Entwicklung eines Verfahrens zur Vorgabe der benötigten Sintereinstellungen von PTC-Bauteilen aus Daten der Granulatfreigabe und den Ergebnissen vorheriger Lose

Masterarbeit
von
Katharina Mertens, BSc

eingereicht am
Lehrstuhl Wirtschafts- und Betriebswissenschaften
der
Montanuniversität Leoben

Leoben, Mai 2017

Aufgabenstellung

Nur bei Master- und Bachelorarbeiten!

Die Aufgabenstellung ist ein einseitiges, bei Masterarbeiten vom Lehrstuhlleiter, bei Bachelorarbeiten vom Betreuer unterschriebenes Dokument. Die Erstellung erfolgt vom Betreuer und wird dem Verfasser noch vor Abschluss der wissenschaftlichen Arbeit ausgehändigt. Die Aufgabenstellung muss direkt nach dem Titelblatt bei jedem Exemplar der Arbeit in Original (Institutsexemplar) bzw. in Kopie mit eingebunden werden. (Siehe Richtlinie zur Erstellung wissenschaftlicher Arbeiten)

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich diese Arbeit selbständig verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und mich auch sonst keiner unerlaubten Hilfsmittel bedient habe.

I declare in lieu of oath, that I wrote this thesis and performed the associated research myself, using only literature cited in this volume.

Leoben, Mai 2017

(Katharina Mertens, BSc)

Gleichheitsgrundsatz

Aus Gründen der Lesbarkeit wurde in dieser Arbeit darauf verzichtet, geschlechtsspezifische Formulierungen zu verwenden. Es wird ausdrücklich festgehalten, dass die bei Personen verwendeten maskulinen Formen für beide Geschlechter zu verstehen sind.

Danksagung

An dieser Stelle möchte ich all jenen meinen Dank aussprechen, die durch ihre fachliche und persönliche Unterstützung zum Gelingen dieser Masterarbeit beigetragen haben.

Danken möchte ich o.Univ.-Prof. Dipl.-Ing. Dr. mont. Hubert Biedermann für die Möglichkeit, meine Masterarbeit am Lehrstuhl für Wirtschafts- und Betriebswissenschaften zu verfassen, und das Bereitstellen der Infrastruktur.

Ebenfalls möchte ich mich bei seinem Mitarbeiter Dipl.-Ing. Robert Bernerstätter bedanken. Vielen Dank für die Zeit, die moralische Unterstützung, die konstruktive Kritik und die Anregungen, die diese Arbeit bereichert haben.

Auch muss ich mich bei der EPCOS OHG und meinem Betreuer Dr. Johann Schmidt bedanken, die mit ihrer Problemstellung den Anstoß zu dieser Arbeit gegeben haben. Vielen Dank für die Unterstützung.

Zusätzlich möchte ich mich bei O.Univ.-Prof. Dipl.-Ing. Dr.techn Paul O'Leary und seinen Mitarbeitern vom Lehrstuhl Automation für ihr offenes Ohr, Anregungen und die Unterstützung beim Erlernen von Python und LaTeX bedanken.

Während meiner Zeit an dieser Universität habe ich viele unterschiedliche Menschen kennen und schätzen gelernt. Einige wurden Wegbegleiter, denen ich an dieser Stelle danken möchte.

Mein besonderer Dank gilt meinen Eltern und vor allem meiner Mutter. Ohne sie hätte ich dieses Studium nach dem Tod meines Vaters nicht weiterführen können. Ich danke ihr für ihre emotionale und finanzielle Unterstützung.

Ein großes Danke an meine Familie und Freunde, die immer für mich da waren und mich in meinem Weg bestärkt haben.

Kurzfassung

Diese Masterarbeit behandelt Möglichkeiten der produktionswirtschaftlichen Verbesserungen durch Data Mining. Anstoß hierfür ist die Problemstellung der EPCOS OHG, die in einem mehrstufigen Produktionsprozess PTC-Bauteile (Positive Temperature Coefficient, Kaltleiter) produziert. Das als Produktionsengpass identifizierte Sintern beeinflusst den elektrischen Widerstand des fertigen Bauteils. Um den gewünschten Widerstand zu erhalten, ist das wiederholte Anpassen der Sintereinstellungen notwendig. Dies führt zu einer schlechten Auslastung der Fertigungskapazität und Verlängerung der Durchlaufzeit. Um den Anpassungsaufwand zu reduzieren, wird ein Vorhersagemodell für den Widerstandswert auf Basis der vom Unternehmen bereitgestellten Daten erstellt. Die Daten beziehen sich auf verschiedene Produktionsabschnitte und Freigabemessungen innerhalb des Produktionsprozesses, die in zwei unterschiedlichen Standorten stattfinden.

Der Prozessstandard CRISP-DM (Cross Industry Standard Process for Data Mining) bietet den Rahmen für die Durchführung der Arbeit. Für die Datenanalyse werden Methoden wie die Berechnung statistischer Größen, die Ausreißeranalyse, die Analyse der fehlenden Daten, die Gruppierung und die Dimensionalitätsreduktion, verwendet. Zur Vorhersage des Widerstandswerts finden die Methoden der Klassifikation Anwendung, zu denen Entscheidungsbäume, Diskriminanzanalyse, Support Vector Machines, k-Nearest Neighbor und Ensemblemethoden zählen. Das Ergebnis des praktischen Teils sind vier aufbereitete und analysierte Datensets, je eines mit und ohne Ausreißer je Standort. Die Evaluation erfolgt durch die Anwendung der Modelle auf unabhängige Testdaten und dem Vergleich der Resultate. Die nicht zufriedenstellenden Ergebnisse beruhen auf den zu kleinen Verhältnissen zwischen der Anzahl der Datensätze und der Anzahl der Attribute. Somit kann keine Implementierung innerhalb des Unternehmens erfolgen. Allerdings würde ein funktionierendes Modell zu 13 % freier Kapazität im betrachteten Zeitraum führen und hätte eine Verkürzung der Durchlaufzeit zur Folge. Um diese in der Zukunft mit Hilfe des Data Minings auszunutzen, werden am Ende der Arbeit Möglichkeiten zur Verbesserung der Datenbasis und der Datenaufbereitung präsentiert.

Abstract

This master thesis deals with production-related improvements by data mining. The problem is posed by EPCOS OHG, which produces PTC (Positive Temperature Coefficient) components in a multi-stage production process. Sintering is the bottleneck of production and influences the electrical properties (resistance value) of the finished component. In order to obtain the desired resistance, repeated adjustments of the sintering settings are necessary. These setting times lead to a poor utilization of the production capacity and an increase in the lead time. For this reason, a prediction model for the resistance value is established based on the data provided by the company. The data comes from two different production sites which include different production stages and release measurements.

The process standard CRISP-DM (Cross Industry Standard Process for Data Mining) provides the framework for this thesis. To analyse data methods like calculating statistical parameters, outlier analysis and analysis of missing data, grouping and dimensionality reduction are applied to the data. For the prediction of the resistance value, methods of classification are used, including decision trees, discriminant analysis, support vector machines, k-nearest neighbor and ensemble methods. The results of the practical part are four prepared and analysed data sets, one each with and without outliers per location. The models are applied to independent test data and their results are compared. The unsatisfactory output is the result of small ratios between the number of attribute values and the number of attributes. Therefore, the implementation within the company won’t be promising. A working model results in 13 % free capacity during the period in question and shortens the lead time. In order to utilise this in the future while using data mining, possibilities to improve the data and the data processing are presented at the end of the thesis.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AVG	Arbeitsvorgang
BaT	Bagged Trees
BG	Beschäftigungsgrad
BoT	Boosted Trees
bzw.	beziehungsweise
CGSVM	Coarse Gaussian Support Vector Machine
CL	Classification Learner
CRISP-DM	Cross-Industry Standard Process for Data Mining
CSVM	Cubic Support Vector Machine
CT	Complex Tree
d. h.	das heißt
DIN	Deutsches Institut für Normung
DL	Deutschlandsberg
ebda	ebenda
engl.	englisch
etc.	et cetera
exkl.	exklusive
FGSVM	Fine Gaussian Support Vector Machine
ggf.	gegebenenfalls
IDC	International Data Corporation
inkl.	inklusive
IQR	Interquartilsabstand
IT	Informationstechnik
KDD	Knowledge Discovery in Databases
KMO	Kaiser-Meyer-Olkin
KNN	k-Nächste-Nachbarn
LSVM	Linear Support Vector Machine
MAR	Missing At Random
MCAR	Missing At Completely At Random
MGSVM	Medium Gaussian Support Vector Machine
MSA	measure of sampling adequacy
MT	Medium Tree
MW	mittlerer gesinterter Widerstand
NM	Nachmahlung

PCA	Hauptkomponentenanalyse (Principal Component Analysis)
PCs	Hauptkompontenten (Principal Components)
PTC	Positive Temperature Coefficient
QSVM	Quadratic Support Vector Machine
RUSBT	RUSBoosted Trees
SEMMA	Sample, Explore, Modify, Model, Assess
ST	Simple Tree
SVD	Singulärwertzerlegung (Singular Value Decompositon)
SVM	Support Vector Machine
TBE	Bearbeitungsende
TBEV	Bearbeitungsende Vorgänger
TRA	Rüstanfang
u. a.	unter anderem
UD	Ofen
VM	Vormahlung
z. B.	zum Beispiel
ZDF	Durchführungszeit
ZDL	Durchlaufzeit
ZUE	Übergangszeit

1 Einleitung

Das erste Kapitel dieser Masterarbeit betrachet die Motivation und die Vorgehensweise zur Erstellung dieser Arbeit. Diese gliedern sich in die Beschreibung der Ausgangssituation und die damit verbundene Problemstellung, in die daraus ableitbare Zielsetzung, in die methodische Vorgehensweise und in die Gliederung der gesamten Arbeit. Das folgende Kapitel hat das Ziel, einen Überblick über die Inhalte und Erwartungen an diese Arbeit zu geben.

1.1 Ausgangsituation und Problemstellung

Anstoß zu dieser Masterarbeit ist die Problemstellung des Unternehmen EPCOS OHG. Die EPCOS OHG entwickelt, fertigt und vertreibt elektronische Bauelemente und gehört zur TDK Corporation, einem Elektronikkonzern mit Sitz in Japan. Der Standort in Deutschlandsberg ist der größte Standort in Europa und das Kompetenzzentrum für keramische Bauelemente der Division Piezo and Protection Device Business Group. Von hier aus werden mehrere Standorte weltweit gesteuert und unterstützt.[1] Dazu gehört auch der Standort in Šumperk, Tschechien, der gemeinsam mit dem Standort in Deutschlandsberg an der Produktion von PTC-Widerständen (Kaltleiter, engl. Positive Temperature Coefficient), im Weiteren mit PTC-Bauteile bezeichnet, beteiligt ist. PTC-Bauteile sind elektronische Bauelemente, deren Widerstand mit zunehmender Temperatur ansteigt. Aus diesem Grund werden sie sowohl als Heizelemente als auch als Temperaturfühler eingesetzt.[2]

Das Erreichen des spezifizierten Sollwert für den gesinterten Widerstand (R_{25}) ist derzeit in der Produktion nicht ohne mehrere Versuche möglich. Entlang des im Abbildung 1 dargestellten mehrstufigen Produktionsprozesses der Herstellung von PTC-Bauteilen gibt es daher mehrere Freigaben, die den R_{25} überprüfen, um ggf. eine Anpassung vornehmen zu können.

[1] Vgl. EPCOS HR (2016), S. 3
[2] Vgl. Reisch, M. (2007), S. 298 f.

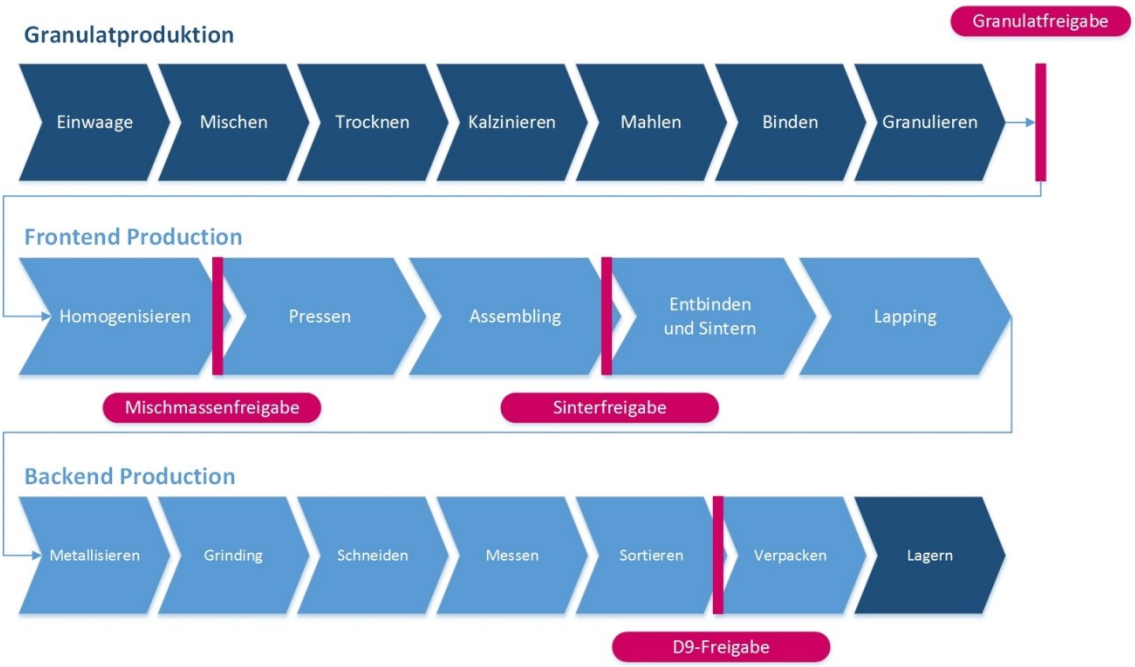


Abbildung 1: Produktionsprozess der PTC-Bauteile[3]

Die Granulatproduktion in Deutschlandsberg (in Abbildung 1 dunkelblau hinterlegt) beinhaltet das Mischen der Komponenten, die in mehreren Schritten zu einem Granulat (Grundmasse) verarbeitet werden, das abschließend mit einer Granulatfreigabe überprüft wird. Dieses Granulat wird anschließend nach Šumperk transportiert. Die in Šumperk stattfindenden Prozessschritte sind in Abbildung 1 hellblau hinterlegt. Hier wird entweder aus mehreren Grundmassen eine neue Mischmasse homogenisiert oder die Grundmasse direkt weiterverarbeitet. Sollte eine Mischmasse entstehen, erfolgt eine Mischmassenfreigabe, wenn nicht, werden sowohl Mischmasse als auch Grundmasse bei der Sinterfreigabe erneut überprüft. Hierbei werden die für das Entbinden und Sintern verwendeten Durchstoßöfen unmittelbar vor der tatsächlichen Produktion bestückt. Allerdings wird dabei zunächst nur jeder fünfzigste Platz ausgelastet, um eine Fehlproduktion zu vermeiden. Anschließend wird der Ofen abhängig vom Resultat eingestellt und es gibt laufende Kontrollen der Sinterqualität. Für Granulat- und Mischmassenfreigabe werden aus den Massen standardisierte Scheiben gepresst, die sich in ihrem Verhalten von dem fertigen Bauteil aufgrund ihrer Geometrie unterscheiden können. Die Sinterfreigabe erfolgt mit der Geometrie des Endproduktes. Alle drei Freigaben unterscheiden sich in der Bearbeitung und Messung der PTC-Bauteile von der Endfreigabe (D9-Freigabe), da die Messteile nicht den gesamten Prozess durchlaufen und unter anderen Voraussetzungen behandelt werden. Sie dauern deutlich länger, da die Sinterung bei der D9-Freigabe bereits abgeschlossen ist.

Die Freigabemessungen, wobei die Sinterfreigabe mehrmals wiederholt werden muss, um die Einstellungen entsprechend anzupassen, sind zeitaufwendig. Durch diese Wiederholungen werden die Fertigungskapazitäten schlecht ausgelastet und die Durchlaufzeit des Fertigungsprozesses wird in die Länge gezogen.

[3] Quelle: in Anlehnung an EPCOS PPD PTC PD (2016), S. 10

1.2 Zielsetzung und Forschungsfrage

Ziel dieser Arbeit ist es, die im Prozess erhobenen Werte für einzelne Produktionschargen, wie Komponenten und Prozessparameter, in Bezug zu den jeweiligen bei den Freigaben gemessenen Zielwerten R_{25} zu setzen. Dies soll dazu führen, Vorhersagen für künftige Lose treffen zu können und den Prozess des Sinterns dementsprechend abzuändern. Zusätzlich soll dies einen positiven Effekt auf die Kapazitätsauslastung haben, da die Sinterfreigabe nicht wiederholt werden müsste bzw. ganz auf sie verzichtet werden und somit direkt mit der Sinterung begonnen werden könnte.

Aus diesen Zielen und der in Kapitel 1.1 erläuterten Ausgangssituation können folgende Forschungsfragen abgeleitet werden:

- Ist es möglich, den Widerstandswert R_{25} der PTC-Bauteile für zukünftige Lose auf Basis der bereits gefertigten vorherzusagen?
- Wie wirkt sich eine exakte Vorhersage auf die Kapazitätsauslastung aus?

1.3 Vorgehensweise und Aufbau der Arbeit

Um die Forschungsfragen aus Kapitel 1.2 bestmöglich zu beantworten, wurden verschiedene Schritte unternommen. Nach einer Besichtigung der Produktionsstandorte Deutschlandsberg und Šumperk wurden die gesammelten Daten gesichtet und eine Literaturrecherche hinsichtlich möglicher bereits anderorts verwendeter Ansätze durchgeführt. Aus dieser entstand die Grundlage für die Vorgehensweise zur Lösung des Problems. Dieses erste Konzept ist in Abbildung 2 ersichtlich. Durch das Schaffen einer einheitlichen Datenbasis, einer anschließenden Dimensionalitätsreduktion und durch Anwenden einer oder mehrerer Methoden aus dem Data Mining werden Modelle zur Vorhersage der Zielwerte der Granulat- und Sinterfreigabe entwickelt.

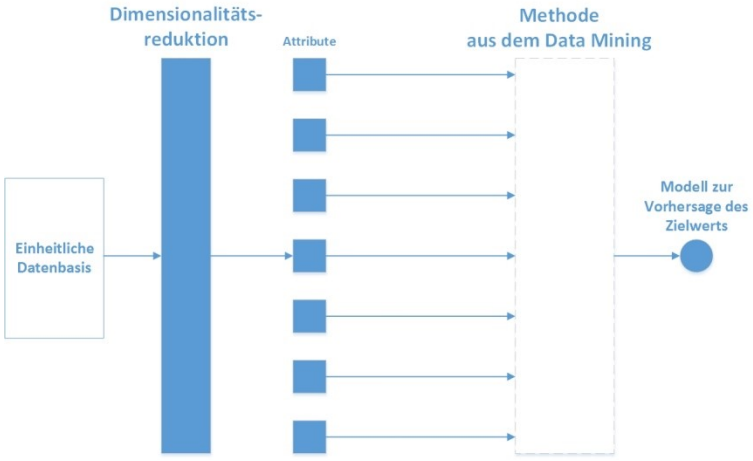


Abbildung 2: Konzept zur Lösung der Forschungsfragen[4]

[4] Quelle: in Anlehnung an Yuan, B. et al. (2000), S. 679

An diesem Konzept orientiert sich der Aufbau der Arbeit, der in Abbildung 3 dargestellt wird. Dunkelblaue Elemente kennzeichnen die grundlegenden Elemente der wissenschaftlichen Arbeit, mittelblaue für die zur Beantwortung der Aufgabenstellung notwendigen Theorie und das hellblaue für die praktische Umsetzung dessen.

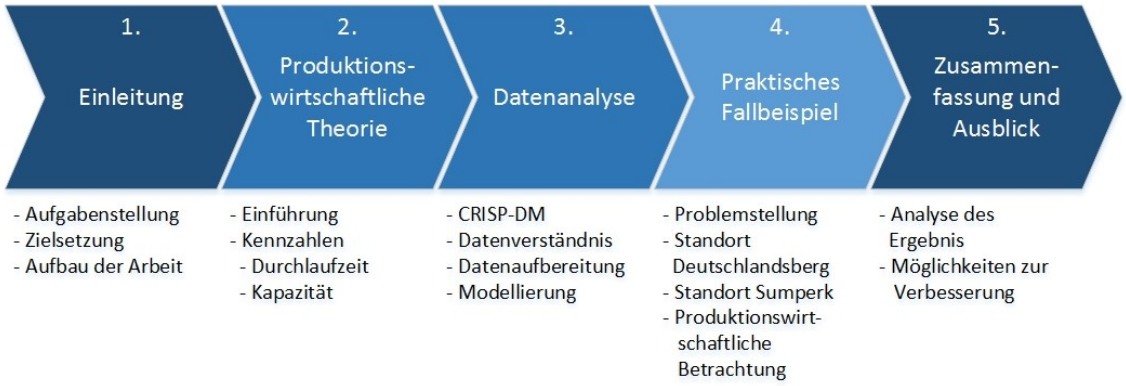


Abbildung 3: Aufbau der Arbeit[5]

Kapitel 1 dient zur Einführung in die Aufgabenstellung. Dabei soll das Problem verstanden, das Ziel definiert und der Aufbau der Arbeit vorgestellt werden.

Um das Ergebnis messen zu können, werden in Kapitel 2 die produktionswirtschaftliche Theorie und die Kennzahlen definiert.

Innerhalb von Kapitel 3 werden mit dem CRISP-DM als Referenzmodell Methoden und Ansätze zum Verständnis und der Aufbereitung der Daten und zur Modellierung auf Basis derer aufgezeigt.

Die praktische Umsetzung erfolgt in Kapitel 4. Hierbei werden die in den beiden vorherigen Kapitel Methoden und Ansätze angewendet und Ergebnisse präsentiert.

Abschließend werden die Arbeit und deren Ergebnisse in Kapitel 5 zusammengefasst. Zusätzlich werden der in Kapitel 4 präsentierte Lösungsweg kritisch hinterfragt und andere Ansätze in Betracht gezogen.

[5] Quelle: eigene Darstellung

2 Produktionswirtschaftliche Theorie

In der Aufgabenstellung in Kapitel 1.2 werden lange Durchlaufzeiten und schlechte Auslastung der Fertigungskapazitäten in der Produktion als Grund für diese Arbeit angegeben. Die daraus abgeleitete Forschungsfrage, wie sich eine exakte Vorhersage auf die Kapazitätsauslastung auswirkt, muss in einen messbaren Kontext gebracht werden. Ziel dieses Kapitels ist es, die betriebswissenschaftliche Theorie und Kennzahlen zu diesen Vorgaben zu betrachten.

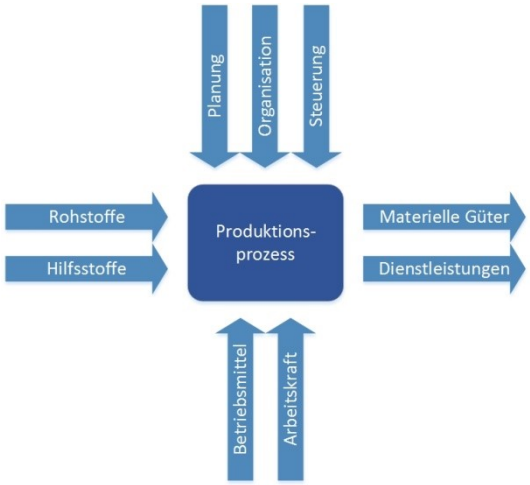


Abbildung 4: Produktion als Kombinations- und Transformationsprozess[6]

Die Grundlage für diese Betrachtung ist die Produktion als Einheit. Eine Produktion als solche ist ein Kombinations- und Transformationsprozess. In diesem werden, wie in Abbildung 4 dargestellt, verschiedene Produktionsfaktoren (Rohstoffe, Hilfsstoffe, Betriebsmittel, u. a.) zu Produkten kombiniert bzw. transformiert. Das Produktionssystem kann in Subsysteme aufgeteilt werden, die kleinsten werden als Leistungsstelle, die aus Maschinen oder technischen Anlagen und Maschinenbedienern bestehen.[7]

[6] Quelle: Zsifkovits, H. E. (2013), S. 106 (leicht modifiziert)
[7] Vgl. ebda., S. 106 ff.

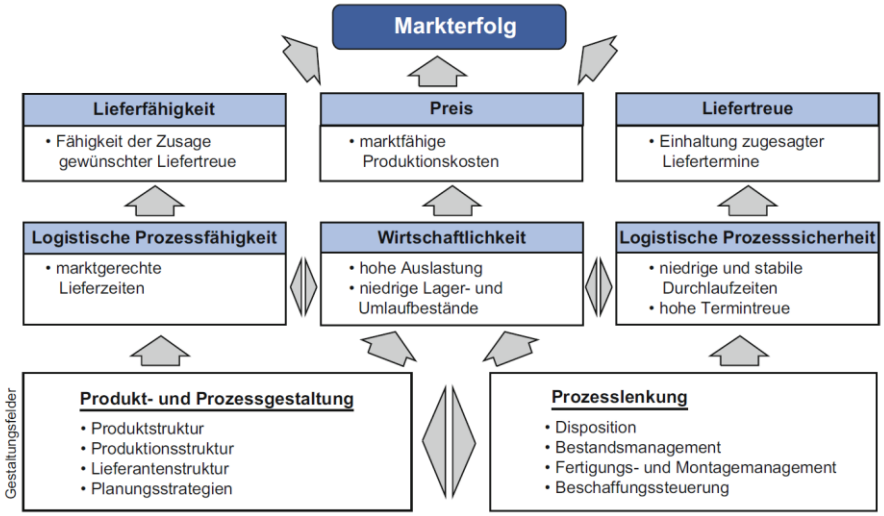


Abbildung 5: Logistische Erfolgsfaktoren von Produktionsunternehmen[8]

Die logistischen Erfolgsfaktoren, mit denen ein Produktionsunternehmen beschrieben werden kann, sind in Abbildung 5 dargestellt. Für das oberste Ziel, den Markterfolg, sind die Lieferfähigkeit, die Liefertreue und der Preis von Bedeutung. Die Lieferfähigkeit wird über die Produkt-, Produktions- und Lieferantenstrukturen beeinflusst. Bei optimaler Gestaltung dieser wird eine logistische Prozessfähigkeit und in weiterer Folge die Prozesssicherheit des Unternehmens gewährleistet. Die dadurch erreichten niedrigen Durchlaufzeiten sollen auf stabilem Niveau und im laufenden Prozess eingehalten werden können. Dies führt zu einer hohen Liefertreue. Die Wechselwirkungen zwischen Leistungs- und Kostenzielen beeinflussen die Wirtschaftlichkeit der Produktion. Marktfähige Produktionskosten können durch eine hohe Auslastung der bereitgestellten Kapazitäten und durch Reduktion der Kapitalbindungskosten, die durch Lager- und Umlaufbeständen anfallen, realisiert werden.[9]

2.1 Das Dilemma der Ablaufplanung

Das Bestreben einzelne logistische Erfolgsfaktoren zu optimieren, führt zu einem Zielkonflikt. Die Zielsetzungen und Anforderungen sind oftmals gegensätzlich und schwanken lokal und temporär. So sind, um eine hohe Auslastung zu erzielen, hohe Bestände notwendig. Dies führt zu langen und schwankenden Durchlaufzeiten, wodurch die Terminsicherheit nicht gewährleistet werden kann. Diese Konstellation der Zielsetzungen wird als Dilemma der Ablaufplanung bezeichnet.[10] Kein einzelner Wert wird optimiert, sondern es muss ein Gleichgewicht zwischen den Teilzielen gefunden werden. Dazu müssen alle berücksichtigt und gewichtet werden.[11]

[8] Quelle: Gläßner, J. (1995) zitiert nach Nyhuis, P.; Wiendahl, H.-P. (2012), S. 3
[9] Vgl. Nyhuis, P.; Wiendahl, H.-P. (2012), S. 3
[10] Vgl. Gutenberg, E. (1971), S. 216
[11] Vgl. Nyhuis, P.; Wiendahl, H.-P. (2012), S. 4

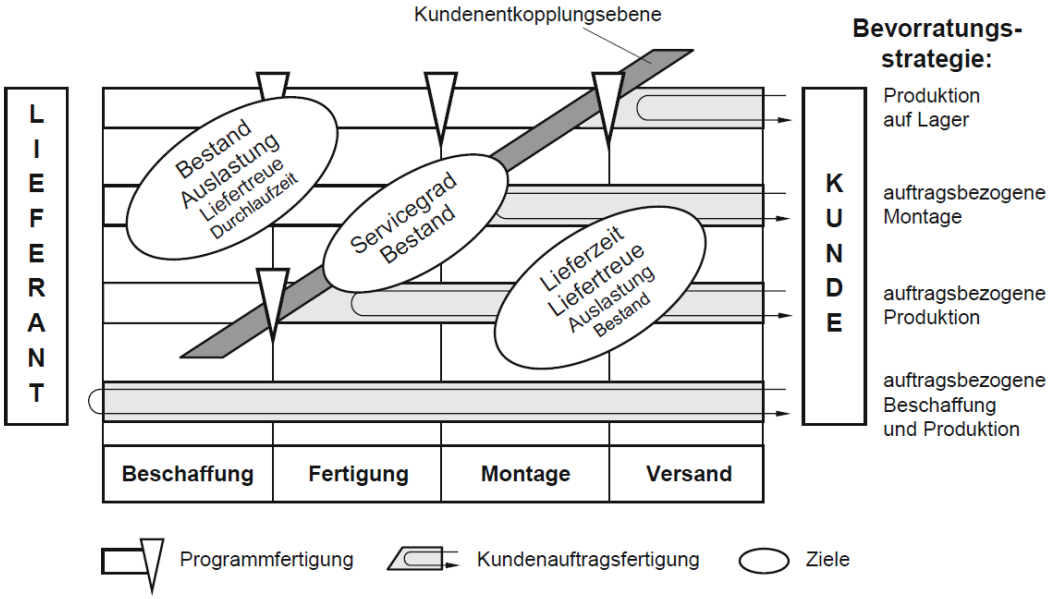


Abbildung 6: Gewichtung von logistischen Zielgrößen[12]

Bei unterschiedlichen Bevorratungsstrategien und unterschiedlichen Kundenentkoppelungspunkten innerhalb der Produktion unterscheidet sich die Gewichtung der Zielgrößen. Bei einer Produktion ohne den konkreten Kundenauftrag stehen geringe Bestände und eine hohe Auslastung als Ziele im Vordergrund, die direkt die Wirtschaftlichkeit der Produktion beeinflussen. Die Ziele der Liefertreue und Durchlaufzeit werden in einem zweiten Schritt betrachtet, da diese die Lagerhaltung betreffen. Das Ziel einer geringen Termintreue und großen Durchlaufzeit bedingt einen großen Lagerbestand. Bei einer kundenbezogenen Produktion verändern sich die Zielgewichtungen, sodass die Lieferzeit und die Liefertreue in den Vordergrund treten. Bei nicht Einhaltung zugesagter Lieferzeiten und -termine ist der Kunde direkt betroffen. Fraglich ist jedoch, ob kurze Lieferzeiten zugunsten von geringer Auslastung der Produktionssysteme sinnvoll sind. Die dadurch notwendigen Kapazitätsausweitungen (Betriebsmittel und/oder Personal) führen zu einem Anstieg der Stückkosten, die an den Kunden weiterverrechnet werden müssen.[13]

Im Hinblick auf das Ziel dieser Arbeit (Kapitel 1.2) wird genauer auf den Faktor Betriebsmittel und dessen Potenzial eingegangen. Die Ziele der Auslastung und der Durchlaufzeit, die in dieser Arbeit im Vordergrund stehen, bedingen eine Auseinandersetzung mit der Kapazität und ihrer Zusammensetzung und der Durchlaufzeit als Kenngröße der Produktionslogistik.

[12] Quelle: in Anlehnung an Eidenmüller, B. (1995) zitiert nach Nyhuis, P.; Wiendahl, H.-P. (2012), S. 4
[13] Vgl. Nyhuis, P.; Wiendahl, H.-P. (2012), S. 4 f.

2.2 Durchlaufzeit

Nyhuis und Wiendahl (2012)[14] unterscheiden zwischen der Durchlaufzeit eines Produktionsauftrags, der Durchlaufzeit eines Fertigungsauftrags, im Weiteren als Auftrag bezeichnet, und dem Durchlaufelement. Diese werden in Abbildung 7 dargestellt. Der obere Teil zeigt den Durchlauf eines Produktionsauftrages, der aus zwei Fertigungsaufträgen und einem Montageauftrag besteht. Handelt es sich um eine losweise Fertigung, folgt nach der Beendigung eines Arbeitsvorgangs auf dem Arbeitssystem mit einer möglichen Liegezeit vor Ort der Transport zum Folgearbeitssystem. Hier kann der Auftrag bearbeitet werden, wenn die Kapazitäten zur Bearbeitung frei sind und die notwendigen Rüstungsmaßnahmen durchgeführt worden sind.[15]

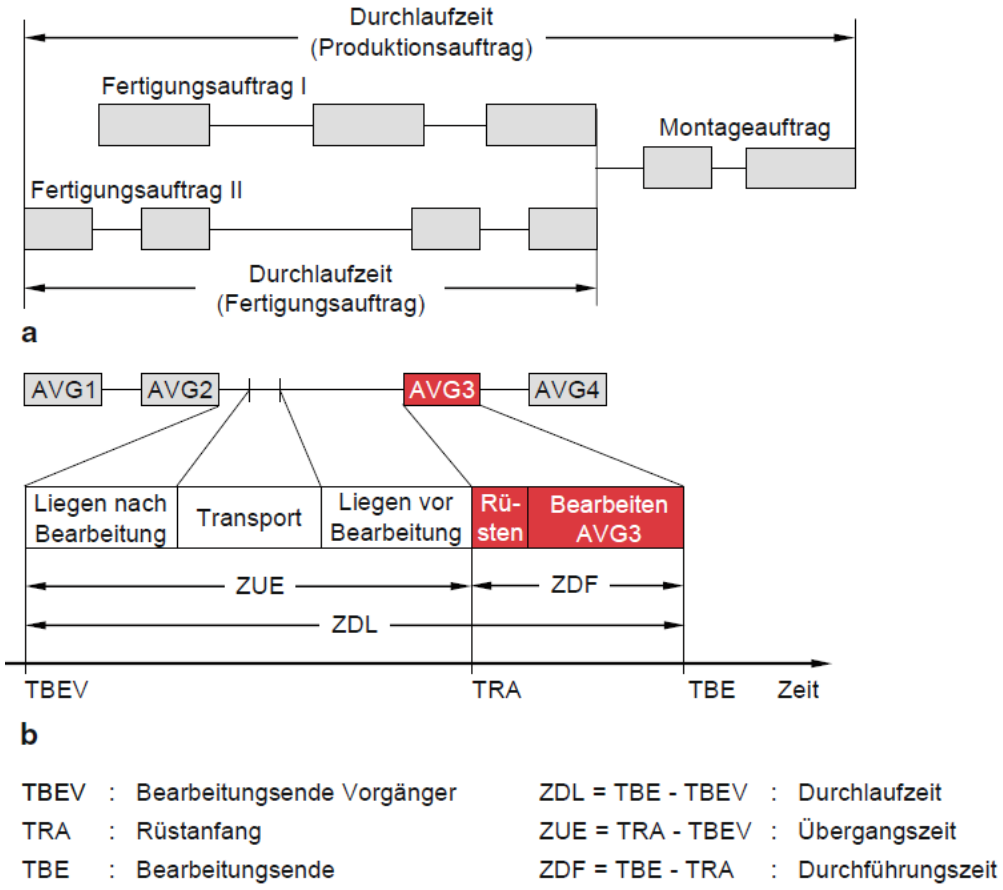


Abbildung 7: Durchlaufzeitanteile und Durchlaufelement[16]

Im unteren Teil der Abbildung 7 sind die Ablaufschritte für einen Arbeitsvorgang, das arbeitsvorgangsbezogenes Durchlaufelement, dargestellt.[17] Als die Durchlaufzeit ZDL eines Arbeitsvorgangs gilt die Zeitspanne, „die ein Auftrag von der Beendigung des vorhergehenden Arbeitsvorgangs bzw. vom Einstoßzeitpunkt des Auftrages (beim

[14] Vgl. Nyhuis, P.; Wiendahl, H.-P. (2012), S. 21 f.
[15] Vgl. ebda., S. 21
[16] Quelle: ebda., S. 22
[17] Vgl. ebda., S. 21

ersten Arbeitsvorgang) $TBEV$ bis zum Bearbeitungsende des betrachteten Arbeitsvorgangs TBE selbst benötigt“ [18] (2.1).

$$ZDL = TBE - TBEV \tag{2.1}$$

Nach Nyhuis und Wiendahls (2012)[19] Definition werden Liegen nach Bearbeiten, Transportzeit und das Liegen vor Bearbeitung dem jeweiligen untersuchten Arbeitsvorgang zugeordnet. Diese werden zur Übergangszeit ZUE zusammengefasst werden, wobei diese durch die Differenz der Durchlaufzeit ZDL und Durchfühgszeit ZDF berechnet wird (2.2).[20]

$$ZUE = ZDL - ZDF \tag{2.2}$$

Ein relatives Maß für die Durchlaufzeit ist der mittlere Flussgrad FG_m. Dieser stellt das Verhältnis der mittleren Durchlaufzeit ZDL_m zur mittleren Durchführungszeit ZDF_m dar (2.3):[21]

$$FG_m = \frac{ZDL_m}{ZDF_m} \tag{2.3}$$

Gudehus (2010)[22] definiert die Aufteilung der Zeitanteile anders. Er unterscheidet in Wartezeit, Rüstzeit, Leistungszeit und Verfahrenszeit (2.4).[23]

$$\text{Auftragsdurchlaufzeit} = \text{Wartezeit} + \text{Rüstzeit} + \text{Leistungszeit} + \text{Verfahrenszeit} \tag{2.4}$$

So gelten jene Zeiten als Rüstzeiten, die vor und nach der Leistungszeit anfallen. Dazu zählen u. a. folgende:[24]

- Auftragsannahmezeiten
- Vor- und Nachbereitungszeiten
- Materialbereitstellungszeiten
- Umschalt- und Räumzeiten
- Be- und Entladezeit
- Datenerfassungszeiten

Die Leistungszeit hingegen ist die Zeit, die eine einsatzbereite Leistungsstelle für die Erbringung der im Auftrag geforderten Leistung benötigt und ist in der Regel von der Auftragsgröße abhängig. Typische Beispiele für diese sind:[25]

- Fertigungszeiten
- Montagezeiten
- Bearbeitungszeiten
- Reparaturzeiten

[18] Nyhuis, P.; Wiendahl, H.-P. (2012), S. 21
[19] Vgl. ebda., S. 21 f.
[20] Vgl. ebda., S. 23
[21] Vgl. ebda., S. 23
[22] Vgl. Gudehus, T. (2010), S. 216
[23] Vgl. ebda., S. 216
[24] Vgl. ebda., S. 216
[25] Vgl. ebda., S. 216

Die Verfahrenszeit ist verfahrenstechnisch bedingt und muss verstreichen, bevor am Auftragsgegenstand der nächste Bearbeitungsschritt durchgeführt werden darf. Beispiele hierfür sind:[26]

- Trocknungszeiten
- Aushärtungszeiten
- Ablagerungszeiten
- Reifezeiten
- Gärungszeiten

Aus den Minima der Rüst-, Leistungs- und Verfahrenszeit kann die minimale Durchlaufzeit berechnet werden. Die Differenz zwischen der tatsächlichen und minimalen Durchlaufzeit wird als Wartezeit bezeichnet. Diese ergibt sich aus:[27]

- Ausfallzeiten durch Betriebsunterbrechung, Störung oder fehlende Personalbesetzung
- Totzeiten wegen fehlender Daten, Informationen, Entscheidungen oder Anweisungen
- Nachbearbeitungszeiten zur Beseitigung von Fehlern und Mängeln am Auftragsgegenstand
- Stauzeiten infolge der Belegung einer Leistungsstelle durch vorherige Aufträge
- Blockierzeiten durch einen Rückstau in einer nachfolgenden Leistungsstelle
- Materialbeschaffungszeiten, die durch fehlendes Material verursacht werden
- Unterbrechungszeiten infolge eines Ausfalls oder Nichtverfügbarkeit einer vorangehenden Leistungsstelle
- Pufferzeiten, die disponiert werden, um Aufträge anzusammeln, um die optimale Auslastung zu erreichen

Auffällig ist bei Gudehus (2010)[28] Einteilung, dass Transportzeiten nicht als solche deklariert werden. Sie könnten der Materialbereitstellungszeit zugeordnet werden, und somit in die Rüstzeit fallen. Nyhuis und Wiendahl (2012)[29] unterscheiden hingegen nicht, ob eine Liegezeit verfahrensbedingt ist oder nicht. Bei einer Betrachtung der Durchlaufzeit ist daher zu überprüfen, wie genau einzelne Zeitanteile definiert werden müssen und welche Definition als Referenz dienen soll.

[26] Vgl. Gudehus, T. (2010), S. 217
[27] Vgl. ebda., S. 217 f.
[28] Vgl. ebda., S. 216
[29] Vgl. Nyhuis, P.; Wiendahl, H.-P. (2012), S. 21 f.

Tabelle 1: Kenngrößen der Produktionslogistik[30]

Zeit	Qualität
• Durchlaufzeit/Lieferzeit • Warte-/Liegezeiten vs. produktive Zeiten • Wiederbeschaffungszeit • Kapitalbindung	• Kundenzufriedenheit • Kundennutzen • Produkt-/Dienstleistungsqualität • Lieferqualität
Kosten/Wirtschaftlichkeit	**Flexibilität**
• Input: Material, Personal, Kapital, Overhead • Output: Umsätze • Produktivität, Rentabilität, Effizienzkennzahlen	• Anpassungsumfang • Kosten/Ertrag aus Anpassungen • Zeitbedarf für Anpassungen

Die Wechselwirkungen der Durchlaufzeit mit den Kenngrößen der Produktionslogistik, wie sie in Tabelle 1 dargestellt sind, sind eine Ausprägung des in Kapitel 2.1 vorgestellten Dilemmas der Ablaufplanung. Auch hier wird ein Zielkonflikt beim Optimierungswunsch mehrere Zielgrößen verursacht. So kann eine Veränderung der Durchlaufzeit die Veränderung anderer Größen, wie Bestände, Kundenzufriedenheit, Produktionskosten und Flexibilität, beeinflussen.[31]

Eine Beeinträchtigung in negative Richtung liegt dann vor, wenn die Durchlaufzeit ansteigt. Dadurch nimmt die Planungsqualität ab und Risiken und Kosten steigen. Unvorhergesehene Entwicklungen und Störungen können eintreten und lange Lieferzeiten können Kundenunzufriedenheit und Auftragsänderungen mit sich bringen.[32]

Eine Verkürzung der Durchlaufzeit hat verschiedene Auswirkung. So wird die Logistikleistung verbessert. Dies führt zu einer Verkürzung von Lieferzeiten und einer Erhöhung der Liefertreue, -fähigkeit, Flexibilität, Lieferqualität und Informationsbereitschaft. Ebenso erfolgt eine Reduktion der Logistikkosten, die durch eine Verringerung der Lenkungs-, Handlings- und Lagerkosten und der Verringerung der Kapitalbindung und Wagnissen erreicht wird. Um dies zu erreichen sollte eine Kernzeitanalyse vorgenommen werden.[33]

Eine Erhöhung der Flexibilität kann nur in dem Maß erfolgen, in dem Produktionskapazität bereit steht. Daher stehen im nächsten Kapitel die Kapazität und die Möglichkeiten der Anpassung im Fokus.

[30] Quelle: Zsifkovits, H. E. (2013), S. 112
[31] Vgl. ebda., S. 112
[32] Vgl. ebda., S. 113
[33] Vgl. ebda., S. 117

2.3 Kapazität

Der Produktionsfaktor Betriebsmittel kann durch verschiedene Größen beschrieben werden. Eine davon ist die Kapazität. In diesem Kapitel soll erklärt werden, wie sie definiert wird und wie Kapazität und Durchlaufzeit in Zusammenhang stehen. Zusätzlich wird die Möglichkeit der Anpassung mit ihren Zielen, Herausforderungen, Voraussetzungen und verschiedenen Formen erläutert.

2.3.1 Einführung

Die Kapazität ist das Leistungsvermögen einer technischen Einheit in einem definierten Zeitabschnitt. Die Ausnutzung in Prozent dieses Leistungsvermögens kann als *Beschäftigungsgrad* (*BG*) (*Kapazitätsnutzungsgrad*) (2.5) angegeben werden.[34]

$$\text{Beschäftigungsgrad} = \frac{\text{tatsächlich in Anspruch genommene Leistung}}{\text{maximal mögliche Leistung}} * 100\% \quad (2.5)$$

Die Kapazität kann in betriebswirtschaftliche und technische Komponenten zerlegt werden:[35]

$$K = C \times I \times D \quad (2.6)$$

Die Multiplikation von Kapazitanz C, Produktionsgeschwindigkeit I, und Produktionsdauer D bildet die Kapazität K (2.6).[36] Die Kapazitanz C bildet die technische Komponente und beschreibt Anzahl und Eigenschaften der Betriebsmittel. Als technisch-wirtschaftliche Komponente wird die Produktionsgeschwindigkeit I (Intensität der Betriebsmittelnutzung) angesehen, die innerhalb bestimmter Grenzen ($I_{min} \leq I \leq I_{max}$) veränderlich ist und die dadurch die Höhe der variablen Kosten und die Durchlaufzeit beeinflussen kann. Die Produktionsdauer D gibt die Länge des Zeitraums, der zur Normierung verwendet wird, an.[37]

Die Ausprägungen des Kapazitätsbegriffs sind in Tabelle 2 zusammengefasst. Soll die Kapazität als betriebliche Kennzahl erfasst werden, muss zuerst der Kapazitätsbegriff als solches definiert werden. Die Kapazität selbst und ihre Inanspruchnahme sind wichtige Einflussfaktoren für Planung, Steuerung und Überwachung der Produktion.[38]

[34] Vgl. Blohm, H. et al. (2008), S. 168
[35] Vgl. Männel, W. (1979) zitiert nach Blohm, H. et al. (2008), S. 168
[36] Vgl. ebda., S. 168
[37] Vgl. Blohm, H. et al. (2008), S. 168
[38] Vgl. ebda., S. 168 f.

Tabelle 2: Ausprägungen des Kapazitätsbegriffs[39]

Unterscheidungs-merkmal	**Kapazitätsbegriff**	**Definition/Charakterisierung**
(1) Teilbarkeit	a) Ganzheitliche Kapazität	Die Kapazität des Betriebes wird als unteilbares Ganzes gesehen
	b) Zusammengesetzte Kapazität	Die Kapazität des Betriebes setzt sich aus den Kapazitäten der einzelnen Aggregate zusammen. Bei Aggregaten, die unabhängig voneinander eingesetzt werden: Addition der Teilkapazitäten. Bei verbundenen Aggregaten: Gesamtkapazität = Kapazität des Engpasses
(2) Menge und Art der Leistung	a) Quantitative Kapazität	Mengenmäßige Leistungsfähigkeit
	b) Qualitative Kapazität	Leistungsfähigkeit in Bezug auf die Eigenschaften der abgegebenen Leistungen.
(3) Technische Gesichtspunkte	a) Technische maximal mögliche Leistungsfähigkeit	Theoretisch mögliche Leistungsfähigkeit unter den günstigsten Bedingungen bei voller Auslastung sämtlicher Produktionssysteme; Vorzug: feste Vergleichsgröße für die erbrachte Mengenleistung
	b) Durchschnittliche Dauerleistungsfähigkeit	Gegenüber (3a) sind hier die normalen Ausfälle infolge Wartung und Reparaturen sowie Abwesenheit von Belegschafts-angehörigen berücksichtigt. Nachteil bei der Verwendung als Bezugsgröße: zeitweise Überschreitung durch tatsächlich erbrachte Leistung möglich (BG > 100%)
	c) Momentane (jeweilige) Leistungsfähigkeit	Leistungsfähigkeit in einem bestimmten, in der Regel kurzen Zeitraum (konkreter Tag, konkrete Woche). Wegen ihrer Schwankungen als Bezugsgröße ungeeignet, aber als Vergleichsmaß für die Beurteilung des auf die technische Höchstleistungsfähigkeit bezogenen BG wertvoll.
(4) Bezugszeitraum	a) Totalkapazität	Leistungsfähigkeit während der gesamten technischen oder wirtschaftlichen Nutzungsdauer
	b) Periodenkapazität	Leistungsfähigkeit während einer Rechnungs- oder Planungsperiode

[39] Quelle: Blohm, H. et al. (2008), S. 169 (leicht modifiziert)

Wird in einem Betrieb, einem definierten Betriebsteil oder auf einer Anlage ein einheitliches Produkt erzeugt, kann die Kapazität einfach gemessen werden. Dies passiert dann über die erzeugten Mengeneinheiten, wie Stück, Liter, Meter, etc. pro Zeiteinheit. Bei einer Sortenfertigung können verschiedene Sorten unter Zuhilfenahme von Äquivalenzzahlen miteinander verglichen werden.[40]

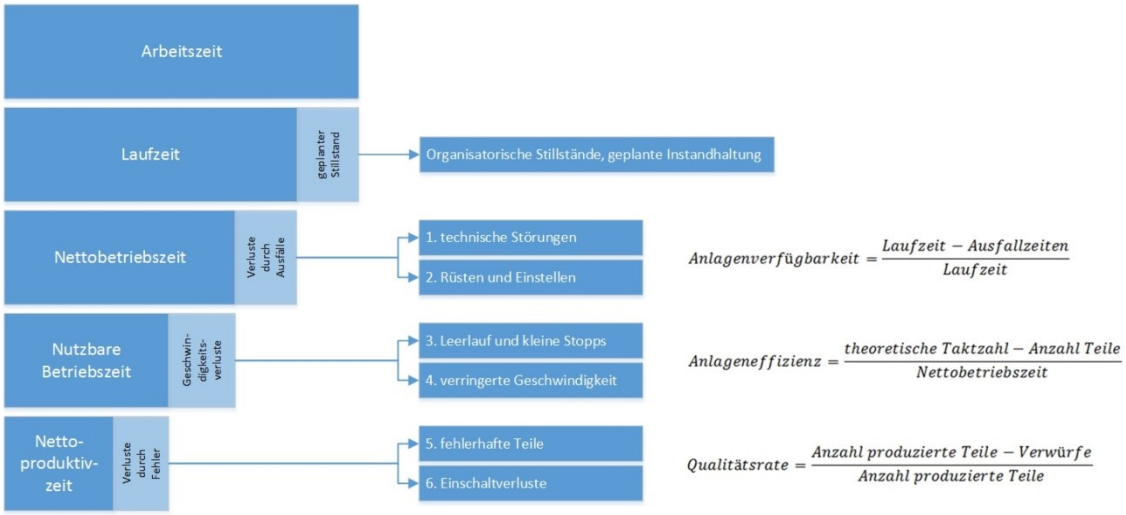


Abbildung 8: Berechnung des OEE[41]

Die Kapazität kann durch ungeplante und geplante Stillstände, die durch Wartung, Urlaub, Krankheit oder technische Ausfälle bedingt sind, geschmälert werden. Eine Möglichkeit, um die verfügbare Kapazität zu bestimmen, ist die Berechnung der *Gesamtanlageneffektivität* OEE (Overall Equipment Efficiency). Diese setzt sich aus den drei Größen Anlagenverfügbarkeit, Anlageneffizienz und der Qualitätsrate zusammen (2.7). Ihre Berechnung wird in Abbildung 8 zusammengefasst.

$$\text{OEE} = \text{Anlagenverfügbarkeit} * \text{Anlageneffizienz} * \text{Qualitätsrate} \qquad (2.7)$$

Die Planung von betrieblichen Teilkapazitäten kann Engpässe aufzeigen. Ein Engpass ist jenes Betriebsmittel bzw. eine Betriebsmittelgruppe, deren Kapazität zur maximalen Gesamtleistung des Betriebs wird, wenn gleichzeitig in anderen Bereichen ungenützte Kapazitäten vorliegen.[42] Dies und eine Differenz zwischen Kapazitätsbedarf und verfügbaren Kapazitäten führen zur Notwendigkeit der Anpassung der Kapazität.[43] Hierzu sollten verschiedene Möglichkeiten und die damit einhergehenden Herausforderungen betrachtet werden, auf die im nächsten Kapitel eingegangen wird.

[40] Vgl. Blohm, H. et al. (2008), S. 171
[41] Quelle: Fraunhofer-Institut Produktionstechnik und Automatisierung IPA zitiert nach Zsifkovits, H. E. (2013), S. 151
[42] Vgl. Blohm, H. et al. (2008), S. 171
[43] Vgl. Zsifkovits, H. E. (2013), S. 152

2.3.2 Anpassungen der Kapazität

Das Ziel der Anpassung der Kapazität liegt „in der Angleichung des Zeitfonds des Kapazitätsangebots und des Kapazitätsbedarfs je Kapazitätseinheit, um eine bestmögliche Bedarfsbefriedigung abzusichern.“[44] Zusätzlich sollten die dabei anfallenden Kosten betrachtet werden. Auszugehen ist von den f*ixen Kosten* Q, die bei Aufrechterhaltung der Betriebsbereitschaft anfallen, und die von der Inanspruchnahme der betrieblichen Anlagen und von der Art der Anpassung an die Beschäftigungsschwankungen unabhängig sind. Die fixen Kosten setzen sich aus Nutzkosten K_n und Leerkosten K_l zusammen. Nutzkosten fallen für die tatsächlich genutzte Kapazität der Anlage an, wohingegen Leerkosten für die nicht genutzte Kapazität stehen.[45]

$$Q = K_l + K_n \tag{2.8}$$

Die Nutz- und Leerkosten können unter Zuhilfenahme der effektiv erzeugten Produktmenge x und dem Maximum der herstellbaren Produktmenge m folgendermaßen ausgedrückt werden:[46]

$$K_l(x) = (m - x)\frac{Q}{m} \tag{2.9}$$

$$K_n(x) = x\frac{Q}{m} \tag{2.10}$$

Der Zusammenhang der Nutz- und Leerkosten wird in Abbildung 9 dargestellt. Bei voller Ausnutzung der verfügbaren Kapazität fallen keine Leerkosten an und die Fixkosten entsprechen den Nutzkosten.[47]

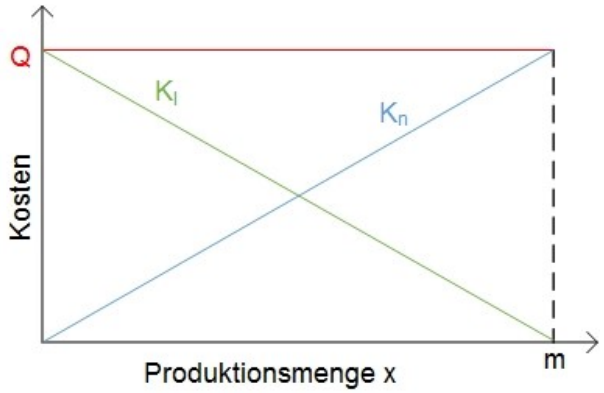


Abbildung 9: Verlauf von Nutz- und Leerkosten[48]

Das Ziel Leerkosten abzubauen kann nicht mit sämtlichen Anpassungsformen erreicht werden. Dies ist der Fall bei der später vorgestellten intensitätsmäßigen Anpassung, da sich hier nur die Inanspruchnahme verändert, jedoch nicht der Faktorbestand.[49]

[44] Vgl. Nebl, T. (2011), S. 224
[45] Vgl. Gutenberg, E. (1971), S. 348 f.
[46] Vgl. ebda., S. 349
[47] Vgl. ebda., S. 349
[48] Vgl. ebda., S. 349
[49] Vgl. ebda., S. 349

Voraussetzungen und Herausforderungen

Die Voraussetzungen und Herausforderungen bei der Anpassung sind vielfältig. Dazu zählen das Anpassungsintervall, die Kostenremanenz, die Anpassungskosten und die Elastizität als Anpassungsfähigkeit, die im Weiteren näher erklärt werden.

Bei dem *Anpassungsintervall* handelt es sich um den Zeitraum zwischen der Anpassungsnotwendigkeit und der erfolgten Ausführung der Anpassungshandlung. Er beinhaltet sowohl Leerzeiten, wie die Zeitspanne des Erkennens der Anpassungsnotwendigkeit oder die Dauer des informationssammelnden, -bearbeitenden und ausführenden Prozess. Die Länge ist abhängig von der Organisation des Entscheidungsprozesses und auch vom Auslöser der Anpassungsnotwendigkeit. Zusätzlich sind Mechanismen zur Kontrolle, ob eine Anpassung notwendig ist, einzusetzen.[50]

Unter *Kostenremanenz* wird der Umstand bezeichnet, dass sich Kosten nicht proportional zum Beschäftigungsgrad verhalten. Bei rückläufiger Beschäftigung laufen die Gesamtkosten nicht auf der gleichen Kostenkurve K_a zurück, die die Zunahme der Kosten bei steigender Beschäftigung beschreibt, sondern sie fallen nach der Kostenkurve K_r, die darüber liegt.[51]

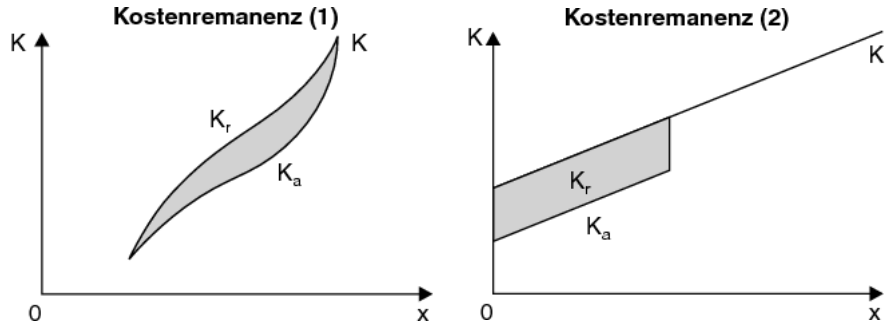


Abbildung 10: Kostenremanenz: (1) Hysteresis-Schleife, (2) remanenter Sprung[52]

Die Kostenremanenz kann sich als Hysteresis-Schleife oder als remanenter Kostensprung manifestieren und ist in Abbildung 10 zu sehen.[53] Dies hat eine Steigerung der Kosten bei gleichem Beschäftigungsgrad aber zu unterschiedlichen Zeitpunkten zur Folge.

Die Anpassungsentscheidung und -ausführung verursacht Kosten. Diese können in drei Gruppen eingeteilt werden: die Elastizitätskosten, die anfallen, um zukünftige Anpassungsmaßnahmen zu erleichtern, jene Kosten, die bei der Wirksamkeit der Anpassungsmaßnahmen entstehen (z. B. Betriebskosten der neuen Anlage), und die *Anpassungskosten*, die gemeinsam mit den Elastizitätskosten durch die Anpassungsentscheidung und -ausführung entstehen. Sie beginnen mit der Erfassung der Anpassungsnotwendigkeit und erreichen ihr Maximum bei der vollen Wirksamkeit der Maßnahme. Die Anpassungskosten gelten als „wertmäßige Trägheitsschwelle“ und

[50] Vgl. Swoboda, P. (1964), S. 56
[51] Vgl. Weber, J.; Steven, M., http://wirtschaftslexikon.gabler.de/Archiv/7702/kostenremanenz-v6.html (Zugriff: 25.04.2017)
[52] Quelle: ebda.
[53] Vgl. ebda.

können ausschlaggebend für die Verzögerung oder Verhinderung einer sonst rentablen Anpassungsmaßnahme sein.[54]

Die *Elastizität* wird als Reagibilität eines produktiven Wirtschaftsgebildes gegenüber veränderten Produktionsaufgaben bezeichnet. Sie kann in der Anpassung der Mittel und Kräfte resultieren und wird durch Verfahren und Ziele des Wirtschaftsgebildes gegenüber Änderungen in der Umwelt und im eigenen Betrieb definiert.[55] Die Elastizität der Anpassungsfähigkeit kann in quantitative, qualitative, zeitliche und räumliche Elastizität unterteilt werden.[56] Die Elastizität schafft die Voraussetzungen für eine Anpassung, wobei jedoch die Wirtschaftlichkeit dieser ausschlaggebend ist.[57]

Arten der Anpassung

Nach der Betrachtung der Voraussetzungen und Herausforderungen besteht die Frage nach den Möglichkeiten zur Anpassung. Gutenberg (1971)[58] hat mit der intensitätsmäßigen, quantitativen und zeitlichen Anpassung die Änderung des Beschäftigungsgrades abgedeckt. Swoboda (1964)[59] ergänzte diese drei Grundtypen mit der qualitativen Anpassung.

Eine *intensitätsmäßige Anpassung* liegt vor, wenn der gesamte fertigungstechnische Apparat bei gleichbleibender Betriebsdauer unterschiedlich stark beschäftigt wird. Es wird die Intensität der Betriebsmittelnutzung angepasst und so die Kapazität verändert (siehe (2.6)). Diese Art der Anpassung kommt bei Anlagen und Anlagenkomplexen zum Einsatz, die nicht in mehrere selbstständige Teileinheiten getrennt werden können.[60] Hierbei muss beachtet werden, dass bei einer rückläufigen Beschäftigung die Leerkosten grundsätzlich konstant bleiben, da nicht der Faktoreinsatz sondern die Inanspruchnahme variiert.[61]

[54] Vgl. Swoboda, P. (1964), S. 59 f.
[55] Vgl. Moxter, A. et al. (1954), S. 87 zitiert nach Swoboda, P. (1964), S. 61
[56] Vgl. Moxter, A. et al. (1954), S. 105 ff. zitiert nach Swoboda, P. (1964), S. 61
[57] Vgl. Swoboda, P. (1964), S. 61
[58] Vgl. Gutenberg, E. (1971), S. 355 f.
[59] Vgl. Swoboda, P. (1964), S. 49
[60] Vgl. Gutenberg, E. (1971), S. 355
[61] Vgl. ebda., S. 349

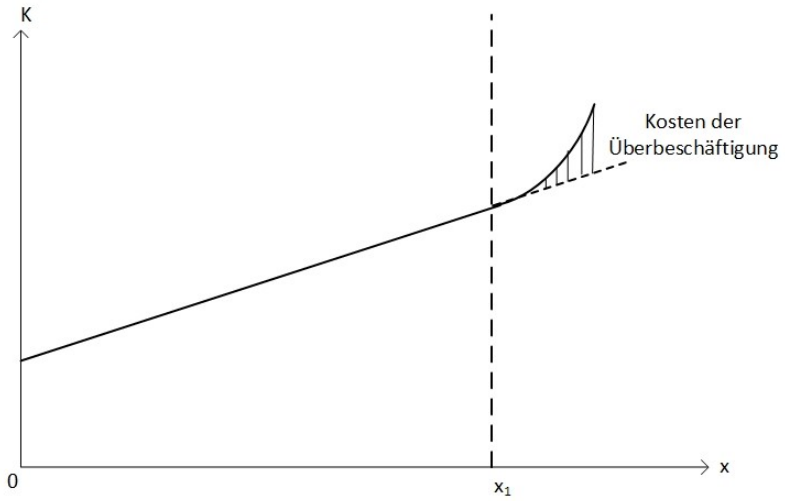


Abbildung 11: Gesamtkostenverlauf beim Übergang[62]

Ein möglicher Kostenverlauf wie in liegt vor, wenn bis zur Ausbringungsmenge x1 innerhalb der maximalen möglichen Betriebszeit und bei optimaler Leistung zeitlich angepasst wird und Mengen, die darüber hinausgehen, durch intensitätsmäßige Anpassungen erzeugt werden.[63]

Werden fertigungstechnische Anpassungen im Sinne einer Stilllegung oder Wiederinbetriebnahme von selbstständigen Teileinheiten an die sich ändernde Beschäftigungslage durchgeführt, wird diese als *quantitative* Anpassung bezeichnet. Es wird die Kapazitanz angepasst (siehe (2.6)). Die Voraussetzungen hierfür ist die Selbstständigkeit der Teileinheiten.[64] Zusätzlich kann die quantitative Anpassung in zwei Formen, die *rein quantitative* und die *selektive* Anpassung unterschieden werden. Bei der rein quantitativen Anpassung hat ein Betrieb über gleichartige Aggregate und ein beliebiges stillgesetzt bzw. in Betrieb genommen. Bei der selektiven Anpassung wird jenes Aggregat ausgewählt, wenn aus einem heterogenen Maschinenbestand hinsichtlich Kriterien, wie Leistungsvermögen und Kostenstrukturen, jene ausgewählt werden, die in Betrieb genommen oder stillgesetzt werden.[65] Diese Anpassungen sind mit Kosten, entweder solche für die Investition oder die für die kurzfristige Bereitstellung von Kapazitäten, wie z. B. durch Anmietung, verbunden.[66]

Bei einer *zeitlichen Anpassung* wird die Betriebszeit je nach Beschäftigungslage verkürzt oder verlängert. Die Betriebsanlagen werden in dieser Zeit voll ausgelastet. In diesem Fall kommt es zu Veränderungen der Produktionsdauer (siehe (2.6)).[67] Dies geschieht durch Überstunden und/oder durch die Anpassung des Schichtmodells, der Pausenregelung, der Fertigungszeiten. Zusätzlich verbessert die Abstimmung der Taktung der einzelnen Aggregate die Betriebszeit.[68] Durch Überstundenzuschläge

[62] Quelle: Blohm, H. et al. (2008) (leicht modifiziert), S. 81
[63] Vgl. Blohm, H. et al. (2008), S. 81
[64] Vgl. Gutenberg, E. (1971), S. 355 f.
[65] Vgl. Blohm, H. et al. (2008), S. 83
[66] Vgl. Zsifkovits, H. E. (2013), S. 152
[67] Vgl. Gutenberg, E. (1971), S. 356
[68] Vgl. Zsifkovits, H. E. (2013), S. 152

steigen die Lohnkosten außerhalb der Normalarbeitszone stärker, jedoch weiterhin linear an.[69]

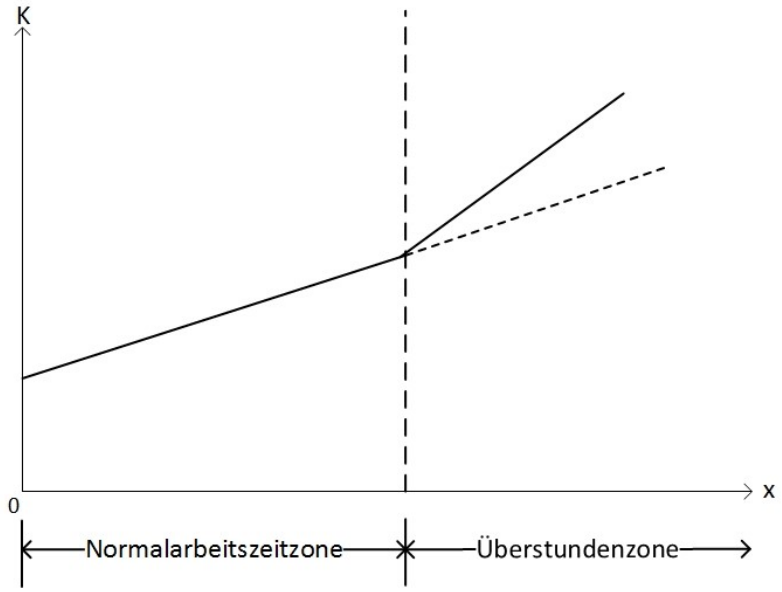


Abbildung 12: Lohnkostenverlauf bei zeitlicher Anpassung[70]

Die *qualitative Anpassung* umfasst u. a. die Änderung der Produktions- und Faktorenqualität (Schulung der Arbeitskräfte, Erneuerung von Maschinen), die Einführung anderer Verfahren und die Substitution von Faktoren, wie z. B. Mensch durch Betriebsmittel – Automatisierung von Abläufen, im Erzeugungsprozess. Je nach Betrachtungspunkt können Anpassungen unterschiedlich kategorisiert werden. So kann die intensitätsmäßige Anpassung einer Maschine zu einer qualitativen Anpassung eines Produkts führen.[71]

Diese Einführung in die produktionswirtschaftliche Theorie soll den Hintergrund für die Datenanalyse und die Motivation zur Durchführung einer solchen beleuchten. Erhöht sich allein die Intensität der Auslastung eines Betriebsmittels, das einen Engpass darstellt, können so die Gesamtdurchlaufzeit und Kosten reduziert werden. Im nächsten Kapitel soll nun die Methodik, mit der diese Ziele erreicht werden sollen, vorgestellt werden.

