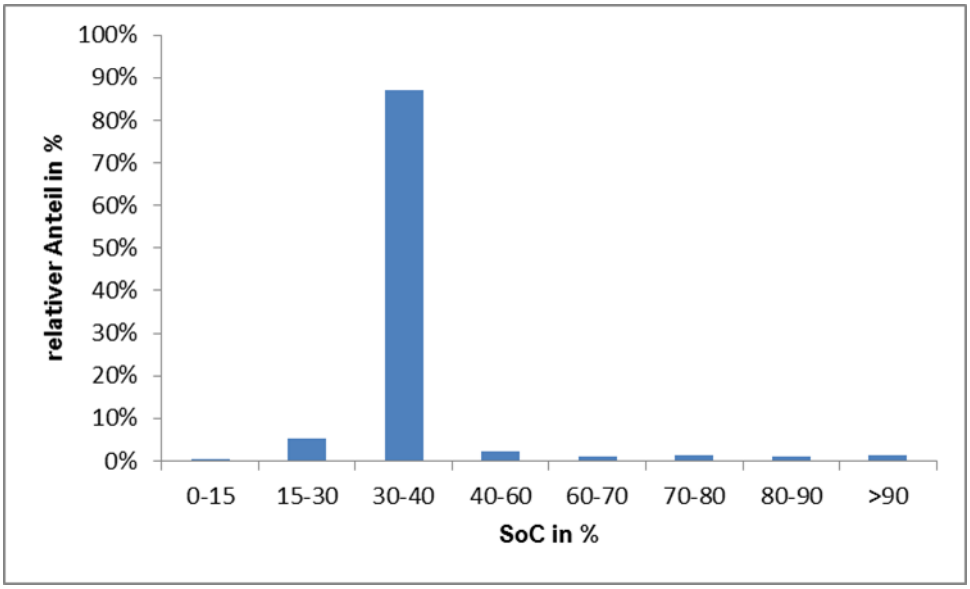

Abbildung 9: Histogramm „SoC“ aus BLK 1: „Temperatur/SoC HV-Batterie“

BLK 2: „Lade-Entladezyklen HV-Batterie“

Die Lade-Entladezyklen der HV-Batterie aus dem BLK 2 stellen die Lastwechsel hinsichtlich der Ladezustände in der Batterie dar. Es ist hier abzulesen, bei welchen Ladezuständen die Lade- und Fahrzyklen gestartet bzw. beendet werden. Wie in 4.1.1 erläutert, wird die Richtung der Lastwechsel vernachlässigt. So ist es z.B. unerheblich, ob ein Lastwechsel von (20 ; 25] SoC zu (85 ; 100] SoC (Ladung) oder von (85 ; 100] SoC zu (20 ; 25] (Fahrt) stattgefunden hat. Das Kollektiv wird nur bei Ladung oder Entladung aufgezeichnet.

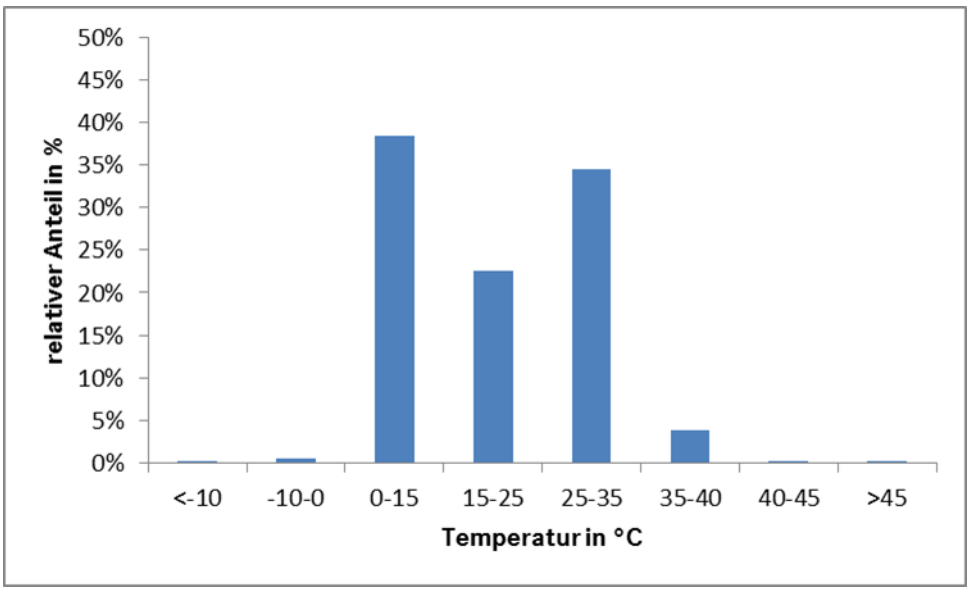

Abbildung 10: Histogramm „Temperatur“ aus BLK 1: „Temperatur/SoC HV-Batterie“

Auffällig in der Heatmap in Abbildung 11 scheint zunächst, dass die Bereiche in der Nähe der beiden Achsen und der Halbierenden stärker besetzt sind als die mittleren Bereiche. So sind z.B. die drei Lastwechsel zwischen [0 ; 15] % SoC und (15 ; 20] % SoC, (15 ; 20] % SoC und (20 ; 25] % SoC sowie (20 ; 25] % SoC und (25 ; 30] % SoC die am häufigsten belegten Klassen. Auch die weiteren Klassen entlang der Halbierenden, die in Relation die geringen Lastwechsel abbilden, sind im Vergleich häufiger besetzt. Lastwechsel im SoC-Bereich von 85-100 % treten kombiniert mit allen anderen SoC-Bereichen auf. So werden dort sowohl geringe (wie oben beschrieben), aber auch sehr hohe Lade-Entladezyklen der HV-Batterie (z.B. zwischen [0 ; 15] % SoC und (85 ; 100] % SoC) beobachtet.

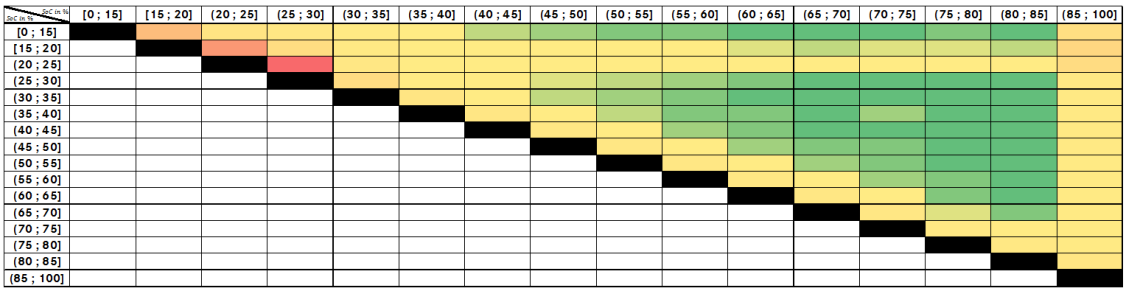

Abbildung 11: Heatmap BLK 2: „Lade-Entladezyklen HV-Batterie“

BLK 3: „Strom der HV-Batterie/Zeit der Pulse“

Das BLK 3 stellt dar, wie lange ein gewisser Strom in der HV-Batterie fließt. Negativer Strom bedeutet hier, dass die Batterie entladen wird, positiver Strom fließt beim Laden der Batterie. Dies kann zum einen an der Ladestation, zum anderen durch Rekuperation während der Fahrt entstehen. Der Bereich von -15 bis 15 Ampere bleibt in dem BLK unbetrachtet, da die minimalen Ströme für die Belastung/Schädigung der Batterie zu vernachlässigen sind. Die Pulslänge beschreibt die Dauer des Stromflusses.

Die Heatmap in Abbildung 12 zeigt, dass am häufigsten kurze Pulse ([0 ; 1] s und (1 ; 5] s) bei Strömen in den Bereichen (-80; -40] A, (-40 ; -15] A und (15 ; 80] A auftreten. Der Eckbereich der längeren Pulse (15 ; 35] s und >35 s in Kombination mit sehr hohen Strömen (sowohl negativ als auch positiv) ist in der relativen Verteilung nahezu gar nicht belegt.

In den Histogrammen der beiden Größen in Abbildung 13 und Abbildung 14 stellt sich das Bild so dar, dass die Eckbereiche des Stromes, ≤ -200 A mit ca. 1 % und >80 A mit ca. 2 % nur sehr mäßig, alle vier weiteren Klassen mit mindestens 10 % relativem Anteil besetzt sind. Die größten Anteile der aufgezeichneten, geflossenen Ströme liegen mit jeweils ca. 32 % zwischen (-40 ; -15] A und (15 ; 80] A.

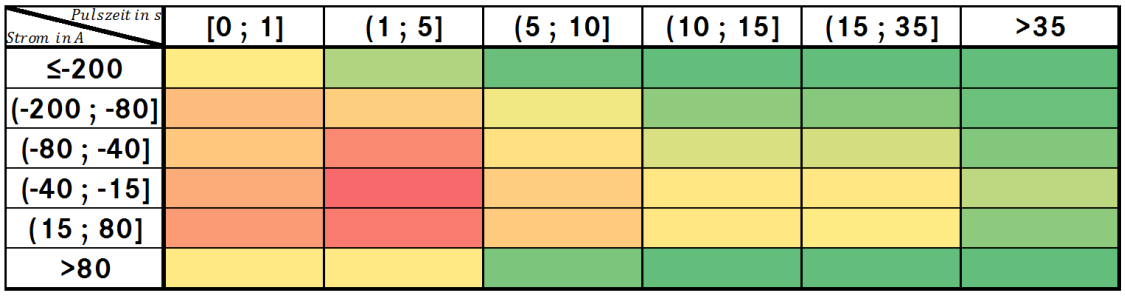

Abbildung 12: Heatmap BLK 3: „Strom der HV-Batterie/Zeit der Pulse“

Verschleißgröße: „Kapazität der HV-Batterie“

Die Kapazität beschreibt, vereinfacht gesagt, den Gesundheitszustand der HV - Batterie. Physikalisch ist es die Menge an Energie, die von der Komponente aufgenommen werden kann. Diese nimmt durch Alterung/Verschleiß über die Zeit (abhängig von der Nutzung/Belastung) ab. Eine Schädigung (Kapazitätsverlust)