[69] Vgl. Blohm, H. et al. (2008), S. 82
[70] Quelle: ebda. (leicht modifiziert), S. 82
[71] Vgl. Swoboda, P. (1964), S. 49

3 Datenanalyse

Das in Abbildung 2 vorgestellte Konzept beruht auf den Grundlagen und Methoden der Datenanalyse. Daher ist das Ziel dieses Kapitels, ein grundlegendes Verständnis und die theoretische Basis zu schaffen, die zur Lösung der Aufgabenstellung aus Kapitel 1 und zur Beantwortung der Forschungsfrage benötigt wird.

Zu diesem Zweck werden eine allgemeine Einführung gegeben und einige Begriffsabgrenzungen vorgenommen. Es wird eine Einführung in den Begriff Data Science gegeben, der als Überbegriff für die Methoden und Ansätze im Umgang mit Daten und daraus zu gewinnenden Erkenntnisse im Kontext von Informatik, Mathematik und den einzelnen Anwendungsgebieten steht. Anhand des CRISP-DM soll aufgezeigt werden, wie ein solches Projekt strukturell aufgebaut werden kann. Dessen Unterteilung in Phasen ist verantwortlich für die Struktur dieses Kapitels. Es werden die Grundlagen aus der Statistik behandelt, bevor das Schaffen einer einheitlichen Datenbasis in der Theorie behandelt wird. Mit dem abschließenden Vorstellen der Methoden, die verwendet werden können, endet dieses Kapitel.

3.1 Einführung

Die Verknüpfung datengestützter Produktion und Logistik sind in der 4. industriellen Revolution auf eine hohe Vernetzung und zielorientierte Auswertung von Daten zurückzuführen.[72] Dies soll zu einer effektiveren Datenverarbeitung und -analyse führen, welche wiederum neue Optimierungspotenziale aufzeigen. Durch diese Fähigkeit erhält die Datenanalyse einen immer höheren Stellenwert in den Bereichen Produktion und Logistik.[73]

[72] Vgl. Bauernhansl, T. (2014), S. 14
[73] Vgl. Freitag, M. et al. (2015), S. 22 f.

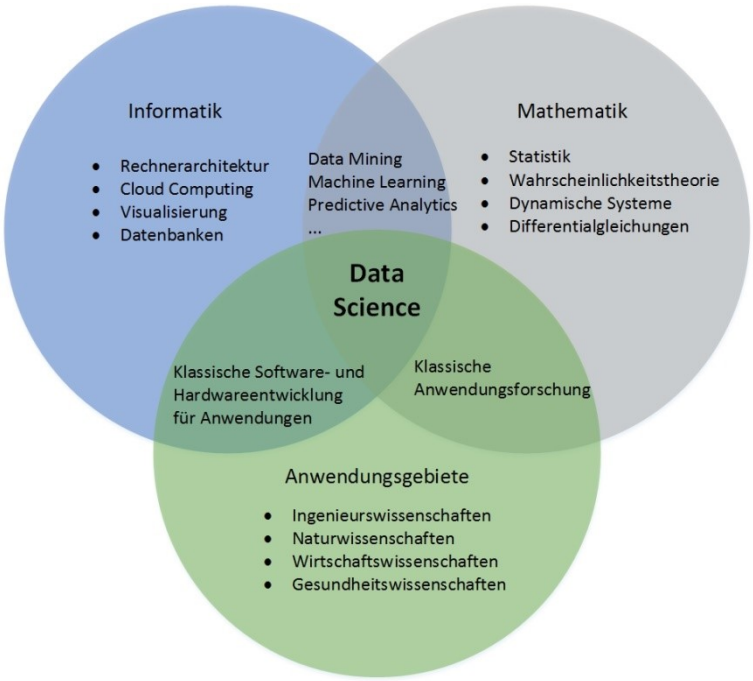


Abbildung 13: Komponenten von Data Science[74]

Der seit 40 Jahren existierende Begriff Data Science gibt der Kombination von Mathematik, Informatik und den jeweiligen Anwendungsgebieten einen Überbegriff.[75,76] Dies wird in Abbildung 13 dargestellt. Durch eben die Anwendungsgebiete ergeben sich neue Potenziale für die Datenanalyse, bestehende Ansätze zu verbessern. Es werden Ansätze verfolgt, wie Daten die Transparenz von Prozessen erhöhen und Entscheidungen zu deren Planung und Steuerung auf Basis dieser getroffen werden können.[77]

Es gibt eine Fülle an Theorien und Ansätzen in den Bereichen Mathematik und Informatik sowie in deren Zusammenspiel. Dazu gehören u. a. Data Mining, Statistik, explorative Datenanalyse, Predictive Analytics und Big Data. Unterschiedliches Vorgehen und Ziele kennzeichnen diese, dennoch sind die Übergänge oft fließend. So kann eine Methode unterschiedlichen Bereichen zugeordnet werden und ein anderes Ziel in diesem Kontext verfolgen. Dennoch sind diese Überbegriffe oftmals anders definiert.[78] So bezeichnen die Begriffe Data Mining[79,80], explorative Datenanalyse[81,82] und Predicitve Analytics[83] ein Konzept, ohne jedoch die in diesem Zusammenhang verwendeten Methoden zur Erreichung eines Ziels zu spezifizieren. Maschinelle Lernverfahren hingegen werden als eine Klasse von Verfahren beschrieben, die in Modell (Algorithmus), Verlustfunktion und Vorgehensweise zur Parameteroptimierung

[74] Quelle: in Anlehnung an Freitag, M. et al. (2015), S. 23 und Schutt, R.; O'Neil, C. (2013), S. 7
[75] Vgl. Freitag, M. et al. (2015), S. 22
[76] Vgl. Schutt, R.; O'Neil, C. (2013), S. 7
[77] Vgl. Freitag, M. et al. (2015), S. 23
[78] Vgl. ebda., S. 23
[79] Vgl. Witten, I. H. et al. (2011), S.4
[80] Vgl. Fayyad, U. et al. (1996), S. 37
[81] Vgl. Fahrmeir, L. et al. (2004), S. 12
[82] Vgl. Shmueli, G. (2010), S. 297
[83] Vgl. Abbott, D. (2014), S. 3

unterschieden werden. Eine Aneinanderreihung und Anwendung dieser Verfahren mit unterschiedlichen Ergebnissen wird als Data Mining Prozess bezeichnet.[84]

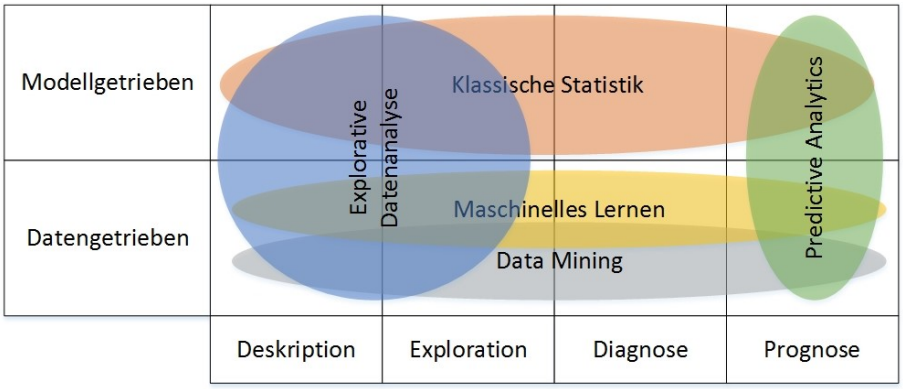


Abbildung 14: Qualitative Unterscheidung verschiedener Ansätze zur Datenanalyse[85]

Es gibt verschiedene Ansätze zur Klassifizierung der Methoden zur Datenanalyse. So definiert IBM über die Eigenschaften der Daten[86], wohingegen Abbildung 14 die Herangehensweise an die Problemstellung und das Ziel der Anwendung zur Unterscheidung heranzieht. Bei der Herangehensweise kann in modellgetrieben, Einsatz der klassischen Statistik, und datengetrieben, Data Mining und maschinelles Lernen, differenziert werden. Beim modellgetriebenen Ansatz liegt die Annahme einem stochastischen Modell zugrunde, wohingegen beim datengetrieben Ansatz der datengenerierende Prozess als unbekannt betrachtet und keine Annahme über ein stochastisches Modell getroffen wird.[87]

Ebenfalls in Abbildung 14 ist die Differenzierung durch das Ziel der Anwendung ersichtlich. Hierbei wird nach Datenanalyse zur Beschreibung (Deskription), Erforschung (Exploration), Erklärung (Diagnose) und Prognose der vorhandenen Daten unterschieden. Die Deskription beschreibt anhand vorliegender Daten historische Systemzustände und Ereignisse. Die Deskription ist eng verknüpft mit der Exploration, so dass diese häufig zusammen ausgeführt werden. Die Diagnose baut auf der Frage auf, warum etwas passiert ist, und die Prognose darauf, was passieren wird.[88]

Sämtliche in Abbildung 14 dargestellten Analyseziele können mit unterschiedlichen Methoden der klassischen Statistik, des Data Minings und des maschinellen Lernens erreicht werden. Daneben befasst sich die explorative Datenanalyse mit der Erforschung der Daten, um ein Verständnis für diese zu schaffen.[89] Predictive Analytics beschäftigt sich mit der Prognose der Daten, wobei dieser unterstellt wird, dass die modell- und datengetriebenen Ansätze zur Prognose ungesehener Daten, eingesetzt werden kann. Die Annahme, dass diese geeignete Ergebnisse liefert, ist oftmals falsch.

[84] Vgl. Alpaydin, E. (2010), S. 2 f.
[85] Quelle: Freitag, M. et al. (2015), S. 24 (leicht modifiziert)
[86] IBM Big Data & Analytic Hub, http://www.ibmbigdatahub.com/infographic/four-vs-big-data (Zugriff: 09.02.2017)
[87] Vgl. Breiman, L. (2001), S. 1
[88] Vgl. Fahrmeir, L. et al. (2004), S. 11
[89] Vgl. Freitag, M. et al. (2015), S. 24

Hierbei muss der erwartete Prognosefehler, der sich in die Komponenten zerlegen lässt, betrachtet werden.[90]

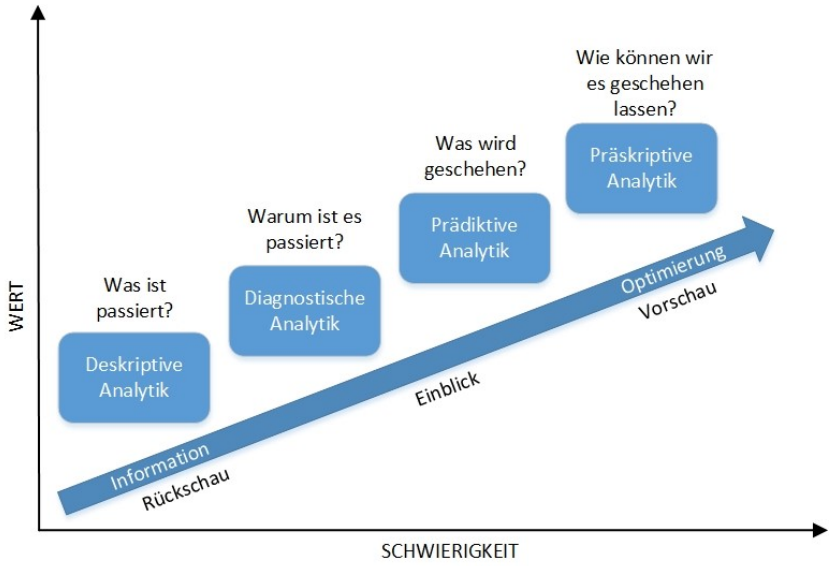


Abbildung 15: Die evolutionstechnischen Stufen der Analytik[91]

Neben den bereits genannten Ansätzen ergibt sich aus der Betrachtung der in Abbildung 15 dargestellten evolutionstechnischen Stufen der Analytik ein weiterer Ansatz. Bei der präskriptiven Analytik sind das Ziel sowohl eine Prognose als auch eine Handlungsanleitung, um Prognosen Realität werden zu lassen bzw. diese ggf. zu verändern.[92]

All diese Ansätze können auf die Daten, die in Produktions- und Logistiksystemen anfallen, angewendet werden. Durch die Wahl der geeigneten Methoden kann ein klares Verständnis des Systems und des OptimierungsPotenzials aufgezeigt werden. Dies soll mit den Beispielen in Tabelle 3 veranschaulicht werden. Die aufgezählten Aufgaben, Ausreißererkennung, Assoziation, Clustering (Segmentierung), Klassifikation, Prognose und Regression, sind typische Aufgaben der Datenanalyse. Jeweils einer Aufgabe werden verschiedenen Methoden zugewiesen, um das Ziel, im Anwendungsbeispiel dargestellt, zu erreichen.[93]

[90] Vgl. Hastie, T. et al. (2009), S. 38
[91] Quelle: in Anlehnung an Gartner, http://ww.gartner.com/it-glossary/predictive-analytics (Zugriff: 17.05.2015) zitiert nach Freitag, M. et al. (2015)
[92] Vgl. Freitag, M. et al. (2015), S. 25
[93] Vgl. ebda., S. 26

Tabelle 3: Beispiele für die Aufgaben der Datenanalyse[94]

Aufgabe	Beschreibung	Methodenbeispiele	Anwendungsbeispiele
Ausreißer-erkennung	Entdeckung auffälliger Objekte oder Werte	• Ausreißertests • Lineare Regression	• Filterung von Sensordaten • Identifikation von Kreditkarten-betrügern
Assoziation	Untersuchung der Zusammenhänge und durch Abhängigkeiten durch Wenn-dann-Regeln	• Assoziations-regeln • Bayessche Netze	• Warenkorbanalyse • Analyse von Kaufverhalten und Kundenbedürfnissen
Clustering/ Segmentierung	Bildung von a priori unbekannten Klassen aufgrund von Ähnlichkeiten	• Clusteranalyse • Neuronale Netze • Selbst-organisierende Karten	• Bildung verschiedener Liefergebiete • Bildung von Produktklassen ähnlicher Eigenschaften
Klassifikation	Zuordnung von Objekten durch Vergleiche von Objekteigenschaften mit den Eigenschaften vorgegebener Klassen	• Diskriminanz-analyse • Entscheidungs-bäume • Neuronale Netze	• Prozess- und Qualitätsanalyse • Zuordnung von neuen Produkten zu vorgegebenen Produktklassen
Prognose	Berechnung zukünftiger Werte auf Basis historischer Daten	• ARIMA • Exponentielle Glättung • Neuronale Netze • Prognose-verfahren der nichtlinearen Dynamik	• Bedarfsprognosen für die Produktionsplanung • Prognosen der Restlebensdauer von Produktions-anlagen
Regression	Modellierung der Beziehung zwischen verschiedenen Variablen	• Lineare Regression • Neuronale Netze • Support Vector Regression	• Bestimmung des Zusammenhangs zwischen Absatz und Marketing • Bestimmung des Zusammenhangs zwischen Maschineneigen-schaften und Lebensdauern

[94] Quelle: Freitag, M. et al. (2015), S. 26

Diese Einführung in Data Science und dessen Umfeld, in dem sich diese Arbeit bewegt, gibt erste Einblicke in die Möglichkeiten zur Lösung der Aufgabenstellung. Daran schließt das Kapitel 3.5 an und geht auf geeignete Modelle des maschinellen Lernens ein. Zunächst wird das CRISP-DM-Modell vorgestellt. Dieses gibt dem Prozess der Datenanalyse bis hin zur Modellanwendung einen standardisierten Rahmen.

3.2 Cross-Industry Standard Process for Data Mining

1996 entstand die Idee für ein nicht eigentumsrechtlich geschütztes, dokumentiertes, branchenunabhängiges und frei verfügbares Data Mining Modell. Dies wird mit dem *CRISP-DM (Cross-Industry Standard Process for Data Mining)* verwirklicht. Das umfassende Modell gliedert den Prozess des Data Minings in die sechs Phasen *Geschäftsmodell verstehen*, *Daten verstehen*, *Daten aufbereiten*, *Modellierung*, *Evaluation* und *Einsatz*. Diese und ihre Beziehungen zueinander sind in Abbildung 16 dargestellt.[95]

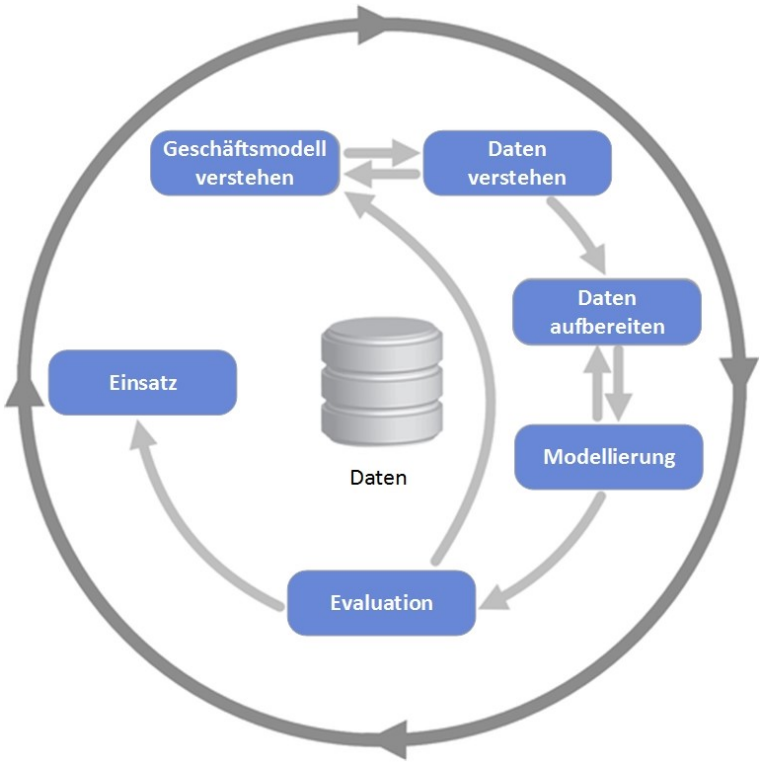


Abbildung 16: Die sechs Phasen des CRISP-DM[96]

In Abbildung 17 wird aufgelistet, welche Aufgaben bzw. Ergebnisse jede Phase beinhalten. So beschäftigt sich die erste Phase mit dem Umfeld und den generellen Zielen des Prozesses. Die zweite Phase schafft die Grundlage für spätere Phasen, indem die Daten gesammelt, beschrieben und ihre Qualität beurteilt wird. Um in der vierten Phase ein Modell erstellen zu können, müssen in der dritten Phase die Daten

[95] Vgl. Shearer, C. (2000), S. 13
[96] Quelle: in Anlehnung an ebda., S. 14

aufbereitet werden. Hierbei werden Daten gelöscht, Attribute abgeleitet und neuen Datensets erstellt. In der fünften Phase erfolgen die Evaluation der Ergebnisse und das daraus abgeleitete weitere Vorgehen. Das Modell schließt mit der Phase sechs, in die sich mit dem Einsatz der Erkenntnisse bzw. des Modells und der Sicherstellung der weiteren Gültigkeit der Ergebnisse befasst. [97]

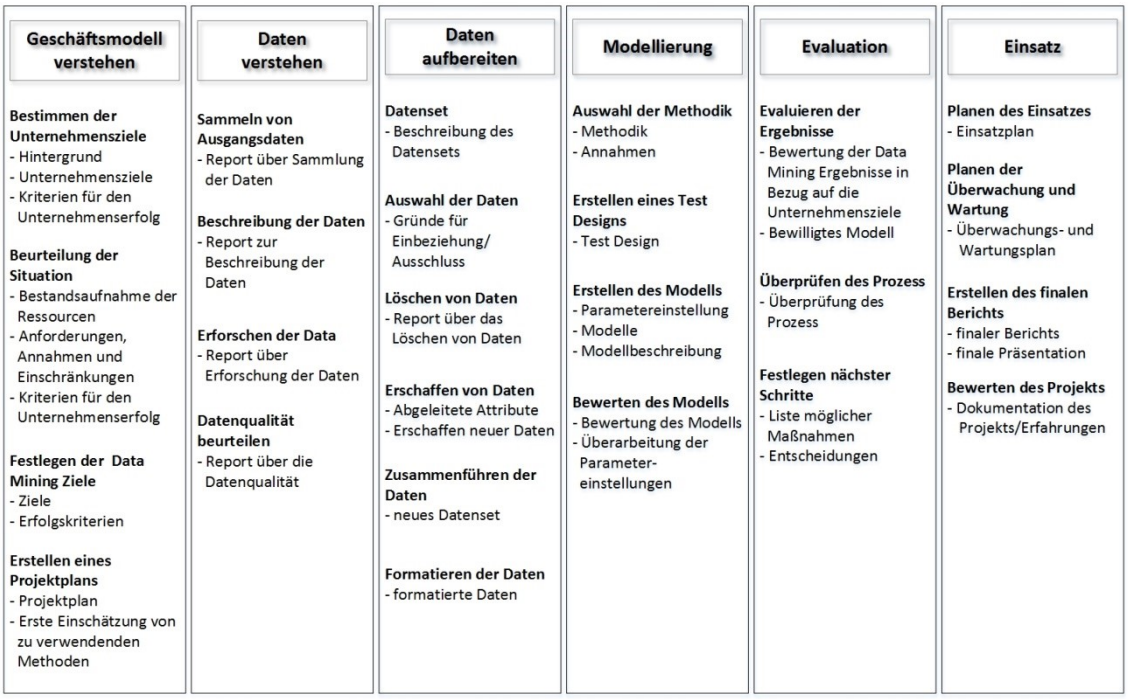


Abbildung 17: Aufgaben und Output des CRISP-DM Modells[98]

Das CRISP-DM Modell basiert auf dem *KDD*-Modell[99] *(Knowledge Discovery in Databases)*, das ebenso für das *SEMMA*-Modell[100] *(Sample, Explore, Modify, Model, Assess)* als Grundlage dient. Die beiden Modelle, CRIPS-DM und SEMMA, unterscheiden sich nur marginal. Ein Vergleich der einzelnen Phasen ist in Tabelle 4 zu finden. Alle drei dienen dem Zweck den Anwendern bei ihren Data Mining Projekten eine Struktur zu geben.[101]

Tabelle 4: Vergleich der Schritte von KDD, SEMMA und CRISP-DM[102]

KDD	SEMMA	CRISP-DM
Pre KDD	—	Geschäftsmodell verstehen
Selektion	Datenauswahl	Daten verstehen
Datenaufbereitung	Exploration	
Transformation	Modifizierung	Daten aufbereiten
Data Mining	Modellieren	Modellierung
Interpretation/Evaluation	Bewertung	Evaluation
Post KDD	—	Einsatz

[97] Vgl. Shearer, C. (2000), S. 13
[98] Quelle: in Anlehnung an ebda., S. 14
[99] Vgl. Fayyad, U. et al. (1996), S. 37
[100] Vgl. Fernandez, G. (2003), S. 5 f.
[101] Vgl. Azevedo, A.; Santos, M. F. (2008), S. 185
[102] Quelle: in Anlehnung an ebda., S. 185

CRISP-DM ist vollständig und dokumentiert. Sämtliche Stufen wurden nachvollziehbar strukturiert und definiert. Dies führt zu einem guten Verständnis und erleichtert eine spätere Überarbeitung.[103] Daher wurde entschieden, dass sich die folgenden Kapitel am Aufbau des CRISP-DM orientieren, und dieses als Referenz für das praktische Fallbeispiel einzusetzen. Als erstes werden die Grundlagen der Datenaufbereitung und ihre Methoden aufgezeigt.

3.3 Datenverständnis

Das Verständnis des Rohmaterials - der Daten - wird, wie in Kapitel 3.2 erklärt, in der Phase zwei "Daten verstehen" erzielt.

"Daten werden stets an gewissen Objekten beobachtet."[104] Diese Objekte werden als statistische Einheiten bezeichnet. An ihnen werden Merkmale beobachtet, die in weiterer Folge als Attribute bezeichnet werden, und deren konkreter Wert für eine bestimmte statistische Einheit Ausprägung genannt wird.[105] Eine inhaltlich zusammenhängende Menge von Ausprägungen der Attribute einer statistischen Einheit wird sowohl logisch als auch physisch zu Datensätzen zusammengefasst und kann einen Index zur eindeutigen Zuordnung enthalten.[106]

Im Folgenden werden mehrere Ansätze, Methoden und Kriterien vorgestellt, Daten zu analysieren und für die Aufbereitung wichtige Erkenntnisse zu gewinnen. Hierbei liegt der inhaltliche Fokus auf den verschiedenen Attributtypen, in denen Daten vorliegen können, und den Maßen der uni- und multivariaten Deskription. Hierbei handelt es sich um statistische Maße, die für ein Attribut berechnet werden können, sowie um Zusammenhangsmaße, die die Beziehung zwischen den Attributen beschreiben können. Mit Hilfe dieser können Schlüsse über die Verteilung und Zusammenhänge innerhalb der Daten erhalten werden.

3.3.1 Attributtypen

Ausprägungen von Attributen werden durch Messen bzw. Beobachten nach festgelegten Regeln einen Zahlenwert zugewiesen. Grundsätzlich gilt dabei, dass die Messbarkeit der jeweiligen Ausprägung eines Attributs von der Art der Eigenschaft abhängig ist. Das bedeutet, wie gut sie in Zahlen ausgedrückt werden kann. So ist der Durchmesser eines Werkstücks leicht messbar, allerdings die subjektive Wahrnehmung eines Menschen nicht. Um diesen Ausprägungen einen Rahmen zu geben, gibt es Skalen. Diese können durch ihre unterschiedlichen Niveaus, den Skalenniveaus, voneinander unterschieden werden und geben an, in welcher Art und Weise die Eigenschaften eines Objekts gemessen worden sind. [107]

[103] Vgl. Azevedo, C. S.; Santos, M. F. (2005) zitiert nach Azevedo, A.; Santos, M. F. (2008), S. 185
[104] Fahrmeir, L. et al. (2004), S. 14
[105] Vgl. ebda., S. 14
[106] Vgl. Mertens, P. et al. (2012), S. 39
[107] Vgl. Backhaus, K. et al. (2011), S. 10

In Tabelle 5 sind die Nominalskala, die Ordinalskala, die Intervallskala und die Ratioskala mit ihren Merkmalen sowie die Anwendbarkeit von Rechenoperationen auf die Daten angegeben. Es ist zu bedenken, dass die Daten immer eine Information darstellen. Somit können Ausprägungen, die unterschiedlich skaliert sind, nicht wahllos miteinander verglichen werden.[108]

Tabelle 5: Skalenniveau[109]

Skala		**Merkmale**	**Mögliche rechnerische Handhabung**
Nicht-metrische Skalen	*Nominalskala*	Klassifizierung qualitativer Eigenschaftsausprägungen	Bildung von Häufigkeiten
	Ordinalskala	Rangwert mit Ordinalskalen	Median, Quantile
Metrische Skalen	*Intervallskala*	Skala mit gleichgroßen Abschnitten ohne natürlichen Nullpunkt	Subtraktion, Mittelwerte
	Ratioskala	Skala mit gleichgroßen Abschnitten und natürlichem Nullpunkt	Summe, Division, Multiplikation

So stellt die *Nominalskala* die primitivste Grundlage des Messens da. Die Klassifizierung von qualitativen Eigenschaften, wie z. B. die Bestimmung der Farbe (rot – blau – grün – ...), kann anschließend in Zahlen übersetzt werden; dennoch sind keine arithmetischen Operationen, wie Addition, Subtraktion, Multiplikation oder Division, erlaubt. Einzig die Häufigkeit des Auftretens der einzelnen Ausprägungen, die Häufigkeit, ist zählbar.[110]

Die zweite nicht-metrische Skala ist die *Ordinalskala*. Bei ihr wird eine Rangordnung mit Hilfe von Rangwerten (d. h. ordinalen Zahlen) ermittelt. Ein Beispiel für eine Ordinalskala ist der Vergleich von Produkt A mit Produkt B, wobei Produkt A besser als Produkt B ist. Dieser Vergleich sagt nichts über die Abstände zwischen den einzelnen Objekten aus. Auch bei dieser Skala sind daher keine arithmetischen Operationen erlaubt. Neben der Häufigkeit sind der Median oder die Quantile erlaubte statistische Größen.[111]

Die *Intervallskala* stellt das nächsthöhere Messniveau dar. Gleichgroße definierte Skalenabschnitte, wie auf z. B. auf der Celsius-Skala zur Temperaturmessung, erlauben erstmals die Anwendung von arithmetischen Operationen (Addition und Subtraktion) sowie statistische Größen wie das arithmetische Mittel und die Standardabweichung. So kann der Temperaturunterschied zwischen zwei Messungen durch dessen Differenz berechnet werden, die wiederum einen eigenen Informationsgehalt hat.[112]

[108] Vgl. Backhaus, K. et al. (2011), S. 10
[109] Quelle: ebda., S. 12
[110] Vgl. ebda., S. 10
[111] Vgl. ebda., S. 11
[112] Vgl. ebda., S. 11

Das höchste Messniveau ist die *Ratio- (oder Verhältnis-)skala*. Der Unterschied zwischen ihr und der Intervallskala liegt in der Existenz eines natürlichen Nullpunkts, der für das betreffende Attribut die Aussage "nicht vorhanden" wiedergibt. Dies ist bei der oben als Beispiel aufgeführten Celsius-Skala nicht der Fall. Die Ratioskala kann daher auf physische Attribute, wie Länge und Gewicht, angewendet werden. Aus diesen so skalierten Daten können sämtliche arithmetische Operationen und die oben genannten statistischen Größen berechnet werden. Zusätzlich können das geometrische Mittel und der Variationskoeffizient berechnet werden.[113]

Grundsätzlich ist zu erkennen, dass je höher das Skalenniveau ist, desto höher ist der Informationsgehalt der betreffenden Daten, und desto mehr Rechenoperationen und statistische Größen können mit den Daten ausgeführt werden. Die Umwandlung von einem höheren auf ein niedrigeres Skalenniveau ist generell möglich. Sinnvoll ist das, wenn die Daten für die Analyse vereinfacht werden müssen. Allerdings geht diese Transformation mit einem Informationsverlust einher.[114]

Zusätzlich zur Einteilung in die Skalen können Attribute als *diskret* und *stetig* bezeichnet werden. Diese geben darüber Auskunft, ob Ausprägungen eines Attributs endlich viele Werte annehmen können (diskret) oder alle Werte eines Intervalls (stetig). Liegen diskrete Attribute in sehr feinen Abstufungen vor, werden diese als *quasistetig* bezeichnet.[115]

Eine Unterscheidung in *qualitativ* (oder *kategorial*) und *quantitativ* wird dann getroffen, wenn zwischen Qualität und Ausmaß getrennt wird. Bei qualitativen Attributen handelt es sich um Größen mit endlich vielen Ausprägungen, die höchstens ordinalskaliert sind. Quantitative Attribute kennzeichnen sich durch die Angabe der Ausprägung in deren Intensität oder Ausmaß. Ordinalskalierte Attribute haben in dieser Ordnung eine Zwitterstellung; sie besitzen sowohl einen schwach quantitativen als auch einen dominierenden qualitativen Anteil.[116]

Nach dieser Einführung, in welcher Form Merkmale skaliert sein können, beschäftigt sich das nächste Kapitel mit den statistischen Maßen, die von einem Attribut abhängig sind. Ziel ist es, zu vermitteln, wie sie berechnet werden können, welche Bedeutung sie haben und welche Erkenntnisse gewonnen werden können.

3.3.2 Univariate Deskription

Wird bei einer Erhebung vom Umfang n ausgegangen, bei der an n Einheiten die Werte $x_1, \ldots, x_n$ eines Attributes erfasst wurden, so werden diese als Urliste, Roh- oder Primärdaten bezeichnet. Bereits bei kleinem oder mittlerem Umfang kann die Darstellung der Rohdaten unübersichtlich werden. Um eine übersichtliche und zusammenfassende Darstellung zu ermöglichen, können verschiedene Methoden angewendet bzw. Werte errechnet werden.[117]

[113] Vgl. Backhaus, K. et al. (2011), S. 11
[114] Vgl. ebda., S. 11 f.
[115] Vgl. Fahrmeir, L. et al. (2004), S. 16 f.
[116] Vgl. ebda., S. 19
[117] Vgl. ebda., S. 31

In einem ersten Teil werden die Berechnung von Häufigkeit und die Darstellung von Verteilungen betrachtet. Diese kann auch durch Maßzahlen bzw. Parameter ausgedrückt werden.[118] Zu diesem Zweck wird eine kurze Einführung zu Lagemaßen sowie zu den Streuungsmaßen gegeben.

Häufigkeiten und Graphische Darstellungen

Die Urliste kann nach *Häufigkeiten* der Ausprägungen $\mathrm{a}_1, \mathrm{a}_2, \dots, \mathrm{a}_\mathrm{k}, \mathrm{k} \leq \mathrm{n}$ untersucht werden. Die Anwendung auf die Nominalskala bringt keine inhaltliche Bedeutung. Bei kategorialen Attributen ist k die Anzahl der Kategorien und zudem meist maßgeblich kleiner als n. Liegen metrischen Werte in der Urliste vor, ist k meist fast genauso groß wie n. Die relative Häufigkeit f_j ist das Verhältnis zwischen der absoluten Häufigkeit h_j zu n. Häufigkeiten und Ausprägungen zusammen werden als Häufigkeitsdaten bezeichnet.[119]

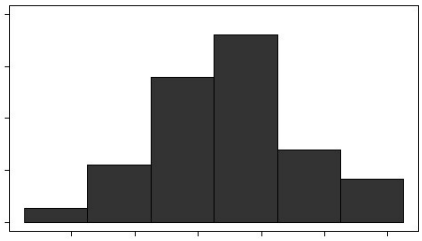


Abbildung 18: Beispiel eines Histogramms[120]

Häufigkeitsdaten können durch ein *Histogramm* visualisiert werden.[121] Zusätzlich können sie in Intervalle zusammengefasst werden. Dieser Vorgang wird gruppieren, auch klassieren oder kategorisieren, genannt.[122] Die dadurch entstandenen Klassen werden auf der Abszisse und die relative oder absolute Häufigkeit auf der Ordinate aufgetragen. Die entstandenen Blöcke sind proportional zu der absoluten oder der relativen Häufigkeit. Eine Darstellung, wie sie in Abbildung 18 zu sehen ist, ist insbesondere von der Klassenbreite bzw. von der Anzahl der Klassen abhängig. Wird die Klassenbreite größer, sinkt die Anzahl der Sprünge. Der Extremfall wäre eine maximale Breite, bei dem nur ein einziger Block dargestellt wird. Es gibt mehrere Ansätze die Klassenanzahl zu berechnen, so wie (3.1), (3.2) und (3.3). Zusätzlich soll der subjektive optische Eindruck des Histogramms berücksichtigt werden.[123]

$$\mathrm{k} = \sqrt{n} \tag{3.1}$$

$$\mathrm{k} = 2\sqrt{n} \tag{3.2}$$

$$\mathrm{k} = 10 \log n \tag{3.3}$$

[118] Vgl. Fahrmeir, L. et al. (2004), S. 53
[119] Vgl. ebda., S. 32
[120] Quelle: ebda., S. 43
[121] Vgl. ebda., S. 41
[122] Vgl. ebda., S. 17
[123] Vgl. ebda., S. 42 f.

Verteilungen können durch diese Darstellung in *unimodal*, *bimodal* und *multimodal* eingeteilt werden. Dies gibt an, ob ein, zwei oder mehrere Gipfel in einer Verteilung vorliegen. Abbildung 18 stellt eine unimodale Verteilung dar.[124]

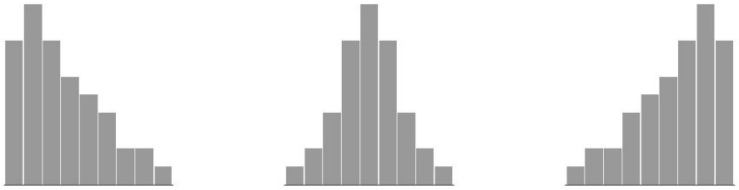


Abbildung 19: Eine linkssteile (a), symmetrische (b) und rechtssteile Verteilung (c)[125]

Sind Verteilungen wie in Abbildung 19 gegeben, wird von *Schiefe* bzw. *Symmetrie* gesprochen. Wenn die rechte und linke Hälfte einer Verteilung annähernd zueinander gleich sind, heißt diese *symmetrisch*. Bei empirischen Verteilungen ist eine exakte Symmetrie eher selten. Sind diese deutlich unsymmetrisch, werden sie als *schief* bezeichnet. Die Verteilung wird als *linkssteil* (oder *rechtsschief*) bezeichnet, wenn der mehrheitliche Teil der Messwert linksseitig konzentriert ist. Fällt die Verteilung nach rechts steiler und links flacher ab, wird diese *rechtssteil* bzw. *linksschief* genannt.[126]

Lagemaße

Lagemaße sind Maßzahlen zur Beschreibung des Schwerpunkts einer Verteilung. Sie werden in numerischen Werten angegeben. Es gibt verschiedene Lagemaße und die Auswahl, welche sinnvoll zur Beantwortung einer Fragestellung ist, hängt von Kontext, der Datensituation und dem Skalenniveau des Attributs ab.[127]

Das *arithmetische Mittel* $\bar{x}$ (3.4) ist das gebräuchlichste Maß für das Festlegen des Zentrums einer Verteilung.[128] Es wird aus der Urliste berechnet.[129]

$$\bar{x}_{arithm} = \frac{1}{n}(\mathrm{x}_1 + \mathrm{x}_2 + \cdots + \mathrm{x}_\mathrm{n}) = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} x_i \tag{3.4}$$

Das arithmetische Mittel wird von extremen Werten und Ausreißern in den Daten beeinflusst. Solche können entstehen, wenn eine empfindliche Einheit betrachtet wird, oder ein Fehler bei der Datenerhebung oder -aufbereitung unterlaufen ist.[130] Die Erkennung solcher Ausreißer wird im Kapitel 3.4.1 behandelt.

Es gibt *robuste* bzw. *resistente* Lagemaße, wie den *Median* (3.5), die den Einfluss von Extremwerten beschränkt. Der Median wird ermittelt, in dem die Urliste mit den Werten $\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{n}$ geordnet wird. In der dadurch enthaltenen geordneten Urliste $x_1 \leq \ldots \leq x_i \leq \ldots \leq x_n$ wird der Median in der Mitte platziert. Daraus folgt, dass mindestens 50 % der

[124] Vgl. Fahrmeir, L. et al. (2004), S. 46
[125] Vgl. ebda., S. 49
[126] Vgl. ebda., S. 48
[127] Vgl. ebda., S. 53
[128] Vgl. Haseloff, O. W.; Hoffmann, H. J. (1965), S.43
[129] Vgl. Fahrmeir, L. et al. (2004), S. 53
[130] Vgl. ebda., S. 55

Daten kleiner oder gleich und mindestens 50 % größer oder gleich x_{med} sind. Dies ist ein weiterer Vorteil des Median gegenüber dem arithmetischen Mittel.[131]

$$x_{med} = \begin{cases} x_{\left(n+\frac{1}{2}\right)} & \text{für } n \text{ ungerade} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2+1}\right)}\right) & \text{für } n \text{ gerade} \end{cases} \tag{3.5}$$

Ebenfalls zu den Lagemaßen wird der *Modus* x_{mod} gezählt. Dieser gibt Auskunft darüber, welche Ausprägung am häufigsten vorkommt. Der Modus ist eindeutig, wenn die Häufigkeitsverteilung ein eindeutiges Maximum aufweist.[132]

Handelt es sich um metrisch skalierte Attribute, kann eine Aussage über Symmetrie und Schiefe mithilfe des arithmetischen Mittels, Median und Modus getroffen werden. Dazu dienen die *Lageregeln*:[133]

- Symmetrische Verteilung: $\bar{x} \approx x_{med} \approx x_{mod}$
- Linkssteile Verteilung: $\bar{x} > x_{med} > x_{mod}$
- Linkssteile Verteilung: $\bar{x} < x_{med} < x_{mod}$

Die größte Bedeutung haben diese Lageregeln für unimodale Verteilungen. Je größer die Unterschiede zwischen $\bar{x}$, x_{med} und x_{mod} sind, desto schiefer sind die Verteilungen. Dieses Wissen kann auf den Umgang mit Situationen, in denen ein "Mittelwert" benötigt wird, Einfluss haben.[134]

Quantile, Standardabweichung und Varianz

Um eine vollständige Beschreibung von Verteilungen zu liefern, sollten die Lagemaße um Angaben zur Streuung ergänzt werden.[135] Hierzu gibt es mehrere Möglichkeiten. Innerhalb dieser Arbeit werden die Quantile, die Standardabweichung und die Varianz betrachtet.

Quantile und deren graphische Darstellung als Boxplots können zur Charakterisierung von Streuungen verwendet werden. Das p-Quantil teilt die Verteilung von Daten in zwei Teile, sodass circa $p * 100\,\%$ der Daten unterhalb und circa $(1-p) * 100\,\%$ oberhalb liegen. Zur Berechnung der Quantile wird die geordnete Urliste als Basis verwendet. Dies kann verallgemeinert werden: Ein p-Quantil wird durch den Wert x_p mit $0 < p < 1$ definiert. p stellt den Anteil der Daten dar, die kleiner/gleich x_p sind (3.6), und der Wert $1-p$ jenen, der größer/gleich x_p sind (3.7). $[np]$ stellt dabei die nächste zu np kleinere Zahl dar.[136]

$$\frac{Anzahl\ x - Werte\ \leq x_p}{n} \geq p \tag{3.6}$$

$$\frac{Anzahl\ x - Werte\ \geq x_p}{n} \geq 1 - p \tag{3.7}$$

[131] Vgl. Fahrmeir, L. et al. (2004), S. 55 f.
[132] Vgl. ebda., S. 57
[133] Vgl. ebda., S. 60
[134] Vgl. ebda., S. 60
[135] Vgl. ebda., S. 64
[136] Vgl. ebda., S. 64

Daher gilt für das p-Quantil (3.8), wenn np nicht ganzzahlig ist, und (3.9), wenn np ganzzahlig ist.[137]

$$x_p = x_{([np]+1)} \tag{3.8}$$

$$x_p \in [x_{(np)}, x_{(np+1)}] \tag{3.9}$$

Das 50 %-Quantil ist der Median. Als *Quartil* werden das *untere Quartil* Q_1 und *obere Quartil* Q_3 bezeichnet. Sie sind das 25 %-Quantil bzw. 75 %-Quantil. Eine weitere Definitionsmöglichkeit wäre, die die Datenhälfte, die über und unter dem Median liegen, wiederum zu halbieren.[138] Gemeinsam mit dem Median sind Quartile Indikatoren für eine schiefe Verteilung. Dies ergibt sich aus der Annahme, dass bei einer symmetrischen Verteilung der Abstand zwischen Median und den beiden Quartilen jeweils gleich groß ist. Ist dies nicht der Fall, weißt dies auf eine steilere Verteilung bei der kleineren Differenz hin. Ebenfalls ist an den Quartilen zu erkennen, wie weit eine Verteilung streut. Daraus lässt sich die Maßzahl der Streuung, der *Interquartilsabstand (IQR)* (3.10) ableiten.[139]

$$d_Q = x_{0,75} - x_{0,25} \tag{3.10}$$

Die Quartile werden durch die Lage der Daten, die kleiner als das untere Quartil und größer als das obere Quartil sind, nicht beeinflusst. Daher ist der Interquartilsabstand resistent gegen Ausreißer.[140] Dieser Umstand wird sich in der Ausreißeranalyse im Kapitel 3.4.1 zu Nutze gemacht.

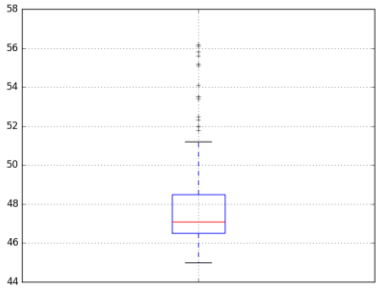


Abbildung 20: Beispiel eines Boxplots[141]

Graphisch werden die Quartile mit Boxplots, wie einer in Abbildung 20 beispielhaft zu sehen ist, dargestellt. Die rote Linie steht für den Median, das untere Ende der Box für das untere Quartil, die obere für das obere Quartil. Die Länge der Linien, die auch „whiskers“ genannt werden, können selbst definiert werden. Diese könnten den minimalen und maximalen Wert der Verteilung einschließen oder eine definierte Länge, wie z. B. $Q_1 - 1{,}5 \times IQR$, $Q_3 + 1{,}5 \times IQR$ haben.[142]

Neben dem IQR ist die Standardabweichung bzw. ihr Quadrat, die Varianz die bekannteste Maßzahl für die Streuung einer Verteilung. Es wird dabei die Streuung der

[137] Vgl. Fahrmeir, L. et al. (2004), S. 64
[138] Vgl. ebda., S. 65
[139] Vgl. ebda., S. 66
[140] Vgl. ebda., S. 66 f.
[141] Quelle: eigene Darstellung
[142] Vgl. Fahrmeir, L. et al. (2004), S. 67

Daten um ihr Mittel $\bar{x}$ gemessen und ist somit nur für metrische Attribute sinnvoll anwendbar. Die *Varianz* $\tilde{s}^2$ (3.11) ist das mittlere arithmetische Mittel der quadratischen Abweichung.[143]

$$\tilde{s}^2 = \frac{1}{n}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.11}$$

Zusätzlich kann die (*korrigierte) Stichprobenvarianz* (3.12) geschätzt werden. Dieser Schätzer gibt darüber Aufschluss, ob eine Stichprobe erwartungstreu, also unverzerrt, ist.[144,145]

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.12}$$

Sofern Varianzen nicht miteinander verglichen werden, sind sie auf Grund der Verwendung der Quadrate der Abweichung kein geeignetes Maß zur direkten Beurteilung der Streuung. Zu diesem Zweck wird die *Standardabweichung* $\tilde{s}$ (3.13) herangezogen.[146]

$$\tilde{s} = +\sqrt{\tilde{s}^2} = \frac{1}{x}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.13}$$

Maßzahlen für Schiefe und Wölbung

Neben den Merkmalen Lage und Streuung können Verteilungen auch in Bezug auf ihre Symmetrie bzw. Schiefe und die Wölbung unterschieden werden. Durch die Wölbung wird gezeigt, ob sich die Daten im Zentrum der Verteilung häufen, oder ob die Verteilung eher flach ist. Das bedeutet, dass die Daten um das Zentrum streuen.[147]

Neben der Ermittlung der Schiefe oder Symmetrie einer Verteilung durch die Lageregeln, gibt es die Möglichkeit für metrische Attribute den *Momentenkoeffizienten* g_m (3.14) zur Beurteilung der Schiefe zu berechnen.[148]

$$g_m = \frac{m_3}{\tilde{s}^4} \text{ mit } m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 \tag{3.14}$$

g_m ist maßstabsunabhängig und durch den Erhalt der positiven bzw. negativen Abweichungen vom Mittelwert kann nach folgenden Regeln die Schiefe, wie sie bereits in Abbildung 19 vorgestellt wurde, beurteilt werden:[149]

- $g_m = 0$ für symmetrische Verteilungen
- $g_m > 0$ für linkssteile Verteilungen
- $g_m < 0$ für rechtssteile Verteilungen

[143] Vgl. Haseloff, O. W.; Hoffmann, H. J. (1965), S. 53
[144] Vgl. Czado, C.; Schmidt, T. (2011), S. 5
[145] Vgl. ebda., S. 104
[146] Vgl. Haseloff, O. W.; Hoffmann, H. J. (1965), S. 54
[147] Vgl. Fahrmeir, L. et al. (2004), S. 74
[148] Vgl. ebda., S. 74 f.
[149] Vgl. ebda., S. 75

Wie stark oder wie schwach der zentrale Bereich bzw. die Randbereiche der Daten besetzt sind, kann durch das *Wölbungsmaß nach Fischer* (3.15) angezeigt werden. Bei spitzen Verteilungen sind die Randbereiche stärker besetzt als bei flachen Verteilungen. Die Normalverteilung dient als Vergleichsverteilung.[150]

$$\gamma = \frac{m_4}{\tilde{s}^4} \text{ mit } m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4 \tag{3.15}$$

Dabei gilt:[151]

- $\gamma = 0$ bei Normalverteilung
- $\gamma > 0$ bei spitzeren Verteilungen
- $\gamma < 0$ bei flacheren Verteilungen

Dichtekurven und Normalverteilung

Bei stetigen Merkmalen kommt es vor, dass Histogramme Informationen unterdrücken. Dies geschieht u. a. durch die Wahl der Klassenbreite. Diese empirische Verteilung verhindert ein Erkennen von Besonderheiten in der Verteilung. Daher kann das Histogramm durch eine glatte Kurve, die sogenannte *Dichtekurve*, approximiert werden.[152]

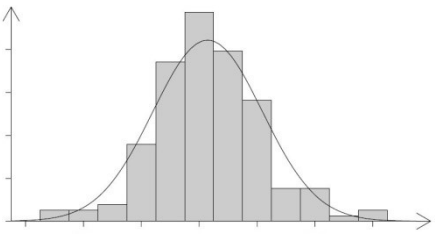


Abbildung 21: Beispiel für eine Dichtefunktion in einem Histogramm[153]

In Abbildung 21 ist ersichtlich, dass diese stetige Kurve einige Details glättet. Ihren Zweck, eine kompakte optische Visualisierung zu liefern, erfüllt sie und darf als Dichtekurve bezeichnet werden, wenn sie die Anforderung (3.16) erfüllt. Also $f(x) \geq 0$ und die Gesamtfläche unter $f(x)$ gleich 1 ist.[154]

$$\int f(x)\,\mathrm{d}x = 1 \tag{3.16}$$

Bei Dichtekurven kann wieder in symmetrische und schiefe sowie in unimodale und multimodale Dichtefunktionen unterschieden werden. Genauso können Maßzahlen der Häufigkeitsverteilung, wie Median und Quantil, auf die Dichtefunktion übertragen werden. Hierbei steht die Fläche anstatt der Anzahl der jeweiligen Werte im Vordergrund. So definiert das p-Quantil x_p für $0 < p < 1$ jenen Wert, der die Gesamtfläche unter $f(x)$ in die beiden Flächen $p * 100\,\%$ und Flächen $(1-p) * 100\,\%$ teilt. Somit ist der Median $x_{0,5}$ jener Wert, der die Gesamtfläche in zwei gleich große

[150] Vgl. Fahrmeir, L. et al. (2004), S. 76
[151] Vgl. ebda., S. 76
[152] Vgl. ebda., S. 87
[153] Quelle: ebda., S. 87 (leicht modifiziert)
[154] Vgl. ebda., S. 87 f.

Hälften aufteilt. Handelt es sich um eine symmetrische, unimodale Dichtekurve ist der Median die Mitte und repräsentiert gleichzeitig den Modus und den Mittelwert der Dichtefunktion $f(x)$ bzw. des Attributes X.[155]

Eine wichtige Klasse der Dichtefunktionen sind die *Normalverteilungen* oder auch *Gauss-Verteilungen*, die symmetrisch, unimodal und glockenförmig sind. Die in Abbildung 21 dargestellte Dichtefunktion gehört zu dieser Klasse. Sie werden durch (3.17) definiert, wobei μ der Mittelwert und σ die Standardabweichung ist. Für gegebene Werte dieser beiden Parameter und für die $\mu \in \mathbb{R}$ und $\sigma > 0$ gilt, ist $f(x|\mu,\sigma)$ eindeutig.[156]

$$\int f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{3.17}$$

Die Exponentialfunktion ist verantwortlich für die Form von $f(x)$. Das Maximum der Dichtefunktion ist mit $x = \mu$ erreicht. Links und rechts davon fällt sie symmetrisch und glockenförmig gegen Null ab. Je kleiner der Wert σ ist, desto steiler fällt diese Kurve ab und umso enger ist die Kurve um den Mittelwert μ. Im Gegensatz dazu ist die Kurve umso flacher, je größer σ ist. Dieser Zusammenhang wird in Abbildung 22 veranschaulicht.[157]

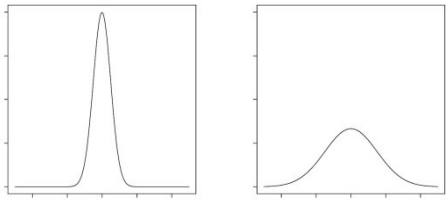


Abbildung 22: Zwei Normalverteilungsdichtekurven mit kleinem und großem σ[158]

Der erste Faktor von (3.17) wird als Normalisierungskonstante bezeichnet. Ihr Zweck ist die Gesamtfläche unter $f(x)$ auf 1 zu setzen und sie hat keine Auswirkung auf die Form der Kurve.[159]

Der bisher verwendete Mittelwert μ und die Standardabweichung σ entsprechen dem arithmetischen Mittel $\bar{x}$ und der Standardabweichung $\tilde{s}$ von Beobachtungen $x_1, \ldots, x_n$ eines Attributes X. Durch die Variation von μ und σ können verschiedene Normalverteilungen entstehen. Alle diese können auf die sogenannte *Standardnormalverteilung*, eine Art der Normalverteilung, zurückgeführt werden.[160]

Die Standardnormalverteilung ist als eine Normalverteilung mit Dichtefunktion $f(x)$, Mittelwert μ und Standardabweichung σ definiert, deren Attribut Z, wie in (3.18) beschrieben, standardisiert wurde.[161]

$$Z = \frac{X-\mu}{\sigma} \tag{3.18}$$

[155] Vgl. Fahrmeir, L. et al. (2004), S. 88 f.
[156] Vgl. ebda., S. 90
[157] Vgl. ebda., S. 90 f.
[158] Quelle: ebda., S. 92
[159] Vgl. ebda., S. 91
[160] Vgl. ebda., S. 92
[161] Vgl. ebda., S. 92

Durch die Standardisierung entsteht eine Dichtekurve einer Normalverteilung mit $\mu = 0$ und $\sigma = 1$. Das Attribut Z standardnormalverteilt. Die Dichtekurve wird mit $\phi(z)$ (3.19) bezeichnet.[162]

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \tag{3.19}$$

Durch die Standardisierung eines Attributs werden Beobachtungen als Abweichungen vom Mittel μ mit der Standardabweichung σ als Maßeinheit angesehen. Dadurch können Berechnungen, wie etwa die der Quantile, auf die Standardnormalverteilung zurückgeführt werden. Diese Werte sind in Tabellen gesammelt und somit entfällt die Berechnung.[163]

Kolgomorov-Smirnov-Test

Da manche Methoden nur anwendbar sind, wenn Attribute bzw. deren Werte normalverteilt sind, ist es notwendig, diese auf Normalverteilung zu testen.[164] Diese Tests sind ein wichtiges Instrument der quantitativen Analyse.[165] Ziel einer solchen Untersuchung ist es, die Frage zu klären, ob eine tatsächliche Verteilung mit einer vorgegebenen Verteilung übereinstimmt.[166] Diese Tests werden daher auch Anpassungs- oder Goodness-of-fit-Tests genannt.[167] Es kann daraus die folgende Nullhypothese (3.20) abgeleitet werden.[168]

$$H_0 : X \sim P_\theta \tag{3.20}$$

P_θ steht für die gegebene Verteilung, die durch θ näher erklärt wird. θ kann als Parameter(-vektor) fixiert sein oder nicht. Würde eine Normalverteilung vorliegen, bestünde θ aus den Werten μ und σ.[169]

Einer dieser Tests ist der *Kolgomorov-Smirnov-Test*. Er dient zur Überprüfung, ob die tatsächliche Verteilung mit der einer stetigen Prüfungsverteilung, wie der Normalverteilung, übereinstimmt. Er basiert auf der empirischen Verteilungsfunktion (3.21).[170]

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \tag{3.21}$$

$$F_n(x) = P(X \leq x) \tag{3.22}$$

Mit ihr (3.21) wird jener Teil der Ausprägungen vom Attribut X angegeben, die kleiner oder gleich x_i sind. Dieser wird anschließend der theoretischen Verteilungsfunktion, welche die Prüfungsverteilung (3.22) repräsentiert, gegenübergestellt. Letztere ist die Wahrscheinlichkeit, einen Wert kleiner (oder gleich) x zu erlangen, im Falle, dass die

[162] Vgl. Fahrmeir, L. et al. (2004), S. 92
[163] Vgl. ebda., S. 93
[164] Vgl. Gertheiss, J.; Tutz, G. (2009), S. 448
[165] Vgl. ebda., S. 439
[166] Vgl. ebda., S. 448
[167] Vgl. Fahrmeir, L. et al. (2004), S. 445
[168] Vgl. Gertheiss, J.; Tutz, G. (2009), S. 448
[169] Vgl. ebda., S. 448
[170] Vgl. ebda., S. 450

Zufallsgröße X der zu überprüfenden Verteilung zugrunde liegt. Hierbei ist die Prüfgröße der maximale Abstand D (3.23) von theoretischer und empirischer Verteilungsfunktion.[171]

$$D = \sup_x |F_n(x) - F(x)| \tag{3.23}$$

Ist diese Distanz D hinreichend groß, wird die Nullhypothese H_0 verworfen.[172] Bei der Anwendung der Anpassungstests und auf Hinblick auf die Nullhypothese (3.20) ist zu beachten, dass dieser nur überprüft, ob signifikante Abweichungen von der theoretischen Verteilungsannahme existieren. Das Ergebnis trifft die Aussage, ob etwas dagegen spricht, dass die beobachteten Werte der angenommenen Verteilung zugrunde liegen.[173] Auch ist der Test unabhängig von der jeweiligen zu testenden Verteilungsfamilie.[174]

3.3.3 Multivariate Deskription

Neben der Möglichkeit ein Attribut zu analysieren, werden oftmals mehrere Attribute und deren Zusammenhänge beschrieben bzw. aufgedeckt.[175] Daher werden in diesem Abschnitt der Korrelationskoeffizient als Zusammenhangsmaß bei metrischen Attributen und die Möglichkeit der Darstellung in einem Streudiagramm vorgestellt.

Streudiagramm

Das *Streudiagramm* stellt die einfachste Darstellung der gemeinsamen Messwerte $(x_i, y_i),\ i = 1, \ldots, n$ zweier stetiger Attribute dar. Dabei werden die Messwerte in ein $x - y$-Koordinatensystem eingetragen.[176]

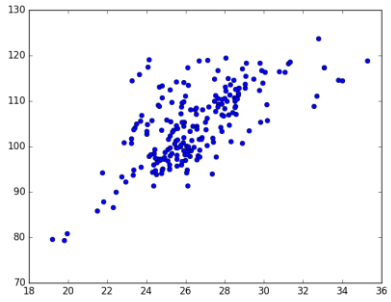


Abbildung 23: Beispiel eines Streudiagramms[177]

Aus einem wie in Abbildung 23 dargestellten Streudiagramm kann ein erster Eindruck über die Zusammenhänge von zwei Attributen gewonnen werden. So ist aus diesem ersichtlich, dass sobald die Werte des einen Attributes ansteigen, die des anderen ebenso ansteigen.