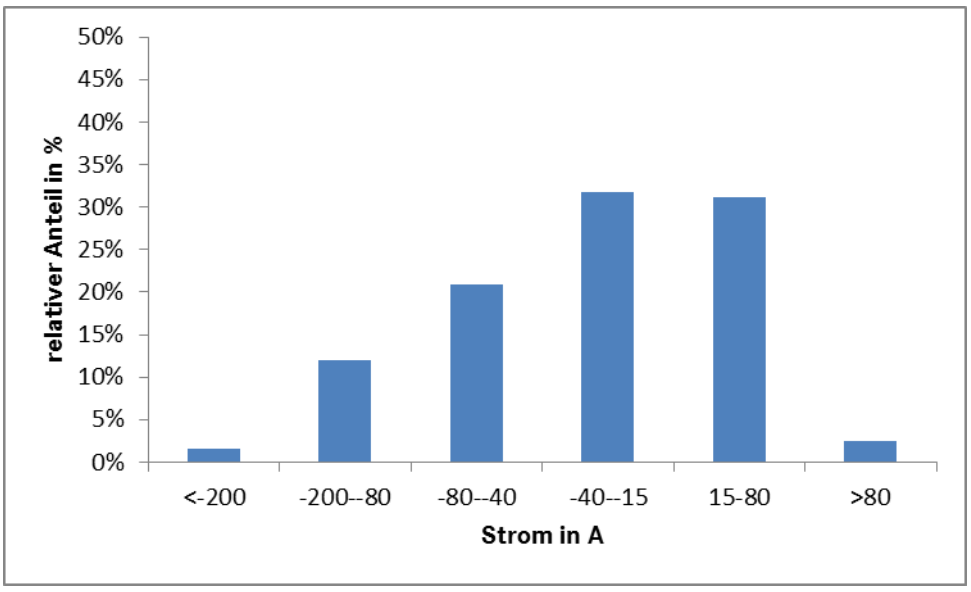

Abbildung 13: Histogramm „Strom der HV-Batterie“ aus BLK 3: „Strom der HV-Batterie/Zeit der Pulse“

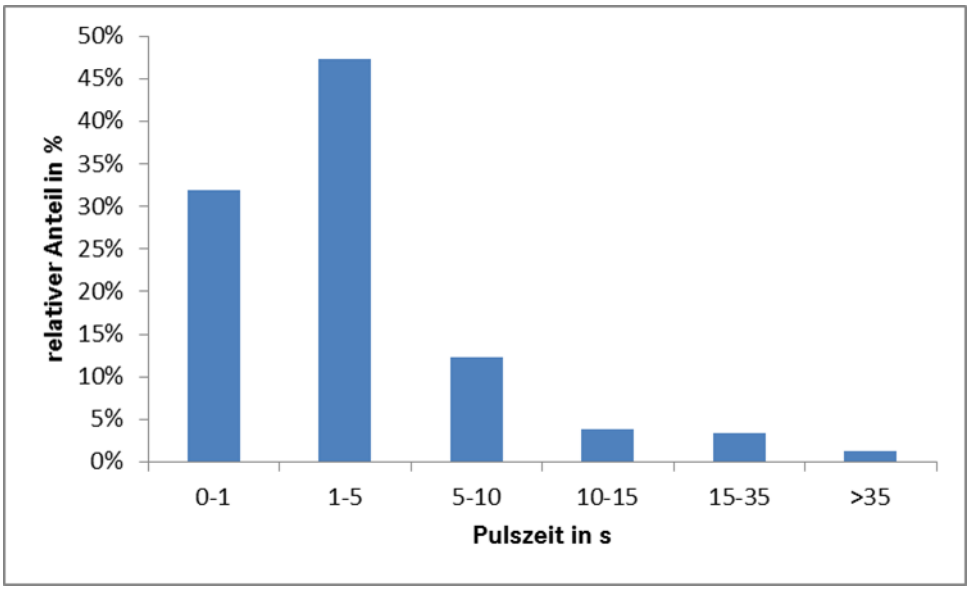

Abbildung 14: Histogramm „Zeit der Pulse“ aus BLK 3: „Strom der HV-Batterie/Zeit der Pulse“

kann seitens der Batterie nicht behoben werden, d.h. ein späteres Ansteigen der Kapazität ist nicht möglich. In diesem Beispiel wird ein HV-Batteriesystem mit einer Nenn-Kapazität von 22 Ah untersucht.

Abbildung 15 zeigt die aktuellen Kapazitäten der in der Stichprobe von ca. 5.000 Fahrzeugen untersuchten HV-Batterien. Auf der x-Achse ist eine Größe aufgetragen, die hoch mit der Zeit korreliert. Jedes Fahrzeug bzw. jede HV-Batterie geht

mit einem Wert und somit einem Punkt in die Grafik ein. Wie in 4.1.1 beschrieben, wird eine gewisse Zeit im Feld für die Fahrzeuge vorausgesetzt; so lässt sich die Lücke nahe der y-Achse erklären. Wie erwartet, ist die Datenlage zu Beginn deutlich gebündelter; je weiter die Zeit voranschreitet, desto weniger Daten von Fahrzeugen liegen vor. So gibt es z.B. offensichtlich mehr Fahrzeuge, die erst seit kürzerer Zeit im Feld sind. Des weiteren war zu erwarten, dass gerade zu Beginn sehr viele Fahrzeuge nahe der 22 Ah Nennkapazität stehen und ein abfallender Trend mit zunehmender Zeit zu beobachten ist.

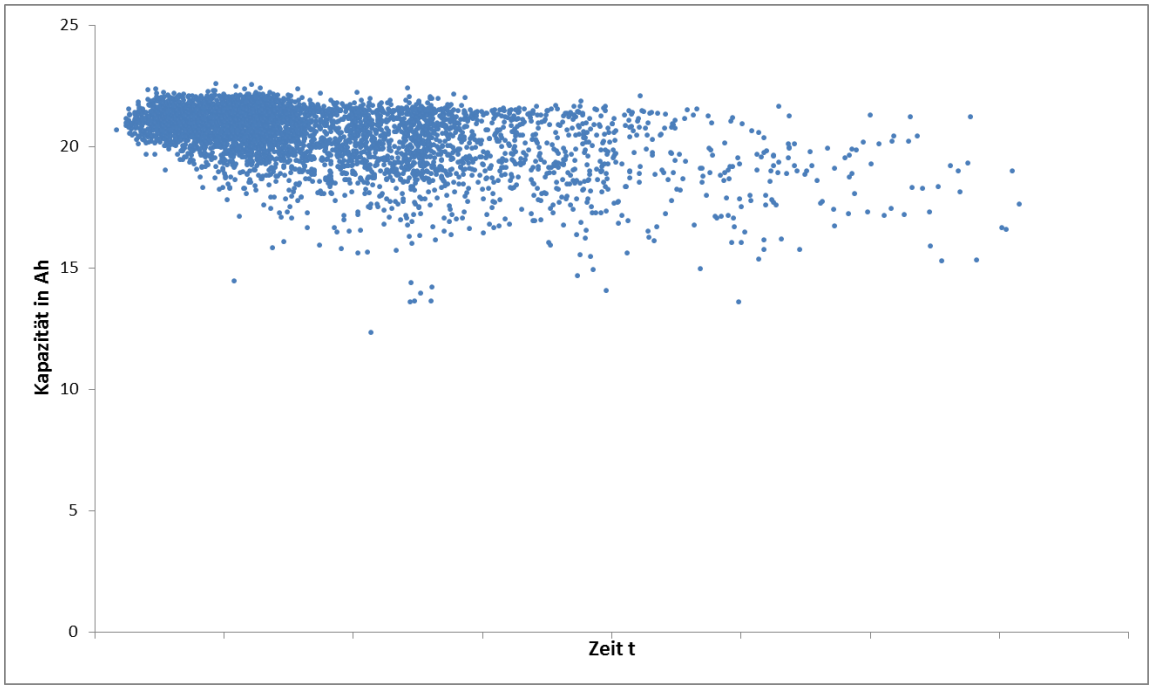

Abbildung 15: Punktdiagramm der Kapazität der HV-Batterien über die Zeit

4.2 Beschreibung der Simulation

4.2.1 Datenabzug und Vorverarbeitung

Wie in 4.1.1 beschrieben, werden hier Daten von ca. 5.000 Fahrzeugen analysiert. Die Daten werden mit der Software IBM SPSS (vgl. IBM Corp. (2015)) aus der internen Daimler-Datenbank abgegriffen und zunächst mit der Open Source Software R (vgl. R Core Team (2013)) vorverarbeitet.

Für den beschriebenen Analyseprozess werden die von der Datenbank abgezogenen Daten zunächst so vorkonditioniert, dass zum einen die Daten vergleichbar, zum anderen die Komplexität hinsichtlich der Rechenbarkeit deutlich reduziert wird. Dies

sind mathematisch triviale Operationen oder Umformungen der ursprünglichen Datenstruktur. Die für die Analyse Relevanten werden in diesem Unterkapitel kurz erläutert.

Die drei zweidimensionalen BLKs werden auf sechs eindimensionale Größen reduziert. BLK *1: Temperatur HV-Batterie/SoC HV-Batterie* und BLK *3: Strom der HV-Batterie/Zeit der Pulse* werden durch die Mittelwerte ihrer Randverteilungen repräsentiert, d.h. jedes Fahrzeug geht jeweils mit einem Mittelwert pro Größe ein. Der Mittelwert wird dabei mithilfe der relativen Anteile in den gegebenen Klassen und den Klassenmittelwerten bestimmt. In den Klassen am Rand ($\leq$ bzw. $\geq$) fließt der Randpunkt ein. Dies wird kurz in einem Beispiel an der Temperatur aus BLK *1: Temperatur HV-Batterie/SoC HV-Batterie* in Tabelle 5 und Gleichung 70 erläutert. Das Beispiel-Fahrzeug würde also mit einer mittleren Temperatur von 25,5°C bewertet. Die Berechnungen der übrigen drei Mittelwerte aus den beiden BLKs 1 und 3 geschieht analog.

Tabelle 5: Beispiel der Mittelwertberechnung aus Randverteilung anhand der Temperatur [°C]

Klasse	$\leq$ -10	(-10 ; 0]	(0 ; 15]	(15 ; 25]	(25 ; 35]	(35 ; 40]	(40 ; 45]	$>$ 45
Mittelwert	-10	-5	7,5	20	30	37,5	42,5	45
Anteil	0,02	0,08	0,12	0,10	0,41	0,12	0,14	0,01

$$\begin{aligned}\bar{t} &= (0,02 \cdot (-10)) + (0,08 \cdot (-5)) + (0,12 \cdot 7,5) + (0,10 \cdot 20) + (0,41 \cdot 30) \\ &\quad + (0,12 \cdot 37,5) + (0,14 \cdot 42,5) + (0,01 \cdot 45) \\ &= 25,5 \end{aligned} \tag{70}$$

Da das BLK *2: Lade-Entladezyklen HV-Batterie* wie in 4.1.2 in Form einer Rainflow-Klassierung mit einer halbbesetzten Matrix vorliegt, ist es nicht möglich, zwei Randverteilungen für den SoC zu bestimmen. Trotzdem ist es möglich, hier zwei interessante Größen zu extrahieren: zum einen den durchschnittlichen SoC während einer Ladung bzw. Entladung, zum anderen den durchschnittlichen SoC-Hub, der während eines Zyklus erreicht wird. Beides sind Faktoren, die bzgl. Alterung von HV-Batterien eine Rolle spielen (vgl. 1.3.2).

Der durchschnittliche SoC während eines Zyklus ergibt sich durch die beiden Klassen-

Mittelwerte der Start- und Zielklasse (vgl. Tabelle 6). In diesem Beispiel errechnet sich der Mittelwert zwischen den Klassen [0;15]-(20;25] folgendermaßen: $0,5 \cdot [0,5 \cdot (0+15) + 0,5 \cdot (20+25)] = 15$. Diese Zelle in der Matrix wird nun also mit 15 % SoC bewertet. Dieser spezifische Wert wird dann jeweils mit dem relativen Anteil in der Klasse gewichtet. Der durchschnittliche Wert pro Fahrzeug ergibt sich dann analog wie in Gleichung 70 beschrieben.