---

[171] Vgl. Gertheiss, J.; Tutz, G. (2009), S. 450
[172] Vgl. Wilcox, R. (2005), S. 1
[173] Vgl. Gertheiss, J.; Tutz, G. (2009), S. 449
[174] Vgl. ebda., S. 450
[175] Vgl. Fahrmeir, L. et al. (2004), S. 109
[176] Vgl. ebda., S. 128
[177] Quelle: eigene Darstellung

Zusammenhangsmaße bei metrischen Attributen

Das Streudiagramm ist eine Hilfe zur Visualisierung der Anordnung von Beobachtungspunkten. Die aus Abbildung 23 a Tendenz zwischen den Werten des jeweiligen Attributs, lässt vermuten, dass ein Zusammenhang besteht. Diese Stärke kann durch einen *empirischen Korrelationskoeffizient* ausgedrückt werden. Hierbei handelt es sich um ein Maß für die Stärke des Zusammenhangs zwischen zwei Attributen. Es gibt verschiedene Methoden, um diesen zu berechnen.[178]

Einer dieser Korrelationskoeffizienten ist der *Bravais-Pearson-Korrelationskoeffizient* r. Er misst die Stärke des linearen Zusammenhangs und wird aus den Daten (x_i, y_i), $i = 1, \dots, n$ nach (3.24) berechnet.[179]

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})((y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.24}$$

Als Beispiele für unterschiedliche Zusammenhänge sind verschiedene Attributpaare als Punktewolken mit unterschiedlichen Korrelationskoeffizienten in Abbildung 24 visualisiert. Der Korrelationskoeffizient kann Werte im Wertebereich von $-1 \leq r \leq 1$ annehmen, welche folgende Bedeutung haben:[180]

- $r > 0$: positive Korrelation, gleichsinniger linearer Zusammenhang, Tendenz: Werte (x_i, y_i) um eine Gerade positiver Steigung liegend
- $r < 0$: negativer Korrelation, gegensinniger linearer Zusammenhang, Tendenz: Werte (x_i, y_i) um eine Gerade negativer Steigung liegend
- $r = 0$: keine Korrelation, unkorreliert, kein linearer Zusammenhang

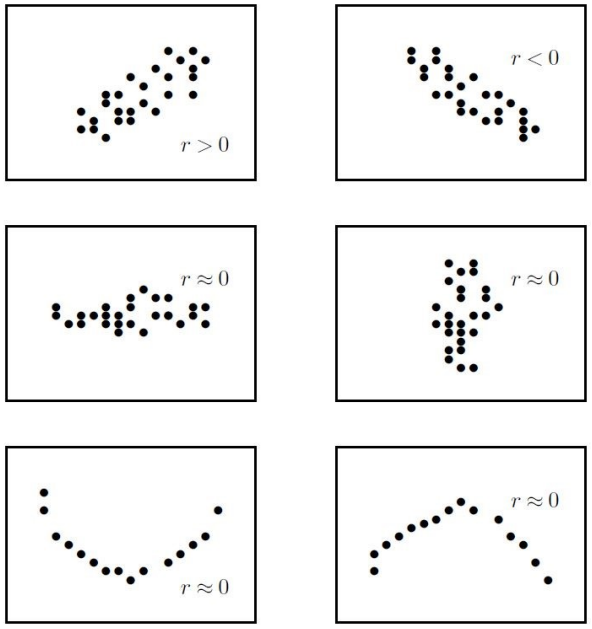


Abbildung 24: Punktekonfiguration und Korrelationskoeffizienten[181]

[178] Vgl. Fahrmeir, L. et al. (2004), S. 135 ff.
[179] Vgl. ebda., S. 139
[180] Vgl. ebda., S. 139
[181] Quelle: ebda., S. 138

Der Bravais-Pearson-Korrelationskoeffizient kann nur bei metrischen Attributen angewendet werden. Sofern nichtmetrische Attribute in nur zwei Ausprägungen, sogenannte dichotome oder binäre Attribute, vorliegen, können diese durch $X, Y \in [0,1]$ ausgedrückt werden.[182]

Um den Korrelationskoeffizienten von metrisch und ordinalskalierten Attributen zu berechnen, kann der *Spearmans Korrelationskoeffizient* herangezogen werden.[183] Bei der Berechnung dieses wird nicht auf die bisher verwendeten Wertepaare (x_i, y_i) zurückgegriffen sondern auf deren Ränge. Dabei werden jeweils unabhängig voneinander jedem x-Wert aus $x_1, \dots, x_n$ und jedem y-Wert aus $y_1, \dots, y_n$ der Rang als Platzzahl zugeordnet, die der jeweilige Wert bei einer Ordnung nach der Größe erhält. Damit ergeben sich aus den ursprünglichen Messpaaren die neuen Rangdaten $rg(x_i), rg(y_i), i = 1, \dots, n$. Treten innerhalb der Werte identische Werte auf, kann eine Rangvergabe nicht eindeutig sein. Daher werden Durchschnittsränge verwendet. Das bedeutet, dass den identischen Messwerten, die als Bindungen oder Ties bezeichnet werden, der Durchschnitt der in Frage kommenden Ränge als Rang zugewiesen wird. Mit diesen Grundlagen wird Spearmans Korrelationskoeffizient nach (3.25) berechnet.[184]

$$r = \frac{\sum_{i=1}^{n}(rg(x_i) - \overline{rg}_X)(rg(y_i) - \overline{rg}_Y)}{\sqrt{\sum_{i=1}^{n}(rg(x_i) - \overline{rg}_X)^2 \sum_{i=1}^{n}(rg(y_i) - \overline{rg}_Y)^2}} \tag{3.25}$$

Der Korrelationskoeffizient kann wieder den Wertebereich $-1 \leq r_{SP} \leq 1$ annehmen. Die Bedeutung ist im Folgenden zusammengefasst, wobei zu beachten ist, dass der Spearmans Korrelationskoeffizient im Gegensatz zum Bravais-Pearson-Korrelationskoeffizient den monotonen Zusammenhang betrachtet. Dies ist in der Abbildung 25 dargestellt, wobei hier die Extremfälle mit $r_{SP} = 1$ (oben) und mit $r_{SP} = -1$ (unten) zu sehen sind.[185]

- $r_{SP} > 0$: gleichsinniger monotoner Zusammenhang,
  Tendenz: x groß $\Leftrightarrow$ y groß, x klein $\Leftrightarrow$ y klein
- $r_{SP} < 0$: gegensinniger monotoner Zusammenhang,
  Tendenz: x groß $\Leftrightarrow$ y klein, x klein $\Leftrightarrow$ y groß
- $r_{SP} \approx 0$: kein monotoner Zusammenhang

[182] Vgl. Fahrmeir, L. et al. (2004), S. 140
[183] Vgl. ebda., S. 144
[184] Vgl. ebda., S. 142
[185] Vgl. ebda., S. 143 f.

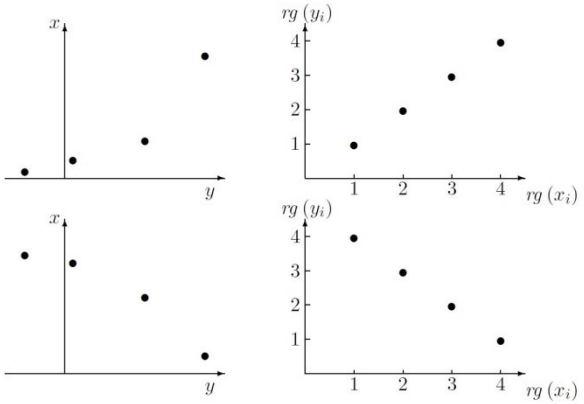


Abbildung 25: Extremfälle für Spearmans Korrelationskoeffizienten[186]

Durch die beschriebenen Korrelationskoeffizienten wird die Richtung der Wirkung, sofern vorhanden, nicht ausgedrückt. Die *Korrelation* ist nur ein Maß für die Stärke des Zusammenhangs zwischen zwei Attributen X, Y.

Auf *Kausalzusammenhänge* kann daher niemals allein durch statistische Analyse bzw. hohe Werte ein Zusammenhangsmaß geschlossen werden. Zu diesem Zweck sind sachlogische Überlegungen einzubeziehen.[187] Dabei kann es passieren, dass wesentliche Merkmale übersehen werden. Das kann zum Auftreten von Scheinkorrelationen aber auch zu verdeckten Korrelationen führen. Eine Scheinkorrelation ist eine solche, die eine hohe Korrelation zwischen zwei Attributen angibt, aber inhaltlich nicht vertretbar ist. Bei einer verdeckten Korrelation kann durch das Auslassen eines Merkmals eine tatsächlich vorhandene Korrelation verschleiert oder hinsichtlich des Vorzeichens umgekehrt werden.[188]

Mit dem Wissen aus diesem Kapitel werden im nachfolgenden die Daten aufbereitet. Die vorgestellten statistischen Maße bilden die Grundlage für einige der Methoden.

[186] Quelle: Fahrmeir, L. et al. (2004), S. 144
[187] Vgl. ebda., S. 148
[188] Vgl. ebda., S. 149 ff.

3.4 Datenaufbereitung

Die Aufbereitung stellt die dritte Phase des CRISP-DM dar. Die Grundlage sind oftmals Daten, die keinerlei Qualitätskontrolle unterliegen, und die aus diesem Grund vor der Verwendung einer Analyse unterzogen und hinsichtlich des Vorhabens aufbereitet werden sollten. Dazu werden im Folgenden die Identifikation von Ausreißern, die Analyse von fehlenden Daten, die Gruppierung und die Dimensionalitätsreduktion thematisiert.

3.4.1 Ausreißeranalyse

Ausreißer sind eine einzigartige Kombination von Charakteristiken, die sich von anderen Beobachtungen klar abgrenzen. Die Auswirkungen, die diese auf die Analyse haben könnten, können von verschiedenen Standpunkten aus betrachtet werden. So stellt sich die Frage, ob die Grundgesamtheit der Daten als solche die Situation vollständig repräsentiert, oder ob es sich dabei schon um einen Ausreißer als solchen handelt. Ebenso ist zu hinterfragen, ob ein einzelner Wert z. B. die Lagemaße so beeinflussen kann, dass sie den Großteil der Daten nicht korrekt wiedergibt.[189]

Ausreißer können somit nicht als „gut“ oder „schlecht“ kategorisiert werden. Vielmehr sollten sie im Kontext der Analyse betrachtet werden und welche Art von Informationen sie repräsentieren. So können Ausreißer, obwohl sie sich vom Großteil der Grundgesamtheit unterscheiden, vorteilhaft für die Analyse sein. Sie können bestimmte Zusammenhänge repräsentieren, die sonst nicht erkannt worden wären. Im Gegensatz dazu stellen sie nicht die Grundgesamtheit dar, können aber statistische Tests verfälschen und das Erreichen das Analyseziels verhindern.[190]

Das Auftreten von Ausreißern kann in vier Klassen eingeteilt werden. Als Merkmal zur Einteilung dieser dient der Grund für ihr Auftreten. Hierbei kann in Verfahrensfehler, außergewöhnliche Ereignisse, außergewöhnliche Beobachtungen und Einzigartigkeit in ihrer Kombination unterschieden werden. Verfahrensfehler treten auf, wenn Daten falsch erfasst wurden. Solche Fehler sollten bereits bei der Aufbereitung der Daten erkannt und eliminiert bzw. korrigiert werden. Außergewöhnliche Ereignisse sind einzigartig und können, falls es dem Ziel der Analyse dient, in der Gesamtmenge belassen werden. Wenn dies nicht der Fall ist, sollte der Datensatz gelöscht werden. Bei außergewöhnlichen Beobachtungen liegt keine Erklärung vor, und es muss individuell entschieden werden, ob diese in der Gesamtmenge belassen werden oder nicht. Datensätze, die sich durch ihre einzigartige Kombination von Werten als Ausreißer repräsentieren, sollten in der Gesamtmenge behalten werden.[191]

Zum Entdecken von Ausreißern gibt es verschiedene Methoden, wie den Ausreißertest nach Grubbs[192] oder den David-Hartley-Pearson-Test[193], die die Normalverteilung der Daten voraussetzen. Eine weitere Möglichkeit ist die graphische Auswertung mit

[189] Vgl. Hair, J. F. et al. (2014), S. 62 f.
[190] Vgl. ebda., S. 63
[191] Vgl. ebda., S. 63
[192] Vgl. Grubbs, F. E. (1950), S. 27
[193] Vgl. David, H. A. et al. (1954), S. 482

Boxplots.[194] Wie bereits in Kapitel 3.3.2 besprochen, kann die Länge der Whisker selbst gewählt werden. Tukey (1977)[195] schlägt die Whiskerlänge von $Q_1 - 1{,}5 \times \mathrm{IQR}, Q_3 + 1{,}5 \times \mathrm{IQR}$ vor, wobei diese dann bei dem letzten innerhalb dieser Grenze gelegenen Wert enden. Dies kann zu unterschiedlich langen Whiskers führen. Alle außerhalb dieser gelegenen Werte werden als „extreme“ Ausreißer identifiziert werden. Werte die außerhalb der Quantile - der Box - liegen, werden als „milde“ Ausreißer bezeichnet.[196]

Ausreißer können nur dann erkannt werden, wenn genügend Daten vorhanden sind. Das Fehlen von Daten und die Möglichkeiten zur Ergänzung werden im nachfolgenden Kapitel betrachtet.

3.4.2 Analyse der Fehlenden Daten

Bei der Exploration der Daten mit denen in Kapitel 3.3 vorgestellten Methoden, waren fehlende Daten kein Grund für die Nichtanwendung. Oftmals ist es aber erforderlich, dass die Daten vollständig ohne Lücken vorliegen. Würden Datensätze aufgrund von fehlenden Daten ausgeschlossen werden, könnte dies die Anzahl der Datensätze so verringern, dass keine Analyse mit diesen Daten als Grundlage durchgeführt werden könnte. Zusätzlich besteht das Problem der Verzerrung der Daten bzw. Maßzahlen der Attribute, sollten die Daten absichtlich oder unabsichtlich verfälscht worden sein. Es sollten daher die Fragen beantwortet werden, ob Daten zufällig oder nach einem bestimmten Muster fehlen.[197]

Fehlende Daten können bei Analysen und vor allem bei multivariaten Analysen eine signifikante Auswirkung auf die Ergebnisse haben. Die zu betrachtenden Beziehungen zwischen den Daten werden komplexer, und die Wahrscheinlichkeit fehlende Daten zu übersehen, wird größer. Daher ist es von Vorteil fehlende Daten vorab zu finden und zu überprüfen. Zu diesem Zweck hat Hair et al. (2014)[198] einen vier Schritte umfassenden Prozess zum Identifizieren und Bereinigen von fehlenden Daten entworfen und wird im Folgenden vorgestellt. Der englische Originaltitel „A Four-Steps Process for Identifying Missing Data and Applying Remedies“ lässt Spielraum bei der Übersetzung des Wort „data“ zu. Im Deutschen wurde daher das Wort „Daten“ gewählt, wobei im späteren praktischen Teil der Prozess für fehlende Werte angewendet wird.

Die folgenden Ausführungen dieses Prozesses wurden, wenn nicht anders angegeben, von Hair et al. (2014)[199] übernommen und in Abbildung 26 zusammengefasst.

[194] Vgl. Fahrmeir, L. et al. (2004), S. 67
[195] Vgl. Tukey, J. W. (1977), S. 44 ff.
[196] Vgl. ebda., S. 44 ff.
[197] Vgl. Hair, J. F. et al. (2014), S. 40
[198] Vgl. ebda., S. 42
[199] Vgl. ebda., S. 42 ff.

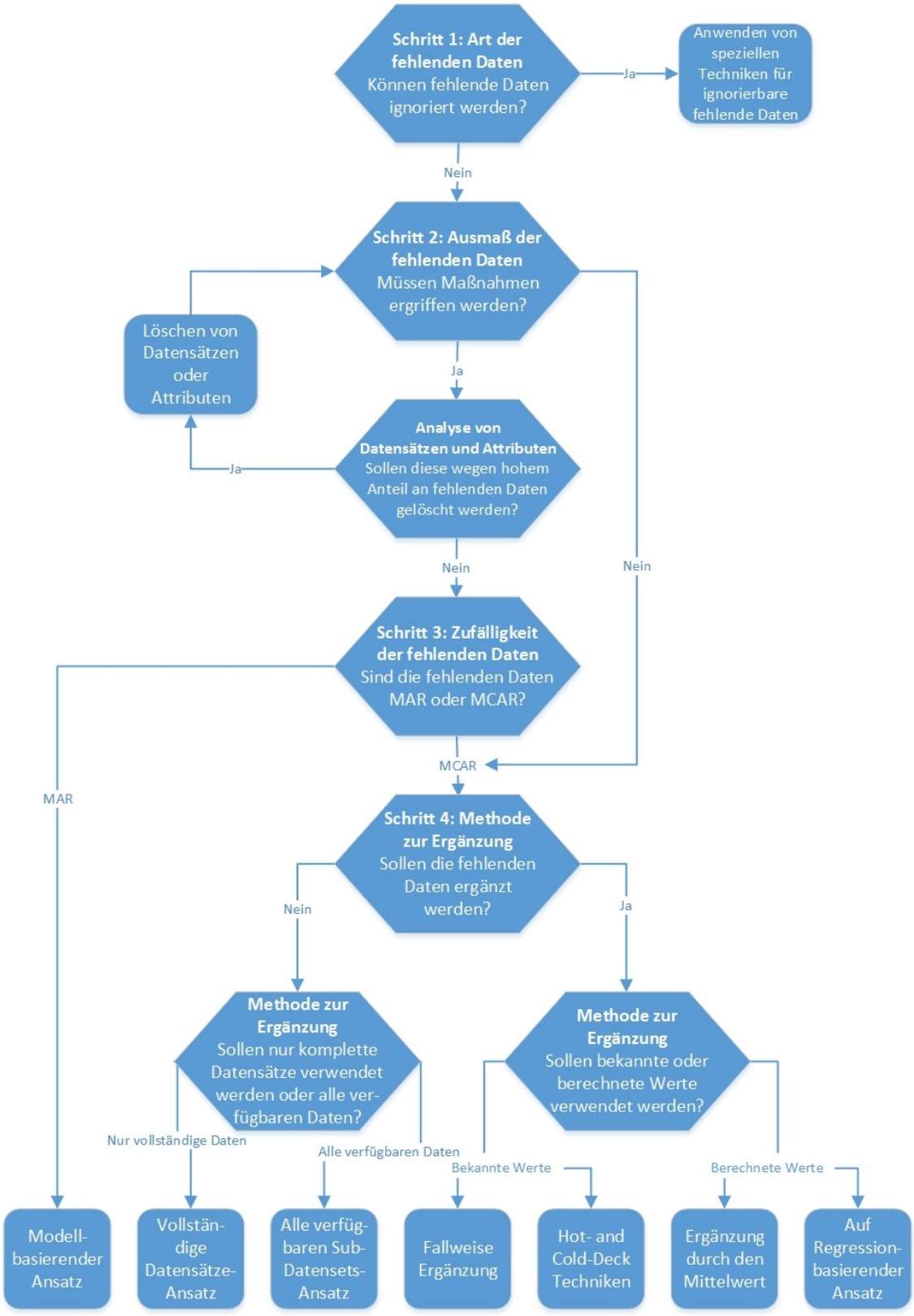


Abbildung 26: Prozess zum Identifizieren und Bereinigen fehlender Daten[200]

[200] Quelle: in Anlehnung an Hair, J. F. et al. (2014), S. 43

Schritt 1: Feststellen der Art der fehlenden Daten

In diesem ersten Schritt soll festgestellt werden, ob Daten, die fehlen, ignorierbar sind oder nicht. Ignorierbare fehlende Daten können folgende sein:

- Daten, die nicht in die Stichprobe einfließen
- Daten, die durch die Struktur der Erhebung entstanden sind
- Zensierte Daten: Diese können aufgrund ihrer Stellung im Erhebungsprozess nicht vollständig sein.

Nicht ignorierbare fehlende Daten können als solche durch mehrere Gründe oder Situationen bezeichnet werden. Im Allgemeinen lassen sie sich in die Kategorien „bekannte“ und „unbekannte“ Ursachen, die während des Datenerhebungsprozess auftreten, einteilen.

Werden die fehlenden Daten als nicht ignorierbar eingeschätzt, wird zu Schritt 2 übergegangen. Ansonsten ist die Analyse beendet.

Schritt 2: Feststellen des Ausmaßes der fehlenden Daten

Die Hauptaufgabe in diesem Schritt ist, festzustellen, ob die Menge der fehlenden Daten gering genug ist, um das Ergebnis nicht zu beeinflussen. Wenn der Anteil gering ist, können die fehlenden Daten durch Ergänzen nach bestimmten Regeln eliminiert werden. Hierbei wird die Frage aufgeworfen, wie das Ausmaß von „gering“ definiert ist. Diese Grenze wird individuell festgelegt. Durch Löschen von Datensätzen oder Attributen kann die Menge von fehlenden Daten verkleinert werden und somit Daten hervorbringen, bei denen das Auffüllen der fehlenden Daten keine Verzerrung der Resultate bewirkt. Hierzu gibt es keine definierten Regeln.

Hair et al. (2014)[201] haben folgende Faustregeln zusammengefasst:

- Grenzen für fehlende Daten
    - Der Anteil fehlender Daten von unter 10 % können generell ignoriert werden. Es sei denn Daten fehlen nach einem bestimmten Muster.[202]
    - Die Anzahl der Datensätze ohne fehlende Daten ist ausreichend für die gewählte Analysemethode, wenn die fehlenden Daten nicht ergänzt wurden.
- Löschen von Attributen und Datensätzen
    - Hair et al. (2014)[203] beziehen sich auf Hertel (1976)[204] mit der Aussage, dass Attribute mit einem Anteil von weniger als 15 % fehlender Daten Kandidaten zum Löschen sind. Bei Anteilen von 20 - 30 % können die Daten oftmals ergänzt werden. Hertel (1976)[205] hingegen verweist auf Sharp und Feldt (1959)[206], dass bei einem Komplettheitsgrad von mindestens 85 % die Daten gelöscht, aber auch andere Möglichkeiten in Betracht gezogen werden sollten. Bei einem geringeren Komplettheitsgrad sollte das Löschen in Betracht gezogen werden.

[201] Vgl. Hair, J. F. et al. (2014), S. 45
[202] Vgl. Malhotra, N. K. (1987), S. 75
[203] Vgl. Hair, J. F. et al. (2014), S. 45
[204] Vgl. Hertel, B. R. (1976), S. 461
[205] Vgl. ebda., S. 461
[206] Vgl. Sharp, H.; Feldt, A. (1959), S. 652

- Es sollte sichergestellt werden, ob die erreichbare Verbesserung des Anteils der fehlenden Daten an der Gesamtmenge durch das Löschen eines Datensatzes oder Attribut groß genug ist.
- Datensätze mit fehlenden Daten in abhängigen Attributen werden üblicherweise gelöscht, um künstliche Verstärkungen in den Beziehungen zwischen unabhängigen Attributen zu vermeiden.
- Wenn ein Attribut gelöscht wird, sollte sichergestellt sein, dass alternative Attribute, die stark miteinander korrelieren und somit die gelöschten Attribute repräsentieren können, zurückbleiben.
- Es sollte in Betracht gezogen werden, die Analyse mit und ohne die fehlenden Daten durchzuführen, um den Unterschied zu erkennen.

Datensätze oder Attribute mit mehr als 50 % fehlender Daten sollten gelöscht werden. Sobald der Prozentsatz darunter ist, muss individuell nach Ziel der Analyse und dem Motto „trial and error“ entschieden werden.

Ist das Ausmaß der fehlenden Daten akzeptabel niedrig und liegt fehlenden Daten kein Muster zugrunde, kann direkt zum Schritt 4 übergegangen werden. Ansonsten wird mit Schritt 3 fortgefahren.

Schritt 3: Feststellen der Zufälligkeit der fehlenden Daten

Sollte beschlossen werden, dass diese fehlenden Daten ergänzt werden sollen, muss der Grad Zufälligkeit der fehlenden Daten bestimmt werden. Hierfür geht man von Daten, die mindestens zwei Attribute repräsentieren, aus. Enthält nun eines fehlende Daten, andere nicht, so kann das Attribut bzw. dessen fehlende Daten als „Missing At Random“ (MAR) oder als „Missing Completely At Random“ (MCAR) definiert werden.[207] *MAR* bedeutet in diesem Fall, dass die fehlenden Werte aus dem einen Attribut mit dem anderen Attribut in Zusammenhang stehen und nicht mit den anderen Werten des eigenen Attributs. *MCAR* hingegen sind solche fehlenden Daten, deren Datensätze sich nicht von anderen Datensätzen unterscheiden lassen. Attribute mit fehlenden Daten können durch Begutachtung bzw. Kenntnisse oder durch vorgefertigte Tests der jeweiligen Kategorie zugeteilt werden.

Schritt 4: Auswahl der Methode zur Ergänzung

Das Ergänzen als solches wird als Prozess zum Schätzen von Daten auf Basis der vorhandenen Daten angesehen. Als Ziel gilt es, Beziehungen zwischen den Daten zu erkennen, die zum Ergänzen verwendet werden können. In diesem Schritt wird die Methode ausgewählt, mit der die fehlenden Daten ergänzt werden. Hierbei wird die in Schritt 3 getroffene Entscheidung, ob es sich um MAR oder MCAR handelt, als Basis verwendet. Unabhängig von dem Grad der Zufälligkeit gibt es verschiedene Methoden zum Ergänzen der Daten.[208,209,210] Wie in Abbildung 26 zuerkennen ist, wird als erstes die Entscheidung getroffen, ob die Daten ergänzt werden sollen. Ist dies nicht der Fall können entweder nur solche Daten verwendet werden, die ohne fehlende Daten sind, oder es wird anstelle der fehlenden Daten ein Lage-, Streuungs- oder

[207] Vgl. Little, R. J. A.; Rubin, D. B. (2002), S. 12
[208] Vgl. ebda., S. 19
[209] Vgl. Heitjan, D. F. (1997), S. 548 ff.
[210] Vgl. Roth, P. L. (1994), S. 539 ff.

Zusammenhangsmaß eingesetzt. Sollte die Daten ersetzt werden, gibt es die Möglichkeit bekannte oder berechnete Werte einzusetzen.

Abschließend ist festzuhalten, dass nur die Ergänzung von metrischen Daten nach diesen Regeln sinnvoll ist. Nicht-metrische Daten sollten anders behandelt werden. Zusätzlich ist die Frage, ob nicht-metrische Attribute für die gewählte Analysemethode überhaupt zulässig sind.

Nach der Analyse der Ausreißer innerhalb der Daten werden im nächsten Kapitel das Zielattribut und seine Eigenschaften betrachtet.

3.4.3 Gruppierung

Die in Kapitel 3.5 verwendeten Methoden sollen klassieren. Daher ist es sinnvoll, das Zielattribut zu gruppieren. Dies bedeutet, die vorhandenen Werte in Klassen mit definierten Grenzen zuzuordnen.[211] Bei der Gruppierung werden jeweils mehrere Ausprägungen des ursprünglichen Attributes zusammengefasst und in eine Klasse bzw. Ausprägung des transformierten Merkmals überführt. Dieser Vorgang geht mit einer Skalentransformation einher. Bei der Klassenbildung wird das Gesamtintervall in Teilintervalle zerlegt, die eindeutig durch ihre Klassenmitte und ihre obere und untere Klassengrenze definiert werden. Bei der Festlegung dieser Klassen gibt es zwei Ansätze: Durch Definition der Klassenbreite, wie schon bei den Histogrammen in Kapitel 3.3.2 erläutert, oder durch die Bestimmung der Klassenanzahl. Das Festlegen der Klassenbreite kann zu Verzerrungen führen, da homogene Intervalle innerhalb der Beobachtungswerte zerschnitten werden könnten. Bei der Bestimmung der Klassenzahl ist darauf zu achten, dass bei zu vielen Klassen nicht die gewünschte Informationsverdichtung erreicht wird, wohingegen zu wenige die ursprüngliche Verteilung der Attributausprägungen verschleiern.[212]

Zum Festlegen der Klassenanzahl gibt es verschiedene Ansätze. Wird eine Einteilung nach Anzahl der Klassen getroffen, so gibt die DIN 55302[213] an, dass bei 100 Beobachtungswerten mindestens 10 Klassen, bei 1000 mindestens 13 und bei 10.000 mindestens 16 Klassen gebildet werden sollten. Daneben sei zu beachten, dass die Anzahl der Klassen bei weniger als 100 Beobachtungswerten nicht größer sein sollte als die Quadratwurzel aus der Zahl der Beobachtungswerte, siehe (3.1). Eine weitere Methode ist die von Sturges aufgestellte Formel (3.26), wobei k die Anzahl der Klassen und n die Anzahl der Beobachtungswerte darstellt. Bei der Bildung der Klassen sollte auf jeden Fall darauf geachtet werden, die Daten nicht zu sehr zu verändern, da so Informationen verloren gehen könnten. [214]

$$\mathrm{k} = 1 + 3.3 \log n \tag{3.26}$$

Das Gruppieren dient der Verdichtung des Zielattributs. Die anderen Attribute können ebenso zusammengefasst werden. Diese Möglichkeit wird im nachfolgenden Kapitel behandelt.

---

[211] Vgl. Fahrmeir, L. et al. (2004), S. 17
[212] Vgl. Degen, H.; Lorscheid, P. (2012), S. 19 ff.
[213] Vgl. DIN 55302 (Blatt 1) zitiert nach Degen, H.; Lorscheid, P. (2012), S. 21
[214] Vgl. Degen, H.; Lorscheid, P. (2012), S. 21

3.4.4 Dimensionalitätsreduktion

In Hinblick auf das Ziel und die Methode der Datenanalyse kann es sein, dass das Verhältnis von Attributen zu Datensätzen nicht optimal ist oder dass die Anzahl von Attributen einen höheren Rechenaufwand nach sich ziehen würde. Um ein optimales Verhältnis zu erhalten, gibt es mehrere Möglichkeiten. Eine davon ist das unwiderrufliche Löschen von Attributen, das zu einem Informationsverlust führt. Die andere ist die Transformation der ursprünglichen Attribute in eine geringere Anzahl neuer Attribute, die annähernd gleiche Informationen enthalten, aber ein besseres Verhältnis von Attributen zu Datensätzen bzw. eine simplere Struktur liefert.

Es gibt verschiedene Methoden zur Dimensionalitätsreduzierung. Dazu gehören die Erstellung eines Sub-Datensets, die Faktorenanalyse oder die multidimensionale Skalierung.[215] Eine der bekanntesten Methoden ist die *Hauptkomponentenanalyse*, die im Folgenden als *PCA (Principal Component Analysis)* bezeichnet wird. Die zentrale Idee der PCA ist das Verringern der Attributanzahl, die in einem Datenset enthalten sind. Zu diesem Zweck werden aus ursprünglichen Attributen neue, die sogenannten *Hauptkomponenten (PCs – Principal Components),* die untereinander nicht korrelieren, gebildet. Dabei soll der Informationsverlust möglichst gering gehalten werden, um die Varianz zu maximieren. Die PCs sind so geordnet, dass die ersten PCs die meisten Variation enthalten, die in allen ursprünglichen Attributen vorhanden ist.[216]

Es gab eine Vielzahl von Mathematikern, die sich seit Ende des 19. Jahrhunderts[217,218] mit der der PCA zugrunde liegenden Singulärwertzerlegung *(SVD – Singular Value Decompositon)* in unterschiedlichsten Kontexten beschäftigt haben. Die PCA als solche wurde als erstes von Pearson (1901)[219] und Hotelling (1933)[220] beschrieben. Dennoch haben sie unterschiedliche Ansätze. Pearsons Ansatz beruht auf der geometrischen Betrachtung der Lage der Datensätze in einem p-dimensionalen Raum und wie diese durch Linien und Ebenen dargestellt werden können.[221] Hotellings Ansatz begründet sich auf der Faktorenanalyse, einem Verfahren der multivariaten Statistik, die die PCA verwendet.[222] Sein Ergebnis unterscheidet sich stark von diesem heute noch genutzten Verfahrens. Seine Idee war, dass es ein kleineres Set von Attributen geben muss, dass die Werte der ursprünglichen p Attribute repräsentiert.[223] Hotellings Methodik ähnelt der in diesem Kapitel vorgestellten allgemeinen Methodik für die PCA. Seit den Anfängen der Forschung im Bereich der PCA, die immer noch anhält, bietet diese ein breites Anwendungsspektrum in den unterschiedlichsten Branchen.[224]

---

[215] Vgl. Alpaydin, E. (2010), S. 110 ff.
[216] Vgl. Jolliffe, I. T. (2002), S. 1
[217] Vgl. Beltrami, E. (1873)
[218] Vgl. Fisher, R. A.; Mackenzie, W. A. (1923)
[219] Vgl. Pearson, K. (1901)
[220] Vgl. Hotelling, H. (1933)
[221] Vgl. Pearson, K. (1901), S. 559
[222] Vgl. Backhaus, K. et al. (2011), S. 356
[223] Vgl. Hotelling, H. (1933), S. 417
[224] Vgl. Jolliffe, I. T. (2002), S. 7 ff.

Für die Anwendung der PCA wird für das Verhältnis von Attributen zu Beobachtungen eine absolute Untergrenze von eins zu fünf gefordert. Ein wünschenswertes Verhältnis wäre ein zu zehn.[225]

Methodik der PCA

Die folgende Beschreibung der Methodik der PCA basiert auf der Eigenvektorzerlegung und zur Gänze auf Shlens (2014)[226] Es gibt andere mathematische Ansätze, die z. B. auf der SVD beruhen. Der Grundgedanke ist, die beste Basis zu finden, um die Attribute in geringerem Ausmaß ausdrücken zu können. Dabei sollen Rauschen herausgefiltert und neue Strukturen entdeckt werden. Hierzu wird ein neues Achsensystem eingeführt, und überprüft, ob eine neue lineare Kombination der Attribute das ursprüngliche Datenset repräsentieren kann. Dazu wird angenommen, dass die Daten linear vorliegen, eine hohe Varianz in den Daten neuen Strukturen verbirgt und die daraus errechneten Hauptkomponenten orthogonal zueinander sind.

Dieser Grundgedanke kann mathematisch formuliert werden: Das Ziel ist es eine orthogonale Matrix $\mathbf{P}$ in $\mathbf{Y} = \mathbf{PX}$, sodass $\mathbf{C_Y} = \frac{1}{n}\mathbf{YY}^T$ eine Diagonalmatrix ist. Die Zeilen von $\mathbf{P}$ sind die Hauptkomponenten von $\mathbf{X}$. Die $m \times n$-Matrix $\mathbf{X}$ stellt in diesem Fall den Datensatz dar, wobei m die Anzahl der Attribute und n die Anzahl der Datensätze ist. $\mathbf{C_Y}$ ist in diesem Fall die Covarianz (3.27), die eine symmetrische $m \times m$-Matrix ist. Ihre Diagonalelemente sind die Varianz der einzelnen Attribute und die Elemente abseits der Diagonalen sind die Kovarianzen zwischen den einzelnen Attributen. Sie spiegeln das Rauschen und die Redundanz der Daten wieder. Große Werte in den Elementen abseits der Diagonalen weisen auf große Redundanz hin.

Das daraus abzuleitende Ziel ist, die Varianz zu maximieren und die Redundanz zu minimieren. Dies geschieht, indem die Zielgleichung nach der unbekannten Kovarianz umgeformt wird:

$$\begin{aligned}\mathbf{C_Y} &= \frac{1}{n}\mathbf{YY}^T \\ &= \frac{1}{n}(\mathbf{PX})(\mathbf{PX})^T \\ &= \frac{1}{n}\mathbf{PXX}^T\mathbf{P}^T \\ &= \mathbf{P}(\frac{1}{n}\mathbf{XX}^T)\mathbf{P}^T\end{aligned}$$

$$\mathbf{C_Y} = \mathbf{PC_X P}^T \tag{3.27}$$

Jede symmetrische Matrix $\mathbf{A}$ wird durch eine orthogonale Matrix ihrer Eigenvektoren diagonalisiert. Für diese gilt: $\mathbf{A} = \mathbf{EDE}^T$, wobei $\mathbf{D}$ die Diagonalmatrix ist. Die Matrix $\mathbf{E}$ besteht aus Eigenvektoren der Matrix $\mathbf{A}$.

[225] Vgl. Rinkenburger, R. (2009), S. 463
[226] Vgl. Shlens, J. (2014), S. 2 ff.

Die Matrix $\mathbf{P}$ wird so gewählt, dass jede Reihe $\mathbf{p_i}$ ein Eigenvektor von $\frac{1}{n}\mathbf{X}\mathbf{X}^T$ ist. Durch diese Wahl gilt $\mathbf{P} \equiv \mathbf{E}^T$. Dies führt zu (3.28):

$$\begin{aligned}\mathbf{C_Y} &= \mathbf{P}\mathbf{C_X}\mathbf{P}^T \\ &= \mathbf{P}(\mathbf{E}^T\mathbf{D}\mathbf{E})\mathbf{P}^T \\ &= \mathbf{P}(\mathbf{P}^T\mathbf{D}\mathbf{P})\mathbf{P}^T \\ &= (\mathbf{P}\mathbf{P}^T)\mathbf{D}(\mathbf{P}\mathbf{P}^T) \\ &= (\mathbf{P}\mathbf{P}^{-1})\mathbf{D}(\mathbf{P}\mathbf{P}^{-1})\end{aligned}$$

$$\mathbf{C_Y} = \mathbf{D} \tag{3.28}$$

Die Ergebnisse der PCA können in den Matrizen $\mathbf{P}$ und $\mathbf{C_Y}$ zusammengefasst werden:

- Die Hauptkomponenten von $\mathbf{X}$ sind die Eigenvektoren von $\mathbf{C_X} = \frac{1}{n}\mathbf{X}\mathbf{X}^T$.
- Das i-te-Diagonalelement von $\mathbf{C_Y}$ ist die Varianz von $\mathbf{X}$ entlang $\mathbf{p_i}$.

Die tatsächliche Anzahl, wie viele Hauptkomponenten dieser Lösung verwendet werden, kann über Kriterien und Tests angegeben werden, wie z. B. den Anteil der erklärten Varianz (Faustregel für Naturwissenschaften 0,95), den Scree-Test oder das Kaiser-Kriterium.[227]

Eignung zur der Daten PCA

Zur Überprüfung der Eignung der vorliegenden Daten für die PCA können verschiedene Kriterien herangezogen werden. In diesem Fall wird die Korrelationsmatrix, die Inverse der Korrelationsmatrix und der Kaiser-Meyer-Olkin-Test vorgestellt.

Die *Korrelationsmatrix* R, die sich aus den in Kapitel 3.3.3 berechneten Korrelationskoeffizienten zusammensetzt, sollte wenige kleine Werte enthalten. Dies würde auf eine heterogene Datenstruktur hindeuten; in diesem Fall wäre die Anwendung in Frage zu stellen.[228]

Die *Inverse der Korrelationsmatrix* R^{-1} sollte eine Diagonalmatrix darstellen. Das bedeutet, dass die Elemente außerhalb der Diagonalen möglichst nahe bei Null liegen. In wie weit diese Werte allerdings von Null abweichen dürfen und wie häufig dies vorkommen darf, ist nicht allgemein geklärt und liegt im eigenen Ermessen.[229]

Guttmann geht davon aus, dass sich die Varianz eines Attributes in zwei Teile zerlegen lässt: das Image und das Anti-Image. Das *Image* stellt jenen Teil der Varianz dar, der durch die restlichen Attribute beschrieben werden kann. Dagegen ist das *Anti-Image* jener Teil, der von den verbleibenden Attributen unabhängig ist.[230] Beim *Kaiser-Meyer-Olkin-Test* wird die Anti-Image-Korrelationsmatrix berechnet. Anhand dieser kann dieser über eine Eignung sämtlicher Attribute und einzelner Attribute für die Faktorenanalyse entschieden werden. Die Gesamtheit der Attribute wird mit dem *MSA-Kriterium (measure of sampling adequacy)*, (3.29) bewertet, das auch *KMO-Kriterium*

[227] Vgl. Rinkenburger, R. (2009), S. 469
[228] Vgl. Backhaus, K. et al. (2011), S. 339
[229] Vgl. ebda., S. 340
[230] Vgl. Guttman, L. (1953), S. 277

(Kaiser-Meyer-Olkin) genannt wird und auch für jedes einzelne Attribut (3.30) berechnet werden kann.[231]

$$\mathrm{MSA} = \frac{\sum_i \sum_{j \neq i} r_{ij}^2}{\sum_i \sum_{j \neq i} r_{ij}^2 + \sum_i \sum_{j \neq i} q_{ij}^2} \tag{3.29}$$

$$\mathrm{MSA(i)} = \frac{\sum_{j \neq i} r_{ij}^2}{\sum_{j \neq i} r_{ij}^2 + \sum_{j \neq i} q_{ij}^2} \tag{3.30}$$

Bei der Berechnung ist darauf zu achten, dass die partielle Korrelationsmatrix Q sich mit $Q = SR^{-1}S$ berechnen lässt, wobei $S = (diag\ R^{-1})^{-1/2}$ ist. Die so berechneten Werte können die Werte im Intervall $[0; 1]$ annehmen. Wünschenswert wäre ein Wert von mindestens 0,8.[232] Und ein $MSA < 0{,}5$ gibt die Untauglichkeit der Daten bzw. des jeweiligen Attributs an.[233] Zudem kann die Eignung der Gesamtheit bzw. einzelner Attribute, wie in Tabelle 6 angegeben, eingestuft werden.[234]

Tabelle 6: Einstufung des MSA-Kriterium hinsichtlich der Eignung für die PCA

$MSA \geq 0{,}9$	marvelous	„erstaunlich“
$MSA \geq 0{,}8$	meritorious	„verdienstvoll“
$MSA \geq 0{,}7$	middling	„ziemlich gut“
$MSA \geq 0{,}6$	mediocre	„mittelmäßig“
$MSA \geq 0{,}5$	miserable	„kläglich“
$MSA < 0{,}5$	unacceptable	„untragbar“

Der KMO-Test gilt als einer der besten Tests zur Bestimmung der Eignung von Daten für die PCA.[235]

Mit der Dimensionalitätsreduktion ist die Vorbereitung der Daten abgeschlossen, und diese können, wie in Abbildung 2 dargestellt, als Basis für die Modellerstellung verwendet werden. Welche Methoden innerhalb des maschinellen Lernens zur Verfügung stehen, wird im nachfolgenden Kapitel vorgestellt.

3.5 Modellierung

Dieses Kapitel soll die in Kapitel 4 angewendeten Methoden des maschinellen Lernens und deren Theorie vorstellen, die im Rahmen der CRISP-DM Phase „Modellierung“ zum Einsatz kommen. Im Rahmen dieser Arbeit liegt der Fokus auf dem Klassieren. Wie bereits aus Tabelle 3 für diese Methode zu entnehmen war, ist das Ziel, einen Datensatz zu klassieren bzw. die Klasse eines Attributes zu ermitteln.

[231] Vgl. Kaiser, H. F. (1970), S. 405
[232] Vgl. ebda., S. 405
[233] Vgl. Edward E. Cureton, R. B. D. (1993), S. 391
[234] Vgl. Backhaus, K. et al. (2011), S. 343
[235] Vgl. Stewart, D. W. (1981), S. 57

3.5.1 Einführung

Um Probleme mit dem Computer zu lösen, werden Algorithmen benötigt. Da die Arbeitsweise nicht mit der eines Menschen vergleichbar ist, ist es ihnen oftmals nicht möglich, den ganzen Prozess zu erfassen. Aber es können Muster und Regelmäßigkeiten identifiziert werden. Das ist der Ansatzpunkt des maschinellen Lernens. Dadurch können Erkenntnisse aus den Daten die Grundlagen für Vorhersagen über die nähere Zukunft sein. Oder sie können die Daten beschreiben und so Wissen über den Prozess generieren.[236]

Die Ausgangsbasis stellen Daten dar. Diese, wie in den vorherigen Kapitel beschriebenen, aufbereiteten Daten werden in zwei Gruppen geteilt: die Trainings- und die Testdaten. *Trainingsdaten* sind jene, von denen gelernt und deren Wissen extrahiert wird. Sie sind die Grundlage für das Modell bzw. dessen Parametrierung. Mit den *Testdaten* wird die Genauigkeit bzw. die Leistung des gelernten Modells überprüft, indem der klassierte Wert mit dem bereits eingetretenen Wert verglichen wird.[237]

Das Lernen von den Trainingsdaten ist mit den im folgenden Kapitel 0 beschriebenen Methoden ein *überwachtes Lernen*. Es wird mit Datensätzen trainiert, die den Zielwert enthalten. Dies beeinflusst den Lernprozess so, dass, wenn etwas falsch gelernt wurde, darauf aufmerksam gemacht wird, und dies bei zukünftigen Klassierungen berücksichtigt wird.[238] Das Problem bei dieser Art des Lernens ist, die Generalisierung des Modells. Sind die Trainingsdaten keine gute Repräsentation der Gesamtmenge der Daten, kann es passieren, dass das Modell nicht generalisierbar ist.[239]

Eine weitere Gefahr ist das die *Überanpassung* (engl. *overfitting*). Dies kann auftreten, wenn das Verhältnis von Attributen zu Datensätzen zu klein ist.[240] Um ein optimales Verhältnis erreichen, kann auf die in Kapitel 3.4.4 Dimensionalitätsreduktion zurückgegriffen werden. Des Weiteren kann zur zusätzlichen Verbesserung des Modells die *Kreuzvalidierung* angewendet werden. Bei der auch K-fache Kreuzvalidierung genannten Methode, werden die X Datensätze in K gleich große Teile, $X_i, i = 1, \dots, K$ geteilt. Dabei wird ein Teil als Testdaten zurückgehalten und die restlichen $K - 1$ Teile sind die Trainingsdaten. Dies wird K-mal wiederholt und so wird immer ein Teil ausgelassen.[241]

Die später im praktischen Fallbeispiel im Kapitel 4 verwendeten Methoden und deren Ansätze werden im folgenden Kapitel vorgestellt.

[236] Vgl. Alpaydin, E. (2010), S. 1 ff.
[237] Vgl. Ertel, W. (2013), S. 181
[238] Vgl. Backhaus, K. et al. (2015), S. 299
[239] Vgl. Alpaydin, E. (2010), S. 24
[240] Vgl. ebda., S. 154
[241] Vgl. ebda., S. 486 f.

3.5.2 Methoden

Alle Methoden haben das Ziel, eine Klassfikation durchzuführen, und somit eine Prognose über den Zielwert durch die Zuordnung eines Datensatzes zu einer Klasse abzugeben. Die Klassifikation beruht auf den Zusammenhängen zwischen den einzelnen Attributen und der Bedeutung ihrer Kombination in Bezug auf den Zielwert.[242]

Tabelle 7: Anwendbarkeit auf Attribute unterschiedlicher Skalenniveaus[243]

Methode	Skala		
	metrisch	**nicht-metrisch**	**gemischt**
Entscheidungsbaum	Ja	Ja	Ja
Diskriminanzanalyse	Ja	Nein	Nein
SVM	Ja	Ja	Ja
KNN	nur euklidische Distanz	Nur Hamming-Distanz	Nein
Ensemblemethoden	Ja	Ja, außer Subspace Diskriminanzanalyse	Ja, außer Subspace Methoden

Nicht jede der nachfolgend vorgestellten Methoden kann mit Attributen unterschiedlichsten Skalenniveaus (siehe Kapitel 3.3.1) arbeiten. Tabelle 7 enthält daher eine Zusammenstellung über die Anwendbarkeit von Methoden auf Attribute in metrischen, nicht-metrischen und gemischten Skalenniveaus.[244]

Entscheidungsbaum

Entscheidungsbäume (engl. *decision trees*) sind äußerst einfache aber effiziente Verfahren, um aus Daten Wissen zu generieren. Der Vorteil ist, dass diese Bäume graphisch dargestellt werden können. Dies erleichtert das Verständnis, die Interpretierbarkeit und Kontrolle.[245] Ein Entscheidungsbaum besteht aus Knoten und Blättern, die die Zielwerte darstellen. Jeder Entscheidungsknoten m hat eine Testfunktion $f_m(x)$, deren jeweiliges Ergebnis in einen der zwei abzweigenden Äste mündet. Beispielhaft ist dies in Abbildung 27 dargestellt, wobei die Knoten mit den Funktionen beschriftet wurden und die Blätter durch die Rechtecke mit den jeweiligen Zielwerten $C1, C2$ dargestellt werden.[246]

[242] Vgl. Alpaydin, E. (2010), S. 5
[243] Quelle: MathWorks, https://de.mathworks.com/help/stats/choose-a-classifier.html (Zugriff: 21.02.2017) (leicht modifiziert)
[244] Vgl. ebda.
[245] Vgl. Ertel, W. (2013), S. 202 f.
[246] Vgl. Alpaydin, E. (2010), S. 185

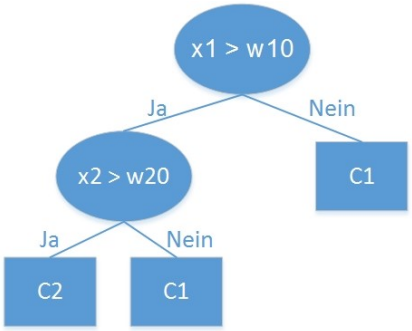


Abbildung 27: Beispiel eines Entscheidungsbaums[247]

Aus einem Entscheidungsbaum können nun Entscheidungsregeln abgelesen werden. Diese sind für das Beispiel in Abbildung 15:

- WENN (x1 > w10) UND (x2 > w20) DANN Zielwert = C2
- WENN (x1 > w10) UND (x2 < w20) DANN Zielwert = C1
- WENN (x1 < w10) DANN Zielwert = C1

Dies macht einen Entscheidungsbaum leicht interpretierbar. Zusätzlich kann dieser durch Experten verifiziert werden, ob dieses Modell bzw. Entscheidungsregeln sinnvoll ist/sind oder nicht.[248] Zusätzlich gibt es die Möglichkeit mehrere Bäume miteinander zu verbinden. Dies fällt unter die Ensemblemethoden.[249]

Diskriminanzanalyse

Bei der *Diskriminanzanalyse* wird angenommen, dass die Datensätze einer Klasse linear von den Datensätzen anderer Klassen getrennt werden können. Dabei wird auf eine Betrachtung der Dichte der Trainingsdaten innerhalb der Klassen verzichtet. Die Methode schätzt die Parameter dieser Trennfunktion durch die Analyse der Trainingsdaten und konzentriert sich einzig auf die Lage dieser Trennfunktion.[250]

Support Vector Machine

Die *Support Vector Machine* (*SVM*) gehört zur Klasse der Kernel Maschinen und basiert auf der Diskrimanten-Methode. Auch hier wird versucht, Datensätze, die als Objekte im Raum vorliegen, durch Trennebene in Klassen zu teilen.[251]

[247] Quelle: in Anlehnung an Alpaydin, E. (2010), S. 186
[248] Vgl. ebda., S. 198
[249] Vgl. Witten, I. H. et al. (2011), S. 356
[250] Vgl. Alpaydin, E. (2010), S. 209 f.
[251] Vgl. ebda., S. 319

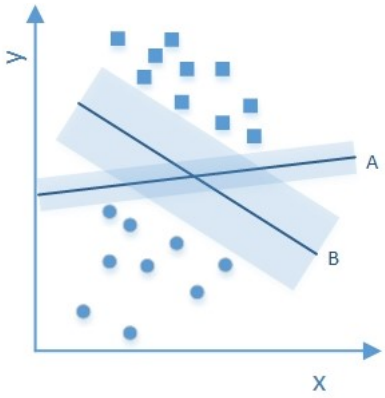


Abbildung 28: Mögliche Platzierungen der Trennebene[252]

Die Möglichkeiten diese zu platzieren sind vielfältig, wie in Abbildung 28 dargestellt wird. Dieses Optimierungsproblem, einen möglichst großen Abstand zwischen den Objekten am Rand der Klassen und der Ebene zu schaffen, wird nicht durch Schätzung sondern durch Berechnung mit Hilfe von Lagrange-Multiplikatoren für die Testdaten gelöst und die daraus abgeleiteten Supportvektoren anschließend zur Verwendung an den Testdaten im Modell gespeichert. Die Abstände zwischen Ebene und Randobjekten soll dazu führen, dass Testdaten richtig klassiert werden.[253] Sind die Daten nicht linear trennbar, kann mit dem Kernel-Trick das lineare Modell in ein neues Achsensystem überführt werden. Durch die Lösung in diesem System und anschließende Rücktransformation wird ein nicht-lineares Problem in einem linearen Modell gelöst.[254]

k-Nearest-Neighbor

Der k-Nearest-Neighbor-Algorithmus (KNN, Nächste-Nachbarn) gehört zur Klasse der parameterfreien Klassifikationsverfahren. Während des Trainings erfolgt nur eine Speicherung der Daten und keine Klassifikationsvorschrift. Erst bei der Anwendung des Modells auf Testdaten, werden diese Datensätze ausgewertet und mit dem zu klassierenden Datensatz abgeglichen. Dabei wird nach dem ähnlichsten gesucht und dem Testdatensatz dessen Klasse zugeordnet. Diese Methode wird auch als „faules Lernen“ (engl. Lazy Learning) bezeichnet.[255]

Ensemblemethoden

Die Kombination von mehreren Klassifikationsmethoden werden Ensemblemethoden genannt. Es gibt verschiedene Ansätze, dazu gehören Bagging und Boosting.[256] *Bagging* kombiniert mehrere Klassenzuordnungen aus verschiedenen Modellen, wobei die einzelnen Klassenzuordnungen gleich gewichtet und am Ende der Durchschnitt bestimmt wird.[257] Beim *Boosting* entsteht aus vielen schwachen Klassifikatoren ein

[252] Quelle: in Anlehnung an Alpaydin, E. (2010), S. 314
[253] Vgl. ebda., S. 309 ff.
[254] Vgl. ebda., S. 319
[255] Vgl. Shakhnarovich, G. et al. (2005), S. 1
[256] Vgl. Witten, I. H. et al. (2011), S. 351 f.
[257] Vgl. Breiman, L. (1996), S. 123

starker Klassifikator. Dieses allgemeine Verfahren hat sehr viele verschiedene Implementierungen, die bekannteste ist AdaBoost.[258]

Im praktischen Fallbeispiel im folgenden Kapitel wird die Entwicklung eines Modells mit allen im Kapitel 3.2 bzw. Kapitel 3 vorgestellten Schritten durchlaufen. Ziel ist es, das Produktionslos durch korrekte Zuordnung der Klasse die Sinterung ohne weitere Einstellung der Öfen durchläuft, und dadurch die Intensität der Betriebsmittelnutzung zu erhöhen. Schlussendlich ist neben der Tauglichkeit des Modells die Kapazität bzw. deren Veränderung die Kennzahl, mit der der Einsatz des Modells begründet wird.

[258] Vgl. Witten, I. H. et al. (2011), S. 358

4 Praktische Fallstudie

Ziel dieses Kapitels ist es, die Vorgehensweise und Ergebnisse von der praktischen Bearbeitung der in Kapitels 1.1 vorgestellten Aufgabenstellung bzw. des Problems zu erläutern. Als erstes wird die Problemstellung wiederholt und ein Grundverständnis für die Art und Aufbau der Datenstruktur geschaffen. Anschließend werden die im Laufe der praktischen Bearbeitung dieser verwendeten Softwareprogramme vorgestellt und die Datenaufzeichnung und deren Zusammenhänge betrachtet.

Danach erfolgt eine Zweiteilung der Problemstellung in die Standorte Deutschlandsberg und Šumperk, wobei in die Bearbeitung des Standort Šumperk die Daten und Erkenntnisse aus Deutschlandsbergs einfließen. Zweck dieser Betrachtung ist die praktische Umsetzung der theoretischen Hintergründe und die Beantwortung der in Kapitel 1.2 formulierten Forschungsfrage, ob und wie der Widerstandswert R_{25} vorausgesagt werden könnte. Dabei orientiert sich die Vorgehensweise am CRISP-DM. Danach wird auf die Forschungsfrage, wie sich eine exakte Vorhersage auf die Kapazitätsauslastung auswirken könnte, eingegangen.

4.1 Einführung

Wie bereits in Kapitel 1.1 erläutert, gibt es mehrere Freigaben entlang des Prozesses zur Herstellung von PTC-Bauteilen. Dieser ist zur besseren Übersicht nochmals in Abbildung 29 dargestellt wird. Sie wird dem Umstand angepasst, dass vorerst nur für einen Produkttypen je ein Modell für die beiden Standorte Deutschlandsberg und Šumperk mit zwei unterschiedlichen Zielwerten erstellt werden soll. Dieser Typ, der im Betrachtungszeitraum eines Jahres am öftesten produziert wurde, besteht aus einer in Deutschlandsberg gefertigten Grundmasse, die als reine Masse weiterverarbeitet wird. Daher entfallen das Homogenisieren und die danach notwendige Mischmassenfreigabe.

Abbildung 29: Produktionsprozess der PTC-Bauteile[259]

Entlang des in Abbildung 29 dargestellten Prozesses fallen eine Vielzahl von Messwerten an. Generell stehen die Daten der Vor- und Nachmahlung der Granulatproduktion, sowie der Granulatfreigabe, und die Daten der Sinterfreigabe für diese Arbeit zur Verfügung. Zusätzlich ist es notwendig, eine Verknüpfung zwischen den Chargen- und Produktionslosnummern der Granulatproduktion und der Frontend Production herzustellen. Die Zusammenhänge und der Umstand, dass mehrere Chargen aus den vorhergehenden Produktionsabschnitten in die nachfolgenden einfließen können, werden in Abbildung 30 dargestellt.

Abbildung 30: Anzahl einfließender Chargen je Produktionsabschnitts[260]

Die Auftragseinplanung erfolgt in Chargen, die beim Start des Loses in der Vormahlung zwischen 1020 und 1200 kg umfasst. Durch Schwankungen im Produktionsprozess werden Endmengen von 490 bis 1010 kg erreicht. Diese werden in den Nachmahlungsprozess eingebracht, wobei sich die Chargen- und Losnummer ändert, da bis zu zwei Chargen aus der Vormahlung in die Nachmahlung einfließen können. In diesem Fall gibt es keine klare Startmenge, wobei sich diese in der Größenordnung von der Liefermenge der Vormahlung bewegt. Die Nachmahlung wird als solches in Šumperk wiederrum mit einer neuen Chargen- und Losnummer versehen, bei denen bis zu vier Chargen der Nachmahlung gemischt werden. Hier beträgt die Startmenge ca. 2000 kg. Die Sinterfreigabemessungen erfolgen fortlaufend während des Sinterns. Für die Beantwortung der Problemstellung ist nur die erste Messung relevant, da diese

[259] Quelle: in Anlehnung an EPCOS PPD PTC PD (2016), S. 10
[260] Quelle: eigene Darstellung

das Zielattribut für Šumperk darstellt. Da bei den nachfolgenden Messungen die Sintereinstellungen an die Ergebnisse angepasst werden, liegt ab diesem Zeitpunkt kein standardisierter Prozess mehr vor.

Vormahlung	Nachmahlung	Materialverwendung	Sinterfreigabe
Stammdaten Komponenten Rückmeldungen Prozesse Messwerte	Stammdaten Komponenten Rückmeldungen Prozesse Messwerte Granulatfreigabe RFA-Analyse	Stammdaten Prozesse	Stammdaten Prozessgrenzen Messwerte
🗝 = Chargennummer	🗝 = Chargennummer	🗝 = Losnummer	🗝 = Lieferschein
Σ = 162 Attribute	Σ = 116 Attribute	Σ = 43 Attribute	Σ = 36 Attribute

Abbildung 31: Aufbau der Ausgangsdaten[261]

Die Ausgangsdaten sind auf 4 einzelne Datensets verteilt. Diese sind in Abbildung 31 übersichtsmäßig dargestellt, da eine vollständige Auflistung aller enthaltenen Attribute zu diesem Zeitpunkt nicht zielführend wäre. Daher werden die Kategorien, in die die Attribute zusammengefasst werden können, das Schlüsselattribut und am unteren Bildrand jeweils Gesamtanzahl der vorliegenden Attribute pro Datenset angegeben. Die Daten der Vor- und Nachmahlung sind bereits aggregierte Daten, die aus unterschiedlichen Quellen für unternehemensinterne Auswertungen zusammengestellt werden. Nach der Begutachtung der Rohdaten für diese Auswertung wird entschieden, mit den aggregierten Daten zu arbeiten, da die Aufbereitung der Rohdaten eben jene aggregierten Daten zur Folge hätte.

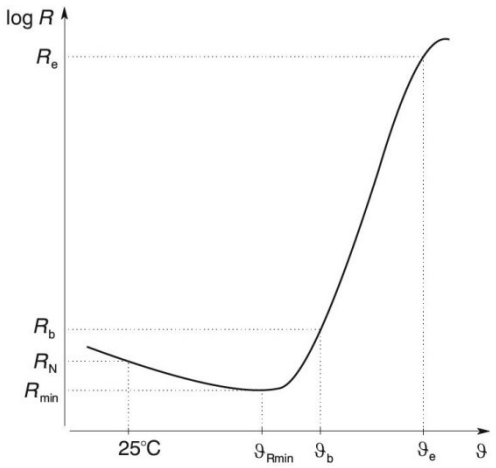


Abbildung 32: R(ϑ)-Kennlinie eines Kaltleiters (schematisch)[262]

Die Zielattribute sind in den Datensets der Nachmahlung und der Sinterfreigabe zu finden. Bei beiden handelt es sich um den Nennwiderstand R_N, der bei bei 25°C

[261] Quelle: eigene Darstellung
[262] Quelle: Reisch, M. (2007), S. 294 (leicht modifiziert)

gemessen wird und daher als R_{25} bezeichnet wird. In Abbildung 32 ist der schematische Verlauf zum Verhalten von Widerstand zu Temperatur eines PTC-Bauteile schematisch dargestellt. Bei Temperaturen ϑ kleiner als ϑ_{min} ist der Temperaturkoeffizient negativ, zudem besitzt das PTC-Bauteil hier den geringsten Widerstand R_{min} und ab diesem Punkt erhöht sich der Widerstandswert mit ansteigender ϑ. Die Bezugstemperatur ϑ_b markiert den Beginn des steilen Anstiegs der Widerstandskurve, die über $R_b = R(\vartheta_b) = 2\,R_{min}$ definiert ist. Die Bezugstemperaturen können in einem Wertebereich von -30°C bis 340°C liegen. Das Ende des steilen Anstiegs liegt bei ϑ_e mit dem Endwiderstand R_e. Bei der Herstellung der PTC-Bauteile kann über die chemische Zusammensetzung die Bezugstemperatur eingestellt werden. Eine Anpassung des R_{25} erfolgt über die im Herstellungsprozess eingesetzten Sintertemperaturen.[263] Durch eine korrekte Vorhersage des zu erwartenden R_{25} können die Sintereinstellungen so angepasst werden, dass der gewünschte R_{25} erreicht wird und keine Nachjustierung der Einstellungen nötig ist.

Die im Folgenden vorgestellten (Zwischen-)Ergebnisse und Berechnungen wird unter Zuhilfenahme der Softwareprogramme Microsoft Excel 2010, PyCharm Community Edition und Matlab 2015b erstellt. Microsoft Excel 2010 als Tabellenkalkulationsprogramm mit der Möglichkeit zur Darstellung der Daten und Erstellen von Diagrammen[264], sowie der Verknüpfung mit Visual Basic for Applications (VBA), der Microsofteigenen Skriptsprache[265], bilden das Grundgerüst für diese Arbeit. PyCharm Community Edition bildet mit der Programmiersprache Python bilden einen kostenlos nutzbare Plattform zur Verarbeitung großer Anzahl von Datensätzen.[266] Dies gilt ebenfalls für Matlab 2015b, wobei die kommerziellen Lizenzen mehrere 1000 Euro kosten können.[267] Daher wurde versucht, Matlab weitestgehend nicht zu verwenden.

Mit diesen Werkzeugen und den Ausgangsdaten wird nun im zweiten Kapitel der Weg zum Erlangen des Verständnisses für die Daten, die Aufbereitung und die Modellierung für die Daten für Deutschlandsberg erläutert.

4.2 Standort Deutschlandsberg

In Deutschlandsberg (DL) wird die Granulatfreigabe betrachtet. Der Input für diese sind Probeteile aus dem freizugebenen Granulat und standardisierte Kontrollteile, die für alle Lose diesen Typs gleich sind. Diese werden gemeinsam gesintert und anschließend deren Widerstände gemessen. Aus dem Verhältnis zueinander wird das Zielattribut „R25 Vergl.%“ berechnet. Ziel ist es, diesen Wert mit einem Modell vorherzusagen.

[263] Vgl. Reisch, M. (2007), S. 293 ff.
[264] Vgl. Microsoft, https://products.office.com/en-us/excel (Zugriff: 13.02.2017)
[265] Vgl. Microsoft, https://msdn.microsoft.com/en-us/library/office/gg264383.aspx (Zugriff: 13.02.2017)
[266] Vgl. JetBrains s.r.o., https://www.jetbrains.com/pycharm/ (Zugriff: 13.02.2017)
[267] Vgl. MathWorks, https://de.mathworks.com/pricing-licensing.html?prodcode=ML (Zugriff: 13.02.2017)

4.2.1 Analyse der Fehlenden Daten

Die Daten vom Standort Deutschlandsberg umfassen die in Abbildung 31 dargestellten Datensets zur Vor- und Nachmahlung im Zeitraum von September 2015 bis August 2016. Wie aus Abbildung 31 hervorgeht, beträgt die Gesamtanzahl der Attribute 278. Sämtliche Attribute, für das Ziel ein Modell wie in Kapitel 0 zu entwickeln, heranzuziehen, ist nicht ratsam. Gründe hierfür können sein, dass diese entweder leer, und somit überflüssig sind, oder zu viele Ausreißer enthalten. Ebenfalls kann das Skalenniveau (siehe Kapitel 3.3.1) unpassend für die gewählten Modelle seine. Genauso müssen die Datensätze überprüft werden, ob Chargen zu Forschungs- und Entwicklungszwecken produziert worden sind. Diese müssen ebenfalls aussortiert werden.

Die Aufbereitung der Daten ist im CRISP-DM (siehe Kapitel 3.2) als dritter Schritt angesiedelt. In diesem Fall macht es Sinn ihn vorzuziehen, um gerade leere Attribute zu entfernen und somit die Übersichtlichkeit zu erhöhen. Außerdem werden solche Datensätze entfernt, die zur weiteren Analyse nicht geeignet sind. Diese Bereinigung wurde mit der in Kapitel 3.4.2 vorgestellten Analyse der fehlenden Daten kombiniert.

Vormahlung

Bei den Daten der Vormahlung (VM) werden sämtliche Forschung- und Entwicklungslose entfernt sowieso solche Datensätze, bei denen kein Ist-Ende angegeben ist, da diese offensichtlich noch nicht fertig gestellt sind, und somit noch nicht in eine Nachmahlung eingeflossen sind. Im anschließenden Schritt 1 werden solche Attribute gelöscht, die Kommentare enthalten oder zu 100 % leer sind. Dazu zählen z. B. Komponenten, die nicht für diesen Produkttyp verwendet werden. Im nächsten Schritt werden Datensätze mit mehr als 69 % fehlender Daten aussortiert. Die 69 %-Schwelle ist als solches nicht in der Literatur zu finden und wird nach Betrachtung der Prozentsätze aller Datensätze gewählt. Abschließend werden solche Attribute eliminiert, die durch das Entfernen der Datensätze keine Werte mehr enthielten.

Tabelle 8: Veränderung der Anzahl und Prozente der fehlenden Werte (VM)[268]

	Anzahl der fehlenden Werte	Prozent der fehlenden Werte
Start	20973	54%
Ende Schritt 1	4478	20%
Ende Schritt 2	3428	15%
Ende Schritt 3	2976	14%

Das Verhältnis von fehlenden zu vorhandenen Daten verbessert sich durch diesen Vorgang von 54 % auf 14 %. Dies ist in Tabelle 8 ersichtlich.