Tabelle 6: Beispiel des durchschnittlichen SoC [%] aus Rainflow Matrix

Klasse	[0;15]-(15;20]	[0;15]-(20;25]	[0;15]-(25;30]	[0;15]-(30;35]	...
durchschn. SoC	12,5	15	17,5	20	...

Der durchschnittlich erreichte SoC-Hub beschreibt die durchschnittliche Differenz zwischen Start- und Ziel SoC. Dazu wird für jede Klasse die durchschnittliche Differenz ebenfalls durch die Klassen-Mittelwerte bestimmt (vgl. Tabelle 7). In diesem Beispiel errechnet sich der mittlere Hub zwischen den beiden Klassen [0;15]-(30;35] folgendermaßen: $|0,5 \cdot (0+15) - 0,5 \cdot (30+35)| = 25$. Diese Zelle in der Matrix wird nun also mit 25 % SoC-Hub bewertet.

Tabelle 7: Beispiel des durchschnittlichen SoC-Hubs [%] aus Rainflow Matrix

Klasse	[0;15]-(15;20]	[0;15]-(20;25]	[0;15]-(25;30]	[0;15]-(30;35]	...
durchschn. Hub	10	15	20	25	...

4.2.2 Normierung der Daten

Um die BLKs hinsichtlich ihrer unterschiedlichen Größenordnungen in der Analyse vergleichbar zu machen, werden sie vor der Simulation normiert, d.h. zentriert und skaliert bzgl. der Varianz. Das Zentrieren geschieht durch Subtraktion des Mittelwertes, das Skalieren dann durch Division des zentrierten Wertes durch die Standardabweichung. Ein normierter Wert x ergibt sich also folgendermaßen:

$$x_{norm} = \frac{x - \bar{x}}{sd(x)} \tag{71}$$

Die hier als Verschleiß- bzw. Zielgröße ausgewählte Kapazität der HV-Batterie muss vor der Analyse ebenso noch normiert werden. Sie liegt in ihrer ursprünglichen

Form zu zwei Zeitpunkten vor: Zu Beginn (Nennkapazität der Batterie seitens des Herstellers) und aktueller Wert bei Datenabgriff (Zeitpunkt analog zum Datenabgriff der BLKs). In dieser Analyse soll der Kapazitätsverlust der HV-Batterien im Feld untersucht werden. Der Kapazitätsverlust wird folgendermaßen hergeleitet:

$$C_{loss} = \frac{\text{Nennkapazität}[Ah] - \text{aktuelle Kapazität}[Ah]}{t} \tag{72}$$

t ist auch hier wieder eine Größe, die mit der Zeit hochkorrelliert.

4.2.3 Durchführung der Simulation des Analyseprozesses in R

Die Durchführung der in Kapitel 3 beschriebenen und Abbildung 5 dargestellten Arbeits- und Simulationsschritte geschieht ebenfalls mithilfe der Open Source Software R.
Die Simulation umfasst 50 Iterationen, also 50 Durchläufe sowohl des Clusterings als auch des gewählten Prognosemodells. Initial wird dem Algorithmus eine Clusteranzahl zwischen 2 und 5 vorgegeben. Das Clustering, das Prognosemodell und die Gütekriterien werden in jeder Iteration berechnet, um die Sensitivität am Ende bewerten zu können. Die Ergebnisse der beispielhaft durchgeführten Analyse werden im folgenden Kapitel beschrieben.

4.3 Auswahl des Prozesspfades und Analyse des Use Cases

An dieser Stelle gilt es, sich anhand des in dieser Arbeit entwickelten statistischen Absicherungsprozesses (dargestellt in Abbildung 5) für einen Pfad im Prozess-Modell zu entscheiden. Dieser wird entsprechend der Struktur der vorliegenden Daten und des Fokus bzw. des Ziels der Analyse gewählt.
Der Fokus in diesem Beispiel soll auf der Analyse der unterschiedlichen Nutzung und Belastung der Komponente liegen. Methodisch betrachtet ist also das Clustering der Belastungskollektive und dessen Güte im Fokus. Für das hier gewählte distanzbasierte Clusteringverfahren (k-means) im Algorithmus des Evidence Accumulation Clusterings (Schritt 1 und 2) gibt es eine Vielzahl von Gütekriterien. Im Analyse-Beispiel wird der Silhouettenkoeffizient (Schritt 3) inkl. seiner Lage- und

Streuungsmaße (Schritt 4) berechnet und interpretiert.
Entsprechend des auf dem Clustering liegenden Fokus wird als Prognosemodell ein hierarchisch lineares Modell (Schritt 5) berechnet. Wie in 3.2.4 beschrieben, eignen sich diese besser, mit hierarchischen Strukturen in den Daten umzugehen (hier gegeben durch das Clustering) als andere Modelle.
Als Gütekriterium für das Regressionsmodell wird klassisch das Bestimmtheitsmaß (hier auf die hierarchisch linearen Modelle angepasst wie in 3.2.5 beschrieben) genutzt (Schritt 6). Im letzten Schritt, der Berechnung und Beschreibung der Sensitivität der Güte des Prognosemodells werden analog zu Schritt 4 anhand der empirischen Verteilung Lage- und Streuungsmaße für das Bestimmtheitsmaß des hierarchisch linearen Modells bestimmt.
Im Rahmen des Evidence Accumulation Clustering findet hier ein Durchlauf über mehrere Iterationen statt. In jeder Iteration werden sowohl das Clustering (hier k-means) als auch die Prognosemodelle (hier das hierarchisch lineare Modell) inklusive der jeweiligen Gütemaße berechnet. Der zu wählende Pfad richtet sich zum einen nach den vorliegenden Daten und zum anderen nach der zu untersuchenden Fragestellung. So wäre es z.B. denkbar, mithilfe modellbasierter Verfahren (z.B. Mixed Effects Modelle) (Schritt 1 und 2) Längsschnittdaten zu clustern und deren Güte zu bewerten (Silhouettenkoeffizient in Schritt 3). Für die Prognose wären dann Zeitreihenverfahren (Schritt 4) die geeigneten Methoden, deren Güte und dessen Sensitivität dann in Schritt 4 und 5 bewertet werden.
Der hier gewählte Analysepfad ist in Abbildung 5 durch die grüne Markierung hervorgehoben. Wie beschrieben, erhält man bei vollständiger Ausführung des Prozesses mehrere (Zwischen-)Ergebnisse:

1. Ergebnis des Clusterings
2. Güte des Clusterings
3. Sensitivität der Güte des Clusterings
4. Ergebnis des Prognosemodells
5. Güte des Prognosemodells
6. Sensitivität der Güte des Prognosemodells

Diese Ergebnisse werden in diesem Unterkapitel anhand des Analyse-Beispiels in der aufgeführten Reihenfolge beschrieben.

4.3.1 Ergebnisse des Clusterings

Nach der Durchführung des Clusterings mithilfe des Evidence Accumulation Clustering Algorithmus werden im Folgenden die Ergebnisse beschrieben und die größten Unterschiede zwischen den Clustern herausgearbeitet. Es wird auf alle drei vorab ausgewählten BLKs der Nutzung im einzelnen eingegangen. Beschrieben wird die durch den Algorithmus als optimal bestimmte Partition von drei Clustern. In jeder Iteration werden jedoch die Gütemaße der jeweiligen Partition „gespeichert“, sodass sich die Sensitivität dessen im weiteren Verlauf quantitativ bewerten lässt. So gehen also auch die Resultate der anderen Clusteranzahlen (vgl. 4.2.3) nicht verloren. Im weiteren Verlauf der Arbeit wird im Kontext der Prognosemodelle darauf nochmals eingegangen.
Cluster 1 zeichnet sich im BLK 1: „Temperatur/SoC HV-Batterie“ dadurch aus, dass SoC-Werte über alle Klassen zwischen 0-15% und 90% erreicht werden. Zu erkennen ist, dass die Klassen 15-30%, 0-15% und 40-60% die mit den höchsten relativen Anteilen sind. Cluster 2 und Cluster 3 hingegen haben eine fast identische, abweichende Verteilung des SoC. So ist fast ausschließlich (relativer Anteil zwischen 80% und 90%) die SoC-Klasse 30-40% belegt. Lediglich die Bereiche 15-30% und 40-60% weisen noch erwähnenswerte relative Anteile auf (s. Abbildung 16).
Die beiden bzgl. der SoC-Verteilung in BLK 1 fast identischen Cluster 2 und 3 unterscheiden sich hinsichtlich der Batterietemperatur deutlich. Cluster 3 zeichnet sich als einzige Gruppe durch im Vergleich höhere Batterietemperaturen aus. So sind fast 100% relative Anteile im Bereich >25°C. Davon liegen über 70% im Temperaturbereich 25-35°C und ca. 25% zwischen 35°C und 40°C. Cluster 1 und Cluster 2 zeigen hier sehr ähnliche Verteilungen. Hauptsächlich die Temperaturbereiche 0-15°C, 25-35°C und 15-25°C (in der Reihenfolge) sind maßgeblich für die relativen Anteile der Nutzung. Lediglich die Klasse 35-40°C weist für Cluster 2 noch einen Anteil im Bereich von ca. 5% auf (s. Abbildung 17).

Aus BLK 2: „Lade-Entladezyklen HV-Batterie“ wurden, wie in 4.2.1 beschrieben,

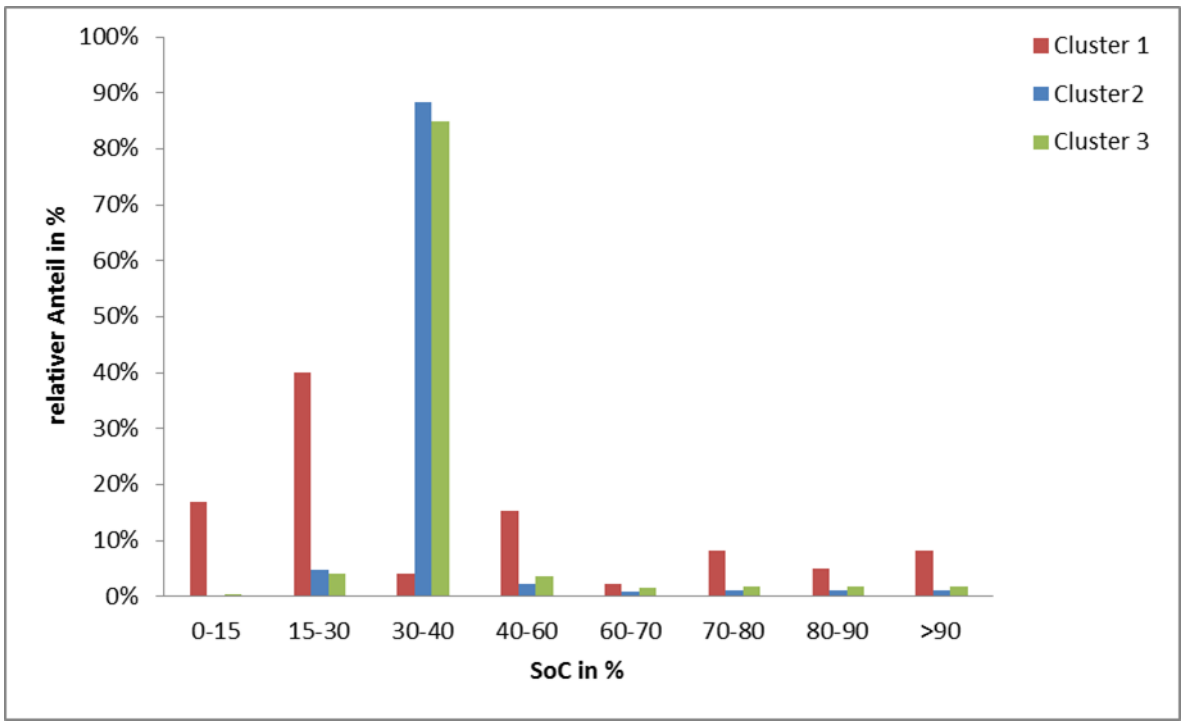

Abbildung 16: Histogramm „SoC“ aus BLK 1: „Temperatur/SoC HV-Batterie“

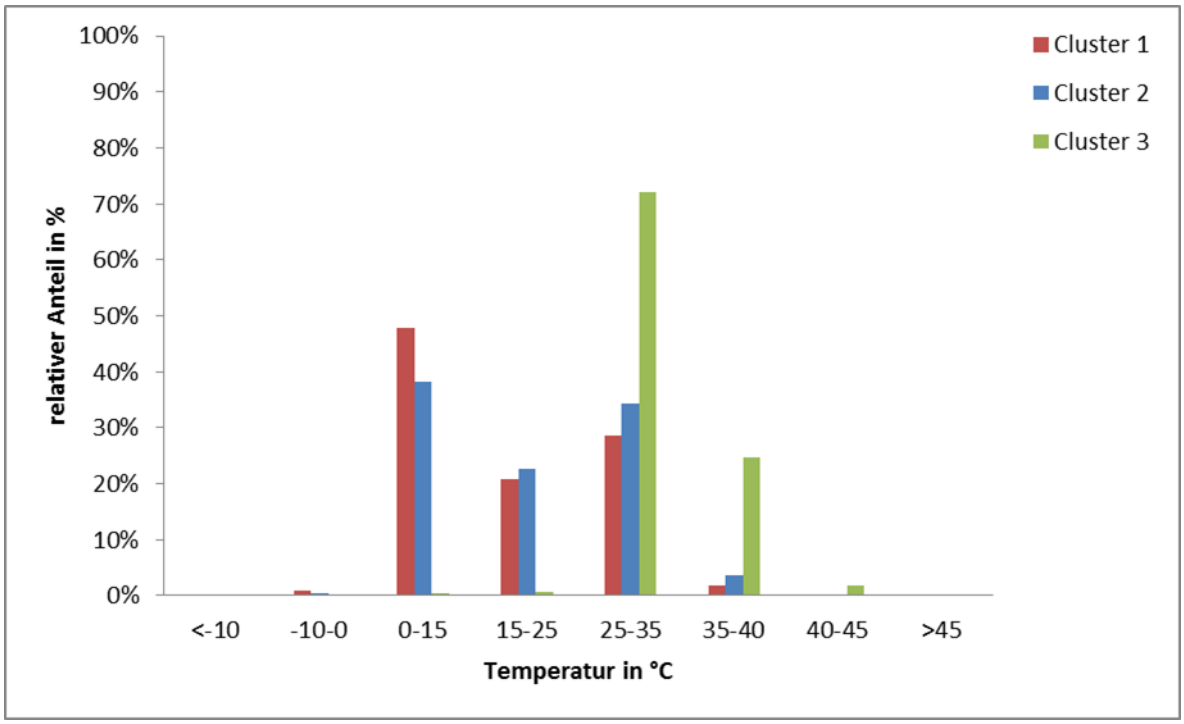

Abbildung 17: Histogramm „Temperatur“ aus BLK 1: „Temperatur/SoC HV-Batterie“

die zwei Größen des durchschnittlichen SoCs während einer Ladung bzw. Entladung und zum anderen des durchschnittlichen SoC-Hubs, der während eines Zyklus erreicht wird, extrahiert und für das Clustering verwendet. Die Ergebnisse der Mit-

telwerte in den 3 Clustern sind Tabelle 8 zu entnehmen.

Nachdem in BLK 1 bereits die SoC-Bereiche während der Fahrt analysiert wurden, beschreibt das BLK 2 das Lade- und Entladeverhalten. Cluster 1 und Cluster 2 zeichnen sich sowohl im mittleren SoC (31% bzw. 33%) als auch im mittleren SoC-Hub (14% bzw. 17%) durch ähnliche Werte aus. An dieser Stelle hebt sich Cluster 3 sowohl durch einen im Vergleich hohen mittleren SoC (40%) als auch durch einen deutlich höheren mittleren SoC-Hub (24%) ab.

Tabelle 8: Durchschnittliche SoC und durchschnittliche SoC-Hübe für die drei Cluster aus BLK 2: „Lade-Entladezyklen HV-Batterie“

	mittlerer SoC [%]	**mittlerer SoC-Hub** [%]
Cluster 1	31	14
Cluster 2	33	17
Cluster 3	40	24

Im Vergleich zu BLK 1 und BLK 2 weist das BLK 3: „Strom der HV-Batterie/Zeit der Pulse“ die geringsten Unterschiede zwischen den drei Clustern auf. Die Verteilung der Länge der Pulszeit des Stroms zeigt über die Cluster hinweg ihre größten relativen Anteile in den Wertebereichen 1-5s, 0-1s und 5-10s (in dieser Reihenfolge). Cluster 3 zeigt am ehesten noch etwas höhere Nutzung durch längere Pulse. So sind dort die Klassen 10-15s und 15-35s durch immerhin ca. 5% relative Anteile belegt. Die drei Histogramme der Cluster bzgl. der Länge der Pulse finden sich in Abbildung 18.

Auch die Verteilung der Belastung der Batterie durch Strom zeigt zwischen den 3 Clustern keine signifikanten Unterschiede auf. Wie in 4.1.2 beschrieben, zeichnet sich hier eine Entladung durch negative und eine Ladung der Batterie durch positive Ströme aus. Die minimalen Ströme zwischen -15 und 15 Ampere werden nicht betrachtet/aufgezeichnet. So sind die Wertebereiche -40 bis -15A, 15 bis 80A und -80 bis -40A die am meisten belegten (s. Abbildung 19).

Da in diesem BLK die Unterschiede zwischen den Clustern relativ gering sind, unterscheiden sich die Verteilungen der beiden betrachteten Größen nicht signifikant von der in Abbildung 13 und Abbildung 14 dargestellten gesamtheitlichen Betrachtung der Flotte.

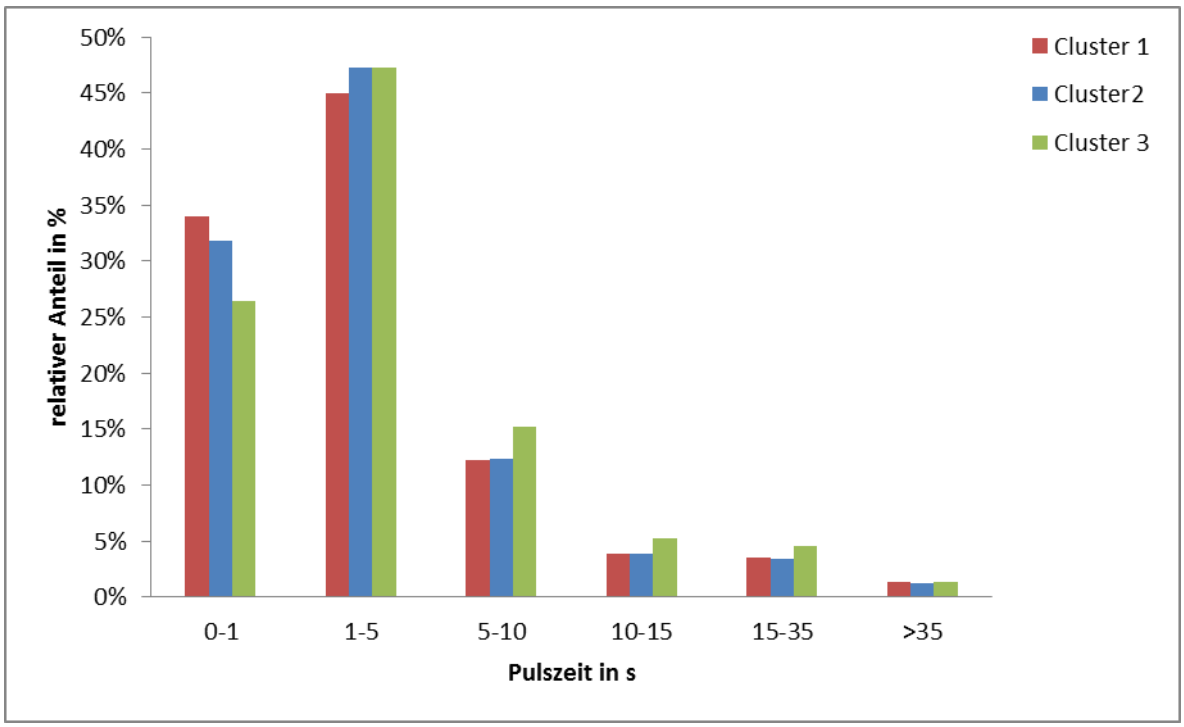

Abbildung 18: Histogramm „Zeit der Pulse“ aus BLK 3: „Strom der HV-Batterie/Zeit der Pulse“

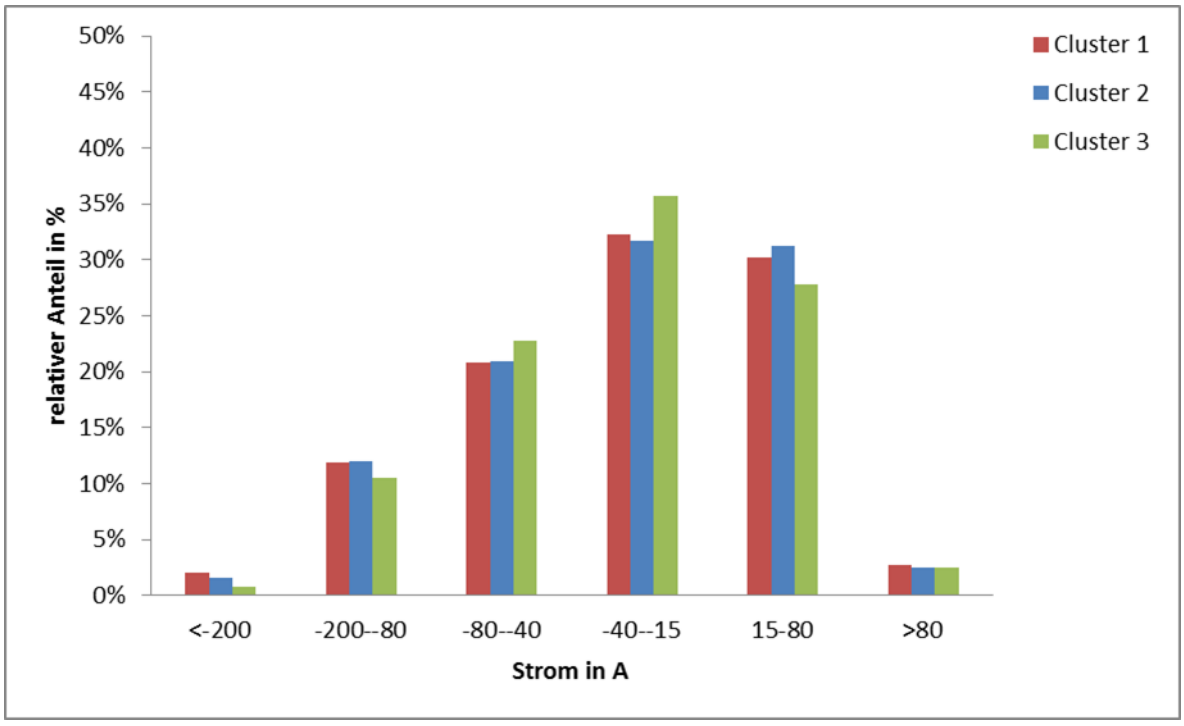

Abbildung 19: Histogramm „Strom der HV-Batterie“ aus BLK 3: „Strom der HV-Batterie/Zeit der Pulse“