Die verbleibenden Attribute werden in die Kategorien MAR und MCAR eingeteilt. Diese Einteilung befindet sich aus Gründen der besseren Darstellung im Anhang Tabelle 27. Die Einteilung wird nach subjektiver Einschätzung vorgenommen. So werden solche Attribute als MAR kategorisiert, die Komponenten repräsentieren. Oftmals gibt es für

[268] Quelle: eigene Darstellung

eine Komponente mehrere Attribute. So ist es möglich, dass bei der Zugabe von z. B. Blei zwei unterschiedliche Rohstoffe verwendet werden, die jeweils mit Menge und Chargennummer vermerkt werden. Dies hat zur Folge, dass der Komponente Blei vier Attribute zugeordnet sind. Sollte allerdings der Bleibedarf mit einer Rohstoffcharge abgedeckt werden können, bleiben die anderen beiden Felder frei. Diese werden in weiterer Folge durch eine Null ergänzt.

Attribute, die als MCAR bezeichnet werden, sind solche, die eigentlich gemessen hätten werden sollen, aber nicht gemessen worden sind. Dies trifft z. B. auf eine fehlenden Vermerk der Bleimenge oder auf eine nicht gemessene oder nicht eingetragene Messung während eines Prozesses. Hier wird mit dem arithmetischen Mittel (siehe Kapitel 3.3.2) des jeweiligen Attributs ergänzt.

Das Ergebnis dieses Vorgangs ist ein Datenset bestehend aus 226 Datensätzen mit 90 Attributen.

Nachmahlung

Die Analyse der fehlenden Werte in den Daten der Nachmahlung (NM) verläuft ähnlich wie bei der Vormahlung. Bevor die Analyse der fehlenden Werte begonnen wird, werden die Forschungs- und Entwicklungslose und solche Lose, die keinen Zielwert in „R25 Vergl.%“ und keinen Eintrag in „Umsatz1“ besitzen, gelöscht. Letzteres ist die Vormahlungscharge, die in der Nachmahlung als Grundlage eingesetzt wird, und ohne die eine Verknüpfung von Vor- und Nachmahlung nicht möglich ist. Im ersten Schritt werden die Attribute mit 100 % fehlenden Werten und die Kommentare gelöscht. In einem zweiten Schritt wird der Prozentsatz der fehlenden Werte der einzelnen Datensätze betrachtet, wobei hier solche Attribute ausgelassen werden, die nach dem „skip lot“-Verfahren gemessen werden. Dabei wird das Attribut nur zu bestimmten Zeiten gemessen und nicht für jedes Los. Daher entstehen hier bewusst Lücken. Diese werden beim Auffüllen anders behandelt. Bei einem Anteil von mehr als 42 % fehlender Werte wurden Datensätze gelöscht.

Tabelle 9: Veränderung der Anzahl und Prozente der fehlenden Werte (NM)[269]

	Anzahl der fehlenden Werte	**Prozent der fehlenden Werte**
Start	8472	33%
Ende Schritt 1	1904	13%
Ende Schritt 2	1860	13%

Die Entwicklung des Verhältnisses von fehlenden zu vorhandenen Werten ist in Tabelle 9 ersichtlich und verbessert sich um 20 %.

Wieder werden die verbleibenden Attribute in die MAR und MCAR aufgeteilt. Die genaue Aufschlüsselung ist im Anhang in Tabelle 28 zu finden. Die Attribute, die als MAR bezeichnet werden, sind die bereits erwähnten nach „skip lot“-Verfahren gemessenen Attribute. Diese werden so aufgefüllt, dass die Datensätze ohne Werte mit dem als letztes gemessenen Wert annehmen. Bei den mit MCAR gekennzeichneten Werten wird, wie bei der Vormahlung, das arithmetische Mittel gemessen über sämtliche Werte des Attributs ergänzt.

[269] Quelle: eigene Darstellung

Das Ergebnis dieses Vorgangs ist ein Datenset bestehend aus 224 Datensätzen mit 63 Attributen.

4.2.2 Datenzusammenführung

Da für die Modellierung die Daten aus Vor- und Nachmahlung verknüpft werden müssen, wird erneut die Abbildung 30 betrachtet. Es können ein oder zwei Chargen der Vormahlung in die Nachmahlung einfließen. Außerdem fließen diese in veränderlichen Gewichtsanteilen ein. Die Vormahlung hat, wie in Kapitel 4.1 erwähnt, unterschiedliche Startgewichte (Ist-Start). Aus diesem Grund müssen vor der Zusammenführung alle Attribute, die sich auf diese Startmenge beziehen, auf diese genormt werden. Gleiches gilt für Attribute der Nachmahlung, die sich auf deren Startmenge beziehen. Bei zwei einfließenden Chargen liegen jeweils zwei Messwerte für ein Attribut vor. Dies wird unterschiedlich gehandhabt:

- Bei Attributen, die gemessene Werte enthalten, wird von beiden der Mittelwert, der mit den Gewichtsanteilen der jeweiligen Chargen gewichtet wird, verwendet.
- Bei Chargennummern der Rohstoffe werden beide beibehalten. (Bei nur einer einfließenden Charge wird der entsprechend zweite Wert freigelassen.)
- Bei Aufzeichnungen über Maschinen (Fryma, Sprühturm, UD (Ofen)) wird jene der ersten einfließenden Charge gewählt.

Zusätzlich werden Rohstoffanteile, wie z. B. die bereits erwähnte Komponente Blei, die durch diesen Prozess der Zusammenführung pro Datensatz bis zu vier unterschiedliche Ausprägungen haben kann, zu einem Wert zusammengefasst. Genauso werden die Ringtemperatur 1 und Ringtemperatur 2 in einem Attribut gemittelt, da diese nicht klar unterschieden werden können.

Hinsichtlich der Modellierung wird entschieden, Chargennummern der Rohstoffe aus der Analyse auszuschließen, da sich diese in Zukunft ändern können und nicht das passende Skalenniveau haben. Genauso werden die Maschinen, die für die Produktion verwendet wurden, vorerst ausgeschlossen und nicht zielführende Attribute wie Materialnummer, Ist-Start, oder Umsatz Menge entfernt.

Durch diese Zusammenführung ergibt sich ein Datenset mit 99 Attributen, der Chargennummer als Index zur Identifizierung der Datensätze und 218 Datensätzen. Sämtliche Attribute sind im Anhang in Tabelle 29 aufgelistet, wobei die hellblauen hinterlegten solche Attribute sind, die aus dem Datenset der Nachmahlung stammen, die dunkelblau hinterlegten aus der Vormahlung (mit dem Suffix „_VM“) und der Zielwert „R25 Vergl.%“ ist rot hinterlegt und stammt aus dem Datenset der Nachmahlung.

Dieses neue Datenset dient als Grundlage für die Trainingsdaten für das Modell für den Standort Deutschlandsberg. Bevor ein solches trainiert werden kann, müssen für die Daten, wie in Kapitel 3.3 beschrieben, verstanden werden.

4.2.3 Auswertung

Aus diesem neu entstandenen Datenset werden die in Kapitel 3.3.2 und Kapitel 3.3.3 vorgestellten statistische Maße mit Hilfe von Python berechnet. Hierbei ist das Ziel, ein Verständnis für die Daten zu entwickeln. Dafür werden der arithmetische Mittelwert $\bar{x}$, der Median x_{med}, das Minimum, das Maximum, die Varianz $\tilde{s}^2$, die Standardabweichung $\tilde{s}$, die Wölbung γ, die Schiefe g_m und der Bravais-Pearson-Korrelationskoeffizient zwischen dem jeweiligen Attribut und dem Zielwert ermittelt. Die Werte für Varianz, Standardabweichung, Schiefe und Korrelationskoeffizient sind aus der Tabelle 30 zu entnehmen.

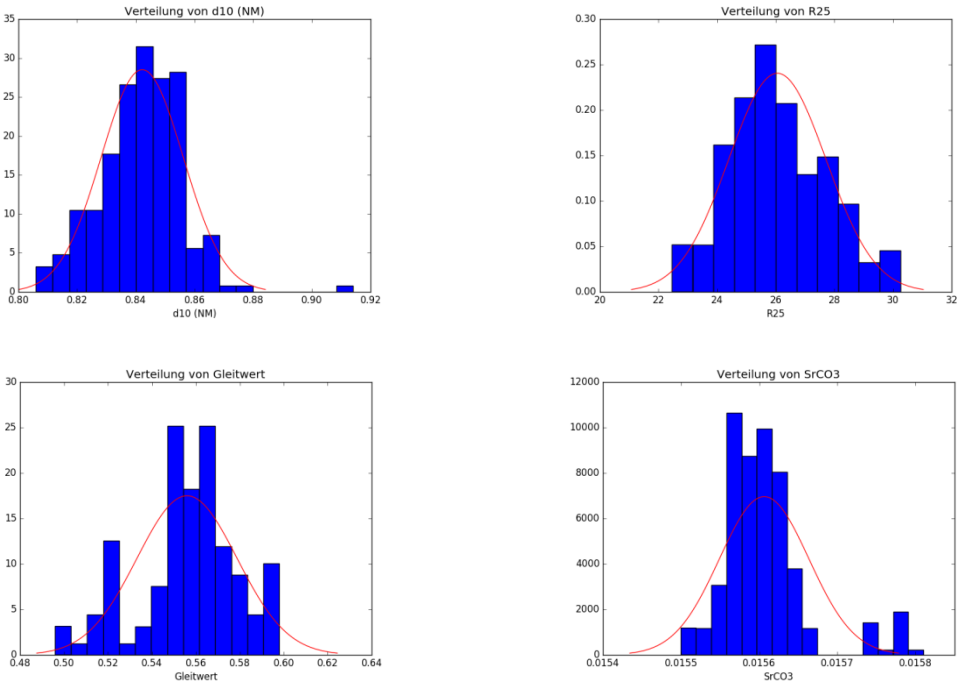


Abbildung 33: Beispiele für Histogramme für DL mit Ausreißer[270]

Zusätzlich werden Histogramme für sämtliche Attribute erstellt, in die die Kurven der Normalverteilung eingefügt wird. In Abbildung 33 sind diese für die Attribute d10 (NM) (links oben), R25 (rechts oben), Gleitwert (links unten) und SrCO3 (rechts unten) zu sehen. Sie weisen unterschiedliche Verteilungen auf, die nicht der Normalverteilung entsprechen. Dies ist ebenfalls in der Tabelle 30 durch die Werte g_m und γ ersichtlich. Diese Erkenntnis gilt auch für den Zielwert R25 Vergl.%, zu sehen in Abbildung 34.

[270] Quelle: eigene Darstellung

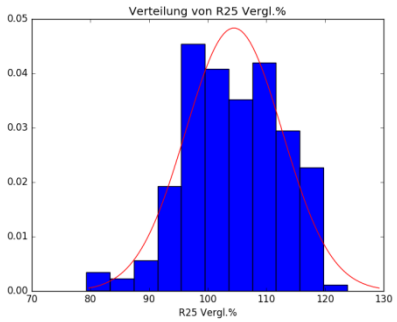


Abbildung 34: Histogramm für R25 Vergl.% mit Ausreißern[271]

In Abbildung 35 sind die Streudiagramme der Wertepaare (Attribut, Zielwert) zu sehen. Als Beispiel werden wieder die Attribute aus Abbildung 33 herangezogen. Für die Attribute d10 (NM), R25 und SrCO3 kann eine positive Korrelation zwischen dem jeweiligen Attribut und dem Zielwert festgestellt werden. Beim Attribut R25 war dies zu erwarten, da dieser als Grundlage für die Berechnung von R25 Vergl.% verwendet wird. Bei dem Attribut Gleitwert gibt es nur eine sehr schwache Korrelation.

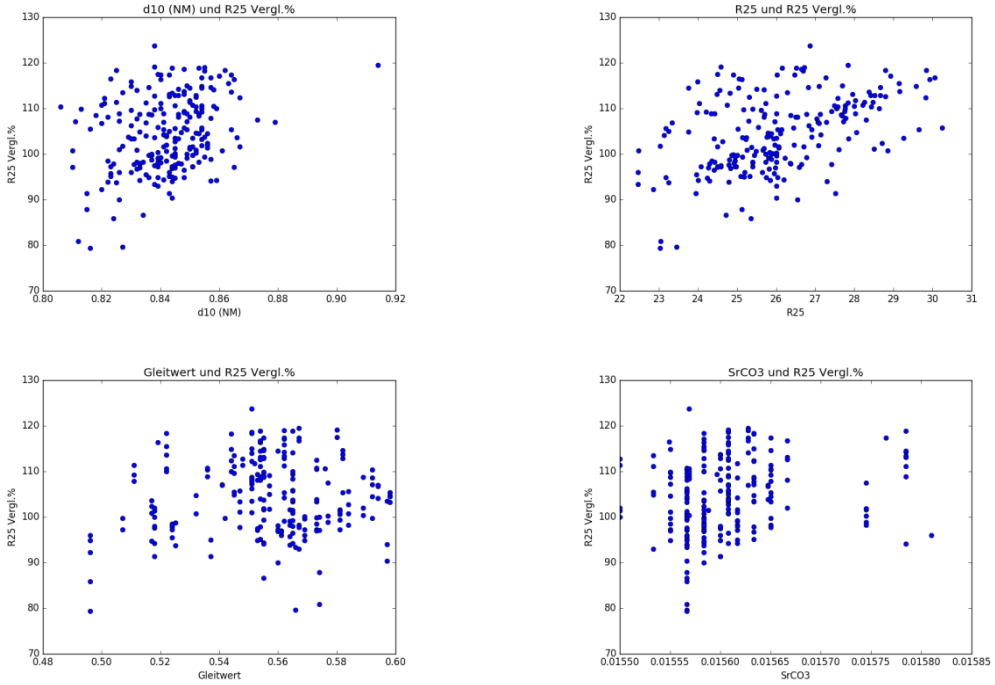


Abbildung 35: Beispiele für Streudiagramme für DL mit Ausreißer[272]

Die Zusammenhänge zwischen den einzelnen Attributen kann ebenso in der Korrelationsmatrix in Abbildung 36 dargestellt werden. Die dunkelroten und dunkelblauen Areale sind solche mit hoher Korrelation. Die helleren Farben sind solche mit geringer Korrelation. Die Achsenbeschriftung erfolgt in Nummern, die die Attribute bezeichnen, in der Reihenfolge wie sie in Tabelle 30 zeilenweise von links nach rechts aufgezählt wurden.

[271] Quelle: eigene Darstellung
[272] Quelle: ebda.

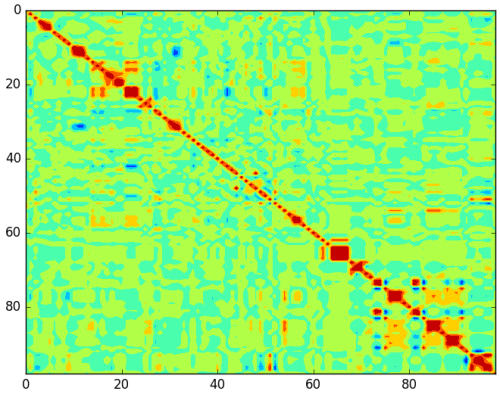


Abbildung 36: Korrelationsmatrix (graphisch) für DL mit Ausreißern[273]

Die hohe Korrelation zwischen den Attributen um den Punkt (10,10) in Abbildung 36 entsteht durch die verschiedenen Durchmesser der Körnung der Rohstoffe und die dadurch resultierende Schlickerdichte für die Nachmahlung. Die Attribute um Punkt (20, 20) werden bei der Granulatfreigabe gemessen. Ebenso korrelieren die gemeinsam bei der Röntgenfloureszensanalyse gemessenen Attribute um Punkt (50,50). Die Korrelation um Punkt (60,60) ist mit der um Punkt (10,10) zu vergleichen, wobei hier die Attribute der Vormahlung aufgetragen sind. Um Punkt (67,67) liegt die Ursache der Korrelation im Zusammenhang zwischen den einzelnen Heizgruppen. Die hohe Korrelation in der rechten unteren Ecke der Korrelationsmatrix folgt aus dem Zusammenhang zwischen den Attributen, die von jeweils einem Attribut das Minimum, Maximum und Standardabweichung angeben.

Zusätzlich wird der Kolmogorov-Smirnov-Test für jedes Attribut durchgeführt, um festzustellen, ob bei diesen eine Normalverteilung vorliegt. Dies ist nicht der Fall. Dies wird auch durch die in Tabelle 30 berechneten Werte bestätigt.

Wie bereits in den Histogrammen in Abbildung 33 ersichtlich, gibt es Ausreißer. Diese befinden sich außerhalb der Normalverteilungskurve bzw. stellen die zweite Spitze der bimodalen Verteilung dar (siehe Attribut Gleitwert). Diese gilt es im nächsten Kapitel zu identifizieren. Zusätzlich werden die Veränderungen, die das Eliminieren von Ausreißern in der Datenstruktur mit sich bringt, anhand der fünf in diesem Kapitel als Beispiele besprochenen Attribute aufgezeigt.

4.2.4 Ausreißeranalyse

Die Identifikation der Ausreißer erfolgte nach der logischen Überlegung, die hinter der Erstellung von Boxplots steht. Dies ist Inhalt des in Kapitel 3.4.1. Die Berechnung des IQR und die Identifikation von Q_1 und Q_3 für jedes Attribut ist erforderlich. Anschließend werden jene Messwerte innerhalb der einzelnen Attribute markiert, die kleiner als $Q_1 - 1{,}5 \times IQR$ oder größer als $Q_3 + 1{,}5 \times IQR$ sind. Lediglich 21 von 99 Attributen und 33 von 218 Datensätzen enthalten keine Ausreißer. Eine Aufschlüsselung hinsichtlich Anzahl bzw. Prozentsatz von Ausreißern je Attribut ist in Tabelle 31 im Anhang zu

[273] Quelle: eigene Darstellung

finden. Die Datensätze mit Ausreißern gänzlich auszuschließen, ist keine Option. Somit wird ab diesem Punkt mit zwei Datensets für Deutschlandsberg weitergearbeitet. Das eine ist das ursprüngliche inklusive der Ausreißer. Bei dem anderen werden die Ausreißer mit dem ohne Ausreißer berechneten arithmetischen Mittelwert der jeweiligen Attribute ersetzt.

In Abbildung 37 sind die Histogramme der gleichen Attribute wie in Abbildung 33 dargestellt. Das Attribut R25 enthielt keine Ausreißer, somit gibt es keine Veränderung im Histogramm. Bei den anderen drei Attributen sind die Veränderungen in den Randbereichen und somit die Erscheinungsbilder klar erkennbar. Die andere Häufigkeitsverteilung von SrCO3 ist die wohl die Folge der automatischen Berechnung der Klassenbreiten, die sich aufgrund der neuen Gesamtheit nun auch ändert.

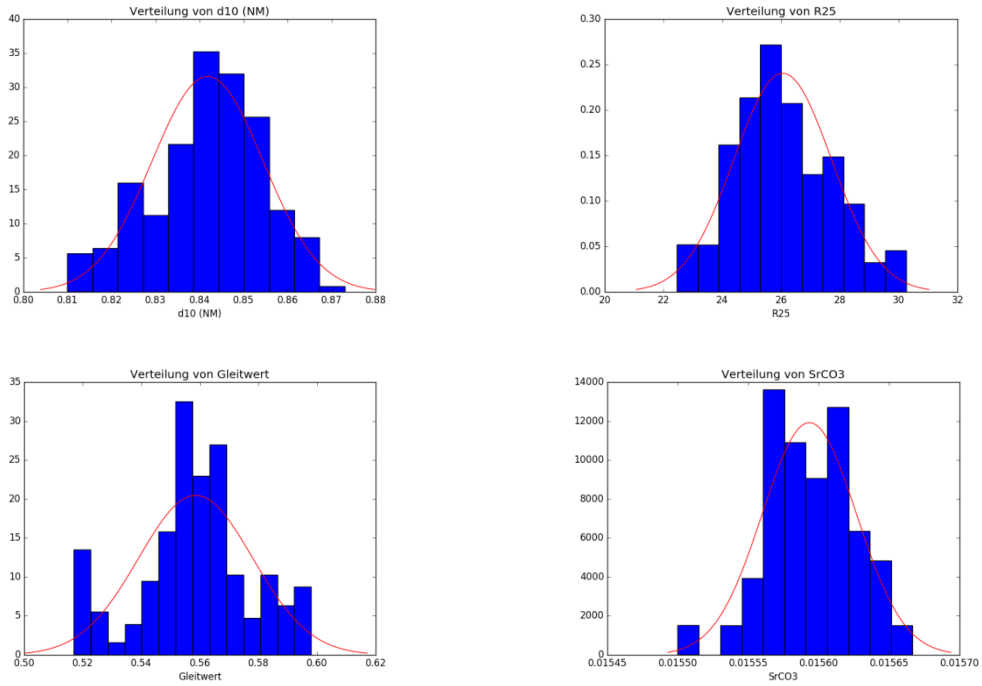


Abbildung 37: Beispiele für Histogramme für DL ohne Ausreißer[274]

Auch der Zielwert R25 Vergl.% enthält Ausreißer. Durch die die Beseitigung dieser wandelt sich das Histogramme in Abbildung 34 und zu jenem in Abbildung 38.

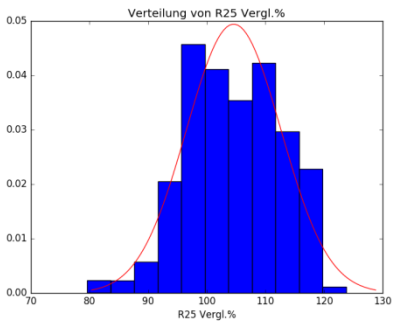


Abbildung 38: Histogramm für R25 Vergl.% ohne Ausreißer[275]

[274] Quelle: eigene Darstellung

Werden die Streudiagramme in Abbildung 35 und Abbildung 39 miteinander verglichen, fällt auf, dass die Randbereiche weniger stark besetzt sind und sich die Wertepaare in der Mitte konzentrieren.

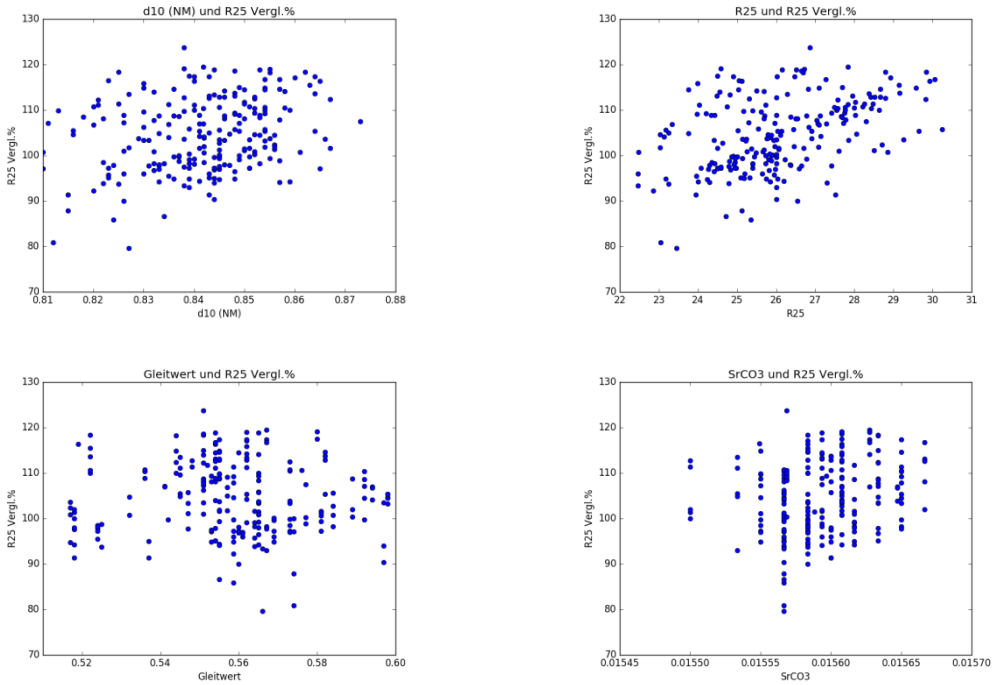


Abbildung 39: Beispiele für Streudiagramme für DL ohne Ausreißer[276]

Auch auf die Korrelationsmatrix hat die Eliminierung der Ausreißer Auswirkungen. Der Vergleich der Korrelationsmatrix mit Ausreißern, in Abbildung 36 und links in der Abbildung 40 zusehen, mit der ohne Ausreißern, rechts in Abbildung 40, zeigt, dass die Korrelation in der rechten unteren Ecke abnimmt. Dies ist auf den Umstand zurückzuführen, dass gerade bei diesen Attributen eine hohe Anzahl von Ausreißern auftritt (siehe Tabelle 31).

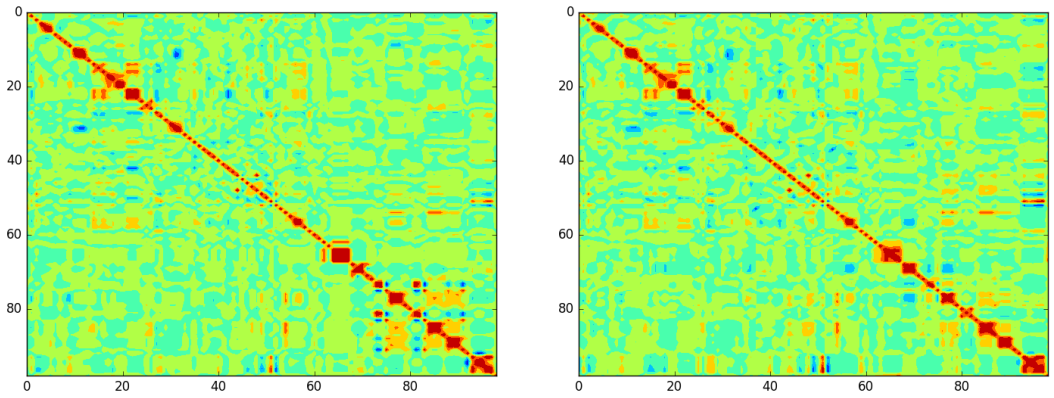


Abbildung 40: Korrelationsmatrizen (graphisch) für DL mit und ohne Ausreißer[277]

[275] Quelle: eigene Darstellung
[276] Quelle: ebda.
[277] Quelle: ebda.

Am Ende dieses Schritts wird entschieden, dass obwohl im Zielattribut ein Ausreißer vorliegt, diesen nicht zu ändern und die anschließende Gruppierung des Zielwerts nur auf Basis der unveränderten Ausprägungen des Attributs vorzunehmen. Dies wird im nächsten Kapitel erläutert.

4.2.5 Gruppierung

In Kapitel 3.4.3 wurden mehrere Ansätze zur Gruppierung erläutert. Für dieses Fallbeispiel wurde die Berechnung der Klassenanzahl nach (3.1) für den Zielwert R25 Vergl.% durchgeführt:

$$k_{DL} = \sqrt{n_{DL}} = \sqrt{218} \approx 15$$

Durch diese Berechnung entstehen 15 Klassen, die durch eine Bezeichnung (arithmetisches Mittel aller in der Klasse vertretenen Werte), die Klassenmitte, die Unter- und Obergrenze definiert werden, siehe Tabelle 10. Zusätzlich wird die Breite angegeben, die in der Skala des Zielattributs, in diesem Fall Prozentpunkte, vorliegt.

Tabelle 10: Einteilung von R25 Vergl.% in Klassen[278]

Bezeichnung	Klassenmitte	Untergrenze	Obergrenze	Anzahl von Datensätzen pro Klasse	Breite
88.62466667	86.57	79.32	93.82	15	14.5
94.87333333	95.02	94.02	96.02	15	2
96.936	96.66	96.02	97.3	15	1.28
98.03066667	98.03	97.53	98.53	15	1
99.516	99.48	98.8	100.17	15	1.37
101.0526667	100.96	100.27	101.66	15	1.39
102.9246667	102.71	101.7	103.72	15	2.02
104.642	104.58	103.73	105.43	15	1.7
106.6053333	106.59	105.44	107.74	15	2.3
108.566	108.51	107.92	109.1	15	1.18
110.0333333	109.99	109.2	110.78	15	1.58
111.9966667	111.86	110.8	112.93	15	2.13
114.066	114.03	113.13	114.93	15	1.8
116.8253333	116.64	114.94	118.34	15	3.4
119.52375	121.095	118.41	123.78	8	5.37

Dieses Kapitel wirdmet sich einzig dem Zielwert und die Grundlage für die Klassifikation künftiger Datensätze geschaffen. Durch die vorgenommene Einteilung gibt es 15 bzw. 8 Datensätze je Klasse, die für das Training des Modells verwendet werden können. Hierbei muss beachtet werden, dass die Einteilung in Klassen an dem Datenset mit Ausreißern vorgenommen und für das bereinigte übernommen wird. Die nachfolgend durchgeführte Überprüfung auf Eignung für die PCA bezieht sich wieder auf alle Attribute und ist der letzte Schritt vor der Erstellung eines Modells.

[278] Quelle: eigene Darstellung

4.2.6 Eignung für die PCA

Die in Kapitel 3.4.4 vorgestellten Vortests, die über die Eignung eines Datensets für die PCA entscheiden, werden für beide Datensets durchgeführt. Es werden die Korrelationsmatrix und ihre Inverse betrachtet. In der Literatur beinhaltet keine aussagekräftigen Grenzwerte. Die Korrelationsmatrix sollte wenige kleine Werte enthalten, wobei hier die Frage ist, in wie weit „klein“ definiert ist und ob klein große negative Zahlen einschließt oder nicht. Ähnliche Überlegungen werden für die Überprüfung der Inversen der Korrelationsmatrix auf die Eigenschaft der Diagonalmatrix angestellt: Was bedeutet klein und in wie weit betrifft dies größere negative Zahlen? Daher wird im Folgenden nur auf die Ergebnisse des KMO-Tests eingegangen, der iterativ durchgeführt wirde, bis zumindest jedes Attribut einen $MSA \geq 0{,}5$ hatte. Um dies zu erreichen, werden jene Attribute gelöscht, die $MSA < 0{,}5$ hatten, und der Test wiederholt. Beide Datensets haben zu Beginn dieser Überprüfung dieselben Attribute, die in Tabelle 29 ersichtlich sind, wobei das Zielattribut R25 Vergl.% ausgeschlossen wird, da dieses nicht in der nachfolgenden PCA eingerechnet werden darf. Zusätzlich werden die Datenwerte nach (3.18) standardisiert.

Datenset mit Ausreißern

Für das Datenset mit Ausreißern wird der KMO-Test viermal durchgeführt, bis für sämtliche Attribute $MSA \geq 0{,}5$ gilt. Im Anhang in Tabelle 32 werden diese genau aufgelistet. Attribute mit $MSA < 0{,}5$ sind rot markiert und gelöscht. Zusätzlich sind sie durchgestrichen. Die Anzahl der Attribute nach dem Löschen ändert sich nach jedem Durchgang, wobei mit 98 gestartet wurde. Die Entwicklung ist in Tabelle 11 zu sehen.

Tabelle 11: Entwicklung der Anzahl der Attribute (VM+NM mit Ausreißern)[279]

	Durchgang			
	1	**2**	**3**	**4**
Anzahl zu löschende Attribute	33	2	1	0
Anzahl Attribute nach Löschen	65	63	62	62

Das MSA-Kriterium für das gesamte Datenset hat sich während der vier Durchgänge wie in Tabelle 12 dargestellt verhalten.

Tabelle 12: MSA für Datenset (VM+NM mit Ausreißern)[280]

	Durchgang			
	1	**2**	**3**	**4**
MSA	0,5957	0,7301	0,7356	0,7382

Das MSA-Kriterium für das Datenset verbessert sich mit jedem Durchgang. Die hohe Anzahl von 36 zu löschenden Attributen nach dem ersten Durchgang hat einen positiven Einfluss auf das das MSA-Kriterium. Dieser Wert ist nach der in Tabelle 6 angegebenen Skala als für die PCA geeignet einzustufen.

Das Ergebnis dieses Schritts ist ein Datenset mit 63 Attributen (Tabelle 11 plus den Zielwert) und 218 Datensätzen.

[279] Quelle: eigene Darstellung
[280] Quelle: ebda.

Datenset ohne Ausreißer

Für das Datenset ohne Ausreißer wird genauso wie bei dem mit Ausreißern vorgegangen. Es gibt vier Durchgänge, wobei beim ersten die Inverse der Korrelationsmatrix nicht berechnet werden kann. Daher kann in weiterer Folge das MSA-Kriterium nicht ermittelt werden. Um dies zu ermöglichen, werden nach dem ersten fehlgeschlagenen Versuch jene fünf Attribute gelöscht, die bei der Berechnung Probleme bereitet haben und der Test wiederholt. Die Auflistung aller Attribute ist im Anhang in Tabelle 33 zu finden. In Tabelle 13 ist die Entwicklung der Anzahl der Attribute zu sehen.

Tabelle 13: Entwicklung der Anzahl der Attribute (VM+NM ohne Ausreißer)[281]

	Durchgang			
	1	**2**	**3**	**4**
Anzahl zu löschende Attribute	5	9	1	0
Anzahl Attribute nach Löschen	93	84	83	83

Der MSA für den gesamten Datensatz verhält sich während der vier Durchgänge wie in Tabelle 14 dargestellt. Auch hier ist eine Veränderung hin zu einem höheren MSA-Kriterium zu sehen.

Tabelle 14: MSA für Datenset (VM+NM ohne Ausreißer)[282]

	Durchgang			
	1	**2**	**3**	**4**
MSA	nan	0,7077	0,7411	0,7441

Aus dieser Abfolge von Durchgängen ergibt sich als Endresultat ein Datenset mit 84 Attributen (Tabelle 13 plus dem Zielwert) und 218 Datensätzen.

Nach der Durchführung des KMO-Tests folgt die Anwendung der PCA. Die Ausführung dieser erfolgte im gleichen Schritt wie das Trainieren der Modelle. Beides wird im nächsten Kapitel besprochen und die Ergebnisse präsentiert.

4.2.7 Modellierung

Das Erstellen der Modelle, die PCA und die Anwendung der Modelle erfolgen in Matlab mittels der App Classification Learner (CL). In dieser App stehen verschiedene Modelle, die mit Hilfe vom überwachten Lernen trainiert werden können, zur Verfügung. Dazu zählen die Kategorie der Entscheidungsbäume, der Diskriminanzanalyse, der SVM, der KNN und der Ensemblemethoden. Jede Kategorie enthält mehrere Variationen der jeweiligen Methode.[283] Zusätzlich können die Einstellungen für die Methode variiert werden, worauf aber in diesem Kontext verzichtet wird und die Standardeinstellungen verwendet werden. In Tabelle 15 und Tabelle 16 wird eine Übersicht über die in Matlab zur Verfügung stehenden Methoden zur Klassifikation gegeben.

[281] Quelle: eigene Darstellung
[282] Quelle: ebda.
[283] Vgl. MathWorks, https://de.mathworks.com/help/stats/choose-a-classifier.html (Zugriff: 21.02.2017)

Tabelle 15: Übersicht über Methoden in Matlab (1)[284]

Methode	Vorhersage-geschwindigkeit	Benötigter Speicherplatz	Interpretier-barkeit	Flexibilität des Modells
Complex Tree	Schnell	Klein	Leicht	Hoch; viele Blätter für eine sehr feine Unterscheidungen zwischen den Klassen (maximale Anzahl der Splits = 100)
Medium Tree	Schnell	Klein	Leicht	Medium; mittlere Anzahl Blätter zur feineren Unterscheidung zwischen Klassen (maximale Anzahl der Splits = 20)
Simple Tree	Schnell	Klein	Leicht	Gering; nur wenige Blätter, um grobe Unterscheidungen zwischen den Klassen zu machen (maximale Anzahl der Splits = 4)
Linear Discriminant	Schnell	Klein	Leicht	Gering; erschafft lineare Grenzen zwischen Klassen
Quadratic Discriminant	Schnell	Groß	Leicht	Gering; erschafft nicht-lineare Grenzen zwischen Klassen (elliptisch, parabolisch oder hyperbolisch)
Linear SVM	binär: schnell; multiclass: medium	Medium	Leicht	Gering; einfache lineare Trennung zwischen den Klassen
Quadratic SVM	binär: schnell; multiclass: langsam	binär: Medium; multiclass: groß	Schwer	Medium
Cubic SVM	binär: schnell; multiclass: langsam	binär: Medium; multiclass: groß	Schwer	Medium
Fine Gaussian SVM	binär: schnell; multiclass: langsam	binär: Medium; multiclass: groß	Schwer	Hoch; Kerneleinstellung = sqrt(P)/4 (P…Anzahl Prädiktoren)
Medium Gaussian SVM	binär: schnell; multiclass: langsam	binär: Medium; multiclass: groß	Schwer	Medium; Kerneleinstellung = sqrt(P) (P…Anzahl Prädiktoren)
Coarse Gaussian SVM	binär: schnell; multiclass: langsam	binär: Medium; multiclass: groß	Schwer	Gering; Kerneleinstellung = sqrt(P)*4 (P…Anzahl Prädiktoren)

[284] Quelle: MathWorks, https://de.mathworks.com/pricing-licensing.html?prodcode=ML (Zugriff: 13.02.2017) (leicht modifiziert)

Für das gegebene Fallbeispiel sind die Geschwindigkeit der Klassifikation und der benötigte Speicherplatz von fast keiner Bedeutung, da die zu analysierenden Datenmenge mit 218 Datensätzen, einem zu klassierenden Attribut mit 15 Klassen und mehr als 60 Prädiktoren eher gering gegenüber der Testmenge (≤ 7.000 Datensätze, 80 Prädiktoren, 50 Klassen) ist, auf der diese Angaben beruhen. Zur Vollständigkeit werden sie dennoch angegeben. Bei der Geschwindigkeit wird in Schnell (0,01 Sekunden), Medium (1 Sekunde) und Langsam (100 Sekunden) unterschieden. Der benötigte Speicherplatz wird in Klein (1 MB), Medium (4 MB) und Groß (100 MB) eingeteilt.[285]

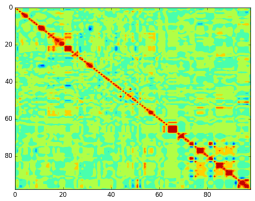
Tabelle 16: Übersicht über Methoden in Matlab (2)[286]

Methode	Vorhersage-geschwindig-keit	Benötigter Speicher-platz	Interpretier-barkeit	Flexibilität des Modells
Fine KNN	Medium	Medium	Schwer	Feine detaillierte Unterscheidung zwischen Klassen. Anzahl der Nachbarn = 1
Medium KNN	Medium	Medium	Schwer	Mittelfeine Unterscheidung zwischen Klassen. Anzahl der Nachbarn = 10
Coarse KNN	Medium	Medium	Schwer	Grobe Unterscheidung zwischen Klassen. Anzahl der Nachbarn = 100
Cosine KNN	Medium	Medium	Schwer	Mittelfeine Unterscheidung zwischen Klassen unter Verwendung einer Cosinus-Distanz-Metrik Anzahl der Nachbarn = 10
Cubic KNN	Langsam	Medium	Schwer	Mittelfeine Unterscheidung zwischen Klassen unter Verwendung einer Kubische-Distanz-Metrik Anzahl der Nachbarn = 10
Weighted KNN	Medium	Medium	Schwer	Mittelfeine Unterscheidung zwischen Klassen unter Verwendung der gewichteten Distanz Anzahl der Nachbarn = 10
Boosted Trees	Schnell	Klein	Schwer	Medium bis Hoch - Methode: AdaBoost mit Decision Tree
Bagged Trees	Medium	Groß	Schwer	Hoch
Subspace Discriminant	Medium	Klein	Schwer	Medium
Subspace KNN	Medium	Medium	Schwer	Medium
RUSBoosted Trees	Schnell	Klein	Schwer	Medium - gut für schiefe Daten

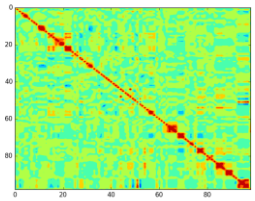
[285] Vgl. MathWorks, https://de.mathworks.com/help/stats/choose-a-classifier.html (Zugriff: 21.02.2017)
[286] Quelle: MathWorks, https://de.mathworks.com/pricing-licensing.html?prodcode=ML (Zugriff: 13.02.2017) (leicht modifiziert)

Die Überlegung bei der Anwendung der App ist, eine erste Abschätzung über die Sinnhaftigkeit des Einsatzes der einzelnen Methoden ohne aufwendige Implementierung treffen zu können. Zu diesem Zweck werden verschiedene Datensets in einem ersten Schritt getestet, und die Ergebnisse miteinander verglichen. In einem zweiten Schritt werden zwei Datensets ausgewählt und mit denen alle zur Verfügung stehenden Modelle trainiert werden. Diese Modelle werden anschließend auf die Testdaten angewendet.

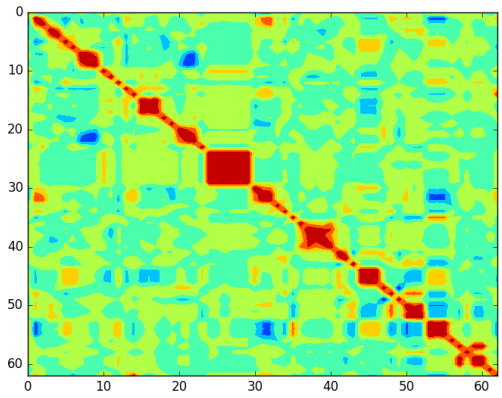
Bei den Testdaten handelt es sich um die 78 Datensätze, die aus Zeitraum von September 2016 bis Anfang Jänner 2017 stammen und nicht in den Trainingsdaten enthalten sind. Die Aufbereitung der Testdaten wird auf das Zusammenführen der Daten (siehe Kapitel 4.2.2) und das Standardisieren beschränkt. Zusätzlich wird das Zielattribut R25 Vergl.% entfernt.



Trainingsdaten mit Ausreißern



Trainingsdaten ohne Ausreißer



Testdaten

Abbildung 41: Korrelationsmatrix der VM+NM[287]

Da im praktischen Einsatz das Modell nur auf einen einzigen Datensatz angewendet wird, wird auf eine umfassende Auswertung verzichtet. Einzig die Korrelationsmatrix für die Trainingsdaten wird berechnet. In Abbildung 41 sind sowohl diese als auch die bereits bekannten Korrelationsmatrizen auf Basis der Trainingsdaten aus Abbildung 36 bzw. Abbildung 40 zu sehen. Aus dem direkten Vergleich wird ersichtlich, dass die Testdaten anders korrelieren als in die Trainingsdaten, wobei dies für die Daten mit und ohne Ausreißer gilt. Diese Erkenntnisse bilden die Grundlage für den Zweifel an der Generalisierbarkeit der auf den Trainingsdaten basierenden Modelle.

[287] Quelle: eigene Darstellung

Tabelle 17: Auswertung der Testdaten (Anzahl je Klasse)[288]

Bezeichnung	Anzahl Datensets je Klasse
88.62	0
94.87	1
96.94	4
98.03	3
98.78	2
99.52	4
101.05	9
102.92	15
104.64	18
106.61	11
107.8	2
108.57	3
110.03	2
112.00	4
114.07	0
116.83	0
119.52	0

Zur späteren Überprüfung der Genauigkeit der Modelle werden die Ausprägungen des Zielattributs in die in Tabelle 10 angeführten Klassen eingeordnet. Dies ist eine Voraussetzung für einen Vergleich der vorhergesagten und tatsächlichen bzw. Ist-Werten. In Tabelle 17 ist eine Aufschlüsselung zu sehen, wie sich die 78 Datensätze auf die 15 Klassen verteilen. Da es zwischen den Klassen Intervalle gibt, die keiner Klasse zugeordnet sind, ist es möglich, dass die in den Testdaten vorhandenen Werte zwischen zwei Klassen liegen. Sie sind in Tabelle 17 rot hinterlegt und müssen als Grenzfälle in der Auswertung gesondert angeführt werden.

Vergleich verschiedener Modelle und Datensets

Für die Daten der Vor- und Nachmahlung liegen bis zu diesem Punkt drei Datensets vor: beide durch den KMO-Test verkleinerten Datensets: jener mit Ausreißern, jener ohne Ausreißer und das ursprüngliche Datenset. Zusätzlich wird die Überlegung seitens des Unternehmens eingebracht, dass das Attribut UD, der Ofen, in dem die PTC-Bauteile gesintert werden, Einfluss auf das Ergebnis hat. Ebenso gibt es die Möglichkeit die Anzahl der Anteile, die bei der Kreuzvalidierung (siehe Kapitel 3.5.1) gebildet werden, zu beeinflussen. Die PCA kann als Option, die bei der Modellerstellung verwendet wird, gewählt werden. Die Varianz wurde hierbei auf 0,95 gesetzt, siehe Kapitel 3.4.4. Aus diesen Kombinationsmöglichkeiten ergaben sich die Datensets in Tabelle 18, die getestet wurden.

[288] Quelle: eigene Darstellung

Tabelle 18: Kombinationen für VM+NM[289]

	Ausreißer	Kreuz-validierung	PCA	UD	Anzahl Prädikatoren
5-fold PCA inkl. Ausreißer (Prädiktoren: 98)	mit	5	ja	nein	98
5-fold PCA inkl. Ausreißer + UD (Prädiktoren: 99)	mit	5	ja	ja	99
5-fold PCA inkl. Ausreißer (Prädiktoren: 62)	mit	5	ja	nein	62
10-fold PCA inkl. Ausreißer (Prädiktoren: 62)	mit	10	ja	nein	62
5-fold PCA exkl. Ausreißer (Prädiktoren: 83)	ohne	5	ja	nein	83
10-fold PCA exkl. Ausreißer (Prädiktoren: 83)	ohne	10	ja	nein	83
5-fold inkl. Ausreißer (Prädiktoren: 62)	mit	5	nein	nein	62
5-fold inkl. Ausreißer + UD (Prädiktoren: 63)	mit	5	nein	ja	63

Die in Tabelle 18 angegebenen Datensets wurden als Trainingsdaten für die Modellerstellung verwendet. In Abbildung 42 sind die Ergebnisse zu sehen. Die angegebenen Prozent auf der Abszisse geben die Genauigkeit des Modells an (siehe Kapitel 3.5.1). Auf der Ordinate sind die in Tabelle 15 und Tabelle 16 vorgestellten Methoden aufgetragen. Die Ergebnisse schwanken in einem Bereich von 7 % bis 23 %. Die Genauigkeit von 0 % steht für die Nicht-Anwendung der Methode. Ursache hierfür ist das Skalenniveau der Daten, das sich durch das Einbeziehen des Attributs UD ändert (siehe Tabelle 7).

[289] Quelle: eigene Darstellung

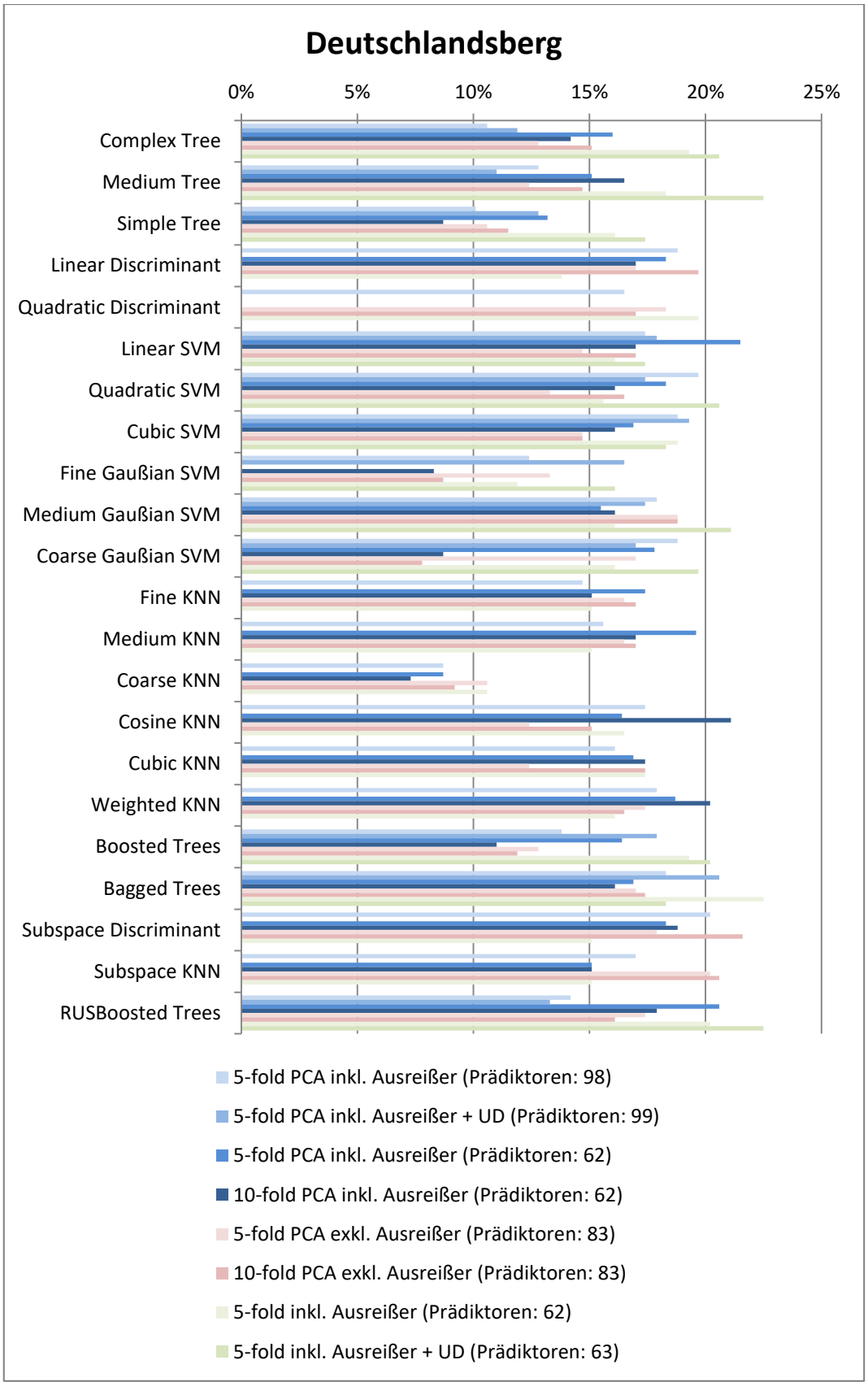


Abbildung 42: Vergleich verschiedener Kombinationsmöglichkeiten[290]

[290] Quelle: eigene Darstellung

Durch diese erste Abschätzung werden folgende Behauptungen aufgestellt:

- Die PCA trägt nicht unbedingt zu einer höheren Genauigkeit bei. Auch die Rechengeschwindigkeit wird nicht wesentlich verbessert, da es sich um so wenige Datensätze handelt.
- Das Attribut UD verändert die Genauigkeit nicht wesentlich; allerdings können nicht alle Methoden verwendet werden.
- Der Unterschied zwischen dem Datenset mit und ohne Ausreißer ist nicht signifikant.
- Die Anzahl der Kreuzvalidierungsteilmengen ist hier nicht ausschlaggebend.

Zusätzlich stellte sich die Frage, ob die beiden Datensets mit und ohne Ausreißer tatsächlich miteinander verglichen werden können, da eine unterschiedliche Anzahl von Attributen vorliegt. Ebenso ist zu bedenken, dass ein neuerliches Durchführen der gleichen Methode zu Modellen mit unterschiedlichen Genauigkeiten führen kann.

Die Abschätzung ist somit nur bedingt maßgebend für die Auswahl der endgültigen Trainingsdaten. Es wird die Entscheidung getroffen, auf die PCA zu verzichten. Dies liegt einerseits daran, dass ähnliche Ergebnisse bei Anwendung bzw. Nicht-Anwendung erzielt wurden, anderseits daran, dass die in Kapitel 3.4.4 empfohlene Untergrenze unterschritten wird. Eine 5-fache Kreuzvalidierung für die Trainingsdaten wird verwendet, um ein mögliches Overfitting zu vermeiden, und es wird mit den 62 Attributen aus dem Datenset mit Ausreißern und dem Attribut UD als Prädikatoren weitergearbeitet. Prädikatoren sind jene Attribute, auf denen die Modelle basieren. Zusätzlich wird das Datenset ohne Ausreißer auf diese 63 Prädiktoren beschränkt, um eine bessere Vergleichbarkeit der Ergebnisse zu gewährleisten. Mit diesen beiden Datensets werden Modelle mit den zur Verfügung stehenden Methoden erstellt und mit den Testdaten getestet. Diese Ergebnisse werden im Folgenden vorgestellt.

Ergebnisse des Datenset mit Ausreißern

Als erster Schritt werden erneut sämtliche Modelle trainiert, deren Methoden Daten in gemischten Skalenniveaus verarbeiten können. Dies führt zu den verschiedenen Methoden der Entscheidungsbäume (Complex Tree, Medium Tree, Simple Tree), den SVM (Linear SVM, Quadratic SVM, Cubic SVM, Fine Gaussian SVM, Medium Gaussian SVM, Coarse Gaussian SVM) und den Ensemblemethoden (Boosted Trees, Bagged Trees, RUSBoosted Trees), die auf den Entscheidungsbäumen basieren. Die Entscheidungsbäume unterscheiden sich im Detaillierungsgrad voneinander, die SVM in der Art der Kernels, die verwendet werden und die Ensemblemethoden in der Art des Kombinierens der einzelnen Bäume. Durch die Anwendung all dieser Methoden wird eine erste Aussage über die Genauigkeit der Modelle getroffen. Diese sind in Abbildung 43 durch die blauen Balken dargestellt.

Bei der Anwendung der Modelle auf die Testdaten wird die Klasse, zu denen ein Datensatz gehört, vorhergesagt. Diese vorhergesagte Klasse wird mit der Ist-Klasse verglichen. Sind diese dieselbe, wird die Vorhersage als Treffer gewertet. Durch die Anzahl der richtig vorhergesagten Klassen je Modell, die in Bezug zur Gesamtanzahl der Testdatensätze gesetzt wird, ergibt sich die Genauigkeit in Prozent je Modell. Diese werden durch die roten Balken in Abbildung 43 dargestellt.

Die grünen Balken in Abbildung 43 symbolisieren die Grenzfälle. Diese entstehen, wenn die tatsächliche Klasse nicht eindeutig zugeordnet werden kann, da der Messwert von R25 Vergl.% zwischen den zwei Klassen liegt. Bei der vorhergesagten Klasse handelt es sich um jene, die unmittelbar an diesen Wert angrenzen. Es kann daher nicht definitiv gesagt werden, ob richtig oder falsch zugeordnet wird. Grenzfälle, die einen größeren Abstand zwischen der vorhergesagten und tatsächlichen Klasse haben, werden nicht berücksichtigt. Daher resultieren die Schwankungen zwischen 0 und 1,3 %. Dieser Prozentsatz entspricht einem Grenzfall.

In Abbildung 43 ist zu sehen, dass die Genauigkeit bei der Anwendung auf die Testdaten im Durchschnitt 9 % niedriger als bei der Abschätzung nach dem Lernen. Ein Grund dafür könnte ein Overfitting der Methode bei der Kreuzvalidierung sein. Die Bagged Trees erreichen im Training die höchste Genauigkeit (24,3 %), und die Fine Gaussian SVM die niedrigste. Bei der Anwendung auf die Testdaten erzielen diese aber die höchste Genauigkeit. Die dem Diagramm zugrunde liegenden Daten sind in Tabelle 34 im Anhang aufgeführt.

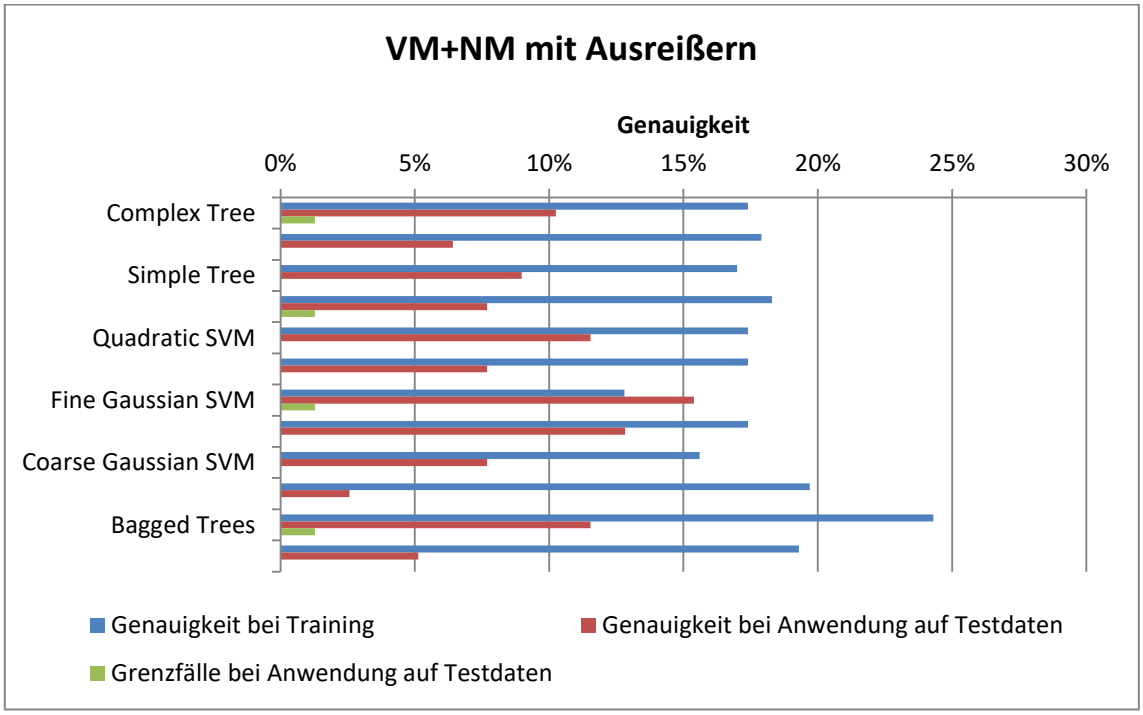


Abbildung 43: Ergebnisse des CL für VM+NM mit Ausreißern[291]

Eine genaue Aufschlüsselung über die vorhergesagte und Ist-Klasse je Datensatz bzw. Charge und Methode ist in Abbildung 62 bis Abbildung 73 im Anhang zu finden. Im Durchschnitt ist die Abweichung von vorhergesagter zur Ist-Klasse 2,47 bis 3,09 Klassen (siehe Tabelle 34), wobei das Maximum von der quadratischen SVM erreicht wird. Dies ist interessant, da die Vorhersagegenauigkeit mit 11,5 % im oberen Drittel liegt. Ebenfalls ist zu bemerken, dass die Fine Gaussian SVM bei der Anwendung auf die Testdaten bessere Ergebnisse als für die Trainingsdaten liefert.

Die Interpretierbarkeit der Modelle wurde bereits in Tabelle 15 und Tabelle 16 angesprochen. Lediglich die Entscheidungsbäume und die Lineare SVM seien leicht zu

[291] Quelle: eigene Darstellung

interpretieren. Die Lineare SVM bezieht sich auf eine lineare Trennung der Datensätze im Raum. Da für die Platzierung dieser 63 Attribute betrachtet werden, ist es dennoch nicht möglich, diese graphisch darzustellen. Die Entscheidungsbäume können wie in Kapitel 0 als Graph oder Code dargestellt werden. Daher wird der Simple Tree als Beispiel für das Ergebnis einer solchen Methode verwendet und für alle Datensets graphisch dargestellt. Zusätzlich werden die Regeln, die der Medium Tree und Complex Tree aufgestellt hat, betrachtet.

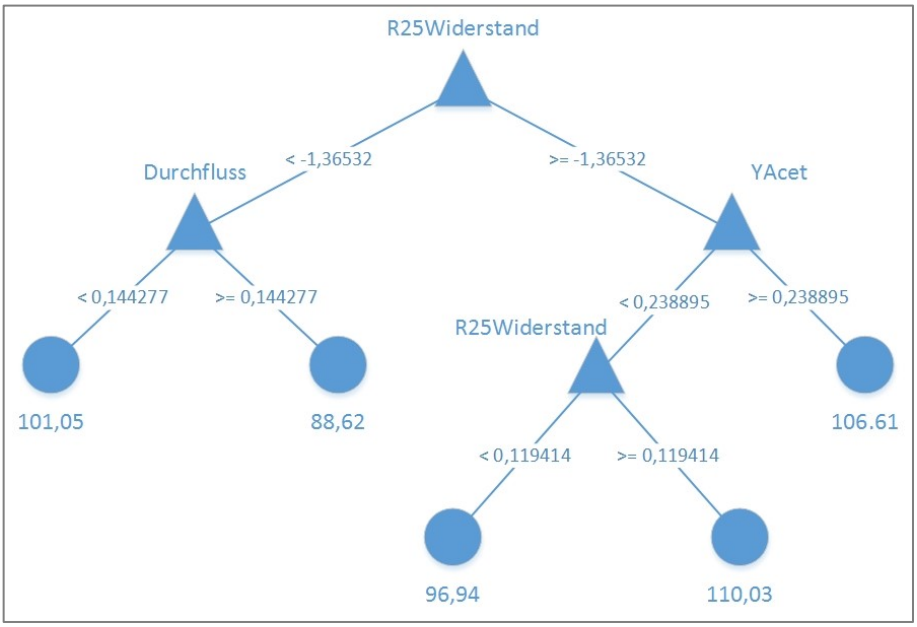


Abbildung 44: Simple Tree für Datenset VM+NM mit Ausreißern[292]

In Abbildung 44 ist der Simple Tree graphisch dargestellt. Es ist ersichtlich, dass die Anzahl der Blätter, also der Endpunkte, nur fünf beträgt. Dies bedeutet, dass nur ein Drittel der Klassen abgebildet werden und somit korrekt klassiert werden können. Grund dafür sind die Standardeinstellungen von Matlab, in denen die Anzahl der Teilungen angegeben werden kann. Das Modell ist somit trotz seiner Genauigkeit von 17 % in der Lernphase nicht für die Testdaten geeignet, da zwei der fünf Endknoten in den Daten nicht enthalten sind. Dieser Umstand wird auch in den Abbildung 64 deutlich, da der Linie der Vorhersage wenig differenziert. Zusätzlich ist die Verwendung für zukünftige Datensätze nicht sinnvoll, da das Modell nicht alle Klassen abbildet.

Ab Seite w im Anhang sind sowohl die Regeln für den Medium Tree als auch die für den Complex Tree angegeben. Diese sind wesentlich detaillierter verzweigt, dennoch liefern sie nur bedingt bessere Ergebnisse.

Ergebnisse des Datensets ohne Ausreißer

Wie bereits bei den Ergebnissen des Datensets mit Ausreißern besprochen, werden sämtliche Methoden auf die Testdaten angewendet. Die Ergebnisse sind nach dem gleichen Prinzip aufbereitet worden und in Abbildung 45 dargestellt.

[292] Quelle: eigene Darstellung

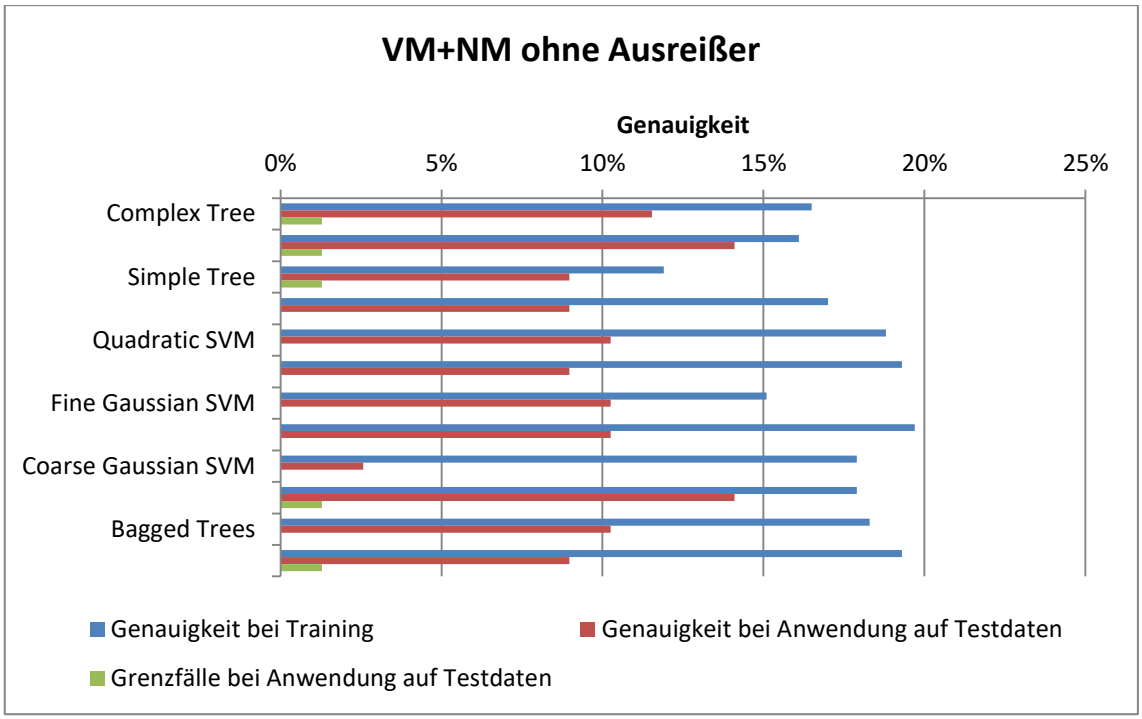


Abbildung 45: Ergebnisse des CL für VM+NM ohne Ausreißer[293]

In Abbildung 45 sind die Ergebnisse des Datensets ohne Ausreißer zu sehen. Sie sind schlechter als für das Datenset mit Ausreißern. Dennoch weicht die Genauigkeit bei der Anwendung mit einem Durchschnitt von 7,5 % weniger vom Ergebnis des Trainings ab. Der durchschnittliche Unterschied zwischen vorhergesagter und Ist-Klasse ist mit 2,06 bis 3,37 Klassen sehr viel größer. Der Trend vom vorher betrachteten Datenset, dass die SVM bessere Ergebnisse liefern, setzt sich bei diesem Test nicht fort und in diesem Fall erzielen die Bäume bzw. die Ensemblemethoden bessere Ergebnisse. Eine genaue Aufschlüsselung ist der Tabelle 35 zu entnehmen. Zusätzlich werden die einzelnen Ergebnisse in den Abbildung 74 bis Abbildung 85 im Anhang über die Chargen als Diagramm dargestellt.