Trotz der Tatsache, dass die Kapazität der Batterien in das Clustering noch nicht eingeflossen ist, sondern erst später als Zielgröße der Prognose genutzt wird, können

in der Punktwolke der aktuellen Kapazitäten der HV-Batterien (Stichprobe von ca. 5.000 Fahrzeugen) die 3 Cluster farblich markiert werden. In Abbildung 20 sind in blau die Werte des mit Abstand größten Clusters 2, die Werte von Cluster 1 in rot und die Cluster 3 zugehörigen Kapazitätswerte in grün eingefärbt. Im Vergleich mit dem großen Cluster 2 zeichnet sich Cluster 3 tendenziell eher durch kürzere Nutzungszeit und durchschnittliche, verbleibende Kapazitäten aus, während es bei Cluster 1 viele Punkte im oberen Kapazitätsbereich über die Nutzungszeit gibt.

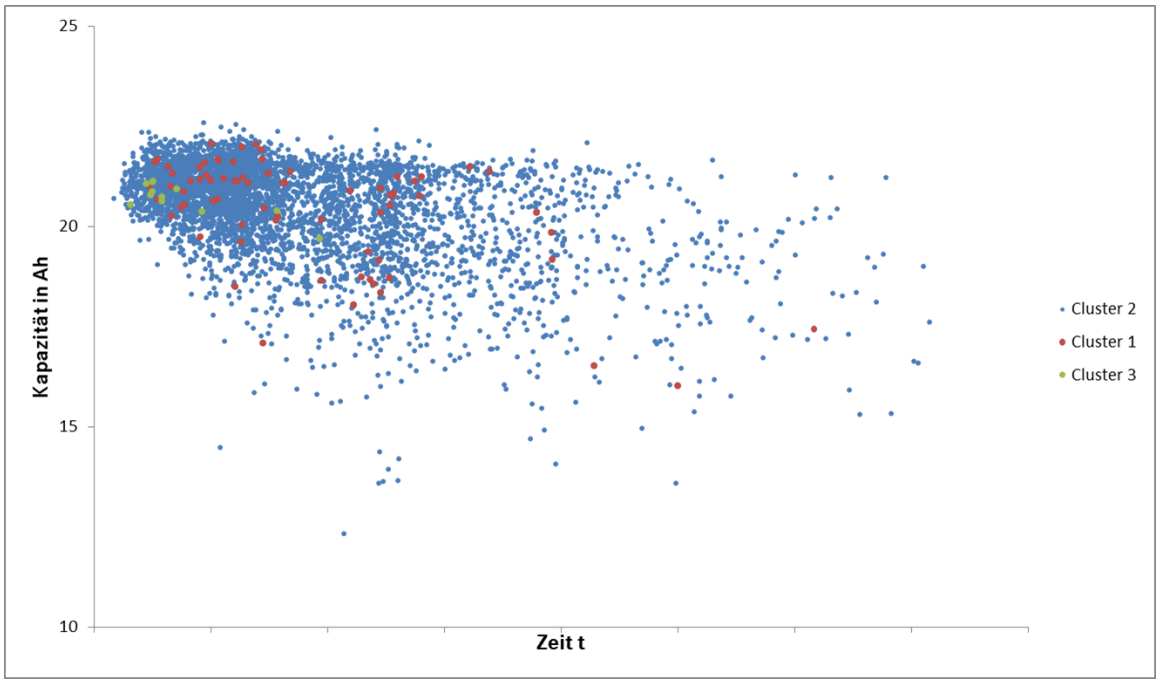

Abbildung 20: Darstellung der drei Cluster in Punktediagramm der Kapazität der HV-Batterien über die Zeit (y-Achse verkürzt)

4.3.2 Sensitivität der Güte des Clusterings

Wie in Kapitel 3 beschrieben und im Methodenbaukasten (Abbildung 5) dargestellt, wird an dieser Stelle die empirische Verteilung des Gütemaße über die Iterationen betrachtet. In diesem Beispiel geschieht dies anhand des in 3.1.3 beschriebenen Silhouettenkoeffizienten. Alle 50 Iterationen mit den unterschiedlichen Clusteranzahlen gehen hier ein. Dadurch kann die Sensitivität der Güte des Clusterings bewertet werden (s. 3.1.4).
Bei dem Durchlauf der 50 Iterationen des Evidence Accumulation Clusterings betrug der Silhouettenkoeffizient mindestens 0,27, im Maximum 0,38 und im Mittel 0,35 (vgl. Tabelle 9). Nach Kaufman, Rousseeuw (1990), Kap. 2.2 lässt sich also hier von einer „schwachen Struktur“ sprechen. Die geringe Standardabweichung von 0,04 spricht für ein relativ konstantes Maß der Silhouette über die Iterationen.
Wie in 4.1.1 beschrieben, wurden die drei BLKs *Temperatur HV-Batterie/SoC HV-Batterie*, *Lade-Entladezyklen HV-Batterie* und *Strom der HV-Batterie/Zeit der Pulse* bereits im Vorfeld ausgewählt. Eine Variablenselektion durch statistische Methoden könnte hier die Güte des Clusterings möglicherweise verbessern. Ziel in diesem Anwendungsbeispiel war es jedoch, die Individuen (hier: Fahrzeuge/Batterien) bestmöglich nach den vorgegebenen Einflussgrößen zu clustern.

Tabelle 9: Lage- und Streuungsmaße der Verteilung des Silhouettenkoeffizienten

Min.	$x_{0,05}$	$x_{0,5}$	$x_{0,95}$	Max.	SD	Var.
0,27	0,27	0,35	0,38	0,38	0,04	0,00

4.3.3 Ergebnisse der Prognose

Wie in 3.2.4 beschrieben, eignen sich hierarchisch lineare Modelle, um Daten zu analysieren, die eine hierarchische Struktur aufweisen. Dies ist durch das vorher beschriebene und durchgeführte Clustering gegeben.
In das Prognosemodell gehen die in 4.1.2 beschriebenen BLKs als erklärende Variablen und die Kapazität als Zielgröße ein. Die in 4.3.1 erhaltenen Cluster bilden die Gruppen im hierarchisch linearen Modell. Hier wird mithilfe des R-Paketes **lme4** und der Funktion **lmer** (vgl. R Core Team (2013)) in jeder Iteration beispielhaft ein Modell mit Intercept und Steigung berechnet. Ein iterativer Prozess zur Bestimmung

des „besten“ Modells (vgl. 3.2.4) gehört nicht zur Zielsetzung dieser Arbeit, ist aber bei weiterführenden Analysen zu empfehlen (s. 5.2). Der Fokus dieser Arbeit liegt auf dem grundsätzlichen, methodischen Prozess und der Quantifizierung der Güte und deren Sensitivität über die verschiedenen Methoden. Das gewählte Prognosemodell soll hier nur einen Eindruck vermitteln, wie das Ergebnis des Gesamt-Prozesses aussehen kann und wie Ergebnisse dann interpretiert werden können. Eine ausführliche Beschreibung des Fittings eines hierarchisch linearen Modells mithilfe des R-Paketes lme4 findet sich z.B. in Bates et al. (2015).

Wie in 4.2.1 und 4.2.2 beschrieben, gehen als erklärende Variablen die normierten Mittelwerte der BLKs und als Zielgröße der normierte Kapazitätsverlust in die Berechnung des Modells ein. Im folgenden sei nun „meanXBlk1“ bzw. „meanYBlk1“ der SoC bzw. die Temperatur aus dem BLK 1:„Temperatur/SoC HV-Batterie“, „meanXBlk“ bzw. „meanYBlk2“ der berechnete, durchschnittliche SoC bzw. SoC-Hub aus BLK 2:„Lade- und Entladezyklen HV-Batterie“, „meanXBlk3“ bzw. „meanYBlk3“ die Temperatur bzw. der Strom aus BLK 3: „Strom der HV-Batterie/Zeit der Pulse“ und „capa“ der Kapazitätsverlust. „clust“ ist hier die Gruppenvariable (Cluster), „df“ ist der Ausgangsdatensatz („data frame“) (vgl. 4.1.1 und 4.2.1). So wird hier beispielhaft ein Modell mit allen erklärenden Variablen sowie zufälliger Steigung und zufälligem Intercept in Form des Modells in Abbildung 21 definiert. Das Ergebnis des Modells mit den quantifizierten Parametern findet sich in Abbildung 22.

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: capa ~ meanXBlk1 + meanYBlk1 + meanXBlk2 + meanYBlk2 + meanXBlk3 +
    meanYBlk3 + (meanXBlk1 + meanYBlk1 + meanXBlk2 + meanYBlk2 +      meanXBlk3 + meanYBlk3 | clust)
   Data: df
```

Abbildung 21: Screenshot: Modelldefinition des ausgewählten hierarchisch linearen Modells