Die andere Datenstruktur und Korrelationsmatrix der Datensets ohne Ausreißer hat ebenso einen anderen Entscheidungsbaum in Abbildung 46 als Ergebnis. Es werden andere Attribute für den Baum gewählt oder andere Grenzwerte für die gleichen Attribute eingesetzt. Ebenso werden andere Klassen als Blätter ausgewählt, wovon eine nicht in den Testdaten vorkommt. Die schlechten Ergebnisse bei der Anwendung des Simple Trees auf die Testdaten sind wieder durch den Umstand der nicht vollständigen Abbildung aller Klassen zu erklären. Die Regeln für den Medium und Complex Tree, die bei diesem Datenset deutlich besser abschnitten und vor allem der Medium Tree die maximale Genauigkeit dieses Datensets liefert, sind im Anhang ab Seite gg zu finden.

[293] Quelle: eigene Darstellung

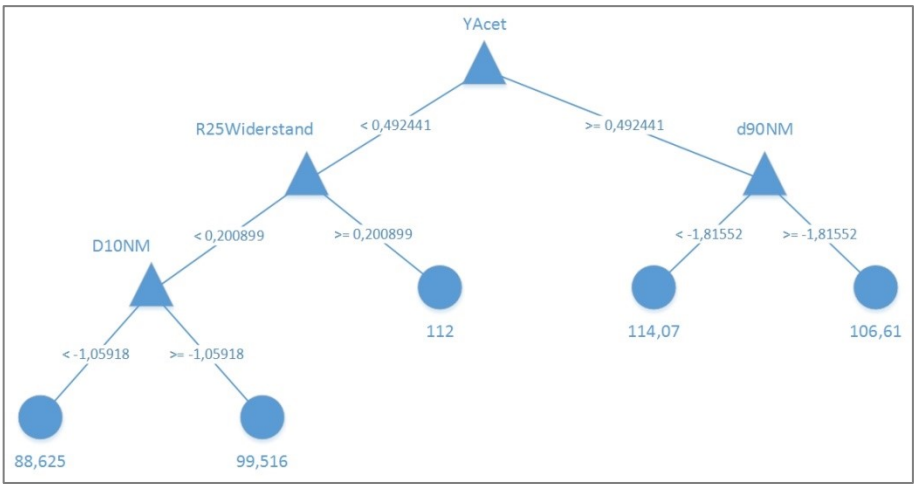


Abbildung 46: Simple Tree für Datenset VM+NM ohne Ausreißer[294]

Die Ergebnisse des Deutschlandsberger Datensets sind auf Grund der geringen Genauigkeit nicht zufriedenstellend. In einem weiteren Schritt werden die in diesem Kapitel entstandenen Datensets als Grundlage für die Erstellung von Modellen für Šumperk bzw. einer Datenbasis für diese verwendet. Die Vorgehensweise wird im nächsten Kapitel erläutert.

4.3 Standort Šumperk

In Šumperk wird die Sinterfreigabe betrachtet. Diese findet unmittelbar vor und während des Sinterprozess im für die Produktion eingesetzten Durchstoßofen statt und bildet die Grundlage für eine eventuelle Änderung der Prozesseinstellungen, um die Produktspezifikationen u. a. hinsichtlich R_{25} zu treffen. Die PTC-Bauteile haben, anders als bei der Granulatfreigabe, die für diesen Produkttypen vorgesehene Geometrie. Durch diese anderen Maße bzw. andere Oberfläche gibt es Änderungen zwischen dem Verhalten bei der Sinterung für die Granulatfreigabe und der Sinterung für Sinterfreigabe. Daher kann es zu anderen Korrelationskoeffizienten zwischen den Attributen und dem neuen Zielattribut, dem gesinterten Widerstand, als zwischen den Attributen und R25 Vergl.% kommen. Das Ziel ist es, den gesinterten Widerstand vorherzusagen, und somit die Sinterfreigabe nicht mehr durchführen zu müssen. Um ein neues Modell, das auf diesem Ziel und diesen Zusammenhängen beruht, erstellen zu können, wird im Folgenden die Aufbereitung des Zielattributs erklärt. Anschließend wird das Schaffen einer neuen Datenbasis für die Modellerstellung beschrieben und die entstandenen Modelle ausgewertet.

[294] Quelle: eigene Darstellung

4.3.1 Aufbereitung des Zielattributs

Das Zielattribut, der gesinterte Widerstand, wird für die Sinterfreigabe aufgezeichnet. Dies ist eine laufende Überprüfung während des Sinterprozesses. Je Freigabemessung werden 15 Bauteile gemessen, die je Charge aufgezeichnet und mit einer laufenden Nummer versehen werden. Für die Erstellung des Modells und das Ziel die Sinterfreigabe obsolet werden zu lassen, ist nur die erste Messung je Charge zu betrachten. Aus diesem Grund liegen je Charge bzw. Datensatz 15 Messwerte für den gesinterten Widerstand vor. Dieser gesinterte Widerstand hat den Sollwert von 15,01.

Als erstes werden diese 15 Messwerte von Ausreißern bereinigt. Es werden jene Messwerte als Ausreißer bezeichnet, die den Wert „-99“ (Messung ungültig) oder einen Wert größer als 50 haben. Da nur ein Wert vorhergesagt werden soll, wird aus diesen 15 Messwerten je Datensatz das arithmetische Mittel gebildet. Aus diesem Grund wird das Zielattribut, der gemittelte gesinterte Widerstand, im Weiteren als MW bezeichnet. Die Verteilung von MW ist in Abbildung 47 zusehen.

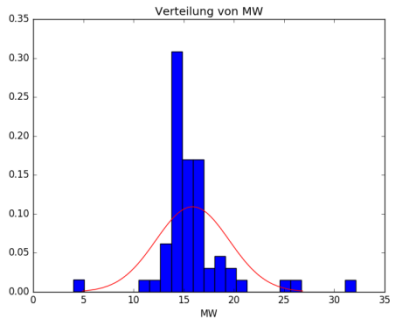


Abbildung 47: Histogramm für gemittelten gesinterten Widerstand (MW)[295]

Die Korrelation von MW mit dem Zielattribut R25 Vergl.% des Kapitels 4.2 ist in Abbildung 48 zu sehen und hat einen Korrelationskoeffizienten von 0,39. Es liegt eine positive Korrelation vor (siehe Kapitel 3.3.3), die im mittleren Ausprägungsbereich liegt.

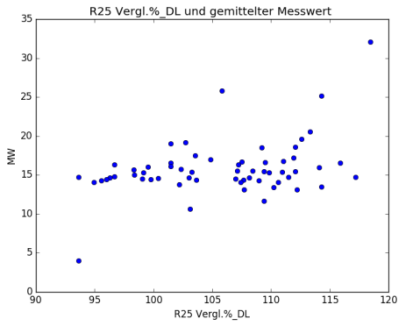


Abbildung 48: Streudiagramm für R25 Vergl.% und MW[296]

Der MW wird keiner weiteren Ausreißeranalyse unterzogen, obwohl die Notwendigkeit in Abbildung 47 deutlich wird. Hierbei wird an der Überlegung aus Kapitel 4.2.4 und

[295] Quelle: eigene Darstellung
[296] Quelle: ebda.

Kapitel 4.2.5 festgehalten, dass auf Basis der unveränderten Ausprägungen des Zielattributs die Gruppierung vorgenommen wird.

Tabelle 19: Einteilung von MW in Klassen[297]

Bezeichnung	Klassenmitte	Untergrenze	Obergrenze	Anzahl von Datensätzen pro Klasse	Breite
11.66375	8.9	4.03	13.77	8	9.74
14.24125	14.25	14.08	14.42	8	0.34
14.58875	14.58	14.45	14.71	8	0.26
15.08875	15.07	14.74	15.4	8	0.66
15.68	15.75	15.45	16.05	8	0.6
16.5	16.44	16.08	16.8	8	0.72
18.33625	18.31	16.97	19.65	8	2.68
25.9075	26.335	20.56	32.11	4	11.55

Die in Tabelle 19 angeführten Klassen beruhen wie die Einteilung in Kapitel 4.2.5 auf (3.1) und der Tatsache, dass 60 Datensätze vorhanden sind:

$$k_{Sumperk} = \sqrt{n_{Sumperk}} = \sqrt{60} \approx 8$$

Nach dieser Gruppierung und Einteilung werden das Zielattribut MW und die Deutschlandsberger Daten der Vor- und Nachmahlung zusammengeführt.

4.3.2 Datenzusammenführung

Die Deutschlandsberger Daten dienen als Grundlage für die Erstellung eines Datensets für Šumperk. In Abbildung 31 wird der Aufbau der Ausgangsdaten beschrieben. Durch die Zusammenführung der Vor- und Nachmahlung stellt sich dieser Aufbau für Šumperk wie in Abbildung 49 dar. Einerseits ist es möglich, nur jene 62 Attribute (ohne UD und R25 Vergl.%) zu verwenden, die in Kapitel 4.2.7 die Basis für die Modellierung sind. Andererseits kann mit sämtlichen Attributen, die nach der Zusammenführung der Daten in Kapitel 4.2.2 (angeführt in Tabelle 29), ein erneuter KMO-Test durchgeführt werden.

[297] Quelle: eigene Darstellung

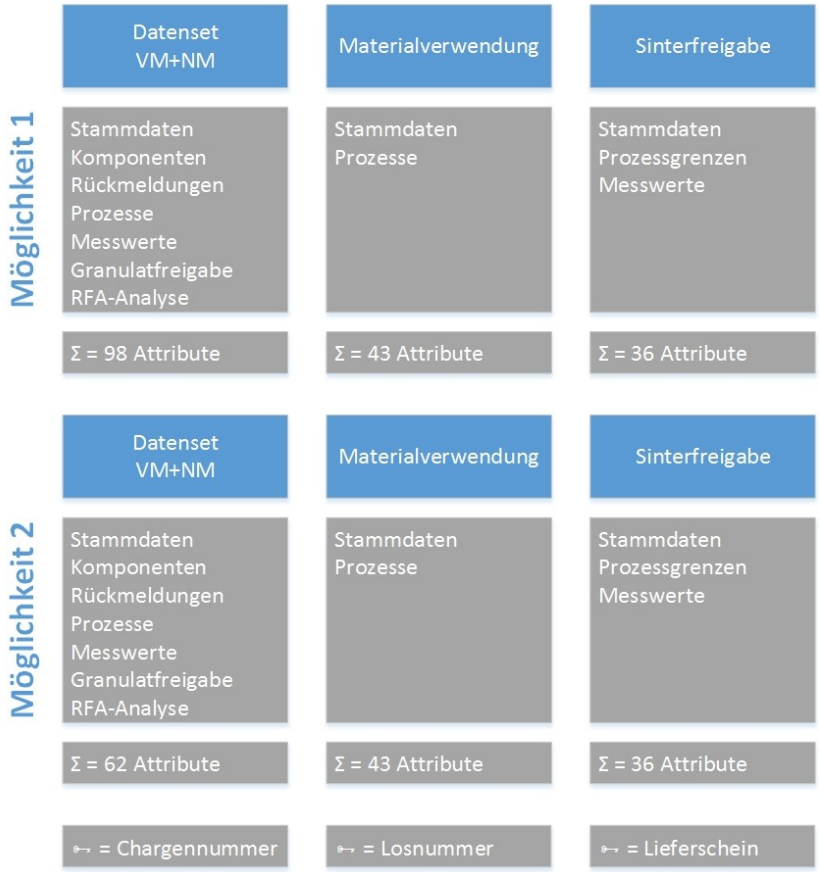


Abbildung 49: Aufbau der Ausgangsdaten für Šumperk[298]

Ebenso muss beachtet werden, ob die Datensets mit oder ohne Ausreißer verwendet werden. Zu diesem Zeitpunkt wird entschieden, dass die Datensets mit Ausreißern verwendet werden, da bis zu vier Chargen aus der Nachmahlung in eine Charge für die Sinterfreigabe einfließen und die Ausreißer dadurch ausgeglichen werden könnten. Die Chargen- und Losnummern werden im Datenset Materialverwendung miteinander verknüpft. Diese Beziehungen werden verwendet, um das arithmetische Mittel aller 98 bzw. 62 Attribute der einfließenden Chargen, das über die Gewichtsanteile gewichtet wird, zu berechnen. Zusätzlich wird der Ofen, auf dem die Freigabe stattgefunden hat, als Attribut aufgenommen.

Die entstandenen Datensets bestehen aus einer veränderlichen Anzahl von Attributen (64 bzw. 100) und 60 Datensätzen. Sämtliche Attribute, die in diesem Datenset das Suffix „_DL“ in ihrer Bezeichnung (siehe Tabelle 36) enthalten, sind jenen im Datenset der Vor- und Nachmahlung ohne Suffix gleichzusetzen. Im nächsten Schritt werden die neuentstandenen Datensets ausgewertet.

[298] Quelle: eigene Darstellung

4.3.3 Auswertung

Wie bereits in Kapitel 4.2.3 wird das Datenset hinsichtlich der statistischen Maße aus Kapitel 3.3.2 und 3.3.3 ausgewertet. Dies wurde zuerst am Datenset mit Ausreißern durchgeführt. Der Vollständigkeit halber wird hier bereits die Auswertung des Datensets ohne Ausreißer, das durch das Zusammenführen der Daten der Vor- und Nachmahlung ohne Ausreißer entstanden ist, angeführt. Allerdings wird erst zu einem späteren Zeitpunkt in dieser Arbeit entschieden, dieses zu verwenden. Auf eine explizite Auswertung des Datensets mit 64 Attributen und 60 Datensätzen wird verzichtet, da dieses ein Subset des 100 Attribute/60 Datensätze-Datensets mit Ausreißern darstellt. Allerdings wird der Ofen nicht mit einbezogen, da dieser ein nicht-metrisches Attribut darstellt und die Maße daher nicht berechnet werden können (siehe Kapitel 3.3.2 und Kapitel 3.3.3).

Datenset mit Ausreißern

In diese Auswertung fließen 99 Attribute und 60 Datensätze ein, deren Grundlage das Deutschlandsberger Datenset mit Ausreißern ist.

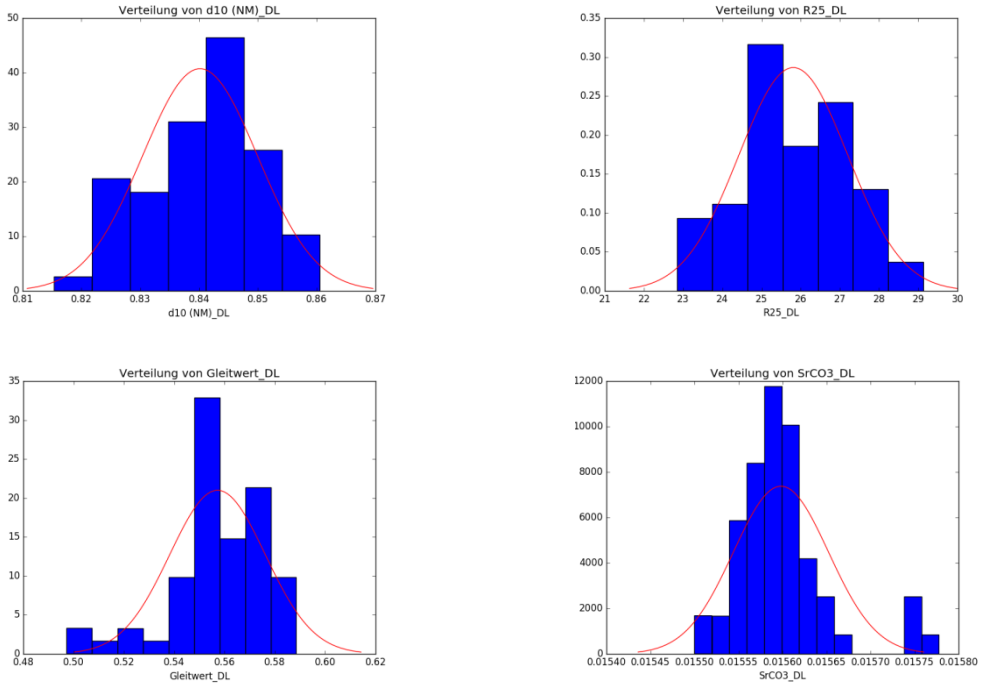


Abbildung 50: Beispiele für Histogramme für Šumperk mit Ausreißern[299]

Aufgrund der geringeren Anzahl von Datensätzen verändert sich die Verteilung der Attribute. In Abbildung 50 werden wieder die vier gleichen Attribute wie in Abbildung 33 und Abbildung 37 dargestellt. Die Ausreißer sind trotz der Mittelwertbildung vor allem im Gleitwert_DL und SrCO3_DL wiederzufinden. Ob es sich um normalverteilte Daten handelt, wird nicht explizit mittels Kolgomorov-Smirnov-Test überprüft. Es werden lediglich die im Anhang in Tabelle 36 angeführte Wölbung und Schiefe betrachtet. Diese lassen den Schluss zu, dass die Daten nicht normalverteilt sind.

[299] Quelle: eigene Darstellung

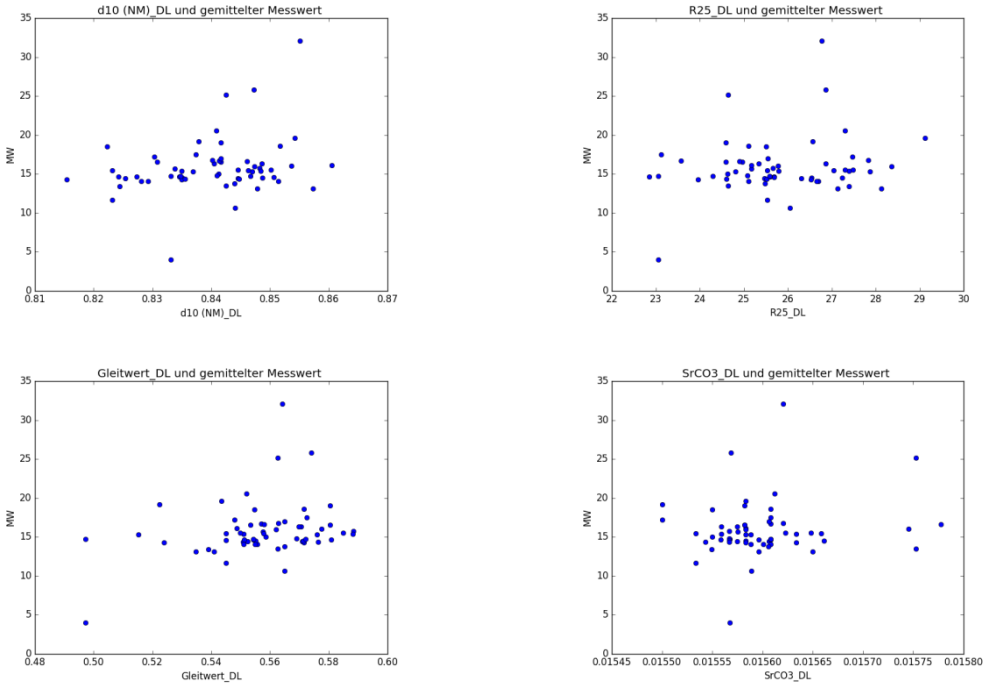


Abbildung 51: Beispiele für Streudiagramme für Šumperk mit Ausreißern[300]

Die Zusammenhänge zwischen MW und den ausgewählten Attributen sind in den Streudiagrammen in Abbildung 51 zu sehen. Die Beziehungen zwischen den MW und den jeweiligen Attributen ist nicht so klar ersichtlich wie in Abbildung 35 und Abbildung 39. Dies liegt u. a. an der geringeren Anzahl von Datensätzen und der nicht stark ausgeprägten Korrelation von R25 Vergl.% und MW. Bei der Betrachtung der Korrelationskoeffizienten in Tabelle 36 wird klar, dass zwischen Gleitwert_DL und MW die am stärksten ausgeprägte Korrelation mit 0,29 innerhalb dieser vier Attribute herrscht.

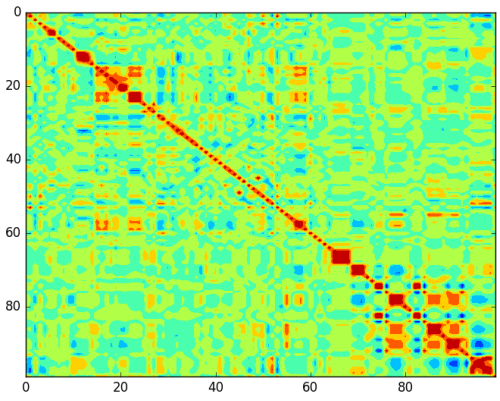


Abbildung 52: Korrelationsmatrix (graphisch) für Šumperk mit Ausreißern[301]

Bei der Betrachtung der Korrelationsmatrix sind Parallelen zu Abbildung 36 zu erkennen. Die Reihenfolge der Attribute ist zu Tabelle 29 bzw. Abbildung 36

[300] Quelle: eigene Darstellung
[301] Quelle: ebda.

unverändert, einzig das Attribut MW steht nun an erste Stelle (siehe Tabelle 36). Hier bestätigt sich, dass die Trends aus dem Datenset der Vor- und Nachmahlung weiter geführt weden (siehe Kapitel 4.2.3). Zusätzlich gibt es eine höhere Anzahl negativer Korrelationskoeffizienten (blau dargestellt). Die Korrelation um Punkt (20,20) sind jene Attribute, die bei der Granulatfreigabe erhoben werden. Diese korrelieren ebenso um den Punkt (60,20) mit den Attributen d50 (VM)_VM_DL, d90 (VM)_VM_DL und pH Wert Schlicker_VM_DL, wobei die ersteren beiden auch hoch miteinander korrelieren. Um den Punkt (68,68) sind wieder die Zusammenhänge zwischen den Heizgruppen ersichtlich. Die diagonal anschließende dunkelrote Fläche steht für die Korrelation der statistischen Maße, die für den Differenzdruck aufgezeichnet wurden. Dieser Trend setzt sich in den nachfolgenden diagonalen auffälligen Flächen fort, da hier jeweils die Maße Minimum, Maximum und Standardabweichung eines Attributs angegeben wurde. Die Korrelation dieser Attribute untereinander ist sehr viel negativer ausgeprägt als bei der Korrelationsmatrix des Datensets der Vor- und Nachmahlung in Abbildung 36.

Datenset ohne Ausreißer

In die folgende Betrachtung fließen 100 Attribute und 60 Datensätze ein, deren Grundlage das Deutschlandsberger Datenset ohne Ausreißer ist.

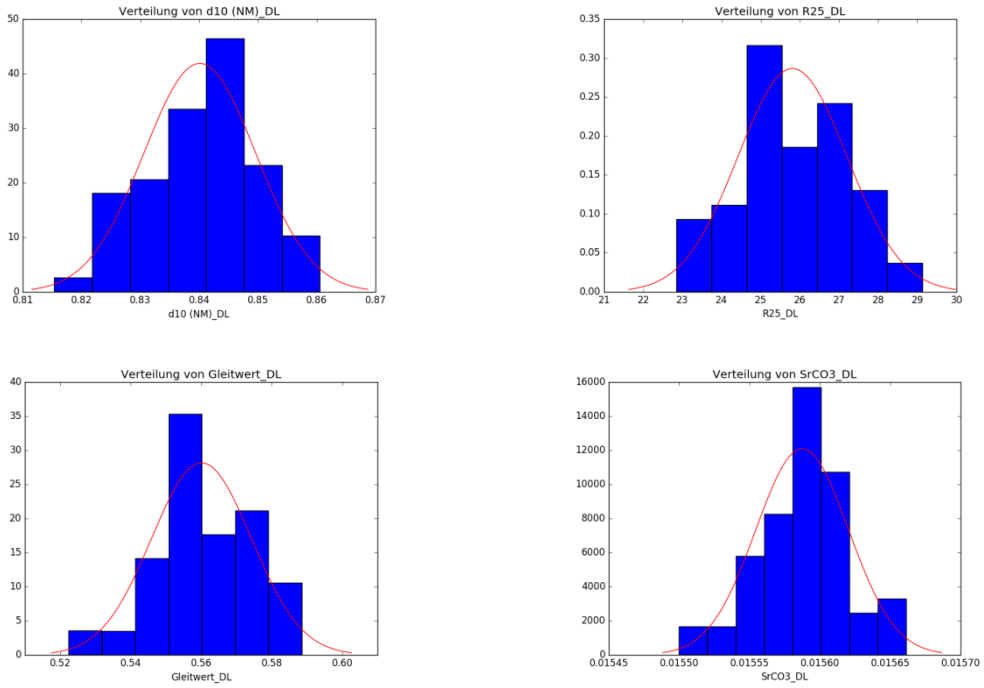


Abbildung 53: Beispiele für Histogramme für Šumperk ohne Ausreißer[302]

In Abbildung 53 sind die Histogramme des neu entstandenen Datensets zu sehen. Im Vergleich mit Abbildung 50 ist eine Veränderung für die Attribute Gleitwert_DL und SrCO3_DL ersichtlich. Die Ursache hierfür ist die Eliminierung der Ausreißer in Kapitel 4.2.4. Gleiches gilt für die in Abbildung 54 zu sehenden Streudiagramme.

[302] Quelle: eigene Darstellung

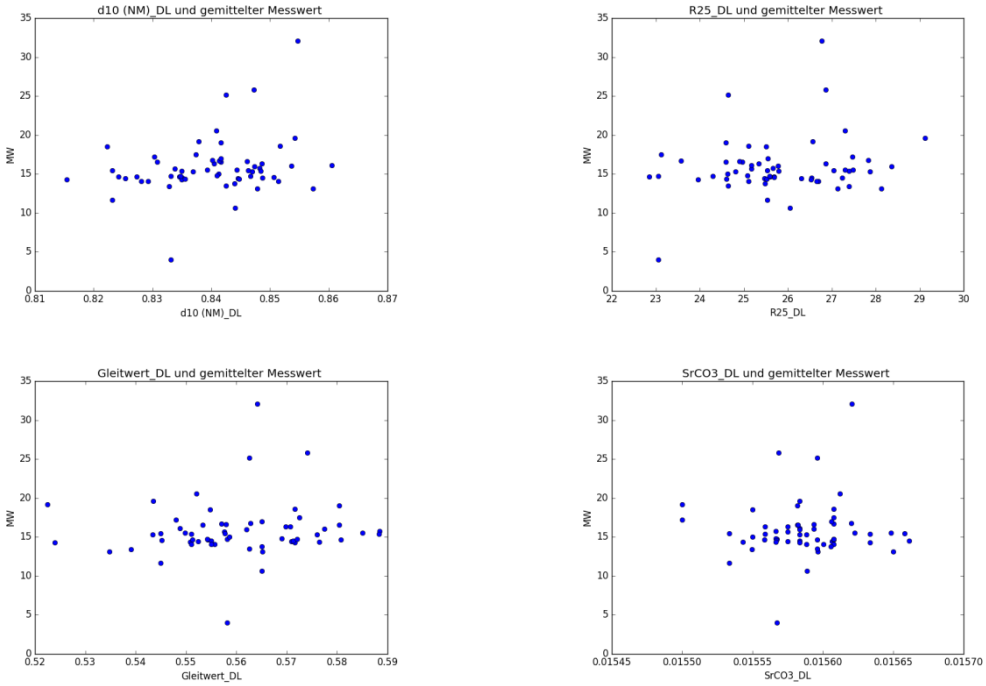


Abbildung 54: Beispiele für Streudiagramme für Šumperk ohne Ausreißer[303]

Die Änderung der Streudiagramme in Abbildung 54 ist ebenso in der Korrelationsmatrix in Abbildung 55 wiederzufinden. Hier ist die links die Korrelationsmatrix mit Ausreißern aus Abbildung 52 und rechts die Korrelationsmatrix ohne Ausreißer zu dargestellt. Die Veränderung der Korrelationsmatrizen in Abbildung 55 folgen demselben Muster wie jene in Abbildung 40. Die ausgeprägten Zusammenhänge in der rechten unteren Ecke werden aufgebrochen.

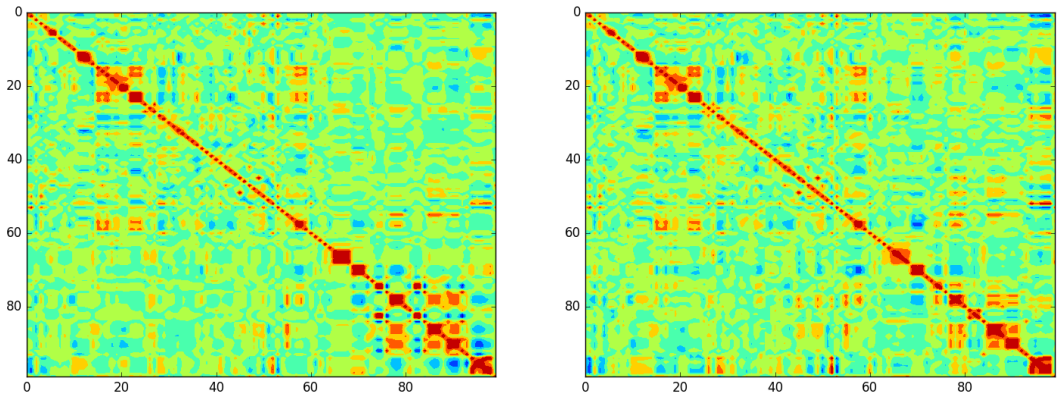


Abbildung 55: Korrelationsmatrizen (graphisch) für Šumperk mit und ohne Ausreißer[304]

Nach dieser kurzen Betrachtung der beiden Datensets wird die Eignung des Datensets mit Ausreißern für die PCA überprüft.

[303] Quelle: eigene Darstellung
[304] Quelle: ebda.

4.3.4 Eignung für PCA

Die Vortests für die PCA wurden in Kapitel 3.4.4 theoretisch betrachtet und in Kapitel 4.2.6 am Beispiel der Deutschlandsberger Daten durchgeführt. Die Vorgehensweise stimmt mit jener in Kapitel 4.2.6 überein. Die Basis für den KMO-Test bildete das Datenset mit 100 Attributen/60 Datensätzen mit Ausreißern, wobei das Attribut Ofen ausgenommen wird, und wird nicht mit dem Datenset ohne Ausreißer durchgeführt. Der Grund hierfür liegt in der in Kapitel 4.2.7 gewonnenen Erkenntnis, dass die Ergebnisse noch schlechter vergleichbar sind, wenn unterschiedliche Attribute für die Modellerstellung herangezogen werden. Für den Zweck der KMO werden der Ofen und das Zielattribut MW ausgeschlossen.

Für das vorliegende Datenset wurde der KMO-Test viermal durchgeführt. Attribute mit $MSA < 0{,}5$ wurden ausgeschlossen. Eine genaue Aufschlüsselung über die Werte sind der Tabelle 37 im Anhang zu entnehmen. Die Entwicklung über die Anzahl der Attribute ist in Tabelle 20 zusammengefasst.

Tabelle 20: Entwicklung der Anzahl der Attribute (Šumperk mit Ausreißern)[305]

	Durchgang			
	1	**2**	**3**	**4**
Anzahl zu löschende Attribute	61	28	1	0
Anzahl Attribute nach Löschen	38	10	9	9

Das MSA-Kriterium für das gesamte Datenset hat sich, wie in Tabelle 21 dargestellt, verändert. Wie bereits beim Datenset der Vor- und Nachmahlung ohne Ausreißer in Kapitel 4.2.6, ist es im ersten Durchgang nicht möglich, das MSA-Kriterium für das gesamte Datenset zu berechnen. Dies wird auf gleiche Art gelöst: Durch Löschen jener Attribute, die die Berechnung der Inversen der Korrelationsmatrix nicht zulassen.

Tabelle 21: MSA für Datenset (Šumperk mit Ausreißern)[306]

	Durchgang			
	1	**2**	**3**	**4**
MSA	nan	0,3730	0,7009	0,7325

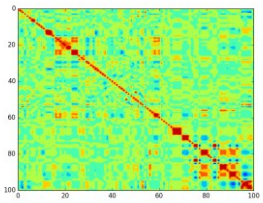
Es ist eine deutliche Verbesserung des MSA-Kriteriums zu sehen und ist nur unwesentlich schlechter als jener der Vor- und Nachmahlung ohne Ausreißer in Tabelle 14. Das Ergebnis dieses Tests ist ein Datenset mit 9 Attributen und 60 Datensätzen, dem die Attribute MW und Ofen wieder hinzugefügt werden. Dadurch entsteht für die anschließende Modellierung ein Datenset mit 10 Prädiktoren, einem vorherzusagenden Zielattribut mit 60 Datensätzen.

4.3.5 Modellierung

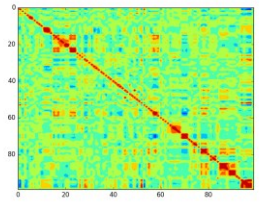
Die Modellierung erfolgte wie in Kapitel 4.2.7 wieder in zwei Schritten und wurde in Matlab in der Classification Learner App durchgeführt. Es stehen dieselben Methoden zur Auswahl (siehe Tabelle 15 und Tabelle 16) und in einem ersten Schritt wird eine generelle Abschätzung getroffen, welches Datenset verwendet werden soll.

[305] Quelle: eigene Darstellung
[306] Quelle: ebda.

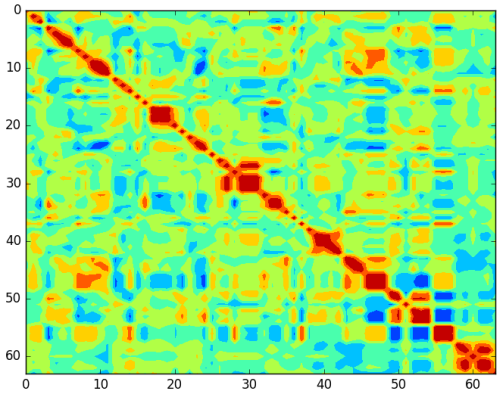
Anschließend werden auf dieses und das passende Gegenstück sämtliche Methoden angewendet, um Modelle zu erstellen. Diese Modelle werden zudem mit den Testdaten verifiziert. Bei den Testdaten handelt es sich um 25 Datensätze, die wie bereits die Testdaten in Kapitel 4.2.7 aus dem Zeitraum von September 2015 bis Anfang Jänner 2017 stammen. Sie werden wie in Kapitel 4.3.2 beschrieben, aufbereitet und standardisiert, wobei auch hier keine Ausreißeranalyse vorgenommen wird.



Trainingsdaten mit Ausreißern



Trainingsdaten ohne Ausreißer



Testdaten

Abbildung 56: Korrelationsmatrizen für Šumperk[307]

Abbildung 56 zeigt die bereits bekannten Korrelationsmatrizen aus Abbildung 52 bzw. Abbildung 55. Wie in Kapitel 4.2.7 fällt die anders strukturierte Korrelationsmatrix der Testdaten gegenüber jenen der Trainingsdaten auf. Auch hier kann die als Indiz dafür gewertet werden, dass die Generalisierbarkeit des Modells, das auf den Trainingsdaten beruht, nicht gegeben ist.

Tabelle 22: Auswertung der Testdaten (Anzahl je Klasse)[308]

Bezeichnung	Anzahl Datensets je Klasse
11.66	2
14.24	1
14.59	3
14.73	1
15.09	7
15.68	1
16.50	5
18.34	4
25.91	1

Ebenso wird das Zielattribut in Klassen eingeteilt. Eine Aufschlüsselung über die Anzahl der Datensets je Klasse sind in Tabelle 22 zu sehen. Wieder gibt es einen Grenzfall (rot hinterlegt), der nicht zugeordnet werden kann.

[307] Quelle: eigene Darstellung
[308] Quelle: ebda.

Vergleich verschiedener Modelle und Datensets

Für die Daten für Šumperk liegen zu diesem Zeitpunkt 2 Datensets vor: jenes, das durch die Ausführung des KMO Tests entsteht, und jenes, dessen Attribute von der Vor- und Nachmahlung übernommen werden. Die Kombinationsmöglichkeiten, mit denen eine erste Abschätzung stattfindet, sind in Tabelle 23 zu sehen. Auf die Durchführung der PCA wird bei dieser Abschätzung verzichtet, da bereits in Kapitel 4.2.7 auf dies verzichtet wurde.

Tabelle 23: Kombinationen für Šumperk[309]

	Datenset	Ausreißer	Kreuz-validierung	PCA	Ofen	Anzahl Prädika-toren
CZ 5-fold inkl. Ausreißer + Ofen (Prädiktoren: 10)	KMO	mit	5	nein	ja	10
CZ 10-fold inkl. Ausreißer + Ofen (Prädiktoren: 10)	KMO	mit	10	nein	ja	10
CZ 5-fold inkl. Ausreißer + Ofen (Prädiktoren: 63)	VM+NM	mit	5	nein	ja	63
CZ 10-fold inkl. Ausreißer + Ofen (Prädiktoren: 63)	VM+NM	mit	10	nein	ja	63

Die in Tabelle 23 angegebenen Datensets werden als Trainingsdaten für die Modellerstellung verwendet. In Abbildung 57 sind die Ergebnisse sehen. Auf der Abszisse sind ist die Genauigkeit der Modelle aufgetragen und auf der Ordinate die verwendeten Modelle. Hierbei wird sich auf die Modelle, die bereits in Kapitel 4.2.7 bei der endgültigen Modellerstellung verwendet werden, konzentriert. Dazu gehören die verschiedenen Bäume, die Varianten der SVM und die Ensemblemethoden.

[309] Quelle: eigene Darstellung

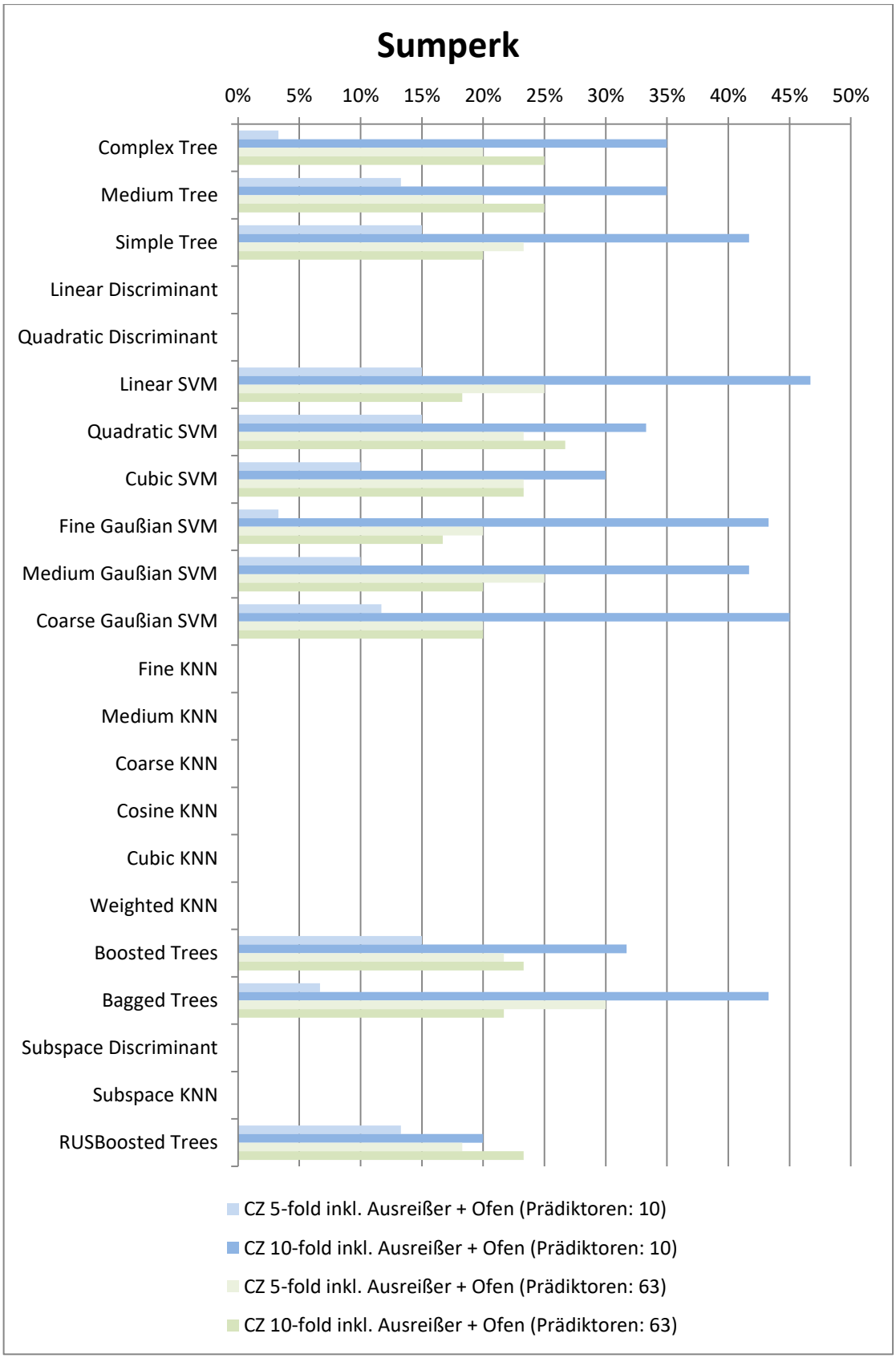


Abbildung 57: Vergleich verschiedener Kombinationsmöglichkeiten für Šumperk[310]

[310] Quelle: eigene Darstellung

Nach dieser ersten Abschätzung können folgende Überlegungen angestellt werden:

- Die Verbesserung zwischen den Datensets, die 10 Prädiktoren aber unterschiedliche Anzahl an Kreuzvalidierungsmengen haben, beruht höchstwahrscheinlich auf Overfitting.
- Die Anzahl der Kreuzvalidierungsmengen hat auf das Datenset mit 63 Prädikatoren keine nennenswerte Auswirkung.
- Im Vergleich beider Datensets mit der 5-fachen Kreuzvalidierung schneidet jenes mit den 63 Prädiktoren besser ab.

Auf Basis dieser Überlegungen und vorrangig auf der letzten wird entschieden, dass das Datenset mit den 63 Prädiktoren für die Modellerstellung und die anschließende Anwendung auf die Testdaten verwendet wird. In Anlehnung an die Vorgehensweise in Kapitel 4.2.7 werden wieder die Ergebnisse für die Datensets mit und ohne Ausreißer gegenüber gestellt. Daher werden die bereits in Kapitel 4.3.3 aufbereiteten Daten in Folge verwendet.

Ergebnisse des Datensets mit Ausreißern

Die Ergebnisse für das Datenset mit Ausreißern für Šumperk sind nach dem gleichen Prinzip wie die Ergebnisse aus Kapitel 4.2.7 in Abbildung 43 und Abbildung 45 aufbereitet.

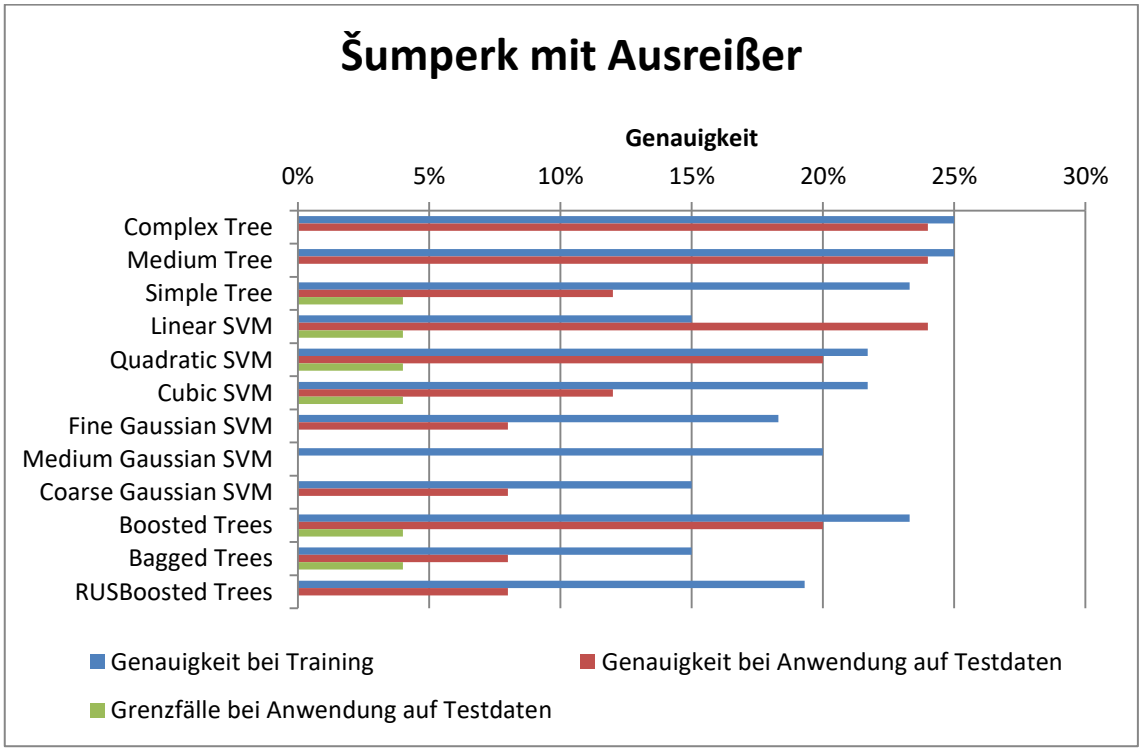


Abbildung 58: Ergebnisse des CL für Šumperk mit Ausreißern[311]

In Abbildung 46 sind die Genauigkeiten beim Training (blau) und bei der Anwendung auf die Testdaten (rot) und deren Grenzfälle (grün) dargestellt. Die genauen Daten hinter dem Diagramm sind der Tabelle 38 zu finden. Complex Tree (25 %), Medium Tree (25 %) und Boosted Trees (23,3 %) liefern die besten Ergebnisse beim Training.

[311] Quelle: eigene Darstellung

Dieser Trend setzt sich bei der Anwendung auf die Testdaten fort (24 %, 24 % und 20 %). Zusätzlich hat die Linear SVM dasselbe Ergebnis (24 %). Die Medium Gaussian SVM versagt trotz 20 % bei den Trainingsdaten bei der Anwendung auf die Testdaten. Die durchschnittliche absolute Abweichung zwischen vorhergesagter und Ist-Klasse liegt bei 1,94 bis 3,42 im ähnlichen Bereich wie die der bereits betrachteten Ergebnisse aus Kapitel 4.2.7. Eine genauere Betrachtung der Ergebnisse je Modell und Charge ist den Abbildung 86 bis Abbildung 97 zu entnehmen.

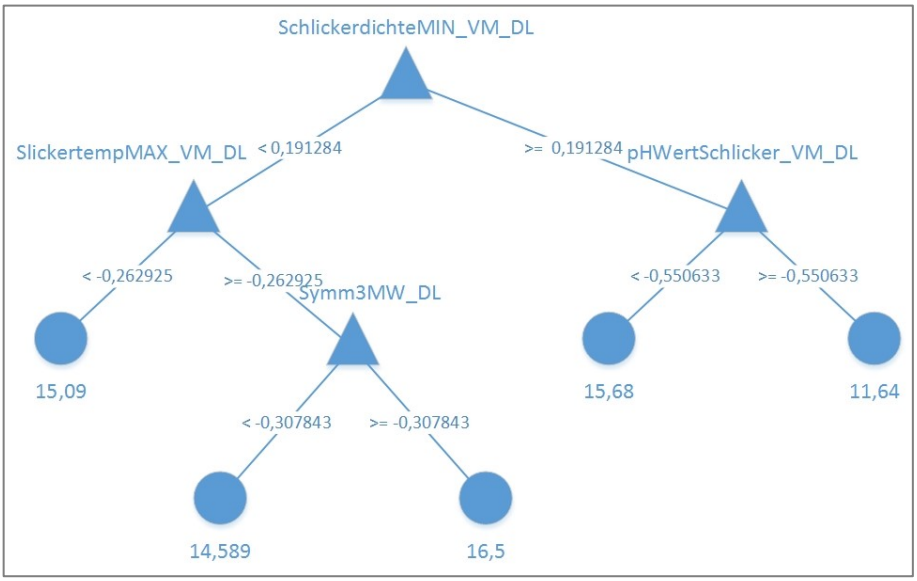


Abbildung 59: Simple Tree für Šumperk mit Ausreißern[312]

Bei der Betrachtung des Simple Tree in Abbildung 59 wird wieder ersichtlich, dass dieser nur 5 Blätter hat, die 5 der 8 im Datenset enthaltenen Klassen repräsentieren. Diese prozentual höhere Abdeckung der Klassen kann zu einem besseren Ergebnis führen als bei den Datensets der Vor- und Nachmahlung. Dennoch wird diese Methode wegen der Unvollständigkeit verworfen.

Bei der Betrachtung der Regeln für Medium Tree und Complex Tree ab Seite ww im Anhang wird ersichtlich, dass diese weniger verzweigt sind als jene für die Vor- und Nachmahlung. Beide enthalten alle Klassen als Blätter.

Ergebnisse des Datensets ohne Ausreißer

Das Datenset ohne Ausreißer wird wie die anderen drei Datensets zuvor aufbereitet. Die Aufbereitung der Ergebnisse des Training, und der Anwendung auf die Testdaten wird in Abbildung 60 zusammengefasst.

[312] Quelle: eigene Darstellung

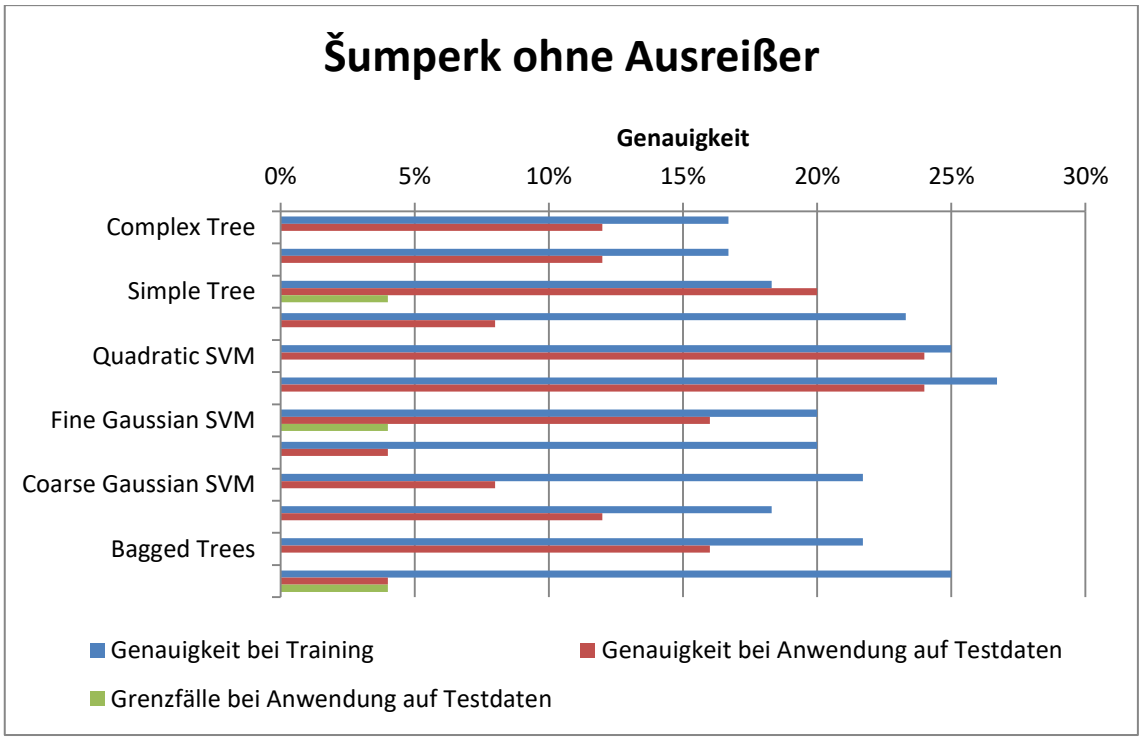


Abbildung 60: Ergebnisse des CL für Šumperk ohne Ausreißer[313]

Die Genauigkeiten beim Training (blau) und bei der Anwendung auf die Testdaten (rot) und deren Grenzfälle (grün) sind in Abbildung 60 sind zu sehen. Die genauen Daten hinter dem Diagramm sind der Tabelle 39 zu finden. Die Ergebnisse beim Training waren relativ konstant, aus denen sich die Linear SVM (23,3 %), Quadratic SVM (25 %) und Cubic SVM (26,7 %) sowie die RUSBoostedTrees (25 %) hervorheben. Dieser Trend wird bei der Anwendung auf die Testdaten nur bei der Quadratic SVM und der Cubic SVM (beide 24 %) beobachtet. Die durchschnittliche absolute Abweichung zwischen vorhergesagter und Ist-Klasse lag bei 1,82 (Cubic SVM) bis 3,62 (Coarse Gaussian SVM) im ähnlichen Bereich wie die der bereits betrachteten Ergebnisse aus Kapitel 4.2.7 und bei den Ergebnissen mit Ausreißer. Eine genauere Betrachtung der Ergebnisse je Modell und Charge ist im Anhang in Abbildung 98 bis Abbildung 109.

[313] Quelle: eigene Darstellung

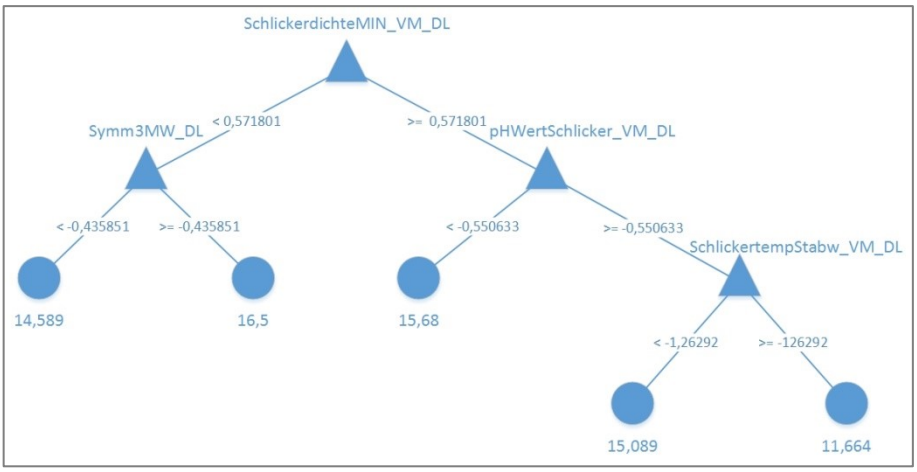


Abbildung 61: Simple Tree für Šumperk ohne Ausreißer[314]

Abbildung 61 zeigt den entstandenen Simple Tree, der wieder nur fünf Blätter hat. Somit werden drei Klassen nicht abgebildet. Dies ist ein Ausschlussgrund für das Modell. Ab Seite fff sind die Regeln für den Medium Tree und Complex Tree zu finden. Wie beim Šumperker Datenset mit Ausreißern ist zu erkennen, dass diese weniger verzweigt sind als jene für die Vor- und Nachmahlung. Beide enthalten alle Klassen als Blätter.

Nach der Aufbereitung der Daten und dem Erstellen der Modelle für Deutschlandsberg und Šumperk wird klar, dass diese nicht die gewünschten Ergebnisse erbringen. Daher ist von einem Einsatz der Modelle abzuraten. Im nächsten Kapitel wird die zusätzliche Fertigungskapazität berechnet, die ein genaues Modell freiwerden lässt.

4.4 Produktionswirtschaftliche Betrachtung

Innerhalb dieses Kapitels soll geklärt werden, welche Auswirkungen ein korrekt vorhergesagter Zielwert auf die Fertigungskapazitäten hätte. Die folgenden Berechnungen konnten nicht in der Praxis überprüft werden.

Im Fall der Granulatfreigabe in Deutschlandsberg würden keine Fertigungskapazitäten freiwerden, da die Sinterung für die Freigabe nicht auf in den Fertigungsprozess integrierten Öfen durchgeführt wird. Einzig könnte es sich auf die Durchlaufzeit auswirken. Allerdings nur unter der Voraussetzung, dass die Granulate direkt nach der Produktion nach Šumperk transportiert und hier ohne Lagerzeit verarbeitet werden.

Im Fall der Sinterfreigabe sind die Voraussetzungen anders. Die Sinterung für die Sinterfreigabe wird auf den Durchstoßöfen durchgeführt, die ebenso in der Produktion eingesetzt werden. Hierbei werden Kapazitäten blockiert, da bis zur korrekten Einstellung der Öfen nur jeder 50. Platz im Ofen besetzt wird. Bei einer korrekten Vorhersage wäre die Sinterfreigabe nicht mehr notwendig und die schlecht ausgelasteten Freigabesinterungen sind hinfällig. Dies wird im Folgenden an Hand der Sinterdaten des bisher betrachteten Produkts auf Basis der Produktionslose, auf denen

[314] Quelle: eigene Darstellung

die Testdaten beruhen, erläutert. Dabei ist der Ofen das einzige betrachtete Aggregat, der gleichzeitig den Engpass darstellt (siehe Punkt 1b) in Tabelle 2).

Tabelle 24: Auswertung der Bestückung des Ofens[315]

	Bestückung des Ofens		
	1	**10**	**50**
Anzahl der Messungen	1967	107	32

In Tabelle 24 werden die Sinterdaten hinsichtlich der Bestückung der Öfen und die Anzahl der Messungen ausgewertet. Die Bestückungen 1, 10 und 50 bedeuten, dass jeder, jeder zehnte oder jeder fünfzigste Platz belegt wird. Bei der Belegung jedes Platzes wird nur jeder zehnte Platz gemessen. Dies bedeutet für die Summe der benötigten Plätze folgendes:

$$\text{Summe der benötigten Plätze} = 1967 * 10 + 107 + 32 = 19809 \text{ Plätze}$$

Wird vom Idealfall ausgegangen, dass ein Modell die korrekte Sinterung vorhersagt, die Sinterfreigabemessung entfällt und somit jeder Platz belegt werden kann, würden folgende Kapazitäten freiwerden:

$$\text{leere Plätze} = (107 * 9) + (32 * 49) = 2531 \text{ Plätze}$$

Diese leeren Plätze haben einen Anteil von 12,78 % an der benötigten Kapazität. Wird davon ausgegangen, dass ein Los im Mittel 792 Plätze (ohne leere Plätze) im Ofen benötigt, hätten 3 Lose mehr im betrachteten Zeitraum (September 2016 bis Anfang Jänner 2017) produziert werden können.

Zusätzlich wirkt sich der Wegfall der leeren Plätze innerhalb der Produktion auch auf die Durchlaufzeit aus. Dies würde zu einer durchschnittlichen Beschleunigung um 101 Plätze je Los führen.

Würde eine Anpassung der Produktion durch ein funktionierendes Modell vorgenommen werden, würde es sich um eine qualitative Anpassung handeln, die wiederum in einer intensitätsmäßigen Kapazitätsanpassung resultieren würde (siehe Kapitel 0).

Nach dieser produktionswirtschaftlichen Betrachtung des Hintergrundes der Arbeit werden im nächsten Kapitel die gewonnenen Erkenntnisse zusammengefasst. Zusätzlich soll ein Ausblick zu weiteren Möglichkeiten gegeben werden.

[315] Quelle: eigene Darstellung

5 Zusammenfassung und Ausblick

Innerhalb dieser Masterarbeit wurde die Thematik der Datenanalyse mit dem produktionswirtschaftlichen Ansatz in Verbindung gesetzt. Anreiz dafür war die Aufgabenstellung der EPCOS OHG, ein Verfahren zur Vorhersage der Sinterergebnisse aus Daten, die entlang der Prozesskette bzw. bei Freigabemessungen gesammelt wurden, zu entwickeln. Hierbei wurden die Produktionsstandorte Deutschlandsberg und Šumperk, die beide am Prozess beteiligt sind, betrachtet. Auf Basis dieser Aufgabenstellung wurden die Forschungsfragen in Kapitel 1.2 abgeleitet, die im Folgenden wiederholt werden:

- Ist es möglich, den Widerstandswert R_{25} der PTC-Bauteile für zukünftige Lose auf Basis der bereits gefertigten vorherzusagen?
- Wie wirkt sich eine exakte Vorhersage auf die Kapazitätsauslastung aus?

Dieses Kapitel dient dazu, eine Zusammenfassung über die theoretischen Hintergründe und praktischen Ergebnisse, die zur Beantwortung der Forschungsfragen führen, zu geben. Eine klare Beantwortung der ersten Frage ist nur beschränkt möglich. Die praktische Fallstudie lieferte kein zufriedenstellendes Ergebnis, so ist es denkbar, dass der theoretisch betrachtete Lösungsweg für eine andere und/oder größere Datenbasis funktioniert. Zusätzlich wird am Ende des Kapitels ein Überblick über die Einflussfaktoren und andere Ansätze geben. Die Antwort der zweiten Forschungsfrage kann in der Praxis nicht überprüft werden, da sie ein funktionierendes Modell voraussetzt.

Die dazu notwendige produktionswirtschaftliche Theorie wurde in Kapitel 2 vorgestellt. Sie umfasste die Grundlagen eines Betriebes und die Kennzahlen Durchlaufzeit und Kapazität. Eine Verbesserung durch ein funktionierendes Modell konnte so quantifiziert werden. Zusätzlich wurden die Anpassungsformen der Kapazität betrachtet.

Eine Abfolge von Methoden, die es möglich machen sollten, das in der ersten Forschungsfrage geforderte Verfahren zu entwickeln, wurden in Kapitel 3 präsentiert. Hierbei wurden zunächst die verschiedenen Ausprägungen der Datenanalyse betrachtet, bevor mit CRISP-DM ein Referenzmodell zur Durchführung einer solchen vorgestellt wurde. Anhand dieses wurde das Kapitel in weiterer Folge aufgebaut. Nach der Theorie zur Untersuchung der Daten in Bezug auf ihre Skala und statistischen Eigenschaften und anschließenden Aufbereitung hinsichtlich Ausreißern, Gruppierung und Dimensionalitätsreduktion, wurde das maschinelle Lernen vorgestellt. Hierbei wurden zuerst das Konzept und die Grundlagen erläutert, bevor die Methoden zum Klassieren der Daten vorgestellt wurden. Diese beruhen auf unterschiedlichen Ansätzen und sind oftmals nicht für dieselbe Datenbasis geeignet.

Das praktische Fallbeispiel übertrug die theoretischen Überlegungen in die Praxis. Hierfür wurden, orientiert am CRISP-DM, eine Datenaufbereitung und -analyse, an Daten die vom Unternehmen in aggregierter Form bereitgestellt wurden, vorgenommen und anschließend Modelle auf Basis der Daten aus Deutschlandsberg und Šumperk erstellt. Für beide Standorte erbrachten die Modelle in der Evaluierungsphase kein

zufriedenstellendes Ergebnis. Somit bleibt die erste Forschungsfrage in der Praxis unbeantwortet. Die veränderte Kapazitätsauslastung, die in der zweiten Forschungsfrage thematisiert wird, kann auf Basis der Daten beantwortet, aber nicht in der Praxis überprüft werden. Allerdings hätten dies im betrachteten Zeitraum freie Kapazitäten von fast 13 % zur Folge gehabt.

Dass das Modell als solches nicht verwendbar ist, kann als Resultat mehrerer Einflussfaktoren gesehen werden. Der vordringlichste Einflussfaktor besteht in einem nicht zufriedenstellenden kleinen Verhältnis der Datensätze zu den Attributen.

Tabelle 25: Verhältnisse Datensätze zu Attributen für VM+NM

	Start	Analyse der fehlende Daten	Zusammen-führung	KMO	Modellierung
VM - Datensätze	266	226	-	-	-
VM - Attribute	162	90	-	-	-
NM - Datensätze	265	224	-	-	-
NM - Attribute	115	62	-	-	-
VM+NM - Datensätze	-	-	218	218	218
VM+NM - Attribute	-	-	98	62	63
Verhältnis VM	1,64	2,51	-	-	-
Verhältnis NM	2,30	3,61	-	-	-
Verhältnis VM+NM	-	-	2,22	3,52	3,46

Für das Deutschlandsberger Datenset wurden die Verhältnisse von Datensätzen zu Attributen für die Vor- und Nachmahlung in Tabelle 25 zusammengefasst. Durch die Aufbereitung konnten diese verbessert werden, allerdings nicht im ausreichenden Maße.

Tabelle 26: Verhältnisse Datensätze zu Attribute für Šumperk

	Start	Modellieren
VM+NM - Datensätze	218	-
VM+NM - Attribute	63	-
Sinterfreigabe - Datensätze	141	-
Sinterfreigabe - Attribute	2	-
Šumperk - Datensätze	-	60
Šumperk - Attribute	-	63
Verhältnis VM	3,46	-
Verhältnis NM	70,50	-
Verhältnis VM+NM	-	0,95

Dasselbe gilt für das Šumperker Datenset. Modelle benötigen große Datenmengen bzw. ein besseres Verhältnis, um ausreichend gut lernen zu können und Generalisierbarkeit gewährleisten zu können. Dies hat sich auf den hier vorgestellten Lösungsweg ausgewirkt und wird sich auch auf zukünftige Ansätze auswirken, die hinsichtlich der Aufbereitung und der Modellwahl vom vorgestellten Ansatz abweichen. Bei der Aufbereitung können andere Ansätze für die Ausreißerelimination gewählt

werden. Dazu gehören die multivariate Ansätze[316], die den gesamten Datensatz betrachten. In deren Folge andere Werte als Ausreißer identifiziert oder mit anderen Werten ersetzt werden. Genauso gibt es andere Ansätze für die Gruppierung, wie z. B. die Gruppenbildung über die Entropie.[317] Andere Verfahren, die zu anderen Klassen führen, können das Ergebnis beeinflussen. Der Einsatz der PCA, der wegen zu geringer Datenmenge verworfen wurde, lieferte den Grund für die Anwendung der KMO und der daraus resultierenden Löschung der Attribute. Daher ist die Frage, ob diese Löschung dennoch gerechtfertigt ist. Der in dieser Arbeit gewählte Ansatz des maschinellen Lernens ist jener der Klassifikation. Es muss in Betracht gezogen werden, ob andere Ansätze, wie z. B. das Prognostizieren mittels Neuronaler Netze oder eine andere Kombination von Bäumen, das angestrebte Resultat erbringt. Da dieses Verfahren fast ausschließlich datengetrieben ist, besteht die Frage, ob ein anderes Ergebnis erzielt werden würde, wenn ein modellgetriebener Ansatz verfolgt wird.

Alle genannten Ansätze können jedoch nur in Betracht gezogen werden, wenn die Datenbasis verbessert wird. Ein besseres Verhältnis von Datensätzen zu Attributen kann entweder durch eine Reduktion der Attribute oder durch eine Erweiterung der Anzahl der Datensätze erzielt werden. Bei einer Reduktion der Attribute muss ein Experte jene Attribute löschen, die von ihm als nicht relevant eingestuft werden. Damit verändert sich der datengetriebene Ansatz zu einem modellgetriebenen Ansatz. Die andere Möglichkeit ist eine Erhöhung der Anzahl der Datensätze. Um mehr Daten als Grundlage zu haben, muss der Betrachtungszeitraum in die Vergangenheit ausgeweitet werden. Eine Vergrößerung des Betrachtungszeitraums und somit Erhöhung der Anzahl der Datensätze wurde mit dem Unternehmen diskutiert, jedoch verworfen, da es ungefähr jährlich zu Technologieumbrüchen kommt und somit die Daten nicht vergleichbar sind. Zusätzlich muss die Möglichkeit einer Ausreißeranalyse vor der Aggregation der Daten in Erwägung gezogen werden, um eventuelle Ausreißer vor ihrem Einfließen in die Datenbasis zu erkennen und zu beseitigen.

[316] Vgl. Hair, J. F. et al. (2014), S. 64
[317] Vgl. Witten, I. H. et al. (2011), S. 316

Literaturverzeichnis

Abbott, D. (2014): Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. Indianapolis, IN: Wiley. ISBN 978-1-118-72796-6.

Alpaydin, E. (2010): Introduction to Machine Learning. 2nd ed, Cambridge, Massachusetts: MIT Press. (Adaptive Computation and Machine Learning). ISBN 978-0-262-01243-0.

Azevedo, A.; Santos, M. F. (2008): KDD, SEMMA and CRISP-DM: A Parallel Overview. Proceedings of IADIS European Conference on Data Mining 2008, Amsterdam, The Netherlands, July 24-26, 2008. Proceedings, januari 1 2008.

Azevedo, C. S.; Santos, M. F. (2005): Data Mining - Descoberta de Conhecimento em Bases de Dados. Capa mole. ISBN 978-972-722-509-5.

Backhaus, K.; Erichson, B.; Plinke, W.; Weiber, R. (2011): Multivariate Analysemethoden: Eine Anwendungsorientierte Einführung. 13., überarbeitete Auflage, Berlin Dordrecht London New York: Springer. (Springer-Lehrbuch). ISBN 978-3-642-16490-3.

Backhaus, K.; Erichson, B.; Weiber, R. (2015): Fortgeschrittene Multivariate Analysemethoden: Eine anwendungsorientierte Einführung. 3., überarbeitete und aktualisierte Auflage, Berlin Heidelberg: Springer Gabler. (Lehrbuch). ISBN 978-3-662-46086-3.

Bauernhansl, T. (2014): Die Vierte Industrielle Revolution – Der Weg in ein wertschaffendes Produktionsparadigma. In: Bauernhansl, T.; ten Hompel, M.; Vogel-Heuser, B. (Hrsg.): Industrie 4.0 in Produktion, Automatisierung und Logistik. Wiesbaden: Springer Fachmedien Wiesbaden. ISBN 978-3-658-04681-1, S. 5–35.

Beltrami, E. (1873): Sulle Funzioni Billineari. In: Giornale di matematiche, Jg. 11, S. 98–106.

Blohm, H.; Beer, T.; Seidenberg, U.; Silber, H. (2008): Produktionswirtschaft. 4., vollst. überarb. Aufl, Herne: Verl. Neue Wirtschafts-Briefe. (NWB-Studium-Betriebswirtschaft Lehrbuch). ISBN 978-3-482-63024-8.

Breiman, L. (1996): Bagging Predictors. In: Machine Learning, Jg. 24, Nr. 2, S. 123–140.

Breiman, L. (2001): Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author). In: Statistical science, Jg. 16, Nr. 3, S. 199–231.

Czado, C.; Schmidt, T. (2011): Mathematische Statistik. Berlin: Springer. (Statistik und ihre Anwendungen). ISBN 978-3-642-17260-1.

David, H. A.; Hartley, H. O.; Pearson, E. S. (1954): The Distribution of the Ratio, in a Single Normal Sample, of Range to Standard Deviation. In: Biometrika, Jg. 41, Nr. 3/4, S. 482–493.

Degen, H.; Lorscheid, P. (2012): Statistik-Lehrbuch: Methoden der Statistik im wirtschaftswissenschaftlichen Bachelor-Studium. 4., korrigierte Aufl, München: Oldenbourg. ISBN 978-3-486-71588-0.

Edward E. Cureton, R. B. D. (1993): Factor Analysis: An Applied Approach. Psychology Press. ISBN 0-8058-1546-5.

Eidenmüller, B. (1995): Die Produktion als Wettbewerbsfaktor: Das Potential der Mitarbeiter nutzen - Herausforderung an das Produktionsmanagement. 3. aktualis. u. erw. Aufl, Köln: TÜV Rheinland. (Leitfaden für Unternehmer und Führungskräfte). ISBN 978-3-8249-0181-4.

EPCOS HR (2016): EPCOS OHG: A TDK Company: Herzlich Willkommen, unveröffentlichtes Dokument zur Information neuer Mitarbeiter, Deutschlandsberg, 2016.

EPCOS PPD PTC PD (2016): 2nd Monozukuri Strategy Meeting: PPD Zero Defect Approach: Reduction of resistance variation at PTC, 2016.

Ertel, W. (2013): Grundkurs Künstliche Intelligenz: Eine praxisorientierte Einführung. 3. Aufl, Wiesbaden: Springer Vieweg. (Lehrbuch). ISBN 978-3-8348-2157-7.

Fahrmeir, L.; Künstler, R.; Pigeot, I.; Tutz, G. (2004): Statistik: Der Weg zur Datenanalyse. 5., Aufl, Berlin: Springer. (Springer-Lehrbuch). ISBN 978-3-540-21232-4.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. (1996): From Data Mining to Knowledge Discovery in Databases. In: AI Magazine, Jg. 17, Nr. 3, S. 37–54.

Fernandez, G. (2003): Data Mining Using SAS Applications: a Case Study Approach. Cary, N.C.: SAS Pub.

Fisher, R. A.; Mackenzie, W. A. (1923): Studies in Crop Variation. II. The Manurial Response of Different Potato Varieties. In: The Journal of Agricultural Science, Jg. 13, Nr. 03, S. 311.

Freitag, M.; Kück, M.; Ait Alla, A.; Lütjen, M. (2015): Potentiale von Data Science in Produktion und Logistik: Teil 1 - Eine Einführung in aktuelle Ansätze der Data Science. In: Industrie 4.0 Management, Jg. 31, Nr. 5, S. 22–26.

Gartner (2015): Gartner: Predictive Analytics. URL: http://ww.gartner.com/it-glossary/predictive-analytics (Zugriff: 17.05.2015).

Gertheiss, J.; Tutz, G. (2009): Statistische Tests. In: Schwaiger, M.; Meyer, A. (Hrsg.): Theorien und Methoden der Betriebswirtschaft: Handbuch für Wissenschaftler und Studierende. München: Vahlen. ISBN 978-3-8006-3613-6, S. 439–454.

Gläßner, J. (1995): Modellgestütztes Controlling der beschaffungslogistischen Prozeßkette. Als Ms. gedr, Düsseldorf: VDI-Verl. (Fortschritt-Berichte VDI Reihe 2, Fertigungstechnik; 337). ISBN 978-3-18-333702-6.

Grubbs, F. E. (1950): Sample Criteria for Testing Outlying Observations. In: The Annals of Mathematical Statistics, S. 27–58.

Gudehus, T. (2010): Logistik: Grundlagen - Strategien - Anwendungen. 4., aktualisierte Aufl, Berlin: Springer. ISBN 978-3-540-89388-2.

Gutenberg, E. (1971): Grundlagen der Betriebswirtschaftslehre Die Produktion. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-61989-2.

Guttman, L. (1953): Image theory for the structure of quantitative variates. In: Psychometrika, Jg. 18, Nr. 4, S. 277–296.

Hair, J. F.; Black, W. C.; Babin, B. J.; Anderson, R. E. (2014): Examining Your Data. In: Hair, J. F. (Hrsg.): Multivariate data analysis. 7. ed., Pearson new internat. ed, Harlow: Pearson. ISBN 978-1-292-02190-4 (Pearson custom library)., S. 31–88.

Haseloff, O. W.; Hoffmann, H. J. (1965): Kleines Lehrbuch der Statistik: Für Naturwissenschaftler, Mediziner, Psychologen, Sozialwissenschaftler und Pädagogen. 2., Berlin: Walter De Gruyter & Co.

Hastie, T.; Tibshirani, R.; Friedman, J. H. (2009): The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed, New York, NY: Springer. (Springer Series in Statistics). ISBN 978-0-387-84857-0.

Heitjan, D. F. (1997): Annotation: What Can Be Done About Missing Data? Approaches to Imputation. In: American Journal of Public Health, Jg. 87, Nr. 4, S. 548–550.

Hertel, B. R. (1976): Minimizing Error Variance Introduced By Missing Data Routines in Survey Analysis. In: Sociological Methods & Research, Jg. 4, Nr. 4, S. 459–474.

Hotelling, H. (1933): Analysis of a Complex of Statistical Variables with Principal Components. In: Journal of Educational Psychology, Jg. 24, S. 417–441.

IBM Big Data & Analytic Hub Infographic: The Four V's of Big Data | IBM Big Data & Analytics Hub. URL: http://www.ibmbigdatahub.com/infographic/four-vs-big-data (Zugriff: 09.02.2017).

JetBrains s.r.o. (2017): PyCharm. JetBrains. URL: https://www.jetbrains.com/pycharm/ (Zugriff: 13.02.2017).

Jolliffe, I. T. (2002): Principal Component Analysis. 2nd ed, New York: Springer. (Springer Series in Statistics). ISBN 978-0-387-95442-4.

Kaiser, H. F. (1970): A Second Generation Little Jiffy. In: Psychometrika, Jg. 35, Nr. 4, S. 401–415.

Little, R. J. A.; Rubin, D. B. (2002): Statistical Analysis with Missing Data. 2nd ed, Hoboken, N.J: Wiley. (Wiley Series in Probability and Statistics). ISBN 978-0-471-18386-0.

Malhotra, N. K. (1987): Analyzing Marketing Research Data with Incomplete Information on the Dependent Variable. In: Journal of Marketing Research, S. 74–84.

MathWorks (2017): Choose Classifier Options - MATLAB & Simulink - MathWorks Deutschland. URL: https://de.mathworks.com/help/stats/choose-a-classifier.html (Zugriff: 21.02.2017).

MathWorks (2017): Pricing and Licensing - MATLAB & Simulink. URL: https://de.mathworks.com/pricing-licensing.html?prodcode=ML (Zugriff: 13.02.2017).

Mertens, P.; Bodendorf, F.; König, W.; Picot, A.; Schumann, M.; Hess, T. (2012): Grundzüge der Wirtschaftsinformatik. 11. Aufl, Berlin: Springer Gabler. (Springer-Lehrbuch). ISBN 978-3-642-30514-6.

Microsoft (2017): Office VBA language reference. URL: https://msdn.microsoft.com/en-us/library/office/gg264383.aspx (Zugriff: 13.02.2017).

Microsoft (2017): Spreadsheet Software Programs | Excel Free Trial. URL: https://products.office.com/en-us/excel (Zugriff: 13.02.2017).

Moxter, A.; Riebel, P.; Wittmann, W.; Muscheid, W. (1954): Die Elastizität des Betriebes: eine produktions- und marktwirtschaftliche Untersuchung. Westdeutscher Verlag.

Männel, W. (1979): Eignung von Produktionsanlagen. In: HWProd (Hrsg) Schäffer-Poeschel, Stuttgart, S. 1465–1481.

Nebl, T. (2011): Produktionswirtschaft. 7., vollständig überarbeitete und erweiterte Auflage, München: Oldenbourg Verlag. (Lehr- und Handbücher der Betriebswirtschaftslehre). ISBN 978-3-486-59669-4.

Nyhuis, P.; Wiendahl, H.-P. (2012): Logistische Kennlinien: Grundlagen, Werkzeuge und Anwendungen. 3. Auflage, Berlin Heidelberg Dordrecht London New York: Springer Vieweg. (VDI-Buch). ISBN 978-3-540-92838-6.

Pearson, K. (1901): LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space*. In: Philosophical Magazine Series 6, Jg. 2, Nr. 11, S. 559–572.

Reisch, M. (2007): Elektronische Bauelemente: Funktion, Grundschaltungen, Modellierung mit SPICE. 2., vollst. neu bearb. Aufl, Berlin: Springer. ISBN 978-3-540-34014-0.

Rinkenburger, R. (2009): Einführung in die Explorative Faktorenanalyse. In: Schwaiger, M.; Meyer, A. (Hrsg.): Theorien und Methoden der Betriebswirtschaft: Handbuch für Wissenschaftler und Studierende. München: Vahlen. ISBN 978-3-8006-3613-6, S. 455–476.

Roth, P. L. (1994): Missing Data: A Conceptual Review for Applied Psychologists. In: Personnel Psychology, Jg. 47, Nr. 3, S. 537–560.

Schutt, R.; O'Neil, C. (2013): Doing Data Science. First edition, Beijing ; Sebastopol: O'Reilly Media. ISBN 978-1-4493-5865-5.

Shakhnarovich, G.; Indyk, P.; Darrell, T. (2005): Introduction. In: Shakhnarovich, G.; Darrell, T.; Indyk, P. (Hrsg.): Nearest-Neighbor Methods in Learning and Vision: Theory and Practice. Cambridge, Mass: MIT Press. ISBN 978-0-262-19547-8 (Neural information processing series)., S. 1–12.

Sharp, H.; Feldt, A. (1959): Some Factors in a Probability Sample Survey of a Metropolitan Community. In: American Sociological Review, Jg. 24, Nr. 5, S. 650.

Shearer, C. (2000): The CRISP-DM Model: The New Blueprint for Data Mining. In: Journal of Data Warehousing, Jg. 5, Nr. 4, S. 13–22.

Shlens, J. (2014): A Tutorial on Principal Component Analysis. In: arXiv preprint arXiv:1404.1100.

Shmueli, G. (2010): To Explain or to Predict? In: Statistical Science, Jg. 25, Nr. 3, S. 289–310.

Stewart, D. W. (1981): The Application and Misapplication of Factor Analysis in Marketing Research. In: Journal of Marketing Research, Jg. 18, Nr. 1, S. 51.

Swoboda, P. (1964): Die betriebliche Anpassung als Problem des betrieblichen Rechnungswesens. Wiesbaden: Gabler Verlag. ISBN 978-3-663-13555-5.

Tukey, J. W. (1977): Exploratory Data Analysis. Reading, Mass: Addison-Wesley Pub. Co. (Addison-Wesley Series in Behavioral Science). ISBN 978-0-201-07616-5.

Weber, J.; Steven, M. (2017): Definition >>Kostenremanenz<< | Gabler Wirtschaftslexikon. URL: http://wirtschaftslexikon.gabler.de/Archiv/7702/kostenremanenz-v6.html (Zugriff: 25.04.2017).

Wilcox, R. (2005): Kolmogorov-Smirnov Test. In: Encyclopedia of Biostatistics. Chichester, UK: John Wiley & Sons, Ltd. ISBN 978-0-470-84907-1.

Witten, I. H.; Frank, E.; Hall, M. A. (2011): Data Mining: Practical Machine Learning Tools and Techniques. 3rd ed, Burlington, MA: Morgan Kaufmann. (Morgan Kaufmann Series in Data Management Systems). ISBN 978-0-12-374856-0.

Yuan, B.; Wang, X. .; Morris, T. (2000): Software Analyser Design Using Data Mining Technology for Toxicity Prediction of Aqueous Effluents. In: Waste Management, Jg. 20, Nr. 8, S. 677–686.

Zsifkovits, H. E. (2013): Logistik. Konstanz: UVK-Verl.-Ges. (UTB Betriebswirtschaftslehre, Technik, Ingenieurwesen; 3673). ISBN 978-3-8252-3673-1.

Anhang

Tabelle 27: VM: Einteilung der Attribute in MAR und MCAR (fehlenden Daten in %)[318]

OK	fehlende Werte					
	MAR (nonrandom)		MCAR (random)			
Matnummer	Pb_2	98,2%	Pb_1	0,4%	Durchfluss MW	5,3%
Mat,- Text	Pb_Menge2	98,2%	BaCO3_1	0,4%	Durchfluss Stabw	5,3%
Losnummer	BaCO3_2	89,8%	SrCO3_1	0,4%	Leistung MAX	5,3%
ChargeSAP	BaCO3_Menge2	89,8%	CaCO3_1	0,4%	Leistung MIN	5,3%
Chargennummer	SrCO3_2	98,7%	Pb3O4_1	0,4%	Leistung MW	5,3%
IstStart	SrCO3_Menge2	98,7%	Y-Acetat_1	0,4%	Leistung Stabw	5,3%
IstEnde	CaCO3_2	96,9%	TiO2 HT05SS_1	0,4%	Schlickerdichte MAX	5,3%
Mstart	CaCO3_Menge2	96,9%	Sprühturm	2,2%	Schlickerdichte MIN	5,3%
Mlief	Pb3O4_2	98,2%	pH Wert Schlicker	0,4%	Schlickerdichte MW	5,3%
Pb_Menge1	Pb3O4_Menge2	98,2%	Schüttdichte Granulat	2,2%	Schlickerdichte Stabw	5,3%
BaCO3_Menge1	TiO2 HT05SS_2	83,2%	Granulatfeuchte	2,2%	Schlickerdruck MAX	5,3%
SrCO3_Menge1	TiO2 HT05SS_Menge2	83,2%	Differenzdruck Filter	2,2%	Schlickerdruck MIN	5,3%
CaCO3_Menge1			Ringtemperatur 1	0,0%	Schlickerdruck MW	5,3%
Pb3O4_Menge1			Ringtemperatur 2	0,0%	Schlickerdruck Stabw	5,3%
Y-Acetat_Menge1			Ablufttemperatur	0,4%	Schlickertemp, MAX	5,3%
TiO2 HT05SS_Menge1			Gewölbeunterdruck	3,5%	Schlickertemp, MIN	5,3%
A40_Menge1			Heizgruppe 7	2,2%	Schlickertemp, MW	5,3%
Optapix_Menge1			Heizgruppe 9	2,2%	Schlickertemp, Stabw	5,3%
Fryma			Heizgruppe 11	2,2%	Anzahl Segmente	5,3%
UD			Heizgruppe 13	2,2%	Segmentzeit MAX	5,3%
Kugelmenge			Differenzdruck MAX	5,3%	Segmentzeit MIN	5,3%
Durchfluss			Differenzdruck MIN	5,3%	MW Dauer je Segment	5,3%
Siebrückstand feucht			Differenzdruck MW	5,3%	Stabw Segmentzeit	5,3%
d10 (VM)			Differenzdruck Stabw	5,3%	Gesamtzeit Segmente	5,3%
d50 (VM)			Durchfluss MAX	5,3%	Stabw Segmentzeit	5,3%
d90 (VM)			Durchfluss MIN	5,3%	Gesamtzeit Segmente	5,3%

[318] Quelle: eigene Darstellung

Tabelle 28: NM: Einteilung der Attribute in MAR und MCAR (fehlenden Daten in %)[319]

OK	fehlende Werte					
	MAR (nonrandom)		MCAR (random)			
Matnummer	Umsatz_2	99,1%	Fryma	0,4%	TiO2	0,4%
Mat,- Text	Umsatz_Menge2	99,1%	Datum	0,4%	Y2O3	0,4%
Losnummer	Methocell A4C_1	99,6%	Kugelmenge	0,4%	SPHT3 MW	0,4%
ChargeSAP	Methocell A4C_Menge2	99,6%	Durchfluss	0,4%	Symm3 MW	0,4%
Chargennummer	Gleitwert	67,9%	Schlickertemperatur	0,4%	b/l3 MW	0,4%
IstStart	Grünteilfestigkeit	67,9%	d10 (NM)	0,4%		
IstEnde	Kohäsion	67,9%	d50 (NM)	0,4%		
Mstart	Pressdruck	67,9%	d90 (NM)	0,4%		
Mlief	Radiale Auffederung	67,9%	Schlickerdichte	0,4%		
Umsatz_1	Schüttwinkel	67,9%	pH Wert Schlicker	1,3%		
Umsatz_Menge1			Glühverlust	0,9%		
Methocell A4C_Menge1			Sinterdichte	0,9%		
Contraspum K1012_Menge1			R25	2,7%		
Sprühturm			Rmin	2,7%		
Datum			T-Rmin	2,7%		
Auslaufzeit			Tb	2,7%		
Schüttdichte Granulat			Tcorr	2,7%		
Granulatfeuchte			BaO	0,4%		
Differenzdruck Filter			CaO	0,4%		
d10 Granulat (NM)			MnO	0,4%		
d50 Granulat (NM)			Na2O	0,4%		
d90 Granulat (NM)			PbO	0,4%		
R25 Widerstand			SiO2	0,4%		
R25 Vergl,%			SrO	0,4%		

[319] Quelle: eigene Darstellung

Tabelle 29: Attribute nach der Zusammenführung (VM+NM)[320]

Kugelmenge	Durchfluss	Schlickertemperatur	d10 (NM)	d50 (NM)
d90 (NM)	Auslaufzeit	Schuettdichte Granulat	Granulatfeuchte	Differenzdruck Filter
d10 Granulat (NM)	d50 Granulat (NM)	d90 Granulat (NM)	Schlickerdichte	pH Wert Schlicker
Gluehverlust	Sinterdichte	R25 Widerstand	R25 Vergl.%	R25
Rmin	T-Rmin	Tb	Tcorr	Gleitwert
Gruenteilfestigkeit	Kohaesion	Pressdruck	Radiale Auffederung	Schuettwinkel
SPHT3 MW	Symm3 MW	b_l3 MW	Contraspum K1012_Menge1_NM_Gehalt	Methocell A4C_NM_Gehalt
BaO_NM_Gehalt	CaO_NM_Gehalt	MnO_NM_Gehalt	Na2O_NM_Gehalt	PbO_NM_Gehalt
SiO2_NM_Gehalt	SrO_NM_Gehalt	TiO2_NM_Gehalt	Y2O3_NM_Gehalt	Pb_Me
BaCO3	SrCO3	CaCO3	Pb3O4	Y-Acet
TiO2	A40_Me	Optapi	Kugelmenge_VM	Siebrueckstand feucht_VM
d10 (VM)_VM	d50 (VM)_VM	d90 (VM)_VM	pH Wert Schlicker_VM	Schuettdichte Granulat_VM
Granulatfeuchte_VM	Differenzdruck Filter_VM	Ablufttemperatur_VM	Gewoelbeunterdruck_VM	Heizgruppe 7_VM
Heizgruppe 9_VM	Heizgruppe 11_VM	Heizgruppe 13_VM	Differenzdruck MAX_VM	Differenzdruck MIN_VM
Differenzdruck MW_VM	Differenzdruck Stabw_VM	Durchfluss MAX_VM	Durchfluss MIN_VM	Durchfluss MW_VM
Durchfluss Stabw_VM	Leistung MAX_VM	Leistung MIN_VM	Leistung MW_VM	Leistung Stabw_VM
Schlickerdichte MAX_VM	Schlickerdichte MIN_VM	Schlickerdichte MW_VM	Schlickerdichte Stabw_VM	Schlickerdruck MAX_VM
Schlickerdruck MIN_VM	Schlickerdruck MW_VM	Schlickerdruck Stabw_VM	Schlickertemp. MAX_VM	Schlickertemp. MIN_VM
Schlickertemp. MW_VM	Schlickertemp. Stabw_VM	Anzahl Segmente_VM	Segmentzeit MAX_VM	Segmentzeit MIN_VM
MW Dauer je Segment_VM	Stabw Segmentzeit_VM	Gesamtzeit Segmente_VM	Ringtemperatur_VM	

[320] Quelle: eigene Darstellung

Tabelle 30: Statistische Maße der VM+NM (mit Ausreißern)[321]

Bezeichnung	$\tilde{s}^2$	$\tilde{s}$	γ	g_m	Korrelations-koeffizient mit R25 Vergl.%
Kugelmenge	0,22	0,47	50	-5,3	0,20
Durchfluss	0,87	0,93	-3	-0,2	-0,33
Schlickertemperatur	7,39	2,72	-3	-0,4	0,21
d10 (NM)	0,00	0,01	0	0,3	0,31
d50 (NM)	0,00	0,03	11	2,0	0,00
d90 (NM)	0,02	0,15	0	1,0	-0,18
Auslaufzeit	4,85	2,20	0	1,7	-0,11
Schuettdichte Granulat	0,00	0,02	0	-0,9	0,08
Granulatfeuchte	0,00	0,01	-3	0,6	-0,23
Differenzdruck Filter	0,00	0,03	-3	-0,2	0,05
d10 Granulat (NM)	10,71	3,27	-3	0,4	-0,15
d50 Granulat (NM)	31,82	5,64	-3	0,1	-0,14
d90 Granulat (NM)	37,95	6,16	-3	0,0	-0,13
Schlickerdichte	0,00	0,02	-1	-1,0	-0,25
pH Wert Schlicker	0,03	0,17	-4	-0,1	0,23
Gluehverlust	0,00	0,04	1	1,1	0,14
Sinterdichte	0,00	0,01	-4	-0,3	0,37
R25 Widerstand	6,19	2,49	-2	0,5	0,68
R25 Vergl,%	68,09	8,25	-3	-0,2	1,00
R25	2,75	1,66	-3	0,3	0,46
Rmin	1,62	1,27	-3	0,3	0,41
T-Rmin	6,21	2,49	-3	-0,2	0,24
Tb	0,62	0,79	-2	0,0	0,07
Tcorr	5,45	2,34	-3	-0,2	0,21
Gleitwert	0,00	0,02	-3	-0,5	0,09
Gruenteilfestigkeit	0,47	0,69	-2	0,6	-0,09
Kohaesion	0,01	0,12	-3	0,1	0,01
Pressdruck	18,36	4,29	-3	0,0	-0,07
Radiale Auffederung	0,00	0,00	-4	0,1	-0,33
Schuettwinkel	1,00	1,00	0	1,2	0,07
SPHT3 MW	0,00	0,00	3	-0,9	-0,07
Symm3 MW	0,00	0,00	-2	-0,7	0,12
b_l3 MW	0,00	0,01	-3	-0,5	0,32
Contraspum K1012_Menge1_NM_Gehalt	0,00	0,00	36	6,4	0,03
Methocell A4C_NM_Gehalt	0,02	0,13	210	14,7	0,04
BaO_NM_Gehalt	0,00	0,04	-3	-0,2	0,22
CaO_NM_Gehalt	0,00	0,02	-2	0,5	0,08
MnO_NM_Gehalt	0,00	0,00	-3	-1,3	0,03
Na2O_NM_Gehalt	0,00	0,00	-4	-0,4	-0,45
PbO_NM_Gehalt	0,00	0,03	2	1,1	-0,14
SiO2_NM_Gehalt	0,00	0,00	-2	0,0	-0,20
SrO_NM_Gehalt	0,00	0,01	8	2,9	0,01
TiO2_NM_Gehalt	0,00	0,03	-3	0,3	-0,16
Y2O3_NM_Gehalt	0,00	0,00	-3	0,2	0,06

[321] Quelle: eigene Darstellung

Fortsetzung von Tabelle 30: Statistische Zusammenhangsmaße der VM+NM[322]

Bezeichnung	$\tilde{s}^2$	$\tilde{s}$	γ	g_m	Korrelationskoeffizient mit R25 Vergl.%
Pb_Me	0,00	0,00	101	9,7	-0,09
BaCO3	0,00	0,00	35	-4,6	-0,06
SrCO3	0,00	0,00	0	1,5	0,21
CaCO3	0,00	0,00	-3	-0,4	0,02
Pb3O4	0,00	0,00	101	9,7	-0,09
Y-Acet	0,00	0,00	-4	0,3	0,44
TiO2	0,00	0,00	4	-1,6	0,11
A40_Me	0,00	0,00	-4	-0,8	-0,37
Optapi	0,00	0,00	-4	0,8	0,37
Kugelmenge_VM	0,01	0,11	65	-8,3	-0,18
Siebrueckstand feucht_VM	3198,41	56,55	-3	0,4	-0,08
d10 (VM)_VM	0,00	0,01	25	3,4	0,14
d50 (VM)_VM	0,00	0,05	-3	0,0	0,22
d90 (VM)_VM	0,02	0,14	-3	0,0	0,28
pH Wert Schlicker_VM	0,02	0,15	12	2,0	0,23
Schuettdichte Granulat_VM	0,00	0,01	0	-0,9	-0,12
Granulatfeuchte_VM	0,00	0,01	2	1,7	-0,02
Differenzdruck Filter_VM	0,00	0,07	-4	-0,4	0,16
Ablufttemperatur_VM	117,64	10,85	20	-3,2	0,12
Gewoelbeunterdruck_VM	0,00	0,05	0	0,7	-0,16
Heizgruppe 7_VM	250,38	15,82	122	-10,0	0,04
Heizgruppe 9_VM	187,01	13,68	131	-10,6	0,04
Heizgruppe 11_VM	152,56	12,35	133	-10,7	0,03
Heizgruppe 13_VM	162,04	12,73	110	-9,3	0,05
Differenzdruck MAX_VM	0,00	0,05	39	1,2	-0,13
Differenzdruck MIN_VM	0,00	0,04	37	-5,6	-0,10
Differenzdruck MW_VM	0,00	0,03	38	-5,2	-0,13
Differenzdruck Stabw_VM	0,00	0,01	161	12,2	-0,03
Durchfluss MAX_VM	0,04	0,20	3	2,0	-0,15
Durchfluss MIN_VM	0,69	0,83	186	-13,4	0,09
Durchfluss MW_VM	0,03	0,17	99	-8,2	0,09
Durchfluss Stabw_VM	0,08	0,27	196	13,9	-0,07
Leistung MAX_VM	495462,94	703,89	-1	1,4	0,08
Leistung MIN_VM	366392,06	605,30	-2	1,3	0,11
Leistung MW_VM	412585,33	642,33	-2	1,2	0,10
Leistung Stabw_VM	4502,54	67,10	41	5,1	0,01
Schlickerdichte MAX_VM	0,00	0,01	100	-8,3	0,09
Schlickerdichte MIN_VM	0,01	0,09	195	-13,9	0,06
Schlickerdichte MW_VM	0,00	0,02	79	-8,6	0,12
Schlickerdichte Stabw_VM	0,00	0,02	197	14,0	-0,05
Schlickerdruck MAX_VM	0,01	0,09	-4	0,1	-0,16
Schlickerdruck MIN_VM	0,00	0,06	-2	0,4	-0,06
Schlickerdruck MW_VM	0,01	0,07	-3	0,2	-0,13
Schlickerdruck Stabw_VM	0,00	0,02	6	2,6	-0,10
Schlickertemp, MAX_VM	2,96	1,72	22	3,7	-0,03
Schlickertemp, MIN_VM	2,04	1,43	1	1,6	-0,06
Schlickertemp, MW_VM	1,95	1,40	1	1,6	-0,04

[322] Quelle: eigene Darstellung

Fortsetzung von Tabelle 30: Statistische Zusammenhangsmaße der VM+NM[323]

Bezeichnung	$\tilde{s}^2$	$\tilde{s}$	γ	g_m	Korrelationskoeffizient mit R25 Vergl.%
Schlickertemp, Stabw_VM	0,05	0,23	93	8,4	0,00
Anzahl Segmente_VM	0,31	0,56	3	2,3	0,04
Segmentzeit MAX_VM	1183399,42	1087,84	2	-2,0	-0,30
Segmentzeit MIN_VM	19803659,69	4450,13	-3	-1,2	-0,03
MW Dauer je Segment_VM	6965942,86	2639,31	-3	-1,2	-0,10
Stabw Segmentzeit_VM	438016*109	20928831	-1	-1,4	-0,32
Gesamtzeit Segmente_VM	845274,61	919,39	4	-2,1	-0,29
Ringtemperatur_VM	8,00	2,83	-4	0,0	0,30

[323] Quelle: eigene Darstellung

Tabelle 31: Ausreißer in Anzahl und Prozent je Attribut[324]

Attribut	Ausreißer	
	Anzahl	Prozent
Kugelmenge	1	0,5%
Durchfluss	0	0,0%
Schlickertemperatur	1	0,5%
d10 (NM)	3	1,4%
d50 (NM)	7	3,2%
d90 (NM)	6	2,8%
Auslaufzeit	15	6,9%
Schuettdichte Granulat	3	1,4%
Granulatfeuchte	0	0,0%
Differenzdruck Filter	0	0,0%
d10 Granulat (NM)	0	0,0%
d50 Granulat (NM)	0	0,0%
d90 Granulat (NM)	3	1,4%
Schlickerdichte	8	3,7%
pH Wert Schlicker	0	0,0%
Gluehverlust	5	2,3%
Sinterdichte	0	0,0%
R25 Widerstand	9	4,1%
R25 Vergl,%	1	0,5%
R25	0	0,0%
Rmin	0	0,0%
T-Rmin	6	2,8%
Tb	11	5,0%
Tcorr	16	7,3%
Gleitwert	10	4,6%
Gruenteilfestigkeit	11	5,0%
Kohaesion	3	1,4%
Pressdruck	3	1,4%
Radiale Auffederung	0	0,0%
Schuettwinkel	0	0,0%
SPHT3 MW	9	4,1%
Symm3 MW	7	3,2%
b_l3 MW	3	1,4%
Contraspum K1012_Menge1_NM_Gehalt	5	2,3%
Methocell A4C_NM_Gehalt	2	0,9%
BaO_NM_Gehalt	2	0,9%
CaO_NM_Gehalt	9	4,1%
MnO_NM_Gehalt	52	23,9%
Na2O_NM_Gehalt	0	0,0%
PbO_NM_Gehalt	13	6,0%
SiO2_NM_Gehalt	1	0,5%
SrO_NM_Gehalt	34	15,6%
TiO2_NM_Gehalt	3	1,4%
Y2O3_NM_Gehalt	12	5,5%
Pb_Me	3	1,4%
BaCO3	4	1,8%

[324] Quelle: eigene Darstellung

Fortsetzung vonTabelle 31: Ausreißer in Anzahl und Prozent je Attribut[325]

Attribut	Ausreißer	
	Anzahl	Prozent
SrCO3	16	7,3%
CaCO3	0	0,0%
Pb3O4	3	1,4%
Y-Acet	0	0,0%
TiO2	24	11,0%
A40_Me	0	0,0%
Optapi	0	0,0%
Kugelmenge_VM	3	1,4%
Siebrueckstand feucht_VM	1	0,5%
d10 (VM)_VM	2	0,9%
d50 (VM)_VM	2	0,9%
d90 (VM)_VM	0	0,0%
pH Wert Schlicker_VM	1	0,5%
Schuettdichte Granulat_VM	3	1,4%
Granulatfeuchte_VM	4	1,8%
Differenzdruck Filter_VM	0	0,0%
Ablufttemperatur_VM	1	0,5%
Gewoelbeunterdruck_VM	1	0,5%
Heizgruppe 7_VM	23	10,6%
Heizgruppe 9_VM	15	6,9%
Heizgruppe 11_VM	28	12,8%
Heizgruppe 13_VM	21	9,6%
Differenzdruck MAX_VM	13	6,0%
Differenzdruck MIN_VM	17	7,8%
Differenzdruck MW_VM	13	6,0%
Differenzdruck Stabw_VM	20	9,2%
Durchfluss MAX_VM	12	5,5%
Durchfluss MIN_VM	5	2,3%
Durchfluss MW_VM	10	4,6%
Durchfluss Stabw_VM	11	5,0%
Leistung MAX_VM	1	0,5%
Leistung MIN_VM	21	9,6%
Leistung MW_VM	3	1,4%
Leistung Stabw_VM	12	5,5%
Schlickerdichte MAX_VM	4	1,8%
Schlickerdichte MIN_VM	11	5,0%
Schlickerdichte MW_VM	3	1,4%
Schlickerdichte Stabw_VM	54	24,8%
Schlickerdruck MAX_VM	0	0,0%
Schlickerdruck MIN_VM	6	2,8%
Schlickerdruck MW_VM	0	0,0%
Schlickerdruck Stabw_VM	18	8,3%
Schlickertemp, MAX_VM	19	8,7%
Schlickertemp, MIN_VM	19	8,7%
Schlickertemp, MW_VM	18	8,3%
Schlickertemp, Stabw_VM	16	7,3%

[325] Quelle: eigene Darstellung

Fortsetzung vonTabelle 31: Ausreißer in Anzahl und Prozent je Attribut[326]

Attribut	Ausreißer	
	Anzahl	Prozent
Anzahl Segmente_VM	47	21,6%
Segmentzeit MAX_VM	4	1,8%
Segmentzeit MIN_VM	47	21,6%
MW Dauer je Segment_VM	49	22,5%
Stabw Segmentzeit_VM	3	1,4%
Gesamtzeit Segmente_VM	2	0,9%
Ringtemperatur_VM	0	0,0%

[326] Quelle: eigene Darstellung

Tabelle 32: MSA für einzelne Attribute (VM+NM mit Ausreißern)[327]

Attribute	Durchgang			
	1	2	3	4
~~Kugelmenge~~	0,49			
Durchfluss	0,76	0,86	0,86	0,86
Schlickertemperatur	0,70	0,80	0,82	0,82
d10 (NM)	0,59	0,61	0,66	0,65
d50 (NM)	0,60	0,53	0,52	0,52
d90 (NM)	0,55	0,62	0,61	0,61
~~Auslaufzeit~~	0,49			
~~Schuettdichte Granulat~~	0,55	0,50	0,48	
Granulatfeuchte	0,64	0,59	0,59	0,59
Differenzdruck Filter	0,67	0,74	0,75	0,76
d10 Granulat (NM)	0,53	0,57	0,56	0,56
d50 Granulat (NM)	0,55	0,59	0,59	0,59
d90 Granulat (NM)	0,60	0,65	0,64	0,64
Schlickerdichte	0,62	0,64	0,64	0,64
pH Wert Schlicker	0,83	0,82	0,83	0,83
Gluehverlust	0,73	0,73	0,73	0,75
Sinterdichte	0,80	0,91	0,91	0,91
R25 Widerstand	0,74	0,64	0,64	0,64
~~R25~~	0,49			
~~Rmin~~	0,48			
T-Rmin	0,81	0,77	0,77	0,78
Tb	0,73	0,75	0,75	0,75
Tcorr	0,64	0,72	0,72	0,72
~~Gleitwert~~	0,40			
~~Gruenteilfestigkeit~~	0,34			
~~Kohaesion~~	0,48			
Pressdruck	0,57	0,84	0,84	0,84
Radiale Auffederung	0,62	0,82	0,82	0,81
~~Schuettwinkel~~	0,46			
SPHT3 MW	0,54	0,57	0,56	0,56
Symm3 MW	0,64	0,68	0,67	0,67
b_l3 MW	0,71	0,69	0,69	0,68
~~Contraspum K1012_Menge1_NM_Gehalt~~	0,34			
~~Methocell A4C_NM_Gehalt~~	0,26			
~~BaO_NM_Gehalt~~	0,36			
~~CaO_NM_Gehalt~~	0,29			
~~MnO_NM_Gehalt~~	0,31			
Na2O_NM_Gehalt	0,66	0,72	0,73	0,73
~~PbO_NM_Gehalt~~	0,20			
~~SiO2_NM_Gehalt~~	0,23			
~~SrO_NM_Gehalt~~	0,12			
~~TiO2_NM_Gehalt~~	0,38			
~~Y2O3_NM_Gehalt~~	0,27			
~~Pb_Me~~	0,45			
~~BaCO3~~	0,47			
SrCO3	0,56	0,75	0,75	0,75

[327] Quelle: eigene Darstellung

Fortsetzung von Tabelle 32: MSA für einzelne Attribute (VM+NM mit Ausreißern)[328]

Attribute	Durchgang			
	1	2	3	4
CaCO3	0,58	0,58	0,58	0,58
~~Pb3O4~~	0,45			
Y-Acet	0,77	0,87	0,86	0,87
TiO2	0,54	0,76	0,76	0,76
A40_Me	0,87	0,85	0,84	0,84
Optapi	0,87	0,85	0,84	0,84
~~Kugelmenge_VM~~	0,50			
Siebrueckstand feucht_VM	0,84	0,89	0,89	0,89
~~d10 (VM)_VM~~	0,40			
d50 (VM)_VM	0,62	0,69	0,69	0,69
d90 (VM)_VM	0,63	0,70	0,70	0,70
pH Wert Schlicker_VM	0,77	0,78	0,79	0,80
Schuettdichte Granulat_VM	0,64	0,75	0,76	0,76
Granulatfeuchte_VM	0,56	0,65	0,65	0,67
~~Differenzdruck Filter_VM~~	0,40			
Ablufttemperatur_VM	0,78	0,79	0,82	0,83
~~Gewoelbeunterdruck_VM~~	0,34			
Heizgruppe 7_VM	0,64	0,66	0,66	0,66
Heizgruppe 9_VM	0,59	0,62	0,62	0,61
Heizgruppe 11_VM	0,59	0,63	0,63	0,63
Heizgruppe 13_VM	0,59	0,64	0,64	0,64
~~Differenzdruck MAX_VM~~	0,47			
~~Differenzdruck MIN_VM~~	0,50			
~~Differenzdruck MW_VM~~	0,46			
~~Differenzdruck Stabw_VM~~	0,57	0,49		
~~Durchfluss MAX_VM~~	0,48			
Durchfluss MIN_VM	0,76	0,77	0,78	0,78
Durchfluss MW_VM	0,74	0,78	0,77	0,77
Durchfluss Stabw_VM	0,78	0,79	0,80	0,80
Leistung MAX_VM	0,76	0,71	0,71	0,71
Leistung MIN_VM	0,70	0,75	0,76	0,76
Leistung MW_VM	0,77	0,67	0,68	0,68
~~Leistung Stabw_VM~~	0,49			
~~Schlickerdichte MAX_VM~~	0,42			
Schlickerdichte MIN_VM	0,72	0,73	0,72	0,72
Schlickerdichte MW_VM	0,65	0,68	0,69	0,69
Schlickerdichte Stabw_VM	0,80	0,78	0,77	0,78
Schlickerdruck MAX_VM	0,76	0,77	0,83	0,83
Schlickerdruck MIN_VM	0,68	0,72	0,76	0,77
Schlickerdruck MW_VM	0,71	0,83	0,75	0,75
~~Schlickerdruck Stabw_VM~~	0,57	0,49		
Schlickertemp, MAX_VM	0,70	0,74	0,74	0,75
Schlickertemp, MIN_VM	0,70	0,73	0,73	0,73
Schlickertemp, MW_VM	0,68	0,71	0,71	0,71
Schlickertemp, Stabw_VM	0,68	0,74	0,74	0,74
~~Anzahl Segmente_VM~~	0,46			
Segmentzeit MAX_VM	0,70	0,75	0,75	0,75

[328] Quelle: eigene Darstellung

Fortsetzung von Tabelle 32: MSA für einzelne Attribute (VM+NM mit Ausreißern)[329]

Attribute	Durchgang			
	1	2	3	4
Segmentzeit MIN_VM	0,48			
MW Dauer je Segment_VM	0,54	0,73	0,75	0,74
Stabw Segmentzeit_VM	0,77	0,75	0,75	0,75
Gesamtzeit Segmente_VM	0,68	0,80	0,83	0,83
Ringtemperatur_VM	0,61	0,64	0,64	0,68

[329] Quelle: eigene Darstellung

Tabelle 33: MSA für einzelne Attribute (VM+NM ohne Ausreißer)[330]

Attribute	Durchgang			
	1	2	3	4
Kugelmenge	nan	0,83	0,86	0,86
Durchfluss	nan	0,85	0,87	0,87
Schlickertemperatur	nan	0,76	0,84	0,84
d10 (NM)	nan	0,63	0,68	0,67
~~d50 (NM)~~	nan	0,47		
d90 (NM)	nan	0,55	0,64	0,65
~~Auslaufzeit~~	nan	0,41		
Schuettdichte Granulat	nan	0,55	0,52	0,54
Granulatfeuchte	nan	0,66	0,65	0,65
Differenzdruck Filter	nan	0,65	0,75	0,75
d10 Granulat (NM)	nan	0,51	0,52	0,52
d50 Granulat (NM)	nan	0,56	0,57	0,58
d90 Granulat (NM)	nan	0,67	0,71	0,71
Schlickerdichte	nan	0,70	0,68	0,67
pH Wert Schlicker	nan	0,89	0,89	0,89
Gluehverlust	nan	0,77	0,79	0,79
Sinterdichte	nan	0,76	0,83	0,83
R25 Widerstand	nan	0,74	0,76	0,79
R25	nan	0,55	0,56	0,56
Rmin	nan	0,52	0,53	0,54
T-Rmin	nan	0,75	0,80	0,80
Tb	nan	0,75	0,78	0,79
Tcorr	nan	0,83	0,83	0,82
Gleitwert	nan	0,57	0,57	0,57
Gruenteilfestigkeit	nan	0,63	0,66	0,66
Kohaesion	nan	0,67	0,64	0,64
Pressdruck	nan	0,79	0,81	0,81
Radiale Auffederung	nan	0,74	0,80	0,80
~~Schuettwinkel~~	nan	0,47		
SPHT3 MW	nan	0,71	0,74	0,74
Symm3 MW	nan	0,59	0,60	0,59
b_l3 MW	nan	0,67	0,71	0,71
~~Contraspum K1012_Menge1_NM_Gehalt~~	nan	0,40		
~~Methocell A4C_NM_Gehalt~~	nan	0,36		
BaO_NM_Gehalt	nan	0,71	0,75	0,75
CaO_NM_Gehalt	nan	0,71	0,74	0,75
~~MnO_NM_Gehalt~~	nan			
Na2O_NM_Gehalt	nan	0,71	0,73	0,72
PbO_NM_Gehalt	nan	0,55	0,58	0,58
SiO2_NM_Gehalt	nan	0,66	0,62	0,63
SrO_NM_Gehalt	nan	0,63	0,62	0,62
TiO2_NM_Gehalt	nan	0,79	0,80	0,80
Y2O3_NM_Gehalt	nan	0,63	0,61	0,61
~~Pb_Me~~	nan			
BaCO3	nan	0,70	0,68	0,68
SrCO3	nan	0,68	0,69	0,69

[330] Quelle: eigene Darstellung

Fortsetzung von Tabelle 33: MSA für einzelne Attribute (VM+NM ohne Ausreißer)[331]

Attribute	Durchgang			
	1	2	3	4
CaCO3	nan	0,53	0,57	0,58
~~Pb3O4~~	nan			
Y-Acet	nan	0,87	0,91	0,91
TiO2	nan	0,68	0,71	0,73
A40_Me	nan	0,88	0,88	0,88
Optapi	nan	0,88	0,88	0,88
~~Kugelmenge_VM~~	nan			
Siebrueckstand feucht_VM	nan	0,78	0,84	0,84
~~d10 (VM)_VM~~	nan	0,43		
d50 (VM)_VM	nan	0,69	0,70	0,70
d90 (VM)_VM	nan	0,68	0,69	0,69
pH Wert Schlicker_VM	nan	0,78	0,84	0,84
Schuettdichte Granulat_VM	nan	0,83	0,82	0,82
Granulatfeuchte_VM	nan	0,74	0,73	0,73
~~Differenzdruck Filter_VM~~	nan	0,45		
Ablufttemperatur_VM	nan	0,62	0,64	0,65
Gewoelbeunterdruck_VM	nan	0,57	0,60	0,58
Heizgruppe 7_VM	nan	0,73	0,75	0,75
Heizgruppe 9_VM	nan	0,72	0,77	0,78
Heizgruppe 11_VM	nan	0,76	0,78	0,79
Heizgruppe 13_VM	nan	0,77	0,77	0,77
Differenzdruck MAX_VM	nan	0,78	0,78	0,80
Differenzdruck MIN_VM	nan	0,73	0,76	0,79
Differenzdruck MW_VM	nan	0,75	0,78	0,78
~~Differenzdruck Stabw_VM~~	nan	0,50	0,50	
Durchfluss MAX_VM	nan	0,50	0,51	0,51
Durchfluss MIN_VM	nan	0,67	0,67	0,66
Durchfluss MW_VM	nan	0,58	0,58	0,58
Durchfluss Stabw_VM	nan	0,69	0,68	0,68
Leistung MAX_VM	nan	0,69	0,72	0,71
Leistung MIN_VM	nan	0,83	0,84	0,85
Leistung MW_VM	nan	0,70	0,74	0,75
Leistung Stabw_VM	nan	0,71	0,72	0,72
Schlickerdichte MAX_VM	nan	0,58	0,58	0,58
Schlickerdichte MIN_VM	nan	0,58	0,65	0,65
Schlickerdichte MW_VM	nan	0,64	0,64	0,64
~~Schlickerdichte Stabw_VM~~	nan	0,37		
Schlickerdruck MAX_VM	nan	0,77	0,77	0,77
Schlickerdruck MIN_VM	nan	0,69	0,71	0,72
Schlickerdruck MW_VM	nan	0,75	0,75	0,76
Schlickerdruck Stabw_VM	nan	0,75	0,76	0,75
Schlickertemp, MAX_VM	nan	0,74	0,76	0,76
Schlickertemp, MIN_VM	nan	0,66	0,75	0,75
Schlickertemp, MW_VM	nan	0,64	0,68	0,68
~~Schlickertemp, Stabw_VM~~	nan	0,33		
Anzahl Segmente_VM	nan	0,57	0,63	0,61
Segmentzeit MAX_VM	nan	0,83	0,83	0,83

[331] Quelle: eigene Darstellung

Fortsetzung von Tabelle 33: MSA für einzelne Attribute (VM+NM ohne Ausreißer)[332]

Attribute	Durchgang			
	1	2	3	4
Segmentzeit MIN_VM	nan	0,66	0,66	0,66
MW Dauer je Segment_VM	nan	0,79	0,79	0,79
Stabw Segmentzeit_VM	nan	0,84	0,84	0,85
~~Gesamtzeit Segmente_VM~~	nan			
Ringtemperatur_VM	nan	0,75	0,76	0,77

Tabelle 34: Ergebnis CL und Testdaten: VM+NM mit Ausreißern[333]

	Genauigkeit bei Training	Genauigkeit bei Anwendung auf Testdaten	Grenzfälle bei Anwendung auf Testdaten	Mittelwert der absoluten Differenz zwischen der vorhergesagten und Ist-Klasse
Complex Tree	17,4%	10,3%	1,3%	2,47
Medium Tree	17,9%	6,4%	0,0%	2,92
Simple Tree	17,0%	9,0%	0,0%	2,85
Linear SVM	18,3%	7,7%	1,3%	3,04
Quadratic SVM	17,4%	11,5%	0,0%	3,09
Cubic SVM	17,4%	7,7%	0,0%	3,14
Fine Gaussian SVM	12,8%	15,4%	1,3%	2,42
Medium Gaussian SVM	17,4%	12,8%	0,0%	2,59
Coarse Gaussian SVM	15,6%	7,7%	0,0%	3,26
Boosted Trees	19,7%	2,6%	0,0%	2,97
Bagged Trees	24,3%	11,5%	1,3%	2,81
RUSBoosted Trees	19,3%	5,1%	0,0%	2,82

[332] Quelle: eigene Darstellung
[333] Quelle: ebda.

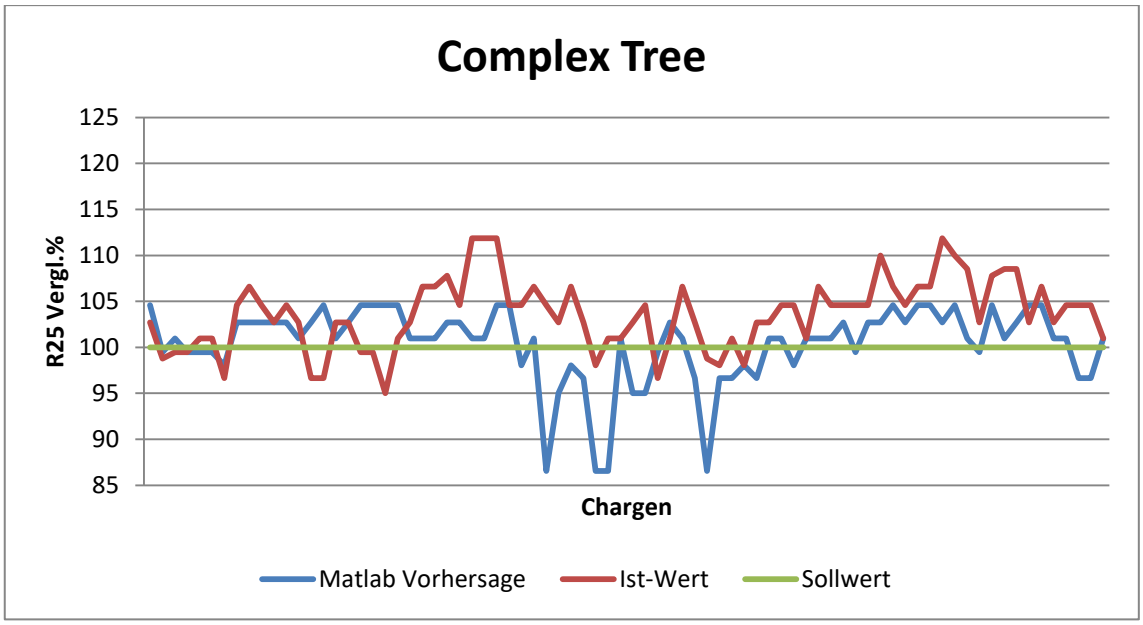


Abbildung 62: CT: Vergleich Vorhersage zu Ist-Wert (VM+NM mit Ausreißern)[334]

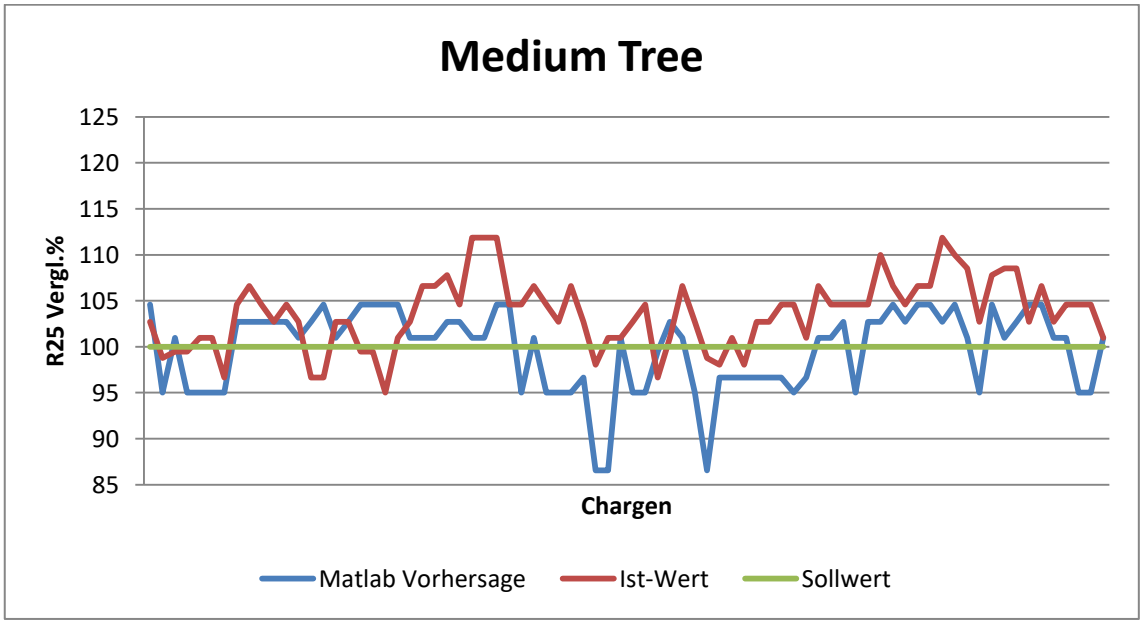


Abbildung 63: MT: Vergleich Vorhersage zu Ist-Wert (VM+NM mit Ausreißern)[335]

[334] Quelle: eigene Darstellung
[335] Quelle: ebda.

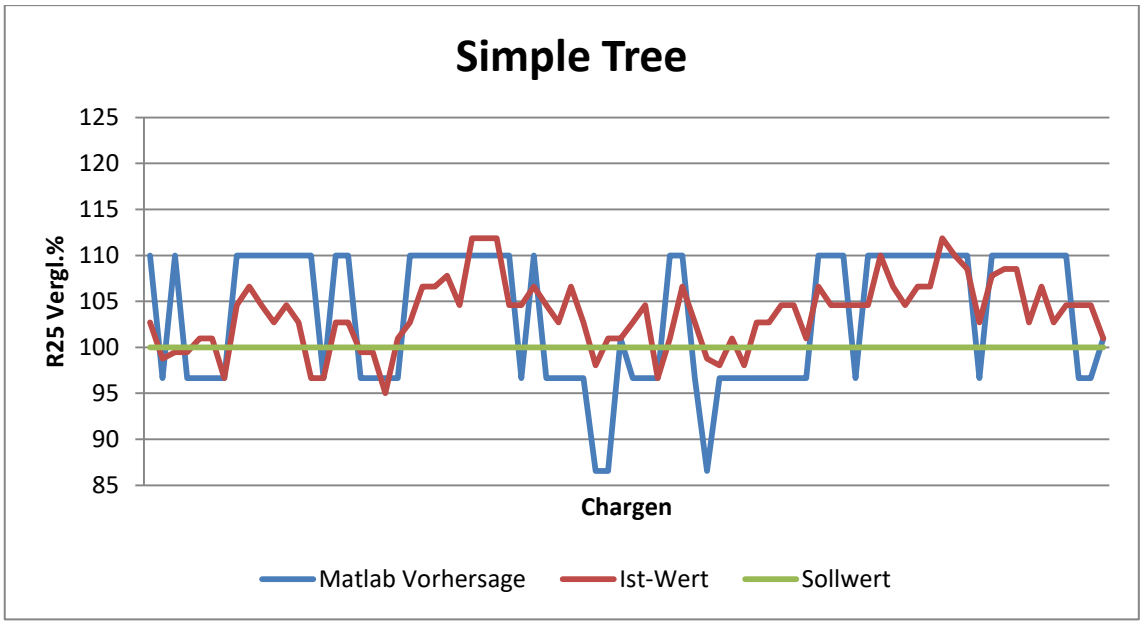


Abbildung 64: ST: Vergleich Vorhersage zu Ist-Wert (VM+NM mit Ausreißern)[336]

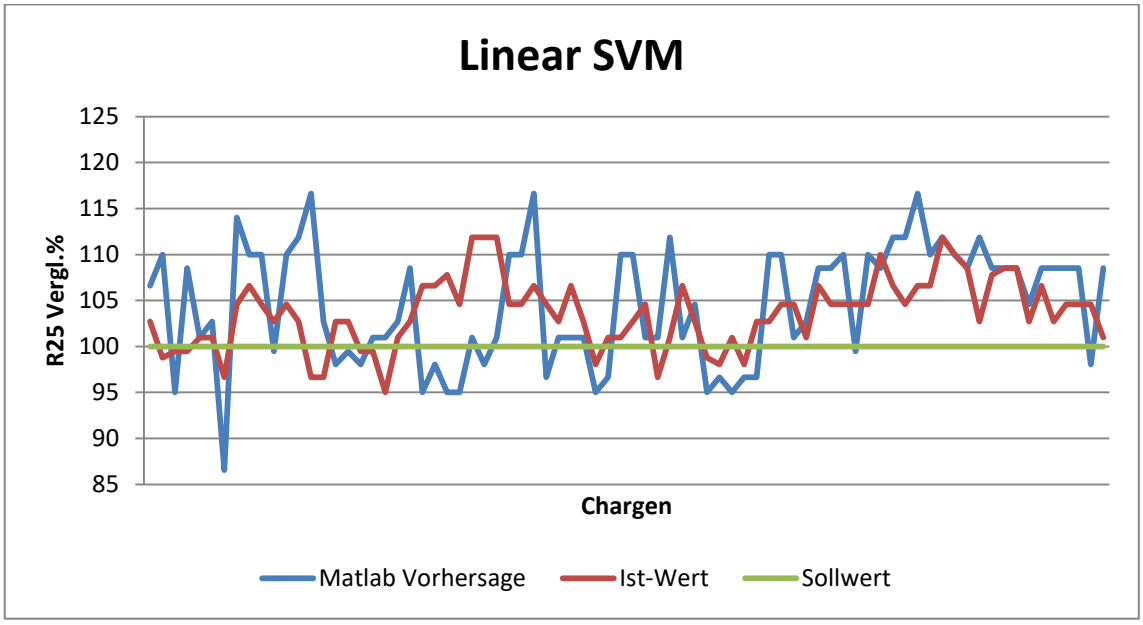


Abbildung 65: LSVM: Vergleich Vorhersage zu Ist-Wert (VM+NM mit Ausreißern)[337]

[336] Quelle: eigene Darstellung
[337] Quelle: ebda.