```
Fixed Effects:
(Intercept)    meanXBlk1    meanYBlk1    meanXBlk2    meanYBlk2    meanXBlk3    meanYBlk3
 -3.103e-04   -1.456e-04    3.111e-04    3.225e-04   -2.703e-04    2.529e-04    4.892e-05
```

Abbildung 22: Screenshot: Ergebnis des ausgewählten hierarchisch linearen Modells

Das Modell in Abbildung 22 beschreibt das mittlere Modell der verschiedenen Cluster. Daraus können dann die unterschiedlichen Steigungen und Intercepts extrahiert und die gruppenspezifischen Verschleißmodelle abgeleitet werden, s. dazu z.B. Nieu-

wenhuis (2018). Da sich die drei Modelle für die verschiedenen Cluster in diesem Analysebeispiel nur marginal unterscheiden, wird an dieser Stelle darauf verzichtet.

4.3.4 Sensitivität der Güte der Prognose

Analog zu 4.3.2 werden hier die in 3.2.5 beschriebenen Gütemaße der hierarchisch linearen Modelle bzgl. ihrer Sensitivität über die Iterationen des Gesamt-Analyseprozesses beleuchtet. Die Lage- und Streuungsmaße der in 3.2.5 definierten marginalen und konditionalen R^2 für das berechnete Modell mit zufälligem Intercept und zufälliger Steigung findet sich in Tabelle 10 und Tabelle 11. Hierbei werden alle Clusterresultate (dem Algorithmus wurde eine minimale Anzahl von 2 und maximale Anzahl von 5 vorgegeben) berücksichtigt (s. 4.2.3).

Tabelle 10: Lage- und Streuungsmaße der Verteilung des marginalen R^2 des Modells

Min.	$x_{0,05}$	$x_{0,5}$	$x_{0,95}$	Max.	SD	Var.
0,01	0,19	0,56	0,71	0,98	0,16	0,03

Tabelle 11: Lage- und Streuungsmaße der Verteilung des konditionalen R^2 des Modells

Min.	$x_{0,05}$	$x_{0,5}$	$x_{0,95}$	Max.	SD	Var.
0,01	0,30	0,67	0,84	0,86	0,18	0,03

Sowohl das marginale als auch das konditionale R^2 weisen ein ähnliches Verhalten bzgl. ihrer Streuung auf. Über die 50 Iterationen innerhalb des durchgeführten Analysebeispiels wird das Modell in jeder Iteration berechnet und jedesmal die beiden Gütemaße abgeleitet. So sind sowohl sehr niedrige Werte (das Minimum liegt jeweils bei 0,01) als auch ziemlich hohe Werte (Maximum bei 0,86 bzw. 0,98) für die Modelle zu beobachten. Die Standardabweichung ist für beide Maße mit 0,16 bzw. 0,18 relativ hoch. Hier liegt also noch kein konstantes Modell mit konstanter Güte über die Iterationen vor.

Der Vergleich der Cluster-Resultate zeigt, dass jeweils die Minima sowohl des marginalen als auch des konditionalen R^2 durch Partitionen auf drei Cluster zustande kommen (bei der Auswahl der Anzahl der Cluster über das Evidence Accumulation Clustering erzielten drei Cluster die besten Ergebnisse). Die durchschnittlichen Werte der Ergebnise mit drei Clustern liegen mit 0,51 (marginal) bzw. 0,63 (konditional)

den Gesamtdurchschnitten von 0,56 bzw. 0,67 immer noch nahe. Die Minimal-Werte stellen also klare Ausreißer nach unten dar.

Die für beide Maße besten Resultate erzielen die Ergebnisse mit vier Clustern. Die beiden Maxima von 0,98 (marginal) und 0,86 (konditional) finden sich dort wieder. Festgemacht an den höchsten Mittelwerten zeigen sich die Ergebnisse mit vier (marginales R^2 im Schnitt bei 0,57) bzw. fünf (konditionales R^2 im Schnitt bei 0,70) am stärksten. Die konstantesten Gütemaße (niedrigste Standardabweichungen) werden durch die Ergebnisse mit fünf Clustern erzeugt.

Der Fokus im Anwendungsbeispiel lag auf der Kombination von Clusteringverfahren mit einem Prognosemodell. Es wird also kein kontinuierlicher Prozess durchgeführt, der das bestmögliche Modell findet. Durch die vorher festgelegte Wahl auf ein hierarchisch lineares Modell überrascht es nicht, dass sich die Gütemaße noch nicht im optimalen Wertbereich bewegen. So wäre es z.B. möglich, in weiteren Analysen spezielle Alterungsmodelle für HV-Batterien zu verwenden.

5 Zusammenfassung und Ausblick

Dieser Teil der Arbeit fasst die Entwicklung des Methodenbaukastens zur Quantifizierung der statistischen Güte von Last- und Verschleißanalyse zusammen. Da die Arbeit im industriellen Umfeld der Automobilindustrie geschrieben wurde, teilt sich dieses Kapitel in zwei Blickwinkel. Zum einen werden die wichtigsten Erkenntnisse aus methodischer Sicht, zum anderen aus Sicht des Managements dargestellt. Im zweiten Teil des Kapitels erfolgt zudem eine kritische Beurteilung der Ergebnisse und ein Ausblick auf weiterführende Ansätze.

5.1 Zusammenfassung

5.1.1 Methodische Zusammenfassung

Der methodische Fokus dieser Arbeit lag auf der Quantifizierung und Fortpflanzung der Unsicherheit von Clusteringverfahren auf die weitere Analyse mithilfe von verschiedenen Prognosemodellen. Dafür wurden alle nötigen statistischen Methoden und dessen Gütekriterien definiert. Dies betrifft in dieser Arbeit in erster Linie Clusteringverfahren, Lineare Modelle und Nichtlineare Modelle. Für alle genannten Verfahren wurden die jeweils gängigsten Gütekriterien eingeführt.
Es wurde ein neuer statistischer Methodenbaukasten entwickelt (vgl. Abbildung 5), der es erlaubt, verschiedene statistische Methoden zum einen zu verknüpfen und zum anderen die Fortpflanzung der Güte bzw. der Unsicherheit über mehrere Analysestufen hinweg zu quantifizieren. Dabei lassen sich sowohl mehrere Methoden (im Rahmen des Clusterings und der Prognose) als auch mehrere Gütekriterien kombinieren. Der entwickelte, iterative Prozess, integriert in den Algorithmus des Evidence Accumulation Clusterings (vgl. 3.1.2), bietet dem Anwender entscheidende methodische Vorteile. Zum einen kann in jedem Schritt die Güte und dessen Sensitivität des jeweiligen Verfahrens bewertet werden, zum anderen wird über die gleichzeitige Durchführung aller Verfahren in jeder Iteration beides über die Analysestufen hinweg quantifiziert.
Das Anwendungsbeispiel zeigt Potentiale auf, sowohl die Einflussgrößen (im Clustering und als erklärende Variablen im Prognosemodell), als auch das zu wählende Prognosemodell bei weiteren Anwendungen genauer zu untersuchen. Die Selektion

der Variablen und die prozessuale Auswahl des geeigneten Modells können die Güte der Modelle steigern und die Sensitivität verringern.

5.1.2 Management Zusammenfassung

Das Ziel dieser Arbeit im Umfeld der Automobilindustrie war es, einen Methodenbaukasten zu entwicklen, um das Belastungs- und Verschleißverhalten von Fahrzeugen prozessual analysieren und vorhersagen zu können. Im Fokus stand dabei die Betrachtung der Unsicherheit der verschiedenen Ergebnisse.
Durch die prozessuale Beschreibung ist es nun einfacher möglich, Fahrzeuge nach ihrem Verschleißverhalten zu clustern, eine vorher festgelegte Verschleißgröße für verschiedene Cluster (Fahrzeuggruppen) über die Zeit zu prognostizieren und die Sensitivität der Güte der jeweiligen Ergebnisse zu bewerten. Mithilfe der Güte und dessen Sensitivität über die verschiedenen Methoden (Clustering und Prognosemodelle) hinweg hat das Management nun Möglichkeiten, die Qualität und in erster Linie die Stabilität der durchgeführten Analysen zu bewerten, um ggf. Handlungsmaßnahmen ableiten zu können. Der im Methodenbaukasten beschriebene iterative Prozess bietet bereits zu frühen Zeitpunkten (geringe Datenbasis) verbesserte Aussagemöglichkeiten.
Der entwickelte statistische Methodenbaukasten wurde im Rahmen der Arbeit an einem Datenbeispiel im Kontext alternativer Antriebe erprobt (vgl. Kapitel 2). Die in diesem Kontext vorliegenden Daten wurden beschrieben (vgl. 4.1.1) und auf die physikalischen Zusammenhänge zwischen den Einfluss- und der Verschleißgröße wurde hingewiesen (vgl. 1.3). Die Aussagekraft der Ergebnisse ist für das gezeigte Anwendungsbeispiel noch nicht optimal, da sowohl die Daten als auch die Modelle vorab definiert wurden. Dies ist der Tatsache geschuldet, dass es lediglich als Demonstration des Prozesses dient. Für jede neue Anwendung sollte es das Ziel sein, die größtmögliche Aussagekraft (z.B. durch den Vergleich von verschiedenen Modellen) zu erzeugen.

5.2 Ausblick

Die Arbeit schließt ab mit dem Ausblick aus methodischer Sicht. Mit Fokus auf dem Clustering und der Prognose werden Ansatzpunkte bzgl. Verbesserung und

Erweiterung des entwickelten statistischen Methodenbaukastens aufgeführt.

5.2.1 Clustering

Das entwickelte Prozessbild beginnt damit, dass sowohl die Nutzungsgrößen als auch die entsprechende Verschleißgröße im Vorfeld definiert werden. An dieser Stelle wäre es in Zukunft möglich, die Variablenselektion durch statistische Methoden durchzuführen (z.B. Korrelationsanalyse, Hauptkomponentenanalyse usw.). Dies ist vor allem dann sinnvoll, wenn eine Analyse durchzuführen ist, bei der die Einflussgrößen auf eine bestimmte Zielgröße im Vorfeld nicht eindeutig bestimmt sind. Das Clustering wäre dann erst Schritt 2 im Analyseprozess.
Bzgl. Evidence Accumulation Clustering ist es sinnvoll, die Weiterentwicklung des Algorithmus zu verfolgen. So wurden bereits am Ende von 3.1.2 interessante Erweiterungen erwähnt. Gerade für größere Datensätze als die in dem Datenbeispiel dieser Arbeit analysierten Stichprobe können die Arbeiten von Lourenco et al. (2010) und Silva (2016) von Interesse sein. So wäre es dann ggf. auch möglich, zwei- oder mehrdimensionale Nutzungsdaten zu analysieren. Desweiteren kann die Methode optimiert werden, indem man innerhalb des Evidence Accumulation Clusterings mehrere Clusteralgorithmen testet bzw. zulässt, statt einen fix vorab zu definieren. Bzgl. der Gütekriterien des Clusterings gibt es in der Literatur deutlich mehr als die in 3.1.3 definierten und vorgestellten Maße.
Bei Vorliegen von Längsschnittdaten müsste der Methodenbaukasten hinsichtlich der vorgestellten Verfahren und Methoden grundsätzlich angepasst werden. Eine methodische Untersuchung von modellbasiertem Clustering mit anschließender Prognose einer Verschleißgröße im automobilen Umfeld findet sich in Kreibich (2014).
Der vorgestellte Analyseprozess ist grundsätzlich offen und anwendbar für andere Cluster-Verfahren, Algorithmen und Maße.

5.2.2 Prognosemodelle

In dieser Arbeit liegt der Fokus auf dem Umgang mit Daten, in denen vorab Gruppen identifiziert wurden. Im Rahmen der Prognosemodelle des Verschleißes wurden detailliert lineare Modelle beschrieben, die gut mit hierarchischen Daten (im Anwendebeispiel: hierarchisch lineares Modell) umgehen können.