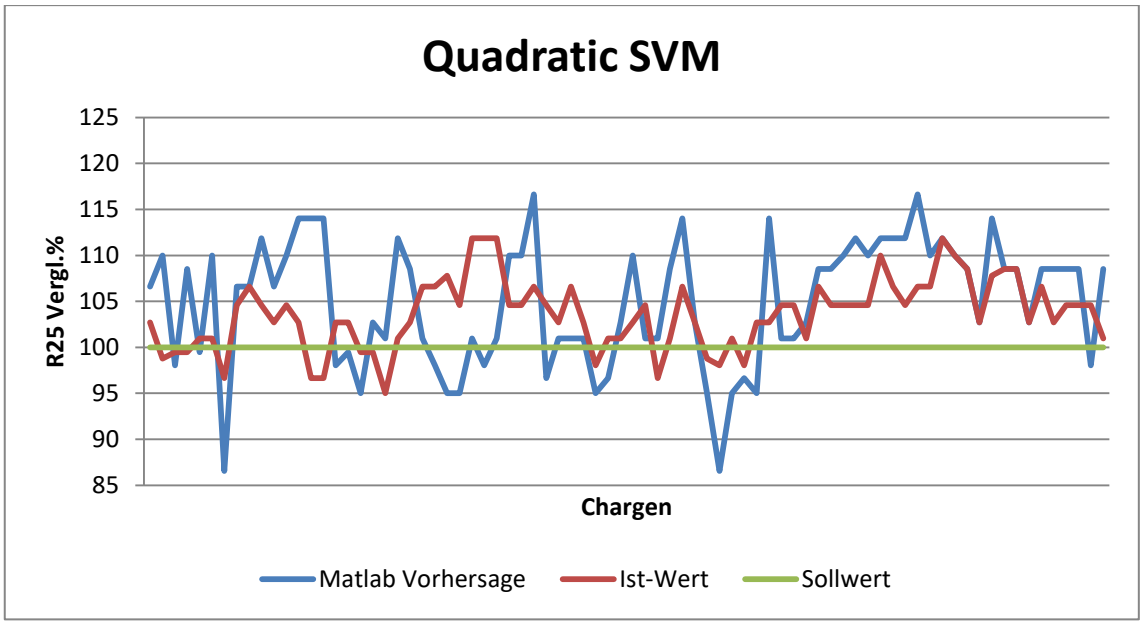


Abbildung 66: QSVM: Vergleich Vorhersage zu Ist-Wert (VM+NM mit Ausreißern)[338]

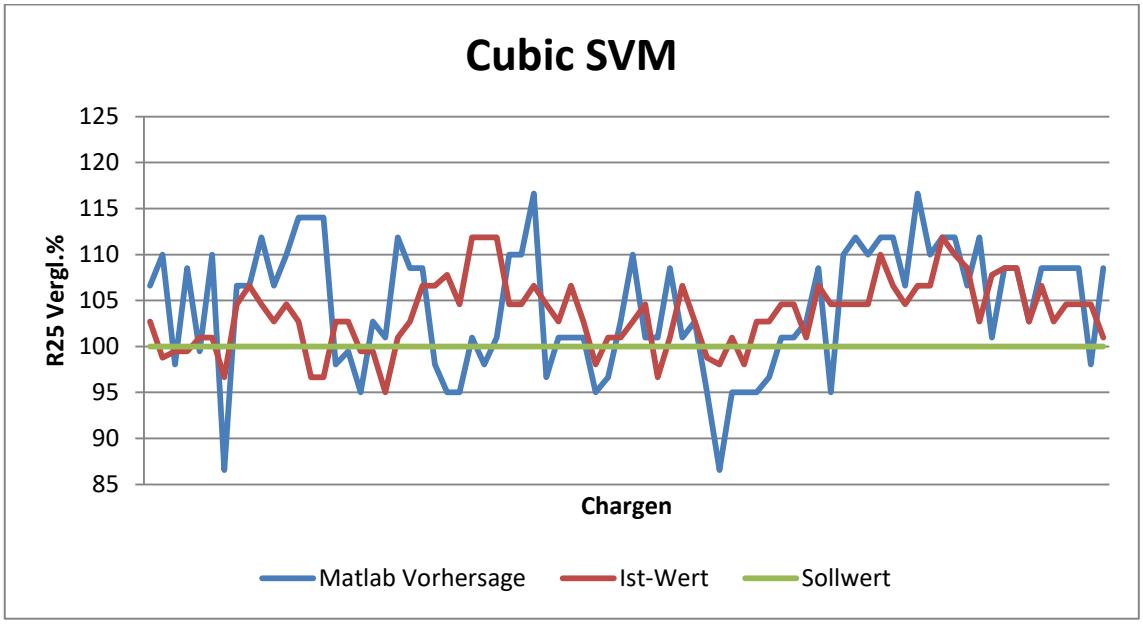


Abbildung 67: CSVM: Vergleich Vorhersage zu Ist-Wert (VM+NM mit Ausreißern)[339]

[338] Quelle: eigene Darstellung
[339] Quelle: ebda.

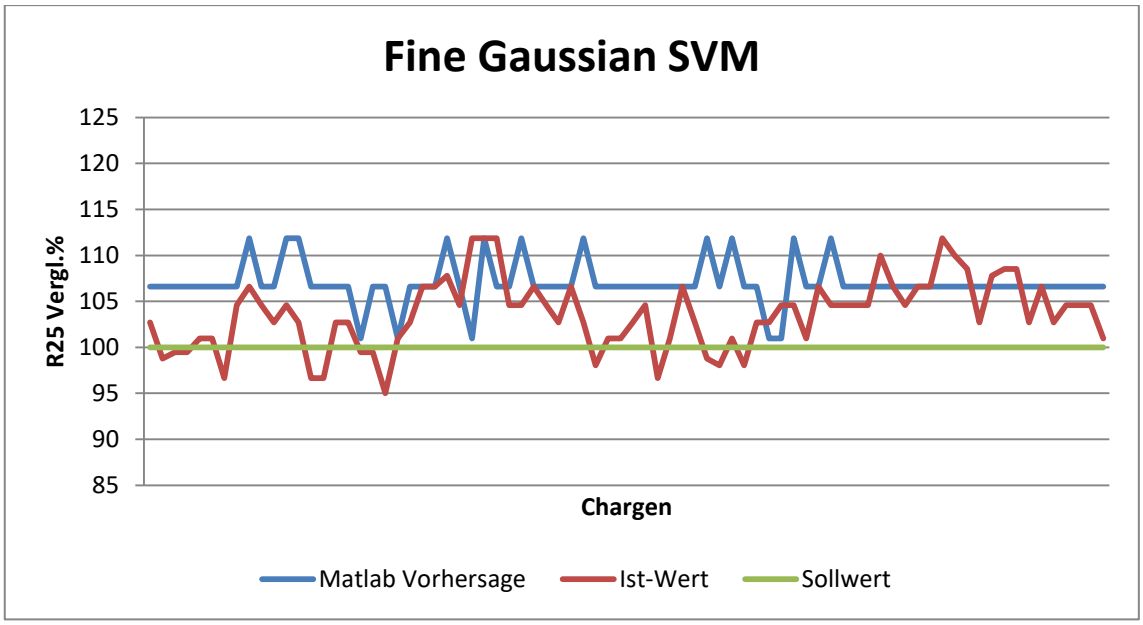


Abbildung 68: FGSVM: Vergleich Vorhersage zu Ist-Wert (VM+NM mit Ausreißern)[340]

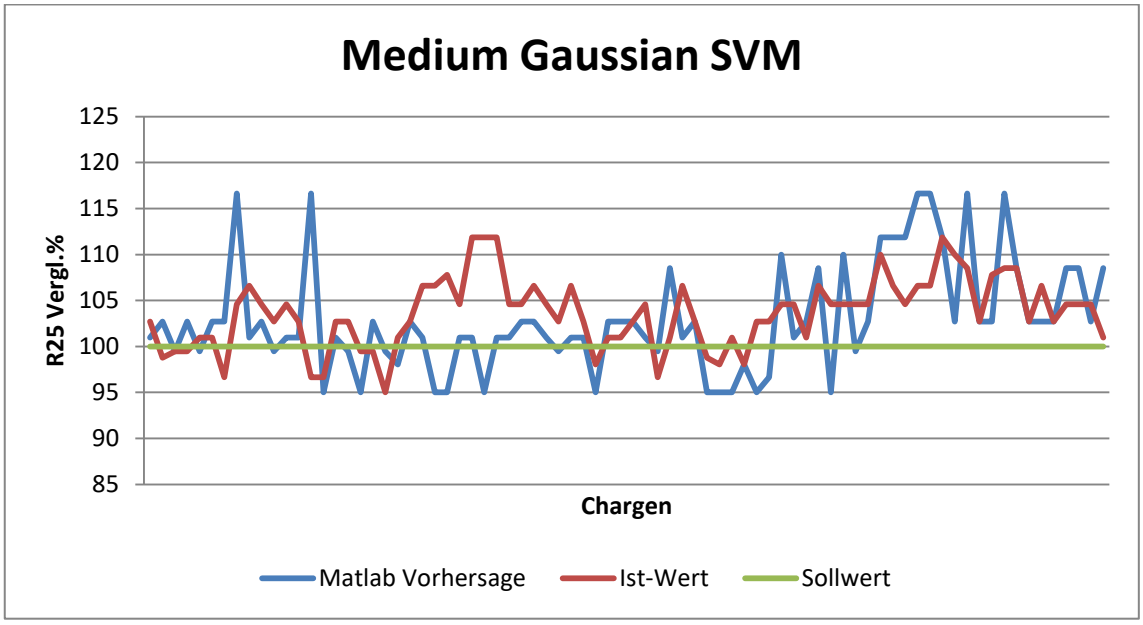


Abbildung 69: MGSVM: Vergleich Vorhersage zu Ist-Wert (VM+NM mit Ausreißern)[341]

[340] Quelle: eigene Darstellung
[341] Quelle: ebda.

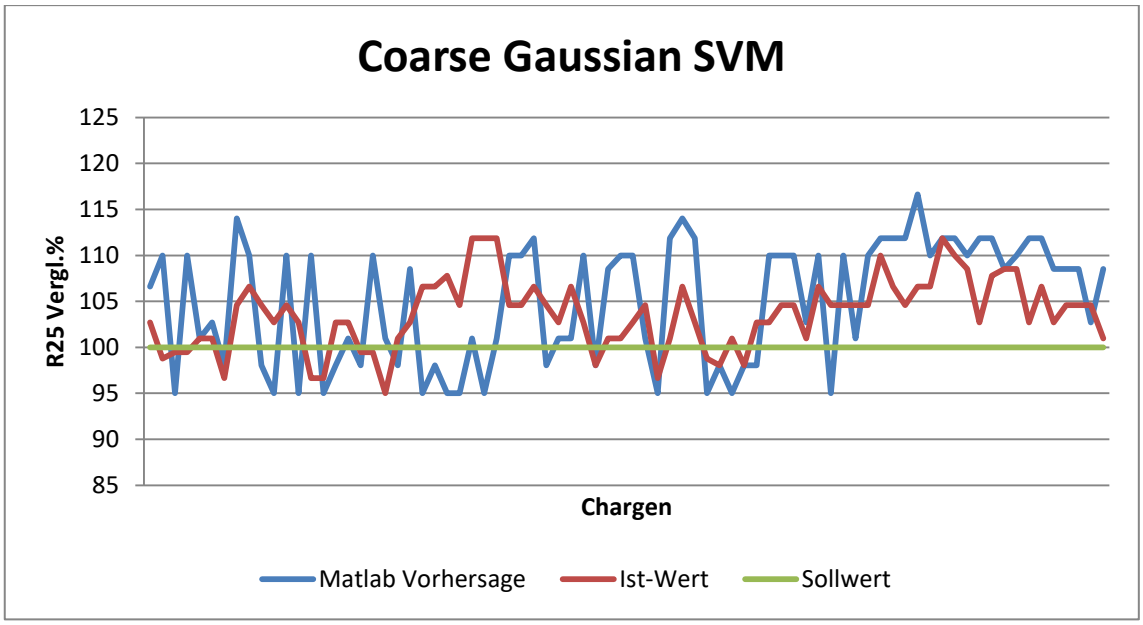


Abbildung 70: CGSVM: Vergleich Vorhersage zu Ist-Wert (VM+NM mit Ausreißern)[342]

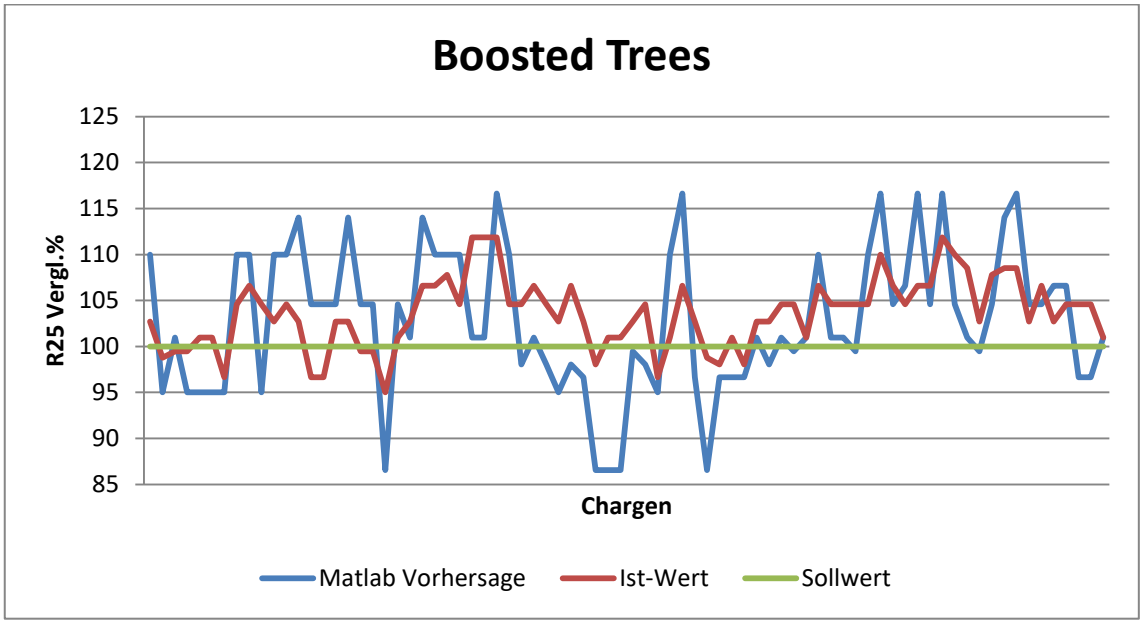


Abbildung 71: BoT: Vergleich Vorhersage zu Ist-Wert (VM+NM mit Ausreißern)[343]

[342] Quelle: eigene Darstellung
[343] Quelle: ebda.

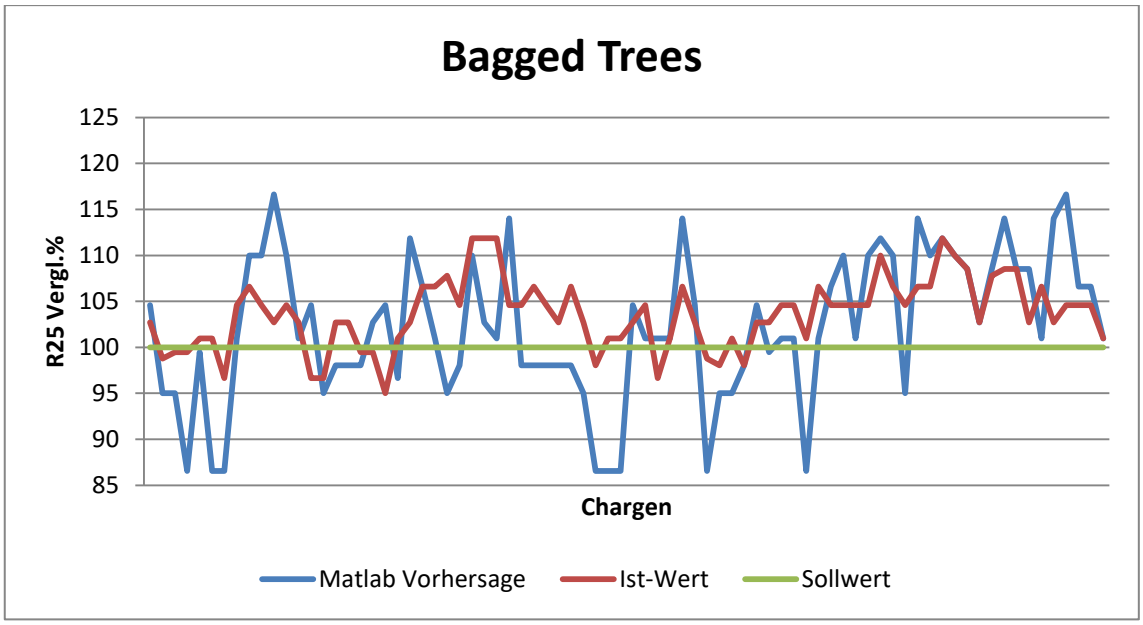


Abbildung 72: BaT: Vergleich Vorhersage zu Ist-Wert (VM+NM mit Ausreißern)[344]

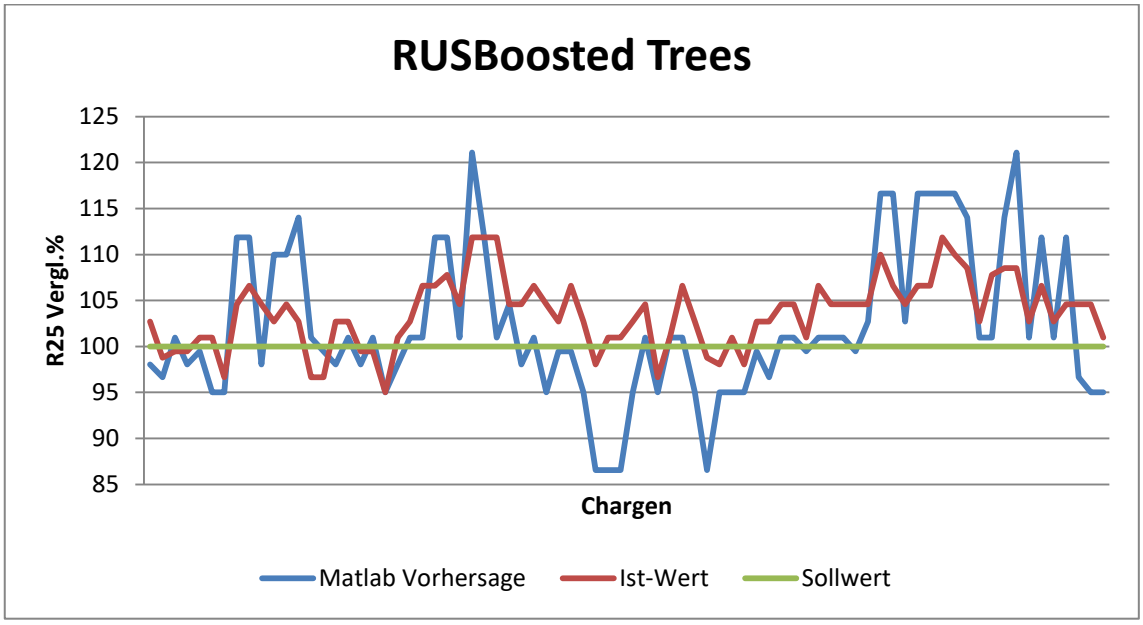


Abbildung 73: RUSBT: Vergleich Vorhersage zu Ist-Wert (VM+NM mit Ausreißern)[345]

[344] Quelle: eigene Darstellung
[345] Quelle: ebda.

Medium Tree VM+NM mit Ausreißern

```
1  if R25Widerstand<-1.36532 then node 2 elseif R25Widerstand>=-
1.36532 then node 3 else 101.0526667
 2  if Durchfluss<0.144277 then node 4 elseif
Durchfluss>=0.144277 then node 5 else 88.62466667
 3  if YAcet<0.238895 then node 6 elseif YAcet>=0.238895 then
node 7 else 102.9246667
 4  class = 101.0526667
 5  class = 88.62466667
 6  if R25Widerstand<0.119414 then node 8 elseif
R25Widerstand>=0.119414 then node 9 else 98.03066667
 7  if SchlickerdruckMW_VM<1.51411 then node 10 elseif
SchlickerdruckMW_VM>=1.51411 then node 11 else 106.6053333
 8  if DifferenzdruckFilter<1.50558 then node 12 elseif
DifferenzdruckFilter>=1.50558 then node 13 else 96.936
 9  if YAcet<-0.57182 then node 14 elseif YAcet>=-0.57182 then
node 15 else 110.0333333
10  if d90NM<-1.49482 then node 16 elseif d90NM>=-1.49482 then
node 17 else 108.566
11  if R25Widerstand<-1.15983 then node 18 elseif
R25Widerstand>=-1.15983 then node 19 else 106.6053333
12  if pHWertSchlicker_VM<-1.38409 then node 20 elseif
pHWertSchlicker_VM>=-1.38409 then node 21 else 96.936
13  class = 104.642
14  if R25Widerstand<0.860772 then node 22 elseif
R25Widerstand>=0.860772 then node 23 else 110.0333333
15  if DurchflussMW_VM<0.282349 then node 24 elseif
DurchflussMW_VM>=0.282349 then node 25 else 101.0526667
16  class = 114.066
17  if R25Widerstand<1.04813 then node 26 elseif
R25Widerstand>=1.04813 then node 27 else 108.566
18  class = 104.642
19  if DifferenzdruckFilter<-0.20924 then node 28 elseif
DifferenzdruckFilter>=-0.20924 then node 29 else 106.6053333
20  class = 99.516
21  if d50VM_VM<-0.978351 then node 30 elseif d50VM_VM>=-
0.978351 then node 31 else 94.87333333
22  if SegmentzeitMAX_VM<0.561374 then node 32 elseif
SegmentzeitMAX_VM>=0.561374 then node 33 else 110.0333333
23  if d10NM<0.345118 then node 34 elseif d10NM>=0.345118 then
node 35 else 111.9966667
24  if DurchflussMW_VM<-0.350973 then node 36 elseif
DurchflussMW_VM>=-0.350973 then node 37 else 101.0526667
25  class = 104.642
26  if UD_VM_1 in {10 11 7} then node 38 elseif UD_VM_1 in {8 9}
then node 39 else 108.566
27  if Granulatfeuchte<-0.635176 then node 40 elseif
Granulatfeuchte>=-0.635176 then node 41 else 116.8253333
28  class = 106.6053333
29  class = 102.9246667
30  class = 96.936
31  class = 94.87333333
32  class = 104.642
33  class = 110.0333333
34  class = 111.9966667
```

```
35  class = 116.8253333
36  class = 101.0526667
37  class = 102.9246667
38  class = 111.9966667
39  class = 106.6053333
40  class = 116.8253333
41  class = 119.52375
```

Complex Tree VM+NM mit Ausreißern

```
  1  if R25Widerstand<-1.36532 then node 2 elseif
R25Widerstand>=-1.36532 then node 3 else 101.0526667
 2  if Durchfluss<0.144277 then node 4 elseif
Durchfluss>=0.144277 then node 5 else 88.62466667
 3  if YAcet<0.238895 then node 6 elseif YAcet>=0.238895 then
node 7 else 102.9246667
 4  class = 101.0526667
 5  class = 88.62466667
 6  if R25Widerstand<0.119414 then node 8 elseif
R25Widerstand>=0.119414 then node 9 else 98.03066667
 7  if SchlickerdruckMW_VM<1.51411 then node 10 elseif
SchlickerdruckMW_VM>=1.51411 then node 11 else 106.6053333
 8  if DifferenzdruckFilter<1.50558 then node 12 elseif
DifferenzdruckFilter>=1.50558 then node 13 else 96.936
 9  if YAcet<-0.57182 then node 14 elseif YAcet>=-0.57182 then
node 15 else 110.0333333
10  if d90NM<-1.49482 then node 16 elseif d90NM>=-1.49482 then
node 17 else 108.566
11  if R25Widerstand<-1.15983 then node 18 elseif
R25Widerstand>=-1.15983 then node 19 else 106.6053333
12  if pHWertSchlicker_VM<-1.38409 then node 20 elseif
pHWertSchlicker_VM>=-1.38409 then node 21 else 96.936
13  class = 104.642
14  if R25Widerstand<0.860772 then node 22 elseif
R25Widerstand>=0.860772 then node 23 else 110.0333333
15  if DurchflussMW_VM<0.282349 then node 24 elseif
DurchflussMW_VM>=0.282349 then node 25 else 101.0526667
16  class = 114.066
17  if R25Widerstand<1.04813 then node 26 elseif
R25Widerstand>=1.04813 then node 27 else 108.566
18  class = 104.642
19  if DifferenzdruckFilter<-0.20924 then node 28 elseif
DifferenzdruckFilter>=-0.20924 then node 29 else 106.6053333
20  class = 99.516
21  if d50VM_VM<-0.978351 then node 30 elseif d50VM_VM>=-
0.978351 then node 31 else 94.87333333
22  if SegmentzeitMAX_VM<0.561374 then node 32 elseif
SegmentzeitMAX_VM>=0.561374 then node 33 else 110.0333333
23  if d10NM<0.345118 then node 34 elseif d10NM>=0.345118 then
node 35 else 111.9966667
24  if DurchflussMW_VM<-0.350973 then node 36 elseif
DurchflussMW_VM>=-0.350973 then node 37 else 101.0526667
25  class = 104.642
26  if UD_VM_1 in {10 11 7} then node 38 elseif UD_VM_1 in {8 9}
then node 39 else 108.566
```

```
27  if Granulatfeuchte<-0.635176 then node 40 elseif
Granulatfeuchte>=-0.635176 then node 41 else 116.8253333
28  class = 106.6053333
29  class = 102.9246667
30  if d90VM_VM<-1.00417 then node 42 elseif d90VM_VM>=-1.00417
then node 43 else 96.936
31  if R25Widerstand<-0.517188 then node 44 elseif
R25Widerstand>=-0.517188 then node 45 else 94.87333333
32  class = 104.642
33  class = 110.0333333
34  class = 111.9966667
35  class = 116.8253333
36  class = 101.0526667
37  class = 102.9246667
38  if d10GranulatNM<-1.40239 then node 46 elseif
d10GranulatNM>=-1.40239 then node 47 else 111.9966667
39  if d10NM<-0.693852 then node 48 elseif d10NM>=-0.693852 then
node 49 else 106.6053333
40  class = 116.8253333
41  class = 119.52375
42  if d90NM<-0.555518 then node 50 elseif d90NM>=-0.555518 then
node 51 else 96.936
43  class = 101.0526667
44  if UD_VM_1 in {11 8 9} then node 52 elseif UD_VM_1 in {10 7}
then node 53 else 94.87333333
45  if d10NM<-0.407239 then node 54 elseif d10NM>=-0.407239 then
node 55 else 98.03066667
46  class = 119.52375
47  if d90NM<-0.636609 then node 56 elseif d90NM>=-0.636609 then
node 57 else 111.9966667
48  class = 106.6053333
49  if SchlickerdichteMIN_VM<0.0450895 then node 58 elseif
SchlickerdichteMIN_VM>=0.0450895 then node 59 else 108.566
50  class = 98.03066667
51  class = 96.936
52  if Gluehverlust<-0.852915 then node 60 elseif
Gluehverlust>=-0.852915 then node 61 else 94.87333333
53  class = 96.936
54  class = 94.87333333
55  if d90NM<0.0188779 then node 62 elseif d90NM>=0.0188779 then
node 63 else 99.516
56  class = 111.9966667
57  if d10GranulatNM<-0.437824 then node 64 elseif
d10GranulatNM>=-0.437824 then node 65 else 102.9246667
58  class = 101.0526667
59  class = 108.566
60  class = 96.936
61  if d50NM<-0.602723 then node 66 elseif d50NM>=-0.602723 then
node 67 else 94.87333333
62  class = 98.03066667
63  if Symm3MW<0.403742 then node 68 elseif Symm3MW>=0.403742
then node 69 else 99.516
64  class = 110.0333333
65  if DurchflussMIN_VM<0.152816 then node 70 elseif
DurchflussMIN_VM>=0.152816 then node 71 else 102.9246667
66  class = 88.62466667
```

```
67  if DifferenzdruckFilter<1.16262 then node 72 elseif
DifferenzdruckFilter>=1.16262 then node 73 else 94.87333333
68  class = 99.516
69  class = 101.0526667
70  if d90NM<-0.187229 then node 74 elseif d90NM>=-0.187229 then
node 75 else 102.9246667
71  if Tb<0.774428 then node 76 elseif Tb>=0.774428 then node 77
else 104.642
72  class = 94.87333333
73  class = 98.03066667
74  class = 108.566
75  class = 102.9246667
76  class = 114.066
77  class = 104.642
```

Tabelle 35: Ergebnis CL und Testdaten: VM+NM ohne Ausreißer[346]

	Genauigkeit bei Training	Genauigkeit bei Anwendung auf Testdaten	Grenzfälle bei Anwendung auf Testdaten	Mittelwert der absoluten Differenz zwischen der vorhergesagten und Ist-Klasse
Complex Tree	16,5%	11,5%	1,3%	2,36
Medium Tree	16,1%	14,1%	1,3%	2,06
Simple Tree	11,9%	9,0%	1,3%	2,71
Linear SVM	17,0%	9,0%	0,0%	3,13
Quadratic SVM	18,8%	10,3%	0,0%	3,14
Cubic SVM	19,3%	9,0%	0,0%	3,37
Fine Gaussian SVM	15,1%	10,3%	0,0%	2,72
Medium Gaussian SVM	19,7%	10,3%	0,0%	3,14
Coarse Gaussian SVM	17,9%	2,6%	0,0%	3,59
Boosted Trees	17,9%	14,1%	1,3%	2,31
Bagged Trees	18,3%	10,3%	0,0%	2,92
RUSBoosted Trees	19,3%	9,0%	1,3%	2,88

[346] Quelle: eigene Darstellung

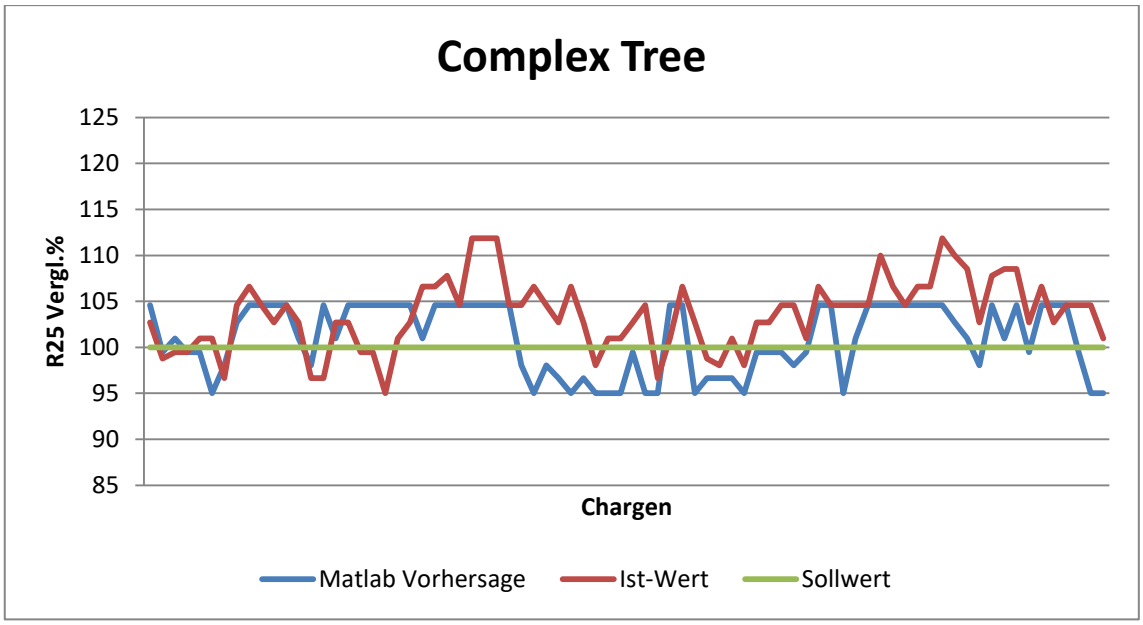


Abbildung 74: CT: Vergleich Vorhersage zu Ist-Wert (VM+NM ohne Ausreißer)[347]

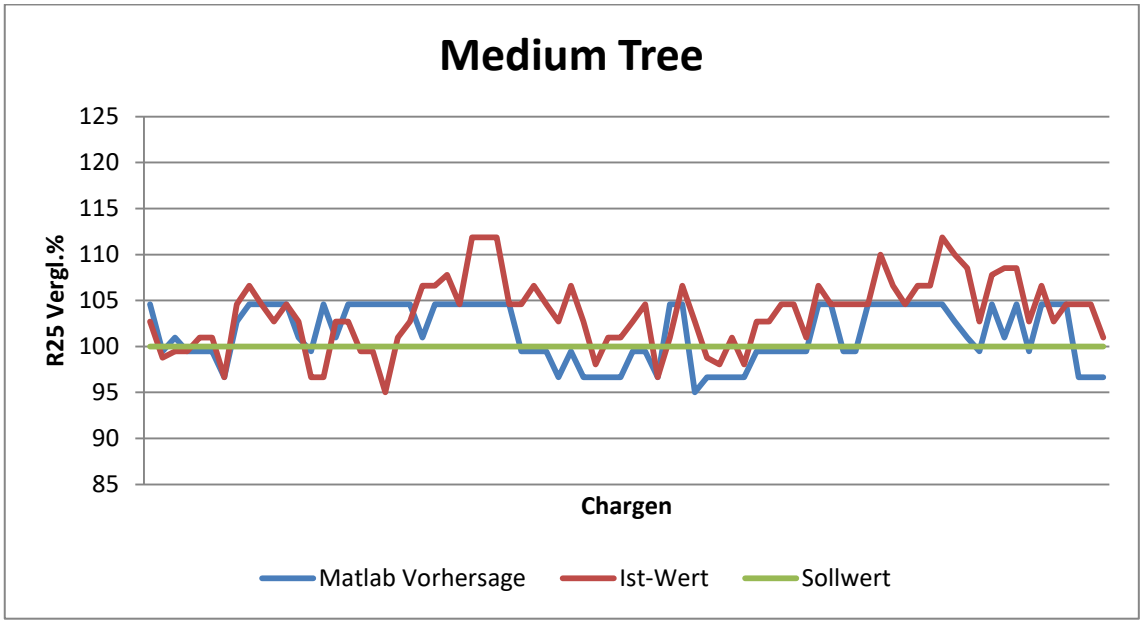


Abbildung 75: MT: Vergleich Vorhersage zu Ist-Wert (VM+NM ohne Ausreißer)[348]

[347] Quelle: eigene Darstellung
[348] Quelle: ebda.

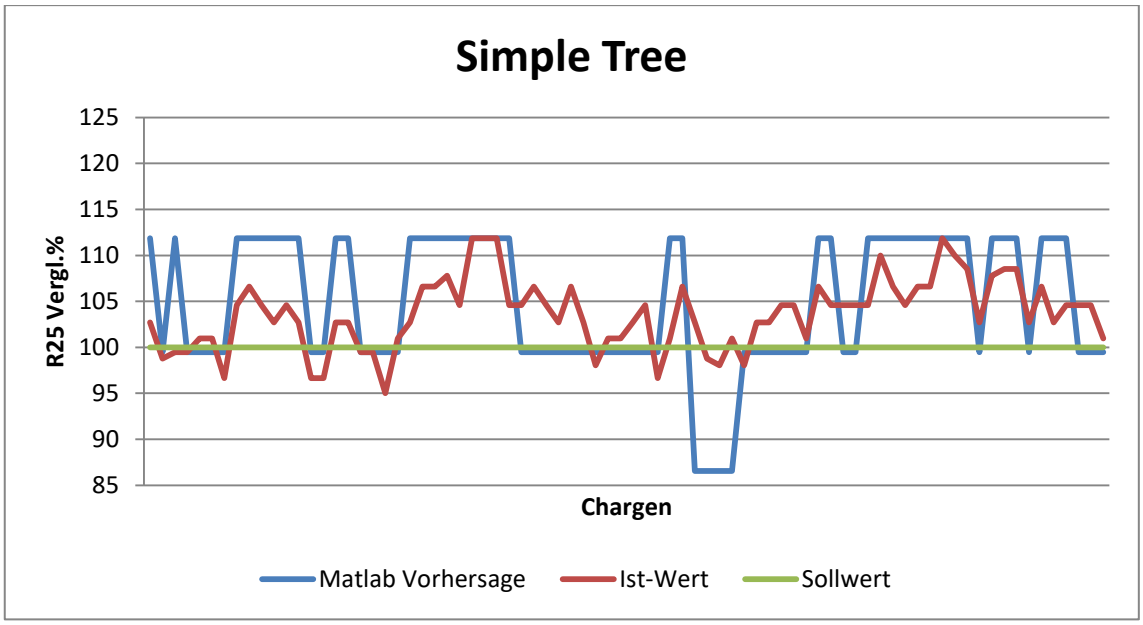


Abbildung 76: ST: Vergleich Vorhersage zu Ist-Wert (VM+NM ohne Ausreißer)[349]

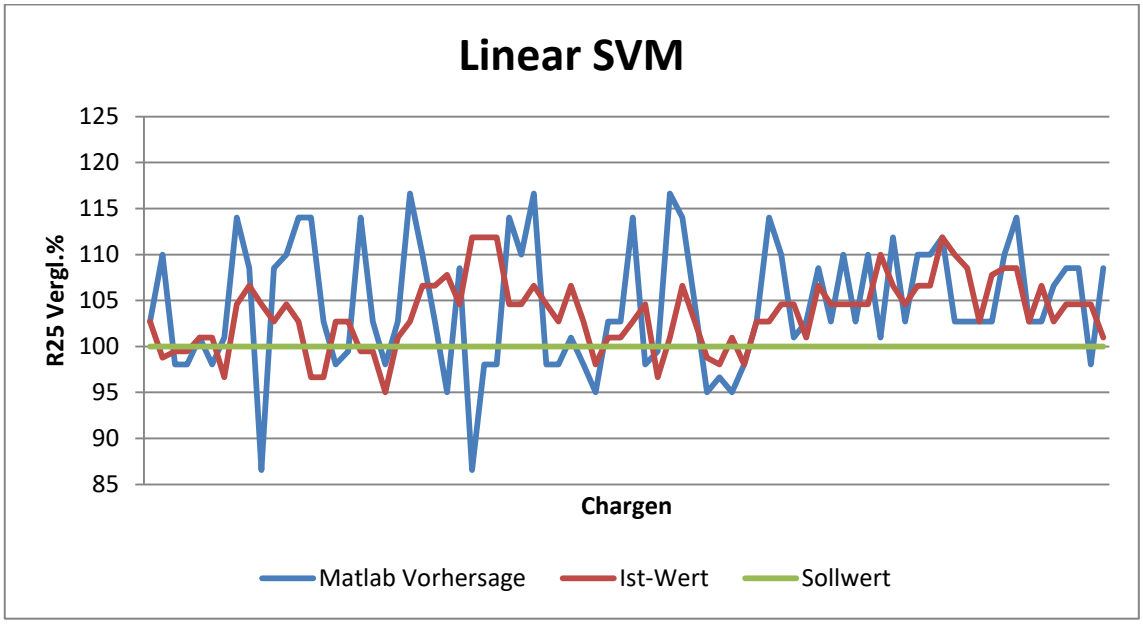


Abbildung 77: LSVM: Vergleich Vorhersage zu Ist-Wert (VM+NM ohne Ausreißer)[350]

[349] Quelle: eigene Darstellung
[350] Quelle: ebda.

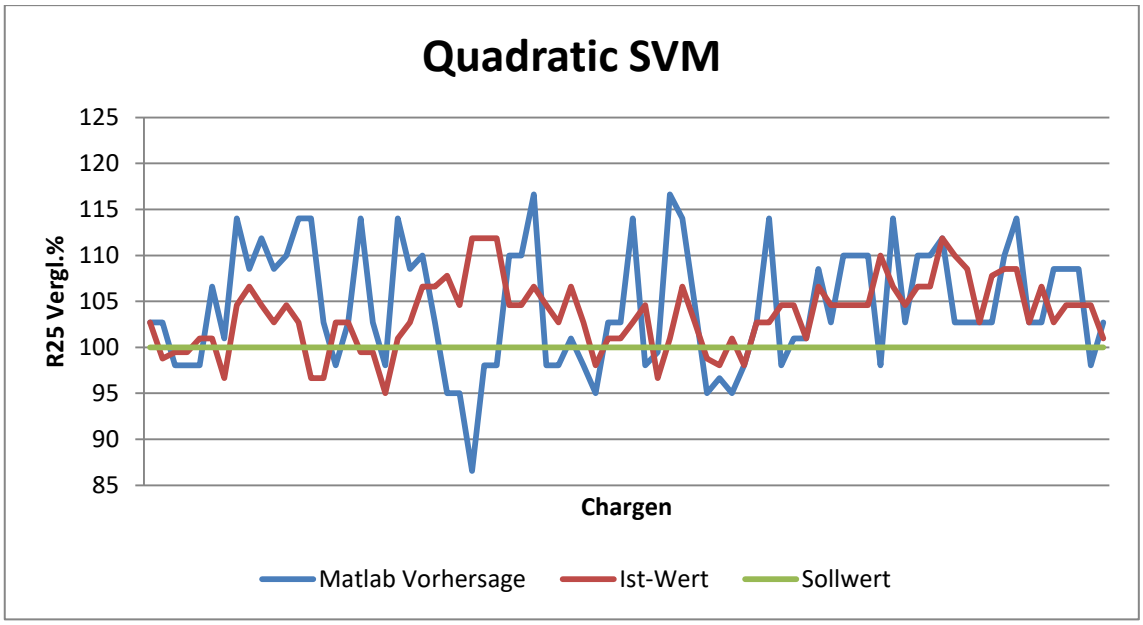


Abbildung 78: QSVM: Vergleich Vorhersage zu Ist-Wert (VM+NM ohne Ausreißer)[351]

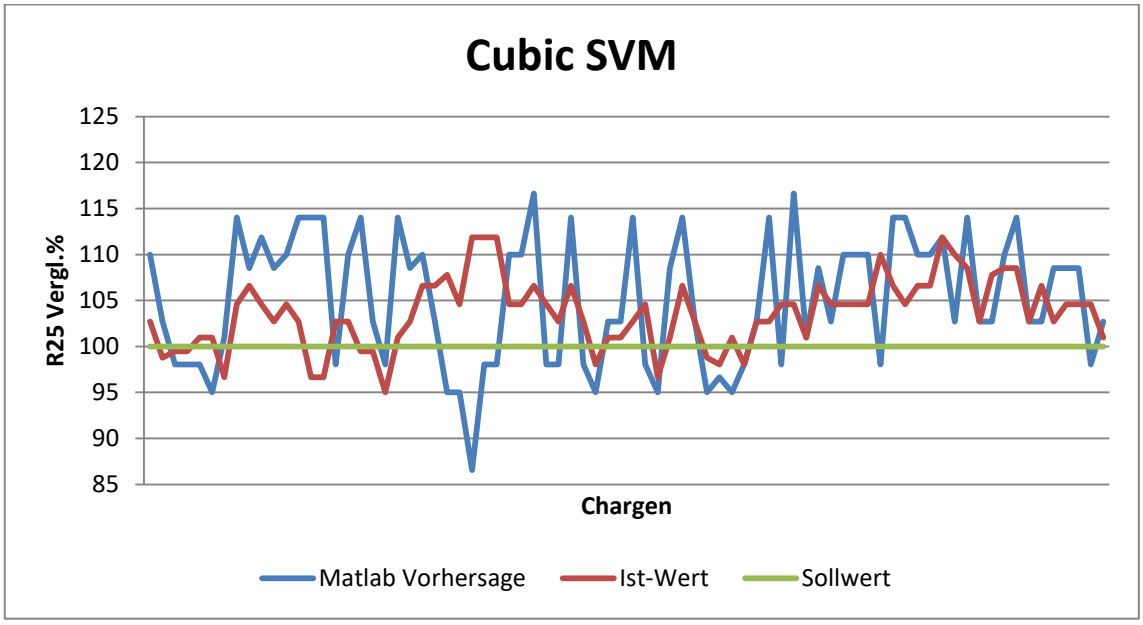


Abbildung 79: CSVM: Vergleich Vorhersage zu Ist-Wert (VM+NM ohne Ausreißer)[352]

[351] Quelle: eigene Darstellung
[352] Quelle: ebda.

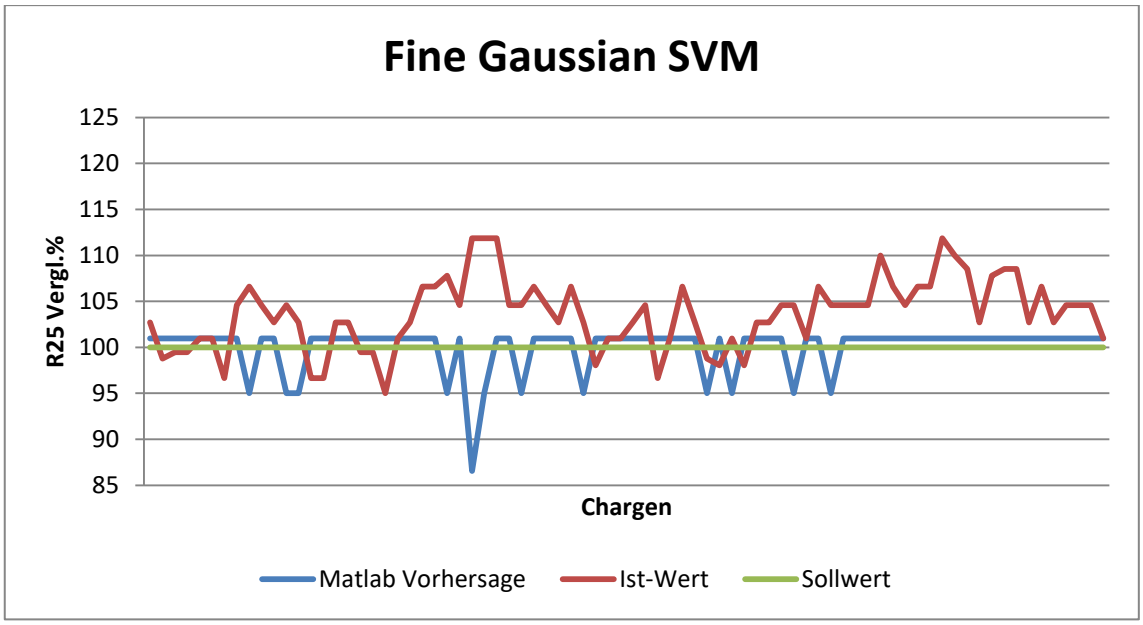


Abbildung 80: FGSVM: Vergleich Vorhersage zu Ist-Wert (VM+NM ohne Ausreißer)[353]

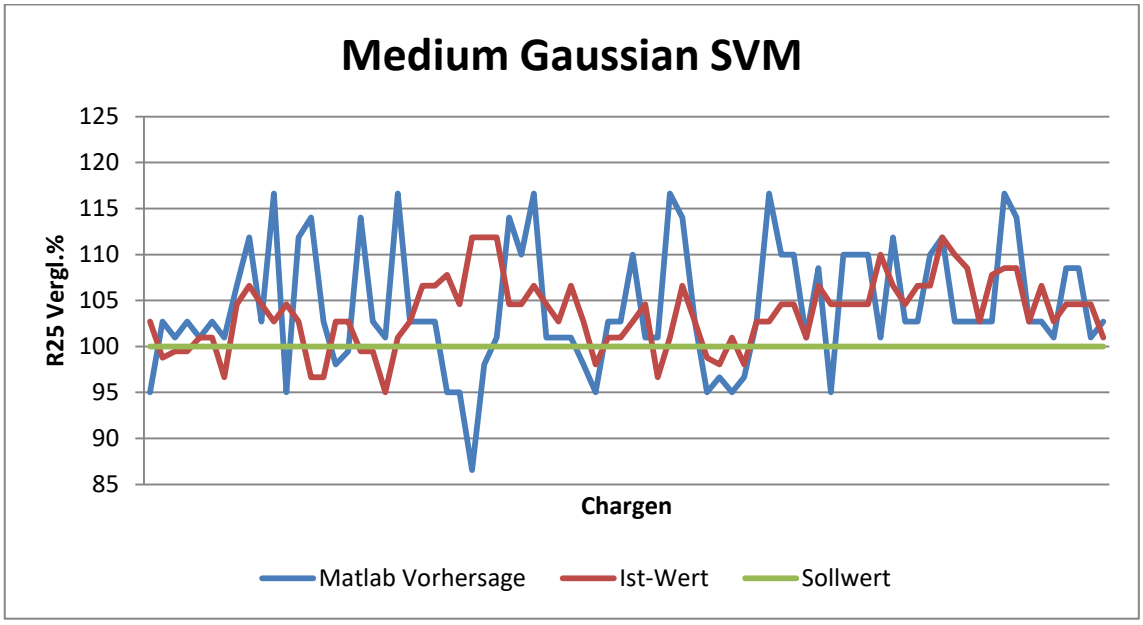


Abbildung 81: MGSVM: Vergleich Vorhersage zu Ist-Wert (VM+NM ohne Ausreißer)[354]

[353] Quelle: eigene Darstellung
[354] Quelle: ebda.

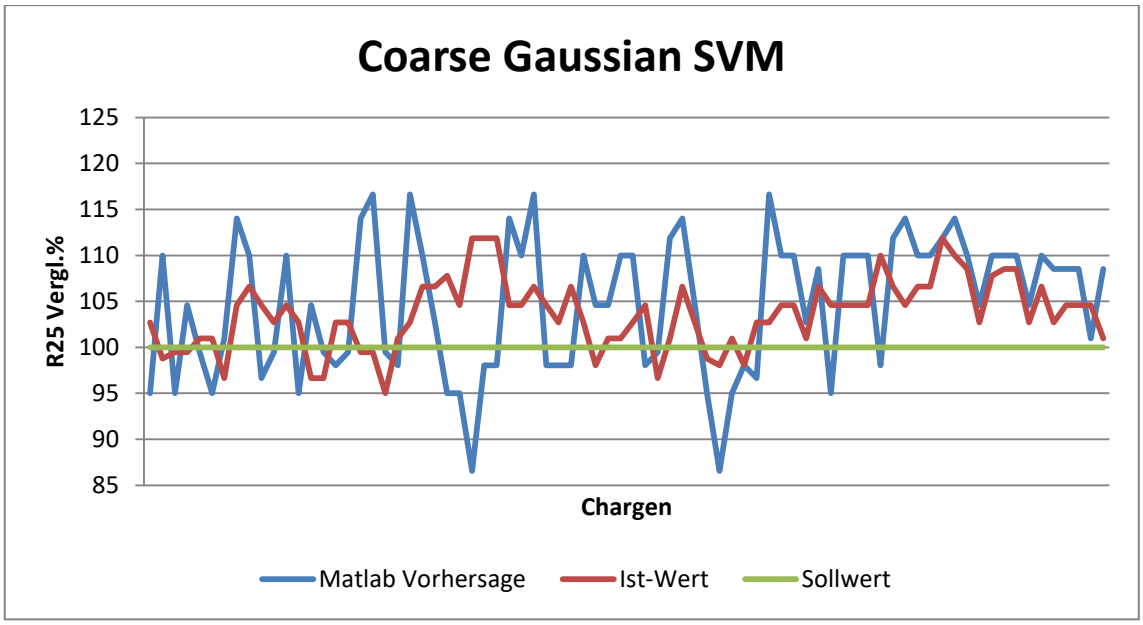


Abbildung 82: CGSVM: Vergleich Vorhersage zu Ist-Wert (VM+NM ohne Ausreißer)[355]

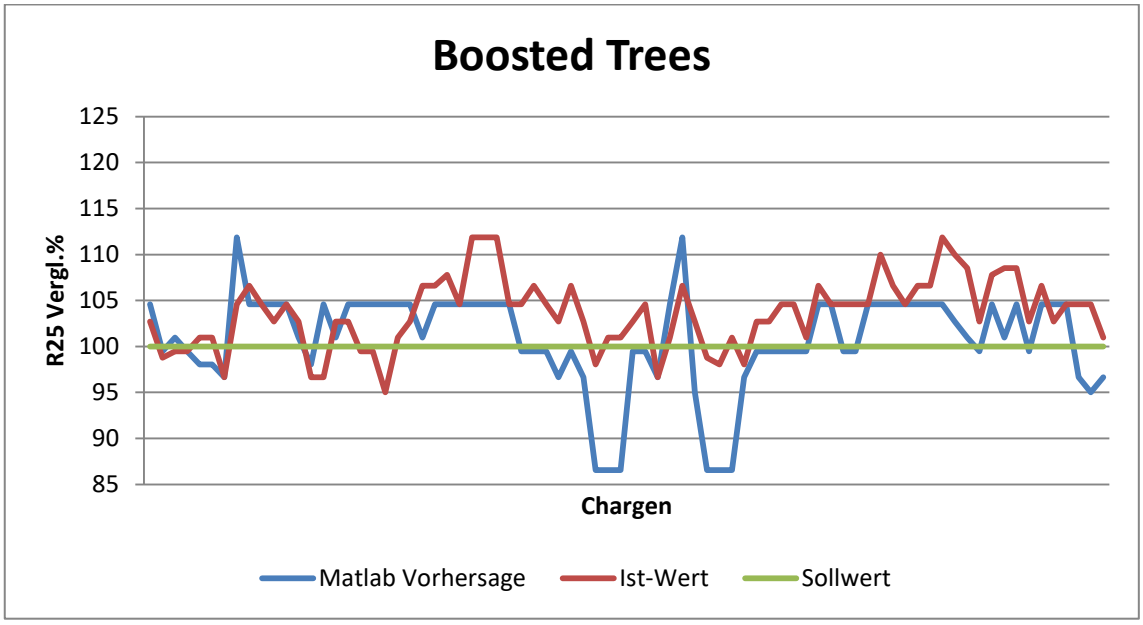


Abbildung 83: BoT: Vergleich Vorhersage zu Ist-Wert (VM+NM ohne Ausreißer)[356]

[355] Quelle: eigene Darstellung
[356] Quelle: ebda.

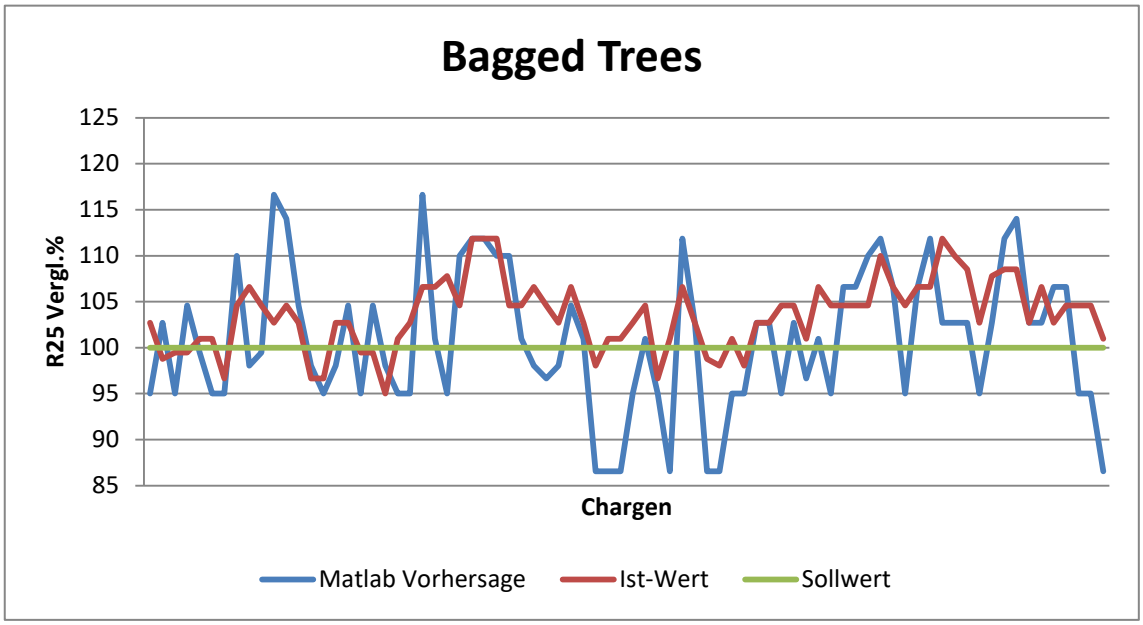


Abbildung 84: BaT: Vergleich Vorhersage zu Ist-Wert (VM+NM ohne Ausreißer)[357]

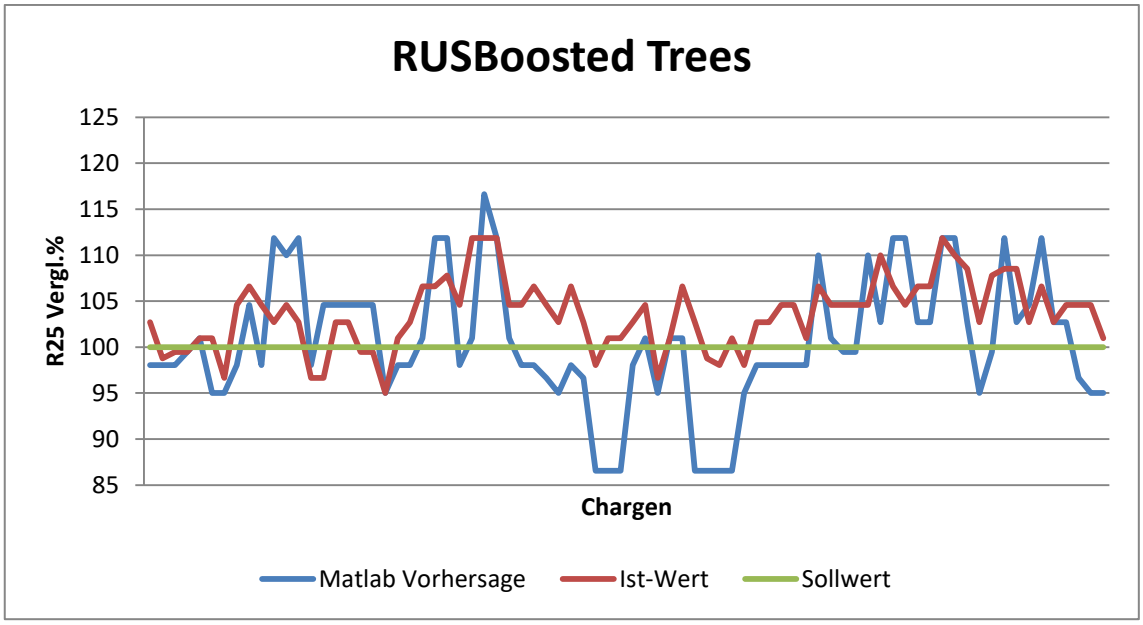


Abbildung 85: RUSBT: Vergleich Vorhersage zu Ist-Wert (VM+NM ohne Ausreißer)[358]

[357] Quelle: eigene Darstellung
[358] Quelle: ebda.