Da der Verschleiß von Bauteilen im Allgemeinen über die Zeit selten lineare Gestalt annimmt, ist es sinnvoll, das Prognosemodell für jede Anwendung des Prozesses individuell zu wählen. Im Kontext von HV-Batterien seien z.B. verschiedene Alterungskurven zu berücksichtigen. Diese finden sich in der Literatur zur Genüge. Eine State-of-the-Art Analyse von Alterungsfunktionen von Lithium-Ionen-Zellen findet sich z.B. in Sedelmaier (2015). Ein methodischer Ansatz, um Ausfallraten von HV-Batteriesystemen anhand von ppm-Raten, einer Alterungsfunktion der Batteriezelle, der Zellverschaltung und einer Annahme der Belastung abzuschätzen, findet sich in Özgen (2016) und Lehmann et al. (2017).

Liegen Alterungsgrößen zu mehreren Zeitpunkten vor, könnten auch Zeitreihenverfahren hilfreich sein. Der Prozess, Längsschnittdaten aus dem automobilen Umfeld sowohl zu clustern als auch durch Zeitreihenverfahren zu prognostizieren, wird in Kreibich (2014) ausführlich beschrieben, eine weitere Ausführung und Anwendung findet sich in Lehmann und Keller (2014).

Wie in 4.3.3 beschrieben, gibt es in dem hier durchgeführten Analysebeispiel keinen iterativen Prozess zur Wahl des besten Modells im Rahmen der hierarchisch linearen Modelle. Es wird beispielhaft ein Modell mit zufälliger Steigung und zufälligem Intercept beschrieben. Hier wird es als sehr sinnvoll angesehen, das Modell iterativ aufzubauen. So können bei Anwendung verschiedene Einflüsse und deren Wechselwirkungen hinzugefügt oder auch wieder entfernt werden. Es können ferner lineare Abhängigkeiten zwischen den Einflussgrößen minimiert und die Modellqualität verbessert werden. Kriterien wie AIC oder BIC können dann dafür genutzt werden.

Desweiteren kann es interessant sein, die Qualitätskriterien der Prognosemodelle bereits in die Entscheidung für das „endgültige“ Clustering einfließen zu lassen. In dieser Arbeit liegt die Entscheidung für das finale Clustering in der Güte des Clusterings (hier: Silhouettenkoeffizient); dieser Fokus kann ebenso auf die Güte des finalen Prognosemodells gesetzt werden.

Der Analyseprozess ist auch hier offen und anwendbar für andere Prognosemodelle und Bewertungskriterien. Auch weitere Analyseschritte sind im Methodenbaukasten denkbar. In Dobry et al. (2018) z.B. wird vorgeschlagen, für die Fahrzeuge im zweiten Schritt (nach dem Evidence-Accumulation-Clustering) durch Klassifikationsverfahren eine Beanstandung bzw. Nicht-Beanstandung zu prognostizieren.

Die vorliegende Arbeit bietet einen Methodenbaukasten an, der für verschiedene

Analysen weiter modifiziert und gemäß der jeweiligen Anforderungen ausgebaut werden kann.

Literaturverzeichnis

Andre, D. (2014): *Systematic Characterization of Ageing Factors for High-Energy Lithium-Ion Cells and Approaches for Lifetime Modelling Regarding an Optimized Operating Strategy in Automotive Applications*. Rheinisch-Westfälische Technische Hochschule Aachen.

Backhaus, K., Erichson, B., Plinke, W. und Weiber, R. (2016): *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung* (14. Auflage). Berlin, Heidelberg: Springer.

Backhaus, K., Erichson, B. und Weiber, R. (2015): *Fortgeschrittene Multivariate Analysemethoden. Eine anwendungsorientierte Einführung* (3. Auflage). Berlin, Heidelberg: Springer.

Baier, D. und Gaul, W. (2013): *Clusteranalyse*, online verfügbar unter: `http://marketing.wiwi.uni-karlsruhe.de/institut/viror/kaiman/kaiman/cluster/index.xml.html`. (Zugriff am 18. Februar 2016)

Baltes-Götz, B. (2013): *Analyse von hierarchischen linearen Modellen mit der SPSS-Prozedur MIXED*. Universität Trier.

Bates, D., Mächler, M., Bolker, B.M. und Walker, S.C. (2015): *Fitting Linear Mixed-Effects Models Using lme4*, erschienen in: Journal of Statistical Software, Volume 67, 1-48.

Bauersachs, B. (2015): *Analyse realer Kundenutzerdaten einer elektrisch angetriebenen Pkw-Flotte*. Universität Stuttgart.

Bayes, T. (1763): *An Essay towards Solving a Problem in the Doctrine of Chances*, erschienen in: Phil. Trans., Vol. 53, 370-418.

Belt, J.R., Duong, T.Q., Ho, C.D., Miller, T.J. und Motloch, C.G. (2003): *A capacity and power fade study of Li-ion cells during life cycle testing*, erschienen in: Journal of Power Sources, Volume 123, 241-246.

Bergmeir, P., Nitsche, C., Nonnast, J. und Bargende, M. (2015): *Methoden des Data Mining zur Visualisierung unterschiedlicher Belastungsmuster einer Hybridfahrzeugflotte auf Basis von Lastkollektivdaten*, Technical Report, Tag des kooperativen Promotionskollegs Hybrid 2015, Universität Stuttgart.

Bergs, S. (1981): *Optimalität bei Clusteranalysen.* Westfälische-Wilhelm-Universität Münster.

Bortz, J. (1993): *Statistik. Für Sozialwissenschaftler.*(4. Auflage). Berlin, Heidelberg, New York: Springer.

Brill, M. (2012): *Entwicklung und Implementierung einer neuen Onboard-Diagnosemethode für Lithium-Ionen-Fahrzeugbatterien.* Universität Ulm.

Broussely, M. ,Biensan, P., Blanchard, P., Bonhomme, F., Herreyre, S., Nechev, K. und Staniewicz, R.J. (2005): *Main aging mechanisms in Li ion batteries*, erschienen in: Journal of Power Sources, Volume 146, 90-96.

Busch, M. (2005): *Analyse dichtebasierter Clusteralgorithmen am Beispiel von DBSCAN und MajorClust.* Universität Paderborn.

Calinski, R.B. and Harabasz, J. (1974): *A Dendrite Method for Cluster Analysis*, erschienen in: Comm. in Statistics, Vol. 3, 1-27.

Cohen, J., Cohen, P., West, S.G. und Aiken, L.S. (2003): *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3. Auflage). Mahwah, N.J.: Lawrence Earlbaum Associates.

Davies, D.L. und Bouldin, D.W. (1979): *A Cluster Separation Measure*, erschienen in: IEEE Transactions on pattern Analysis and Machine Intelligence, Vol. 1, No. 2, April 1979, 224-227.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977): *Maximum Likelihood from Incomplete Data via EM Algorithm*, erschienen in: J R Stat Soc B, Vol. 39, 1-38.

Ditton, H. (1998): *Mehrebenenanalyse: Grundlagen und Anwendungen des Hierarchisch Linearen Modells.* Weinheim: Juventa.

Dobry, P., Lehmann, T. und Bertsche, B. (2018): *Two-Step Data Mining Method To Identify Failure Related Driving Patterns*, erschienen in: 2018 Annual Reliability and Maintainability Symposium (RAMS).

Dunn, J.C. (1973): *A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters*, erschienen in: J. Cybernetics,

Vol. 3, 32-57.

Ecker, M., Dechent, P., Gerschler, J.B., Hust, F., Käbitz, S., Sauer, D.U. und Vogel, J. (2012): *Development of a lifetime prediction model for lithium-ion batteries based on extended accelerated aging test data* erschienen in: Journal of Power Sources, Volume 215, 248-257.

Ecker, M. und Sauer, D.U. (2013): *Batterietechnik. Lithium-Ionen-Batterien*, erschienen in: MTZ. Motortechnische Zeitschrift, Issue 01/2013.

Ester, M., Kriegel, H.P., Sander, J. und Xu, X. (1996): *A Density-Based Algorithm for Discovering Clusters*, erschienen in: Kdd, Vol. 96, No. 34, 226-231.

Ester, M. und Sander, J. (2000): *Knowledge Discovery in Databases. Techniken und Anwendungen*. Berlin, Heidelberg: Springer.

Fahrmeir, L., Kneib, T. und Lang, S. (2007): *Regression-Modelle, Methoden und Anwendungen*. Berlin, Heidelberg: Springer.

FAZ (2013): *Merkel hält an Absatzziel fest*, online verfügbar unter: `http://www.faz.net/aktuell/wirtschaft/wirtschaftspolitik/eine-million-elektroautos-bis-2020-merkel-haelt-an-absatzziel-\fest-12196498.html`. (Zugriff am 01. November 2016)

Fisher, R.A (1928): *The general sampling distribution of the multiple correlation coefficient*, erschienen in: Proceedings of the Royal Society of London, Series A, 654-673.

Fred, A. (2001): *Finding Consistent Clusters in Data Partitions*, erschienen in: Lecture Notes in Computer Science 2096, 309-318.

Fred, A. und Jain, A. (2002): *Evidence Accumulation Clustering Based on the K-Means Algorithm*, erschienen in: Lecture Notes in Computer Science, 442-451.

Fred, A. und Jain, A. (2005): *Combining Multiple Clusterings Using Evidence Accumulation*, erschienen in: IEEE Transactions on pattern Analysis and Machine Intelligence, Vol. 27, No.6, 835-850.

Freialdenhoven, A. (2009): *Stärkung der Wettbewerbsfähigkeit der Automobilindustrie durch Vernetzung von Wissenschaft und Industrie.*

Aachen: Forschungsgesellschaft Kraftfahrwesen.

Freis, E. (2013): *Integrativer Ansatz zur Identifizierung neuer, prognostisch relevanter Metagene mittels Clusteranalyse*. Technische Universität Dortmund.

Gies, S. (2008): *Unkonventionelle Fahrzeugantriebe*, Schriftenreihe Automobiltechnik, Vorlesungsumdruck (Version 4.0), Institut für Kraftfahrzeuge, Rheinisch-Westfälische Technische Hochschule Aachen.

Hartung, J. und Elpelt, B. (2007): *Multivariate Statistik. Lehr- und Handbuch der angewandten Statistik* (7. Auflage). München: Oldenbourg.

Hastie, T., Tibshirani, R. und Friedman, J. (2001): *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. New York: Springer.

Herb, F. (2010): *Alterungsmechanismen in Lithium-Ionen-Batterien und PEM-Brennstoffzellen und deren Einfluss auf die Eigenschaften von daraus bestehenden Hybrid-Systemen*. Universität Ulm.

Hofmann, D.A. (1997): *An Overview of the Logic and Rationale of Hierarchical Linear Models*, erschienen in: Journal of Management, No. 23, 723-744.

Hosoya, G., Koch, T. und Eid, M. (2014): *Längsschnittdaten und Mehrebenenanalyse*, erschienen in: Kölner Zeitschrift für Soziologie und Sozialpsychologie, No. 66, 189-218.

IBM Corp. (2015). IBM SPSS Statistics for Windows. *IBM Corp* (Version 24.0), Armonk, NY.
`http://https://www.ibm.com/de-de/products/spss-statistics`.

Jäckle, S. (2017): *Neue Trends in den Sozialwissenschaften. Innovative Techniken für qualitative und quantitative Forschung*. Wiesbaden: Springer.

Kaufman, L. und Rousseeuw (1990): *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Khatri, C.G. (1966): *A note on a large sample distribution of a transformed multiple correlation coefficient*, erschienen in: Annals of the Institute of Statistical Mathematics, 375-380.

Köhler, M., Jenne, S., Pötter, K. und Zenner, H. (2012): *Zählverfahren und*

Lastannahme in der Betriebsfestigkeit. Heidelberg, Dordrecht, London, New York: Springer.

Köttermann, T., Jacobi, A., Jordan, C. und Bracke, S. (2015): *Anwendung multivariater Methoden auf automobile Daten zur lastbasierten Zuverlässigkeitsanalyse*, erschienen in: VDI-Fachtagung Technische Zuverlässigkeit.

Kreft, I. und de Leeuw, J. (1998): *Introducing Multilevel Modeling*. London: Sage.

Kreibich, C. (2014): *Clustering und Prognose von Längsschnittdaten im Kontext alternativer Antriebskonzepte in der Automobilindustrie. Entwicklung, Validierung und Implementation eines Methodikkonzepts*, Universität Bamberg.

Kriegel, H.P. (2002): *Clustering*, online verfügbar unter: `http://www.dbs.informatik.uni-muenchen.de/Lehre/Hauptseminar/SS02/KDD02/Clustering-Peer.pdf`. (Zugriff am 01. März 2016)

Lamm, A., Friebe, P., Kaufmann, R., Mohrdieck, C., Soczka-Guth, T., Spier, B., Stuis, H., Warthmann, W. (20009): *Lithium-Ionen-Batterie. Erster Serieneinsatz im S400 Hybrid*, erschienen in: ATZ Automobiltech Z, Vol. 111, Issue 7, 490-499.

Langer, W. (2009): *Mehrebenenanalyse. Eine Einführung für Forschung und Praxis.* (2. Auflage). Wiesbaden, VS Verlag für Sozialwissenschaften.

Lee, Y.S. (1971): *Some results on the sampling distribution of the multiple correlation coefficients*, erschienen in: Journal of the Royal Statistical Society, Series B, 117-130.

Lee, Y.S. (1972): *Tables of upper percentage points of the multiple correlation coefficients*, erschienen in: Biometrika, 175-189.

Lehmann, T. (2013): *Methodenentwicklung zur Fahr- und Lastmustererkennung validiert auf Basis SMART ELECTRIC DRIVE*, Technische Universität Dortmund.

Lehmann, T. und Keller, J. (2014): *A Two-Step Method by Clustering and Forecasting a State Variable for Real Data Application*, erschienen in:

European Network for Business and Industrial Statistics (ENBIS).

Lehmann, T., Özgen, O., Keller, J. (2017): *Methodological framework for estimation of failure rates for various high-voltage battery systems*, erschienen in: Risk, Reliability and Safety: Innovating Theory and Practice, 372-374.

Lienkamp, M . (2012): *Elektromobilität. Hype oder Revolution?* Berlin, Heidelberg, Springer.

Lourenco, A., Bulo, S.R., Fred, A. und Pelillo, M. (2013): *Consensus Clustering with Robust Evidence Accumulation*, erschienen in: EMMCVPR 2013 Proceedings of the 9th International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition, Volume 8081, 307-320.

Lourenco, A., Fred, A. und Jain, A. (2010): *On the scalability of Evidence Accumulation Clustering*, erschienen in: Pattern Recognition (ICPR) 2010 20th International Conference on IEEE, 782-785.

MacQueen, J. (1967): *Some methods for Classification and analysis of multivariate observations.*, erschienen in: Lecam, L. M., Neyman, J. (eds.): Proc. 5th Berkely Symp. Math. Stat. Prob. 1965/66, **1**, 281-297.

Maulik, U. und Bandyopadhyay, S. (2002): *Performance Evaluation of Some Clustering Algorithm and Validity Indices*, erschienen in: IEEE Transactions on pattern Analysis and Machine Intelligence, Vol. 24, No.12, 1650-1654.

McLachlan, G. und Peel, D. (2000): *Finite Mixture Models.* New York: Wiley.

Meyna, A. und Pauli, B. (2010): *Zuverlässigkeitstechnik. Quantitative Bewertungsverfahren* (2. Auflage). München, Wien: Hanser.

Müller, R.M. und Lenz, H.-J. (2013): *Business Intelligence.* Berlin, Heidelberg: Springer.

Nakagawa, S., Johnson, P.C.D. und Schielzeth, H. (2017): *The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded* erschienen in: J.R. Soc. Interface, Volume 14: 20170213.

Nakagawa, S. und Schielzeth, H. (2013): *A general and simple method for obtaining*

R^2 *from generalized linear mixed-effects models* erschienen in: Methods in Ecology and Evolution 2013, Volume 4, 133-142.

Nieuwenhuis, R. (2018): *R-Session 16: Multilevel Model Specification (lme4)*, online verfügbar unter: `http://www.rensenieuwenhuis.nl/r-sessions-16-multilevel-model-specification-lme4/`. (Zugriff am 12. Dezember 2018)

Özgen, O. (2016): *Statistische Ausfallratenprognose von HV-Batterien: Abschätzung für unterschiedliche Batteriekonzepte.* Technische Universität Dortmund.

Olkin, I. and Finn, J.D. (1995): *Correlations Redux*, erschienen in: Psychological Bulletin, Vol. 118, Issue 1, 155-164.

Pal, N.R. und Bezdek, J.C. (1995): *On Cluster Validity for the Fuzzy c-Means Model*, erschienen in: IEEE Transactions on Fuzzy Systems, Vol. 3, No.3, 370-379.

Peterson, S.B., Apt, J. und Whitacre, J.F. (2010): *Lithium-ion battery cell degradation resulting from realistic vehicle and vehicle-to-grid utilization* erschienen in: Journal of Power Sources, Volume 195, 2385-2392.

Ploehn, P. (2014): *Kategorisierung von Indizes zur Clustervalidierung*, Technische Universität Darmstadt.

Proff, He., Proff, Ha. und Fojcik, T.M. (2014): *Management des Übergangs in die Elektromobilität. Radikales Umdenken bei tiefgreifenden technologischen Veränderungen.* Wiesbaden: SpringerGabler.

Raudenbush, S.A. und Bryk, A.S. (2002): *Hierarchical Linear Models* (2. Auflage). Thousands Oaks: Sage.

R Core Team (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. `http://www.R-project.org`.

Reif, K., Borgeest, K. und Noreikat, K.E. (2012): *Kraftfahrzeug- Hybridantriebe. Grundlagen, Komponenten, Systeme, Anwendungen.* Wiesbaden: Springer Vieweg.

Rousseeuw, P.J. (1987): *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, erschienen in: Journal of Computational and Applied Mathematics, Vol. 20, 53-65.

Rusch, T. (2011): *Hierarchisch lineare Modelle. Lineare Mixed-Effects Models für hierarchische Datenstrukturen bei metrischen Variablen*, online verfügbar unter: `http://statmath.wu.ac.at/people/trusch/lmm/HLMRusch.pdf`. (Zugriff am 18. Februar 2019)

Sarre, G., Blanchard, P. und Broussely, M. (2004): *Aging of lithium-ion batteries* erschienen in: Journal of Power Sources, Volume 127, 65-71.

Sauer, D.U. und Kowal, J. (2012): *7. Batterietechnik Grundlagen und Übersicht*, erschienen in: MTZ - Motortechnische Zeitschrift, Ausgabe 12/2012. München: Springer Automotive Media.

Schlittgen, R. und Streitberg, B.H.J. (2001): *Zeitreihenanalyse* (9. Auflage). München: Oldenbourg.

Sedelmaier, M. J. (2015). *State-of-the-Art Analyse zu Alterungsfunktionen zur Modellierung des elektrischen Verhaltens von Lithium-Ionen-Zellen.* Technische Universität München.

Silva, D. (2016): *Efficient Evidence Accumulation Clustering for large datasets/big data*, erschienen in: ICPRAM 2016 Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods, 367-374.

Snijders, T.A.B. und Bosker, R.J. (1994): *Modeled Variance in Two-Level Models*, erschienen in: Sociological Methods & Research, Vol. 22, Issue 3, 342-363.

Song, J., Zhu, Z. und Price, C. (2016): *A New Evidence Accumulation Method with Hierarchical Clustering*, erschienen in: IEEE International Conference on Cloud Computing and Big Data Analysis, 122-126.

Soper, D.S. (2016): *R-Square Confidence Interval Calculator [Software]*, online verfügbar unter: `http://www.danielsoper.com/statcalc`. (Zugriff am 24. März 2016)

Stahl, D. und Sallis, H. (2012): *Model-based cluster analysis*, erschienen in: WIREs Comput Stat 2012, Vol. 4, 341-358.

Stan, C. (2015): *Alternative Antriebe für Automobile* (2. Auflage). Berlin, Heidelberg: Springer.

Springer Fachmedien Wiesbaden (2012): *Antriebskonzepte für Die Elektromobilität*, erschienen in: ATZ - Automobiltechnische Zeitschrift, Ausgabe 10/2012.

Steiger, J.H. (2017): R^2, online verfügbar unter: `http://www.statpower.net/Software.html#R2`. (Zugriff am 13. Juli 2017)

Tan, Li Jr. (2012): *Confidence Intervals for Comparison of the Squared Multiple Correlation Coefficients of Non-nested Models*. Electronic Thesis and Dissertation Repository, The University of Western Ontario.

Terzimehic, T. (2012): *Application of the Support Vector Machine to Health Diagnosis and Prognostics of Automotive Lithium-Ion Batteries*. University of Sarajevo.

Vendramin, L., Campello, R.J.G.B. und Hruschka, E.R. (2010): *Relative clustering validity criteria: A comparative overview*, erschienen in: Statistical Analysis and Data Mining, 3(4), 209-235.

Vetter, J., Besenhard, J.O., Hammouche, A., Möller, K.-C., Novák, P., Wagner, M.R., Winter, M., Wohlfahrt-Mehrens, M., Veit, C. und Vogler, C. (2005): *Ageing mechanisms in lithium-ion batteries* erschienen in: Journal of Power Sources, Volume 147, 269-281.

Vogel, F. (2005): *Beschreibende und schließende Statistik* (13. Auflage). München, Wien: Oldenbourg.

Wallentowitz, H., Freialdenhoven, A. und Olschewski, I. (2010): *Strategien zur Elektrifizierung des Antriebstranges* (1. Auflage). Wiesbaden: Vieweg+Teubner.

Walter, S.G. und Rack, O. (2009): *Eine anwendungsbezogene Einführung in die Hierarchische Lineare Modellierung (HLM)*, erschienen in: Albers, S. et al. (2009) *Methodik der empirischen Forschung*. Wiesbaden: Gabler.

Ward, J.H. Jr. (1963): *Hierarchical Grouping to Optimize an Objective Function*, erschienen in: Journal of the American Statistical Association, Volume 58, Issue 301, 236-244.

Wishart, J. (1931): *The Mean and Second Moment Coefficient of the Multiple Correlation Coefficient in Samples from a Normal Population*, erschienen in: Biometrika, 22, 353-361.

Wunder, J. (2014): *Analyse des Verhaltens verschiedener Clusterverfahren nach Imputation fehlender Werte*. Ludwig-Maximilian-Universität München.

Xie, X.L. und Beni, G. (1991): *A Validity Measure for Fuzzy Clustering*, erschienen in: IEEE Transactions on pattern Analysis and Machine Intelligence, Vol. 13, No.8, 841-847.

Zeit Online (2009): *Elektroauto. Zurück in die Zukunft*, online verfügbar unter: `http://www.zeit.de/2009/38/A-Elektroauto`. (Zugriff am 03. November 2018)

Zeit Online (2016): *4.000 Euro Prämie für Kauf eines Elektroautos*, online verfügbar unter: `http://www.zeit.de/politik/deutschland/2016-04/bundesregierung-elektroautos-subvention-kaufpraemie`. (Zugriff am 01. November 2016)