Medium Tree VM+NM ohne Ausreißer

```
1  if YAcet<0.492441 then node 2 elseif YAcet>=0.492441 then
node 3 else 88.625
 2  if R25Widerstand<0.200899 then node 4 elseif
R25Widerstand>=0.200899 then node 5 else 88.625
 3  if d90NM<-1.81552 then node 6 elseif d90NM>=-1.81552 then
node 7 else 106.61
 4  if d10NM<-1.05918 then node 8 elseif d10NM>=-1.05918 then
node 9 else 88.625
 5  if YAcet<-0.697275 then node 10 elseif YAcet>=-0.697275 then
node 11 else 112
 6  class = 114.07
 7  if SchlickerdruckMW_VM<1.51411 then node 12 elseif
SchlickerdruckMW_VM>=1.51411 then node 13 else 106.61
 8  if TRmin<0.338175 then node 14 elseif TRmin>=0.338175 then
node 15 else 88.625
 9  if DifferenzdruckFilter<1.50558 then node 16 elseif
DifferenzdruckFilter>=1.50558 then node 17 else 99.516
10  if R25Widerstand<1.07384 then node 18 elseif
R25Widerstand>=1.07384 then node 19 else 110.03
11  if DurchflussMW_VM<-0.859258 then node 20 elseif
DurchflussMW_VM>=-0.859258 then node 21 else 101.05
12  if UD_VM_1 in {10 11 7} then node 22 elseif UD_VM_1 in {8 9}
then node 23 else 108.57
13  if R25Widerstand<-1.37386 then node 24 elseif
R25Widerstand>=-1.37386 then node 25 else 106.61
14  if d90GranulatNM<0.0729912 then node 26 elseif
d90GranulatNM>=0.0729912 then node 27 else 88.625
15  class = 94.873
16  if R25Widerstand<-0.582763 then node 28 elseif
R25Widerstand>=-0.582763 then node 29 else 99.516
17  class = 104.64
18  if SegmentzeitMAX_VM<0.599055 then node 30 elseif
SegmentzeitMAX_VM>=0.599055 then node 31 else 110.03
19  if d10NM<0.408982 then node 32 elseif d10NM>=0.408982 then
node 33 else 112
20  class = 101.05
21  if Schlickerdichte<-0.0478531 then node 34 elseif
Schlickerdichte>=-0.0478531 then node 35 else 102.92
22  if Tb<0.202637 then node 36 elseif Tb>=0.202637 then node 37
else 116.83
23  if d10NM<-0.741741 then node 38 elseif d10NM>=-0.741741 then
node 39 else 106.61
24  class = 104.64
25  if DifferenzdruckFilter<-0.20924 then node 40 elseif
DifferenzdruckFilter>=-0.20924 then node 41 else 106.61
26  class = 88.625
27  class = 96.936
28  class = 96.936
29  class = 99.516
30  class = 104.64
31  class = 110.03
32  class = 112
33  class = 116.83
34  class = 102.92
```

```
35  class = 104.64
36  class = 116.83
37  class = 112
38  class = 106.61
39  class = 108.57
40  class = 106.61
41  class = 102.92
```

Complex Tree VM+NM ohne Ausreißer

```
1   if YAcet<0.492441 then node 2 elseif YAcet>=0.492441 then
node 3 else 88.625
 2  if R25Widerstand<0.200899 then node 4 elseif
R25Widerstand>=0.200899 then node 5 else 88.625
 3  if d90NM<-1.81552 then node 6 elseif d90NM>=-1.81552 then
node 7 else 106.61
 4  if d10NM<-1.05918 then node 8 elseif d10NM>=-1.05918 then
node 9 else 88.625
 5  if YAcet<-0.697275 then node 10 elseif YAcet>=-0.697275 then
node 11 else 112
 6  class = 114.07
 7  if SchlickerdruckMW_VM<1.51411 then node 12 elseif
SchlickerdruckMW_VM>=1.51411 then node 13 else 106.61
 8  if TRmin<0.338175 then node 14 elseif TRmin>=0.338175 then
node 15 else 88.625
 9  if DifferenzdruckFilter<1.50558 then node 16 elseif
DifferenzdruckFilter>=1.50558 then node 17 else 99.516
10  if R25Widerstand<1.07384 then node 18 elseif
R25Widerstand>=1.07384 then node 19 else 110.03
11  if DurchflussMW_VM<-0.859258 then node 20 elseif
DurchflussMW_VM>=-0.859258 then node 21 else 101.05
12  if UD_VM_1 in {10 11 7} then node 22 elseif UD_VM_1 in {8 9}
then node 23 else 108.57
13  if R25Widerstand<-1.37386 then node 24 elseif
R25Widerstand>=-1.37386 then node 25 else 106.61
14  if d90GranulatNM<0.0729912 then node 26 elseif
d90GranulatNM>=0.0729912 then node 27 else 88.625
15  class = 94.873
16  if R25Widerstand<-0.582763 then node 28 elseif
R25Widerstand>=-0.582763 then node 29 else 99.516
17  class = 104.64
18  if SegmentzeitMAX_VM<0.599055 then node 30 elseif
SegmentzeitMAX_VM>=0.599055 then node 31 else 110.03
19  if d10NM<0.408982 then node 32 elseif d10NM>=0.408982 then
node 33 else 112
20  class = 101.05
21  if Schlickerdichte<-0.0478531 then node 34 elseif
Schlickerdichte>=-0.0478531 then node 35 else 102.92
22  if Tb<0.202637 then node 36 elseif Tb>=0.202637 then node 37
else 116.83
23  if d10NM<-0.741741 then node 38 elseif d10NM>=-0.741741 then
node 39 else 106.61
24  class = 104.64
25  if DifferenzdruckFilter<-0.20924 then node 40 elseif
DifferenzdruckFilter>=-0.20924 then node 41 else 106.61
```

```
26  if Durchfluss<-0.448149 then node 42 elseif Durchfluss>=-
0.448149 then node 43 else 88.625
27  class = 96.936
28  if DurchflussStabw_VM<0.554187 then node 44 elseif
DurchflussStabw_VM>=0.554187 then node 45 else 96.936
29  if Heizgruppe11_VM<-1.23287 then node 46 elseif
Heizgruppe11_VM>=-1.23287 then node 47 else 99.516
30  class = 104.64
31  class = 110.03
32  class = 112
33  class = 116.83
34  class = 102.92
35  class = 104.64
36  if SchuettdichteGranulat_VM<-0.643787 then node 48 elseif
SchuettdichteGranulat_VM>=-0.643787 then node 49 else 116.83
37  if pHWertSchlicker_VM<-0.000115404 then node 50 elseif
pHWertSchlicker_VM>=-0.000115404 then node 51 else 112
38  class = 106.61
39  if TRmin<-0.243261 then node 52 elseif TRmin>=-0.243261 then
node 53 else 108.57
40  class = 106.61
41  class = 102.92
42  class = 98.031
43  class = 88.625
44  if R25Widerstand<-1.04899 then node 54 elseif
R25Widerstand>=-1.04899 then node 55 else 96.936
45  class = 94.873
46  class = 101.05
47  if Ringtemperatur_VM<-0.335616 then node 56 elseif
Ringtemperatur_VM>=-0.335616 then node 57 else 99.516
48  class = 119.52
49  if Durchfluss<-0.986719 then node 58 elseif Durchfluss>=-
0.986719 then node 59 else 116.83
50  if d50NM<0.797421 then node 60 elseif d50NM>=0.797421 then
node 61 else 104.64
51  if Gluehverlust<-0.623386 then node 62 elseif
Gluehverlust>=-0.623386 then node 63 else 112
52  class = 101.05
53  class = 108.57
54  class = 94.873
55  if DifferenzdruckFilter<0.991135 then node 64 elseif
DifferenzdruckFilter>=0.991135 then node 65 else 96.936
56  class = 99.516
57  if Ablufttemperatur_VM<0.313699 then node 66 elseif
Ablufttemperatur_VM>=0.313699 then node 67 else 98.031
58  class = 114.07
59  if SchlickertempStabw_VM<-0.120338 then node 68 elseif
SchlickertempStabw_VM>=-0.120338 then node 69 else 116.83
60  class = 104.64
61  class = 110.03
62  class = 102.92
63  class = 112
64  if DurchflussMW_VM<-0.72149 then node 70 elseif
DurchflussMW_VM>=-0.72149 then node 71 else 96.936
65  class = 98.031
66  class = 98.031
```

```
67  class = 94.873
68  class = 116.83
69  class = 119.52
70  class = 99.516
71  class = 96.936
```

Tabelle 36: Statistische Maße für Šumperk (mit Ausreißern)[359]

Bezeichnung	$\tilde{s}^2$	$\tilde{s}$	γ	g_m	Korrelations-koeffizient mit MW
MW	1,33E+01	3,65E+00	4,54	1,52	1,00
Kugelmenge_DL	8,87E-02	2,98E-01	-4,30	-0,19	0,22
Durchfluss_DL	7,32E-01	8,56E-01	-3,30	-0,38	-0,03
Schlickertemperatur_DL	5,86E+00	2,42E+00	-3,44	-0,37	0,07
d10 (NM)_DL	9,59E-05	9,79E-03	-3,41	-0,34	0,27
d50 (NM)_DL	3,16E-04	1,78E-02	-2,43	-0,99	-0,08
d90 (NM)_DL	1,16E-02	1,08E-01	-2,87	-0,44	-0,15
Auslaufzeit_DL	2,27E+00	1,51E+00	9,48	2,81	-0,09
Schuettdichte Granulat_DL	1,27E-04	1,13E-02	-1,78	-0,61	0,16
Granulatfeuchte_DL	4,94E-05	7,03E-03	-2,88	0,04	0,01
Differenzdruck Filter_DL	6,84E-04	2,61E-02	-2,34	-0,27	-0,04
d10 Granulat (NM)_DL	5,91E+00	2,43E+00	-2,99	0,31	-0,30
d50 Granulat (NM)_DL	1,93E+01	4,39E+00	-3,06	0,22	-0,24
d90 Granulat (NM)_DL	2,25E+01	4,74E+00	-2,71	0,52	-0,18
Schlickerdichte_DL	1,18E-04	1,09E-02	-2,40	-0,01	-0,11
pH Wert Schlicker_DL	2,62E-02	1,62E-01	-3,58	-0,30	0,04
Gluehverlust_DL	7,33E-04	2,71E-02	-3,44	0,43	0,18
Sinterdichte_DL	9,80E-05	9,90E-03	-3,56	-0,40	0,01
R25 Widerstand_DL	3,42E+00	1,85E+00	-2,71	0,55	0,18
R25 Vergl,%_DL	4,07E+01	6,38E+00	-3,93	-0,15	0,39
R25_DL	1,94E+00	1,39E+00	-3,34	-0,02	0,19
Rmin_DL	1,15E+00	1,07E+00	-3,21	0,01	0,23
T-Rmin_DL	3,42E+00	1,85E+00	-2,70	-0,56	-0,15
Tb_DL	3,50E-01	5,92E-01	-2,52	-0,48	-0,25
Tcorr_DL	2,96E+00	1,72E+00	-2,25	-0,62	-0,18
Gleitwert_DL	3,61E-04	1,90E-02	-1,19	-1,10	0,29
Gruenteilfestigkeit_DL	3,46E-01	5,88E-01	-3,38	0,04	-0,25
Kohaesion_DL	8,30E-03	9,11E-02	-1,85	-0,60	0,16
Pressdruck_DL	1,26E+01	3,55E+00	-2,80	0,76	-0,28
Radiale Auffederung_DL	3,50E-07	5,92E-04	-3,62	0,26	-0,19
Schuettwinkel_DL	4,79E-01	6,92E-01	-3,23	0,29	0,19
SPHT3 MW_DL	2,16E-06	1,47E-03	-2,87	-0,05	-0,14
Symm3 MW_DL	1,19E-06	1,09E-03	-2,26	-0,66	0,10
b_l3 MW_DL	1,02E-05	3,19E-03	-2,97	-0,15	0,09
Contraspum K1012_Menge1_NM_Gehalt_DL	8,19E-10	2,86E-05	11,85	3,98	0,16
Methocell A4C_NM_Gehalt_DL	1,93E-13	4,39E-07	-3,85	0,14	0,08
BaO_NM_Gehalt_DL	1,33E-03	3,64E-02	-2,39	0,22	0,00
CaO_NM_Gehalt_DL	2,48E-04	1,57E-02	-3,25	0,68	0,00
MnO_NM_Gehalt_DL	7,59E-08	2,76E-04	-3,58	-0,85	-0,18
Na2O_NM_Gehalt_DL	1,73E-07	4,15E-04	-4,38	-0,57	-0,10
PbO_NM_Gehalt_DL	5,18E-04	2,28E-02	-1,99	-0,42	-0,20
SiO2_NM_Gehalt_DL	1,35E-05	3,67E-03	-3,47	-0,48	-0,10
SrO_NM_Gehalt_DL	5,72E-05	7,56E-03	3,77	2,39	0,44
TiO2_NM_Gehalt_DL	6,94E-04	2,63E-02	-1,83	0,48	0,07

Fortsetzung von Tabelle 36: Statistische Maße für Šumperk (mit Ausreißern)[360]

[359] Quelle: eigene Darstellung

Bezeichnung	$\tilde{s}^2$	$\tilde{s}$	γ	g_m	Korrelations-koeffizient mit MW
Y2O3_NM_Gehalt_DL	6,07E-07	7,79E-04	-1,81	0,71	-0,03
Pb_Me_DL	7,75E-10	2,78E-05	1,10	1,31	-0,14
BaCO3_DL	4,88E-09	6,98E-05	-2,86	-0,23	-0,13
SrCO3_DL	2,92E-09	5,41E-05	-0,07	1,48	0,13
CaCO3_DL	2,09E-10	1,44E-05	-2,89	-0,21	0,04
Pb3O4_DL	7,75E-10	2,78E-05	1,10	1,31	-0,14
Y-Acet_DL	8,73E-07	9,35E-04	-4,44	0,01	0,16
TiO2 _DL	4,73E-09	6,88E-05	-0,77	-0,03	0,07
A40_Me_DL	9,18E-14	3,03E-07	-4,62	-0,59	-0,26
Optapi_DL	1,36E-14	1,17E-07	-4,62	0,59	0,26
Kugelmenge_VM_DL	4,30E-07	6,56E-04	52,02	-7,55	-0,58
Siebrueckstand feucht_VM_DL	2,46E+03	4,96E+01	-3,87	0,22	0,05
d10 (VM)_VM_DL	5,66E-05	7,52E-03	-3,19	0,12	-0,08
d50 (VM)_VM_DL	1,80E-03	4,24E-02	-3,74	0,06	0,00
d90 (VM)_VM_DL	1,99E-02	1,41E-01	-3,66	0,02	0,11
pH Wert Schlicker_VM_DL	1,16E-02	1,08E-01	-3,94	-0,10	-0,07
Schuettdichte Granulat_VM_DL	1,01E-04	1,00E-02	-3,38	0,23	-0,17
Granulatfeuchte_VM_DL	5,51E-05	7,42E-03	1,46	1,77	0,04
Differenzdruck Filter_VM_DL	3,95E-03	6,28E-02	-3,61	-0,36	-0,04
Ablufttemperatur_VM_DL	3,29E+01	5,74E+00	-3,77	-0,13	-0,02
Gewoelbeunterdruck_VM_DL	6,91E-04	2,63E-02	-3,65	0,20	-0,18
Heizgruppe 7_VM_DL	2,98E+01	5,46E+00	2,68	-1,00	-0,06
Heizgruppe 9_VM_DL	1,77E+01	4,21E+00	3,34	-2,11	0,02
Heizgruppe 11_VM_DL	1,48E+01	3,84E+00	4,27	-2,33	0,01
Heizgruppe 13_VM_DL	2,35E+01	4,84E+00	3,73	-0,10	-0,07
Differenzdruck MAX_VM_DL	3,81E-04	1,95E-02	5,99	-1,36	-0,04
Differenzdruck MIN_VM_DL	3,22E-04	1,79E-02	6,00	-2,02	-0,03
Differenzdruck MW_VM_DL	3,04E-04	1,74E-02	8,63	-2,13	-0,04
Differenzdruck Stabw_VM_DL	3,28E-06	1,81E-03	1,53	1,92	0,06
Durchfluss MAX_VM_DL	3,25E-02	1,80E-01	4,30	2,49	-0,22
Durchfluss MIN_VM_DL	5,89E-01	7,68E-01	45,58	-6,91	0,11
Durchfluss MW_VM_DL	2,13E-02	1,46E-01	31,66	-5,34	0,13
Durchfluss Stabw_VM_DL	6,34E-02	2,52E-01	49,56	7,30	-0,07
Leistung MAX_VM_DL	4,79E+05	6,92E+02	-3,72	0,62	0,17
Leistung MIN_VM_DL	3,63E+05	6,03E+02	-3,26	0,86	0,13
Leistung MW_VM_DL	4,07E+05	6,38E+02	-3,57	0,72	0,14
Leistung Stabw_VM_DL	2,65E+03	5,15E+01	5,98	2,49	0,31
Schlickerdichte MAX_VM_DL	1,49E-05	3,86E-03	2,99	0,46	-0,04
Schlickerdichte MIN_VM_DL	6,45E-03	8,03E-02	50,77	-7,43	0,06
Schlickerdichte MW_VM_DL	1,33E-04	1,15E-02	42,26	-6,57	0,03
Schlickerdichte Stabw_VM_DL	4,33E-04	2,08E-02	50,81	7,43	-0,05
Schlickerdruck MAX_VM_DL	6,65E-03	8,16E-02	-3,67	0,10	-0,07
Schlickerdruck MIN_VM_DL	3,50E-03	5,91E-02	-2,78	0,40	-0,01

[360] Quelle: eigene Darstellung

Fortsetzung von Tabelle 36: Statistische Maße für Šumperk (mit Ausreißern)[361]

Bezeichnung	$\tilde{s}^2$	$\tilde{s}$	γ	g_m	Korrelations-koeffizient mit MW
Schlickerdruck MW_VM_DL	4,35E-03	6,59E-02	-3,14	0,17	-0,04
Schlickerdruck Stabw_VM_DL	1,45E-04	1,21E-02	10,42	3,07	-0,10
Schlickertemp, MAX_VM_DL	1,86E+00	1,36E+00	9,13	2,96	0,15
Schlickertemp, MIN_VM_DL	9,84E-01	9,92E-01	-1,73	1,37	0,19
Schlickertemp, MW_VM_DL	9,52E-01	9,76E-01	-2,02	1,25	0,19
Schlickertemp, Stabw_VM_DL	3,44E-02	1,85E-01	29,85	5,29	0,05
Anzahl Segmente_VM_DL	1,41E-01	3,76E-01	-1,19	1,51	0,29
Segmentzeit MAX_VM_DL	1,22E+06	1,10E+03	2,11	-1,87	-0,26
Segmentzeit MIN_VM_DL	1,03E+07	3,21E+03	-3,77	-0,68	-0,21
MW Dauer je Segment_VM_DL	4,10E+06	2,02E+03	-3,64	-0,63	-0,27
Stabw Segmentzeit_VM_DL	4,54E+14	2,13E+07	-0,82	-1,35	-0,28
Gesamtzeit Segmente_VM_DL	1,01E+06	1,01E+03	5,57	-2,28	-0,24
Ringtemperatur_VM_DL	5,41E+00	2,33E+00	-3,78	-0,10	0,24

[361] Quelle: eigene Darstellung

Tabelle 37: MSA für einzelne Attribute (Šumperk mit Ausreißern)[362]

Attribute	Durchgang			
	1	2	3	4
Kugelmenge_DL	nan	0,61	0,72	0,78
~~Durchfluss_DL~~	nan			
Schlickertemperatur_DL	nan	0,64	0,70	0,71
d10 (NM)_DL	nan	0,58	0,53	0,62
~~d50 (NM)_DL~~	nan			
~~d90 (NM)_DL~~	nan	0,32		
~~Auslaufzeit_DL~~	nan	0,46		
~~Schuettdichte Granulat_DL~~	nan			
~~Granulatfeuchte_DL~~	nan			
~~Differenzdruck Filter_DL~~	nan	0,24		
d10 Granulat (NM)_DL	nan	0,51	0,75	0,73
~~d50 Granulat (NM)_DL~~	nan			
~~d90 Granulat (NM)_DL~~	nan			
~~Schlickerdichte_DL~~	nan			
~~pH Wert Schlicker_DL~~	nan			
~~Gluehverlust_DL~~	nan	0,48		
~~Sinterdichte_DL~~	nan			
~~R25 Widerstand_DL~~	nan			
~~R25 Vergl,%_DL~~	nan			
~~R25_DL~~	nan			
~~Rmin_DL~~	nan			
~~T-Rmin_DL~~	nan			
~~Tb_DL~~	nan			
~~Tcorr_DL~~	nan			
~~Gleitwert_DL~~	nan			
~~Gruenteilfestigkeit_DL~~	nan			
~~Kohaesion_DL~~	nan			
Pressdruck_DL	nan	0,60	0,81	0,88
~~Radiale Auffederung_DL~~	nan			
~~Schuettwinkel_DL~~	nan	0,44		
~~SPHT3 MW_DL~~	nan	0,48		
~~Symm3 MW_DL~~	nan	0,20		
~~b_l3 MW_DL~~	nan	0,16		
~~Contraspum K1012_Menge1_NM_Gehalt_DL~~	nan	0,27		
~~Methocell A4C_NM_Gehalt_DL~~	nan	0,15		
~~BaO_NM_Gehalt_DL~~	nan	0,26		
~~CaO_NM_Gehalt_DL~~	nan			
~~MnO_NM_Gehalt_DL~~	nan			
~~Na2O_NM_Gehalt_DL~~	nan			
~~PbO_NM_Gehalt_DL~~	nan			
~~SiO2_NM_Gehalt_DL~~	nan			
~~SrO_NM_Gehalt_DL~~	nan			
~~TiO2_NM_Gehalt_DL~~	nan			
~~Y2O3_NM_Gehalt_DL~~	nan			
~~Pb_Me_DL~~	nan			

[362] Quelle: eigene Darstellung

Fortsetzung von Tabelle 37: MSA für einzelne Attribute (Šumperk mit Ausreißern)[363]

Attribute	Durchgang			
	1	2	3	4
~~BaCO3_DL~~	nan	0,23		
~~SrCO3_DL~~	nan			
~~CaCO3_DL~~	nan			
~~Pb3O4_DL~~	nan	0,17		
Y-Acet_DL	nan	0,70	0,82	0,81
~~TiO2_DL~~	nan			
~~A40_Me_DL~~	nan			
Optapi_DL	nan	0,58	0,76	0,76
~~Kugelmenge_VM_DL~~	nan	0,23		
~~Siebrueckstand feucht_VM_DL~~	nan			
~~d10 (VM)_VM_DL~~	nan			
~~d50 (VM)_VM_DL~~	nan	0,32		
~~d90 (VM)_VM_DL~~	nan	0,37		
~~pH Wert Schlicker_VM_DL~~	nan			
~~Schuettdichte Granulat_VM_DL~~	nan	0,40		
~~Granulatfeuchte_VM_DL~~	nan			
~~Differenzdruck Filter_VM_DL~~	nan			
~~Ablufttemperatur_VM_DL~~	nan	0,15		
~~Gewoelbeunterdruck_VM_DL~~	nan	0,24		
~~Heizgruppe 7_VM_DL~~	nan			
~~Heizgruppe 9_VM_DL~~	nan			
~~Heizgruppe 11_VM_DL~~	nan			
~~Heizgruppe 13_VM_DL~~	nan			
~~Differenzdruck MAX_VM_DL~~	nan	0,32		
~~Differenzdruck MIN_VM_DL~~	nan			
~~Differenzdruck MW_VM_DL~~	nan			
~~Differenzdruck Stabw_VM_DL~~	nan	0,39		
~~Durchfluss MAX_VM_DL~~	nan			
~~Durchfluss MIN_VM_DL~~	nan	0,80	0,34	
~~Durchfluss MW_VM_DL~~	nan			
~~Durchfluss Stabw_VM_DL~~	nan			
~~Leistung MAX_VM_DL~~	nan			
~~Leistung MIN_VM_DL~~	nan			
~~Leistung MW_VM_DL~~	nan			
~~Leistung Stabw_VM_DL~~	nan	0,32		
~~Schlickerdichte MAX_VM_DL~~	nan			
~~Schlickerdichte MIN_VM_DL~~	nan	0,39		
~~Schlickerdichte MW_VM_DL~~	nan			
~~Schlickerdichte Stabw_VM_DL~~	nan			
~~Schlickerdruck MAX_VM_DL~~	nan			
~~Schlickerdruck MIN_VM_DL~~	nan	0,27		
~~Schlickerdruck MW_VM_DL~~	nan			
~~Schlickerdruck Stabw_VM_DL~~	nan	0,22		
~~Schlickertemp, MAX_VM_DL~~	nan	0,37		
~~Schlickertemp, MIN_VM_DL~~	nan			
~~Schlickertemp, MW_VM_DL~~	nan	0,31		
~~Schlickertemp, Stabw_VM_DL~~	nan	0,39		

[363] Quelle: eigene Darstellung

Fortsetzung von Tabelle 37: MSA für einzelne Attribute (Šumperk mit Ausreißern)[364]

Attribute	Durchgang			
	1	2	3	4
Anzahl Segmente_VM_DL	nan	0,60	0,62	0,62
~~Segmentzeit MAX_VM_DL~~	nan			
Segmentzeit MIN_VM_DL	nan	0,57	0,61	0,63
~~MW Dauer je Segment_VM_DL~~	nan			
~~Stabw Segmentzeit_VM_DL~~	nan			
~~Gesamtzeit Segmente_VM_DL~~	nan			
~~Ringtemperatur_VM_DL~~	nan			

Tabelle 38: Ergebnis CL und Testdaten: Šumperk mit Ausreißern[365]

	Genauigkeit bei Training	Genauigkeit bei Anwendung auf Testdaten	Grenzfälle bei Anwendung auf Testdaten	Mittelwert der absoluten Differenz zwischen der vorhergesagten und Ist-Klasse
Complex Tree	25,0%	24,0%	0,0%	1,94
Medium Tree	25,0%	24,0%	0,0%	1,94
Simple Tree	23,3%	12,0%	4,0%	2,38
Linear SVM	15,0%	24,0%	4,0%	2,30
Quadratic SVM	21,7%	20,0%	4,0%	2,34
Cubic SVM	21,7%	12,0%	4,0%	2,62
Fine Gaussian SVM	18,3%	8,0%	0,0%	3,10
Medium Gaussian SVM	20,0%	0,0%	0,0%	3,30
Coarse Gaussian SVM	15,0%	8,0%	0,0%	3,42
Boosted Trees	23,3%	20,0%	4,0%	2,18
Bagged Trees	15,0%	8,0%	4,0%	2,98
RUSBoosted Trees	19,3%	8,0%	0,0%	2,06

[364] Quelle: eigene Darstellung
[365] Quelle: ebda.

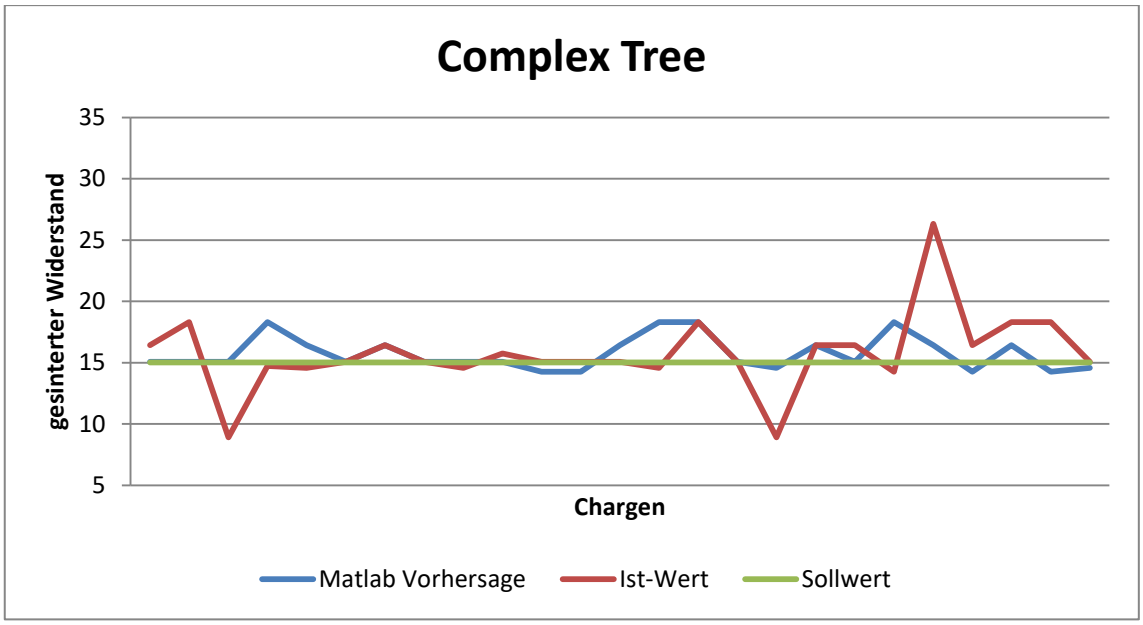


Abbildung 86: CT: Vergleich Vorhersage zu Ist-Wert (Šumperk mit Ausreißern)[366]

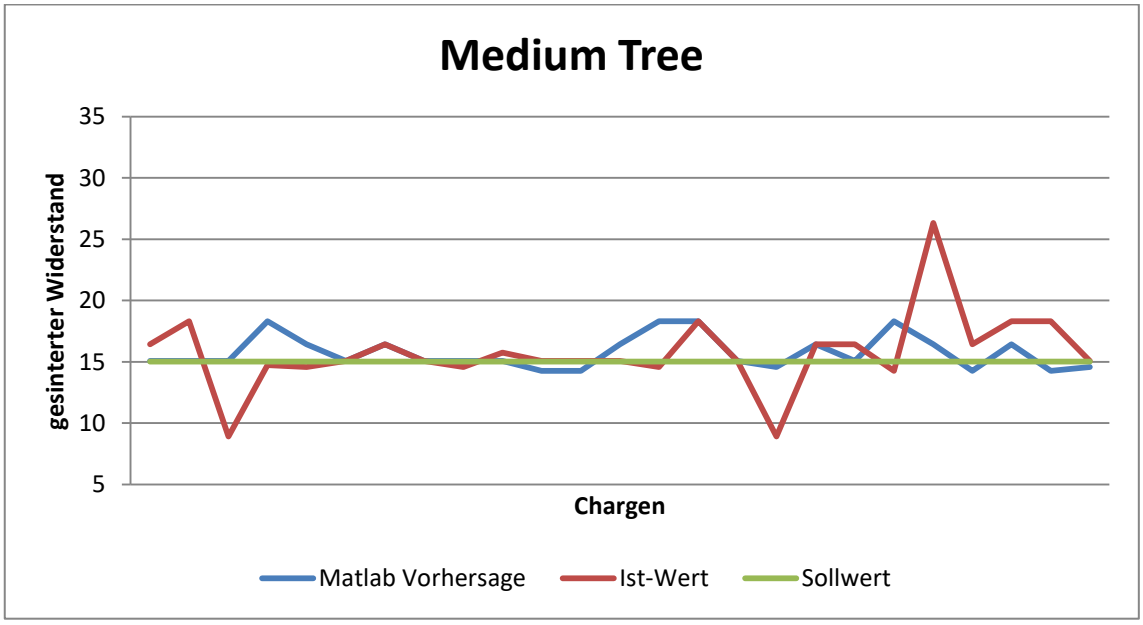


Abbildung 87: MT: Vergleich Vorhersage zu Ist-Wert (Šumperk mit Ausreißern)[367]

[366] Quelle: eigene Darstellung
[367] Quelle: ebda.

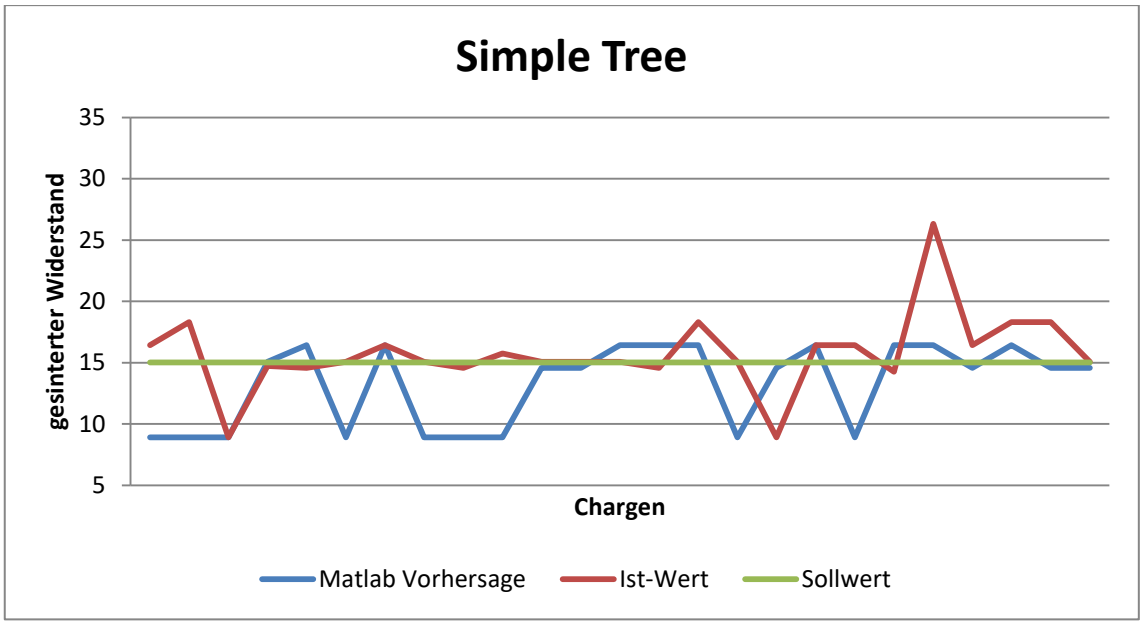


Abbildung 88: ST: Vergleich Vorhersage zu Ist-Wert (Šumperk mit Ausreißern)[368]

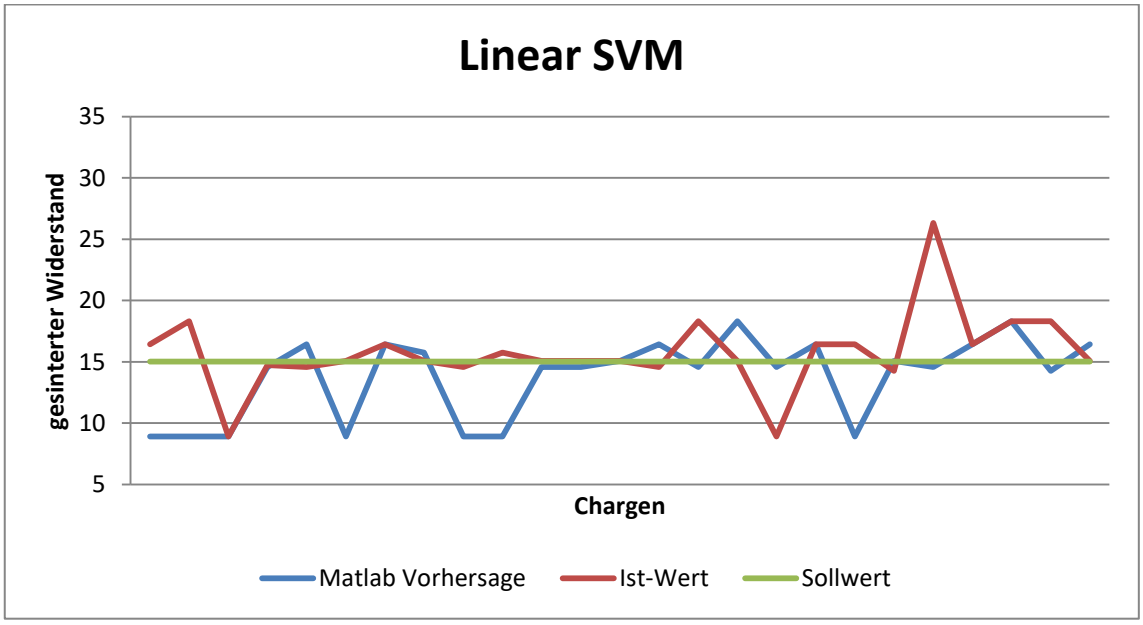


Abbildung 89: LSVM: Vergleich Vorhersage zu Ist-Wert (Šumperk mit Ausreißern)[369]

[368] Quelle: eigene Darstellung
[369] Quelle: ebda.

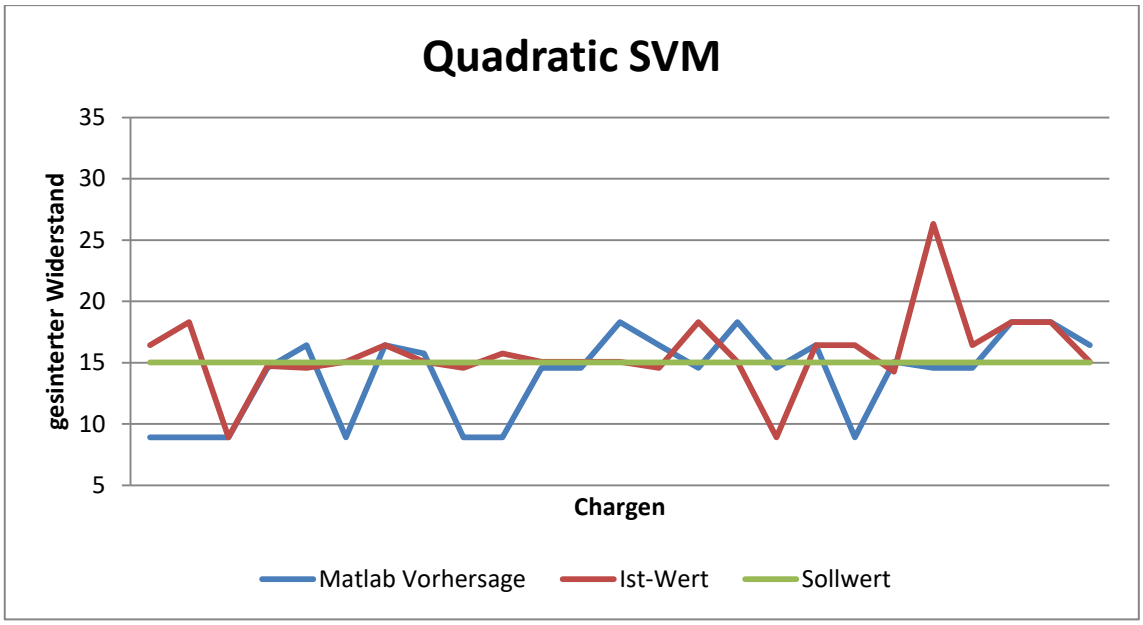


Abbildung 90: QSVM: Vergleich Vorhersage zu Ist-Wert (Šumperk mit Ausreißern)[370]

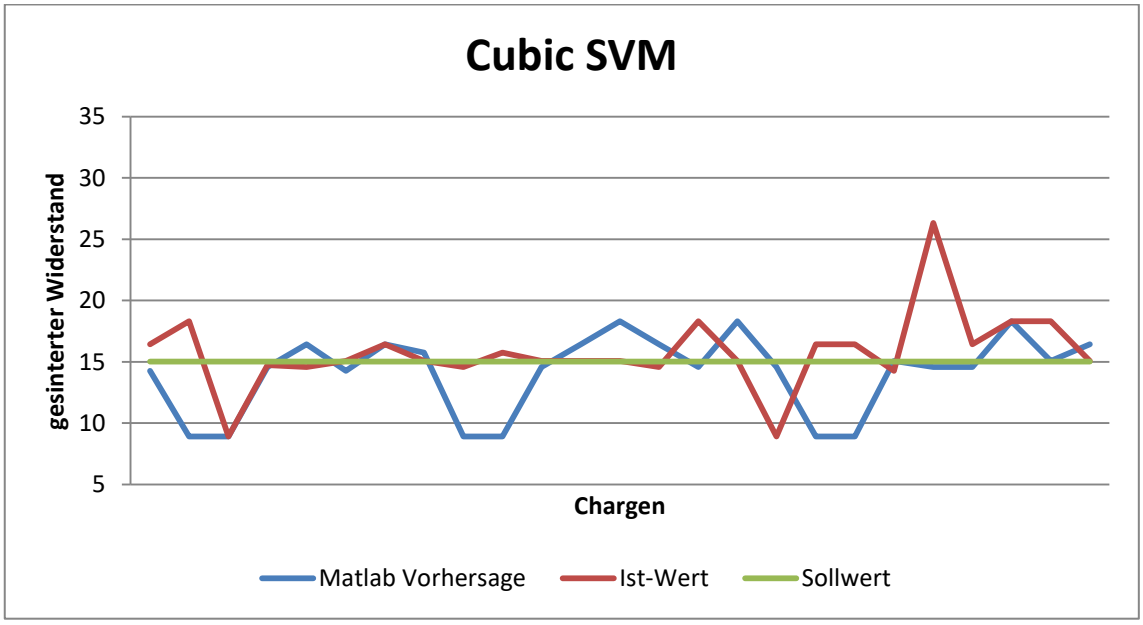


Abbildung 91: CSVM: Vergleich Vorhersage zu Ist-Wert (Šumperk mit Ausreißern)[371]

[370] Quelle: eigene Darstellung
[371] Quelle: ebda.

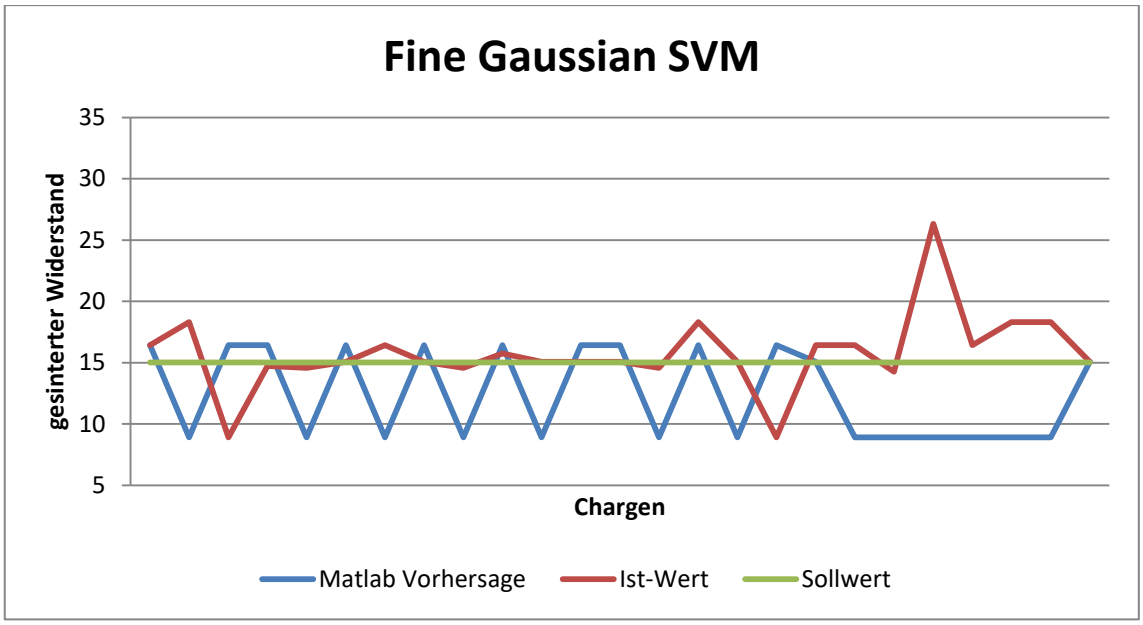


Abbildung 92: FGSVM: Vergleich Vorhersage zu Ist-Wert (Šumperk mit Ausreißern)[372]

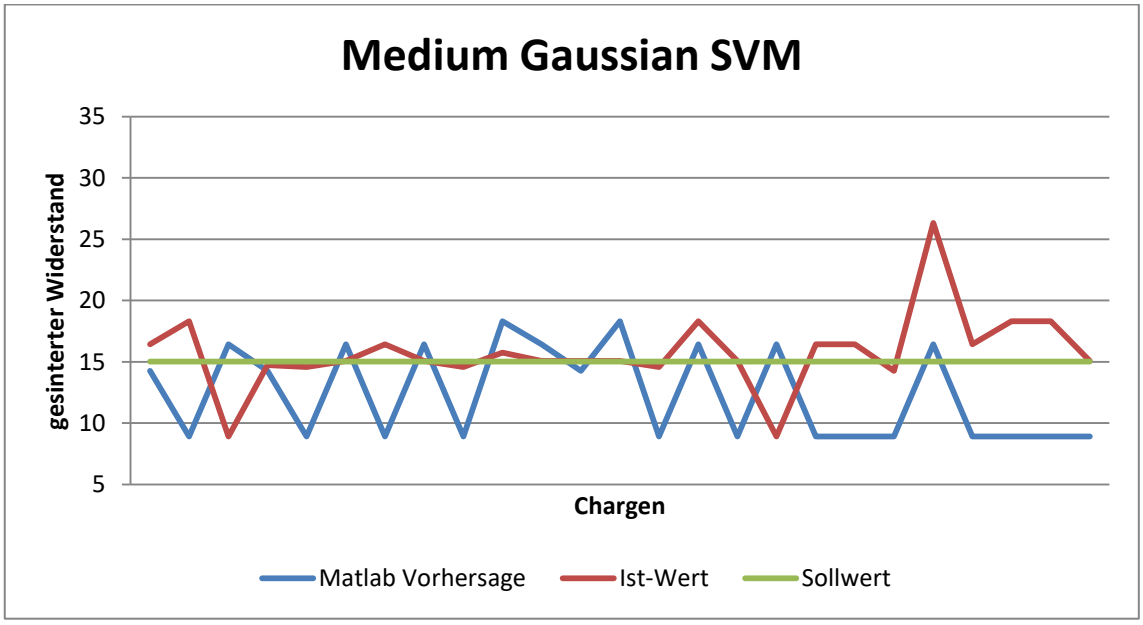


Abbildung 93: MGSVM: Vergleich Vorhersage zu Ist-Wert (Šumperk mit Ausreißern)[373]

[372] Quelle: eigene Darstellung
[373] Quelle: ebda.

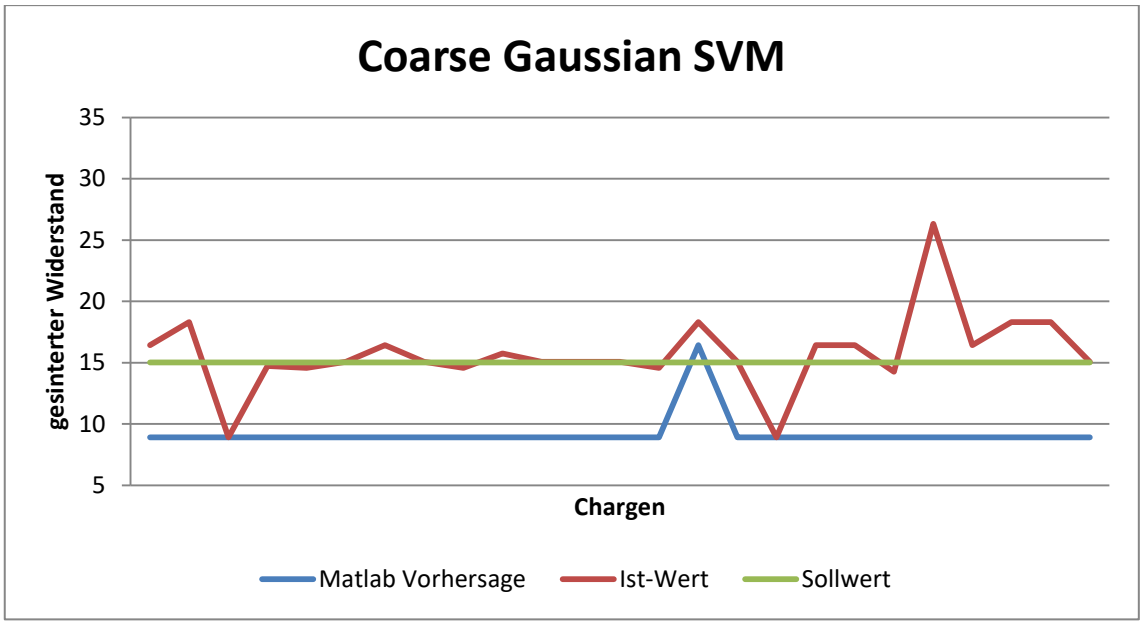


Abbildung 94: CGSVM: Vergleich Vorhersage zu Ist-Wert (Šumperk mit Ausreißern)[374]

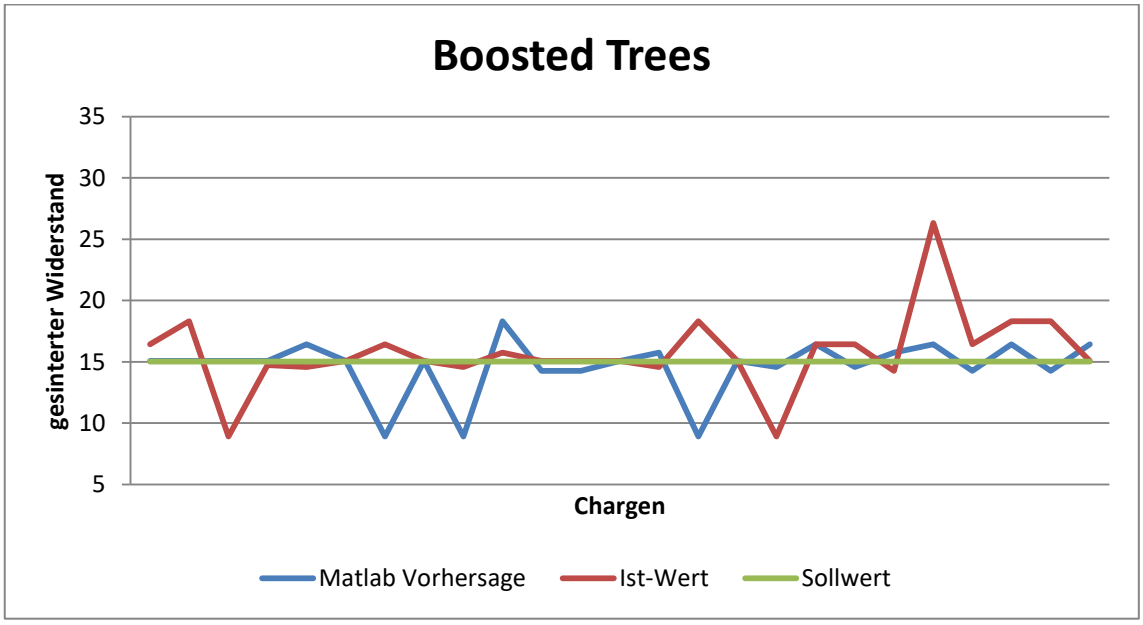


Abbildung 95: BoT: Vergleich Vorhersage zu Ist-Wert (Šumperk mit Ausreißern)[375]

[374] Quelle: eigene Darstellung
[375] Quelle: ebda.

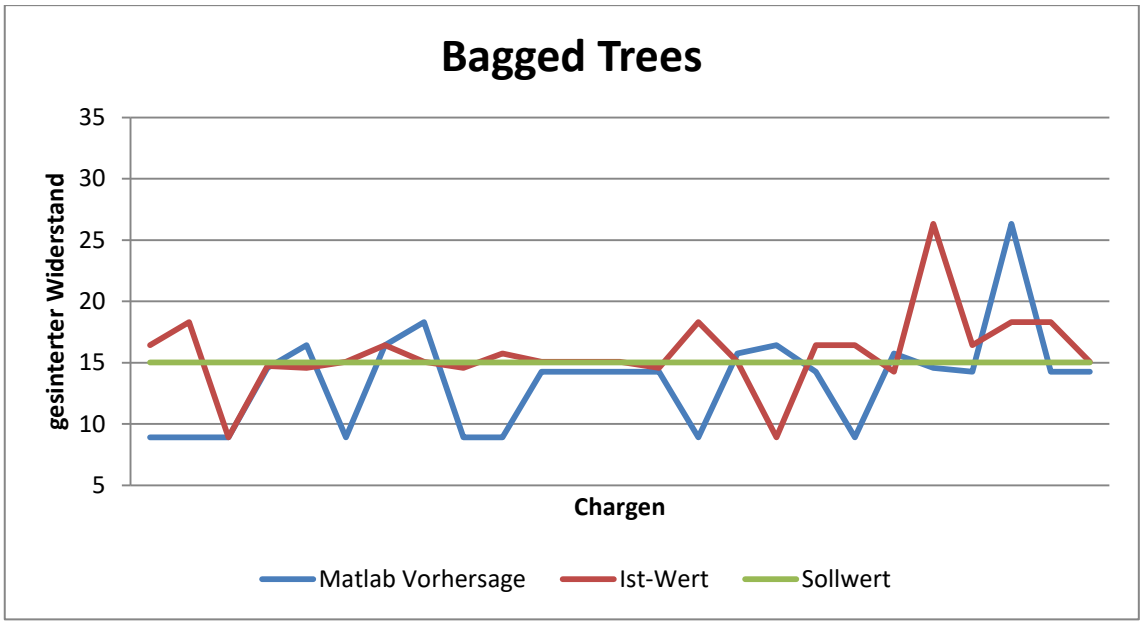


Abbildung 96: BaT: Vergleich Vorhersage zu Ist-Wert (Šumperk mit Ausreißern)[376]

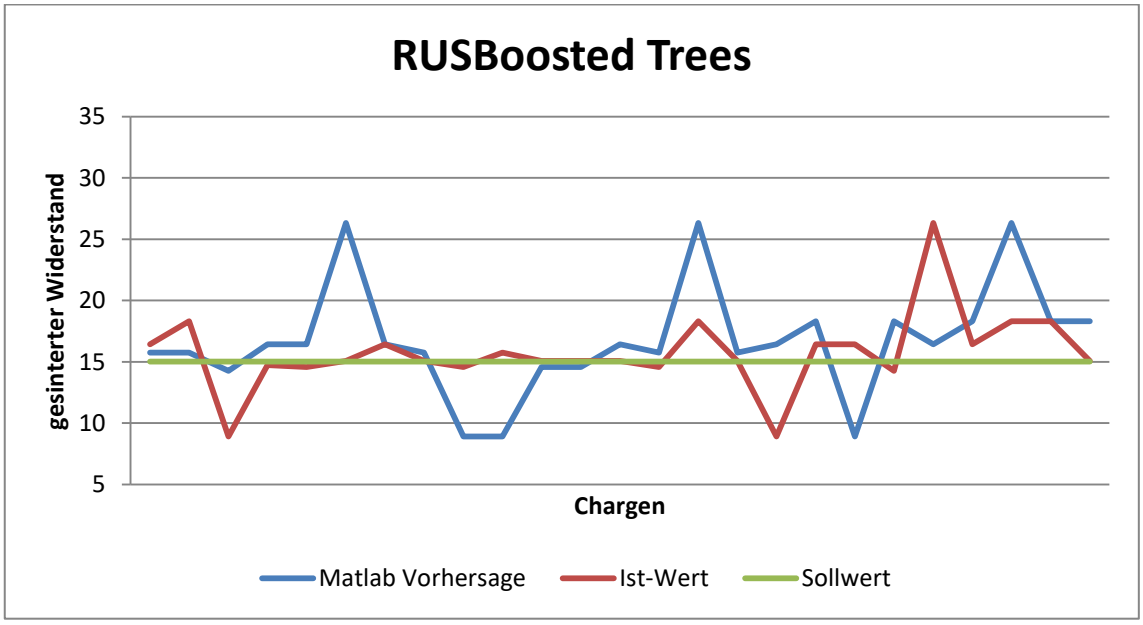


Abbildung 97: RUSBT: Vergleich Vorhersage zu Ist-Wert (Šumperk mit Ausreißern)[377]

[376] Quelle: eigene Darstellung
[377] Quelle: ebda.

Medium Tree Šumperk mit Ausreißern

```
1  if SchlickerdichteMIN_VM_DL<0.191284 then node 2 elseif
SchlickerdichteMIN_VM_DL>=0.191284 then node 3 else 11.664
 2  if SchlickertempMAX_VM_DL<-0.262925 then node 4 elseif
SchlickertempMAX_VM_DL>=-0.262925 then node 5 else 16.5
 3  if pHWertSchlicker_VM_DL<-0.550633 then node 6 elseif
pHWertSchlicker_VM_DL>=-0.550633 then node 7 else 11.664
 4  if d90VM_VM_DL<-0.773474 then node 8 elseif d90VM_VM_DL>=-
0.773474 then node 9 else 15.089
 5  if Symm3MW_DL<-0.307843 then node 10 elseif Symm3MW_DL>=-
0.307843 then node 11 else 16.5
 6  class = 15.68
 7  if RadialeAuffederung_DL<-0.793705 then node 12 elseif
RadialeAuffederung_DL>=-0.793705 then node 13 else 11.664
 8  class = 15.089
 9  class = 18.336
10  if d10GranulatNM_DL<-0.147212 then node 14 elseif
d10GranulatNM_DL>=-0.147212 then node 15 else 14.589
11  if Ringtemperatur_VM_DL<0.336911 then node 16 elseif
Ringtemperatur_VM_DL>=0.336911 then node 17 else 16.5
12  class = 25.907
13  if SchlickerdichteMIN_VM_DL<0.251001 then node 18 elseif
SchlickerdichteMIN_VM_DL>=0.251001 then node 19 else 11.664
14  class = 14.241
15  if Heizgruppe11_VM_DL<0.646222 then node 20 elseif
Heizgruppe11_VM_DL>=0.646222 then node 21 else 14.589
16  class = 16.5
17  class = 18.336
18  class = 11.664
19  class = 15.089
20  class = 14.589
21  class = 14.241
```

Complex Tree

```
1  if SchlickerdichteMIN_VM_DL<0.191284 then node 2 elseif
SchlickerdichteMIN_VM_DL>=0.191284 then node 3 else 11.664
 2  if SchlickertempMAX_VM_DL<-0.262925 then node 4 elseif
SchlickertempMAX_VM_DL>=-0.262925 then node 5 else 16.5
 3  if pHWertSchlicker_VM_DL<-0.550633 then node 6 elseif
pHWertSchlicker_VM_DL>=-0.550633 then node 7 else 11.664
 4  if d90VM_VM_DL<-0.773474 then node 8 elseif d90VM_VM_DL>=-
0.773474 then node 9 else 15.089
 5  if Symm3MW_DL<-0.307843 then node 10 elseif Symm3MW_DL>=-
0.307843 then node 11 else 16.5
 6  class = 15.68
 7  if RadialeAuffederung_DL<-0.793705 then node 12 elseif
RadialeAuffederung_DL>=-0.793705 then node 13 else 11.664
 8  class = 15.089
 9  class = 18.336
10  if d10GranulatNM_DL<-0.147212 then node 14 elseif
d10GranulatNM_DL>=-0.147212 then node 15 else 14.589
11  if Ringtemperatur_VM_DL<0.336911 then node 16 elseif
Ringtemperatur_VM_DL>=0.336911 then node 17 else 16.5
12  class = 25.907
13  if SchlickerdichteMIN_VM_DL<0.251001 then node 18 elseif
SchlickerdichteMIN_VM_DL>=0.251001 then node 19 else 11.664
14  class = 14.241
15  if Heizgruppe11_VM_DL<0.646222 then node 20 elseif
Heizgruppe11_VM_DL>=0.646222 then node 21 else 14.589
16  class = 16.5
17  class = 18.336
18  class = 11.664
19  class = 15.089
20  class = 14.589
21  class = 14.241
```

Tabelle 39: Ergebnis CL und Testdaten: Šumperk ohne Ausreißer[378]

	Genauigkeit bei Training	Genauigkeit bei Anwendung auf Testdaten	Grenzfälle bei Anwendung auf Testdaten	Mittelwert der absoluten Differenz zwischen der vorhergesagten und Ist-Klasse
Complex Tree	25,0%	12,0%	0,0%	2,58
Medium Tree	25,0%	12,0%	0,0%	2,58
Simple Tree	23,3%	20,0%	4,0%	2,02
Linear SVM	15,0%	8,0%	0,0%	2,74
Quadratic SVM	21,7%	24,0%	0,0%	2,14
Cubic SVM	21,7%	24,0%	0,0%	1,82
Fine Gaussian SVM	18,3%	16,0%	4,0%	2,82
Medium Gaussian SVM	20,0%	4,0%	0,0%	3,02
Coarse Gaussian SVM	15,0%	8,0%	0,0%	3,62
Boosted Trees	23,3%	12,0%	0,0%	2,34
Bagged Trees	15,0%	16,0%	0,0%	2,58
RUSBoosted Trees	19,3%	4,0%	4,0%	2,50

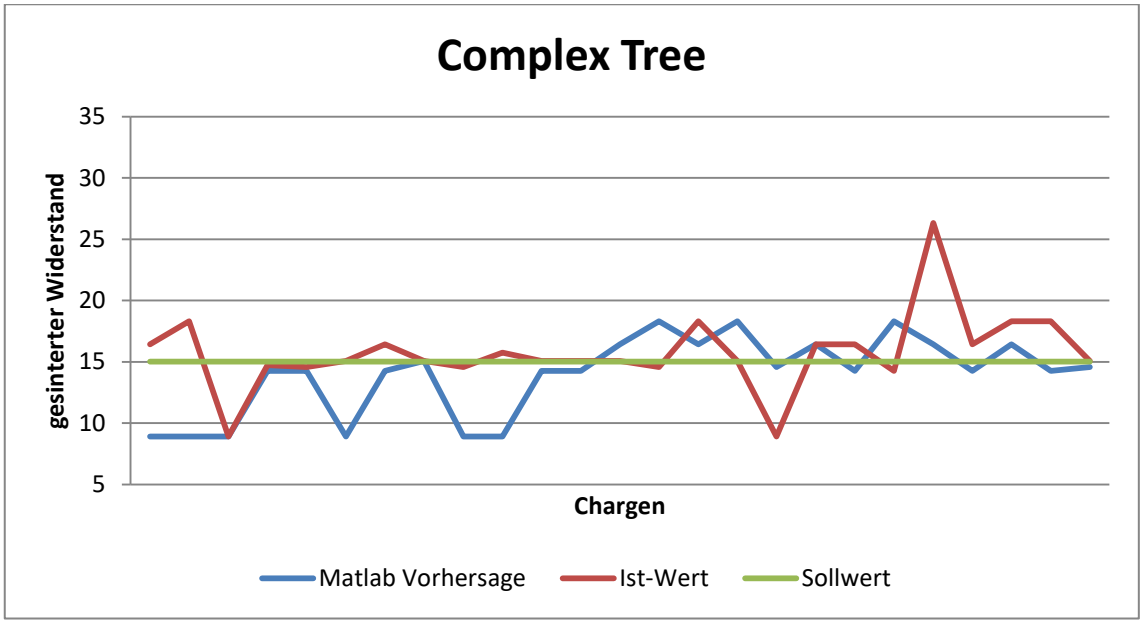


Abbildung 98: CT: Vergleich Vorhersage zu Ist-Wert (Šumperk ohne Ausreißer)[379]

[378] Quelle: eigene Darstellung
[379] Quelle: ebda.

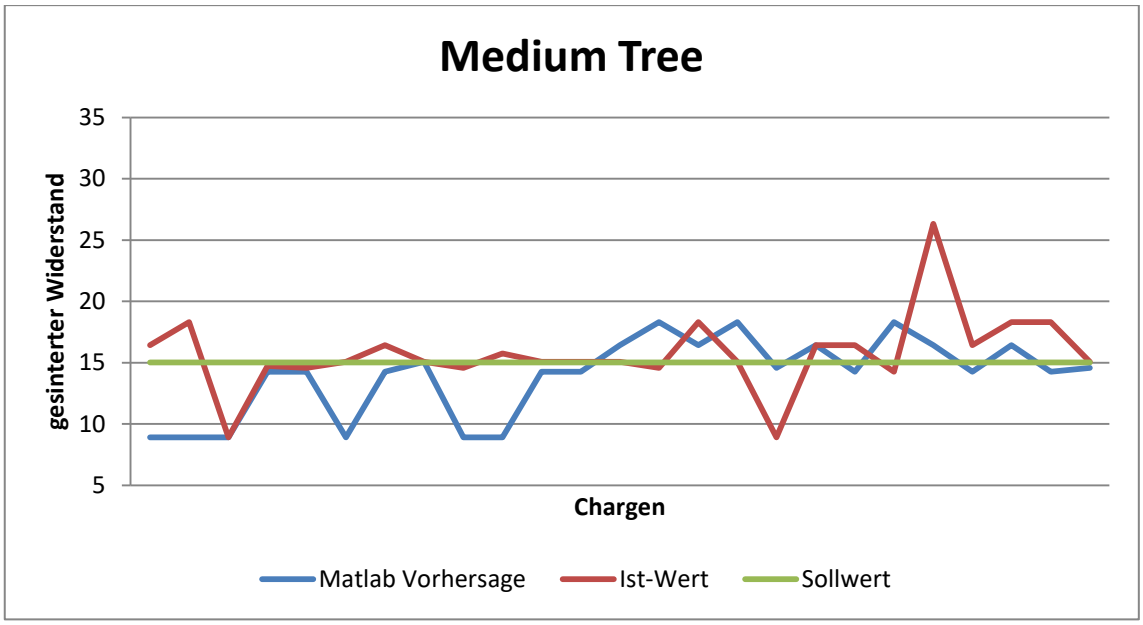


Abbildung 99: MT: Vergleich Vorhersage zu Ist-Wert (Šumperk ohne Ausreißer)[380]

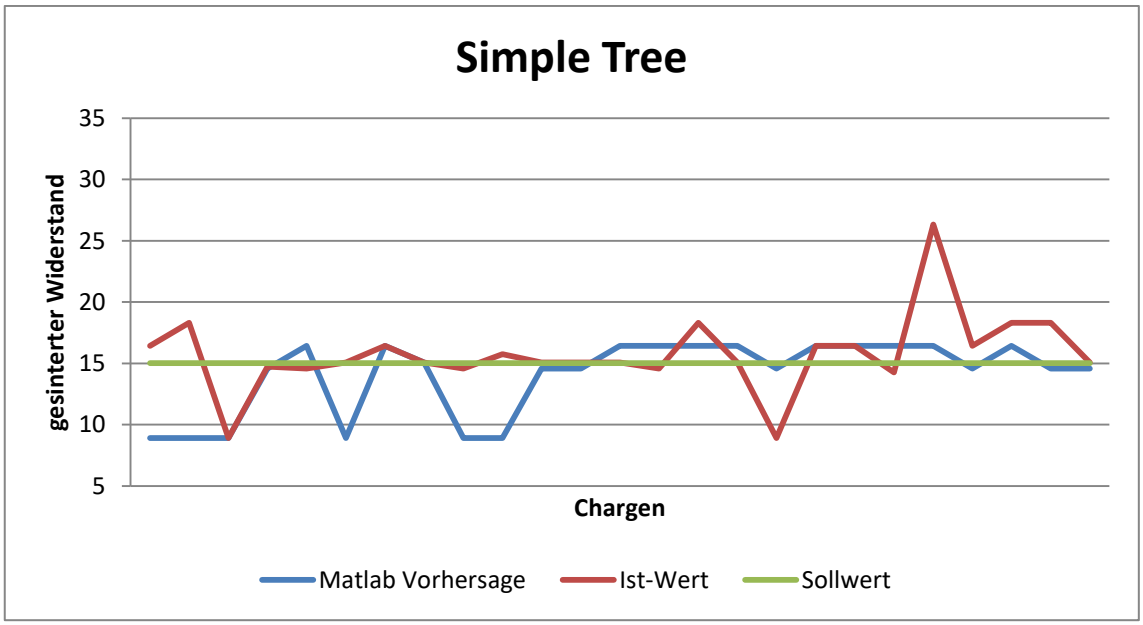


Abbildung 100: ST: Vergleich Vorhersage zu Ist-Wert (Šumperk ohne Ausreißer)[381]

[380] Quelle: eigene Darstellung
[381] Quelle: ebda.

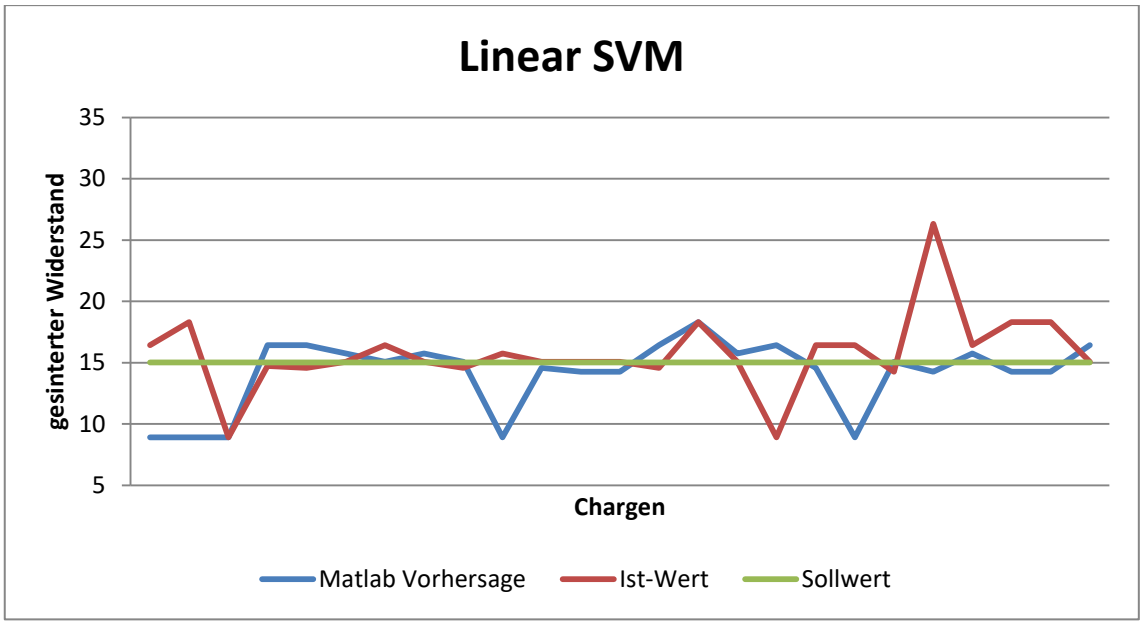


Abbildung 101: LSVM: Vergleich Vorhersage zu Ist-Wert (Šumperk ohne Ausreißer)[382]

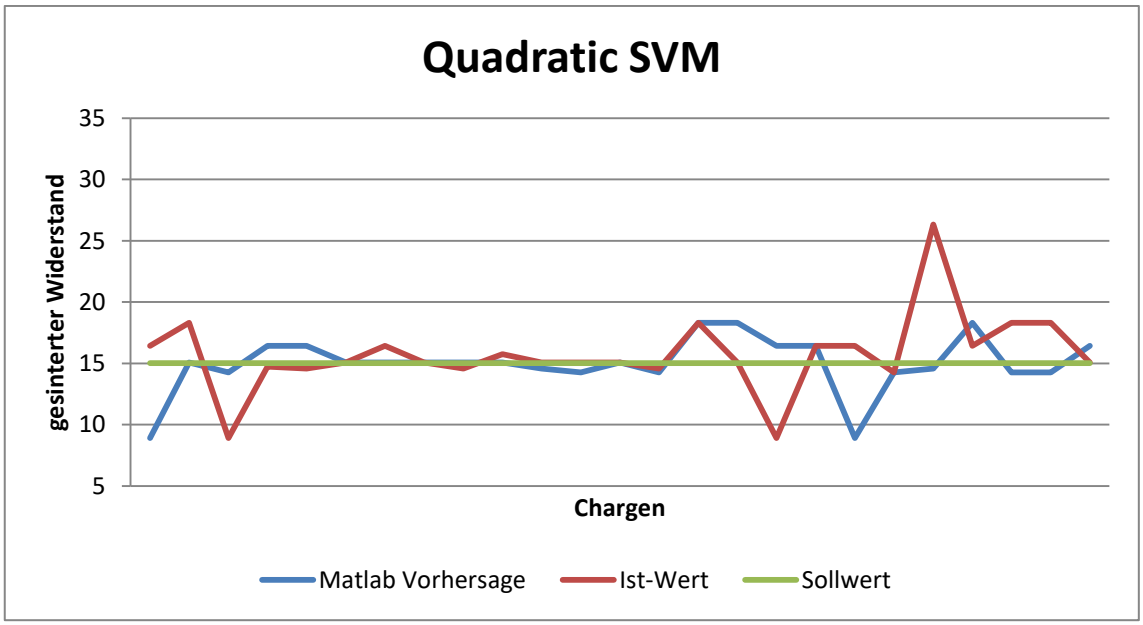


Abbildung 102: QSVM: Vergleich Vorhersage zu Ist-Wert (Šumperk ohne Ausreißer)[383]

[382] Quelle: eigene Darstellung
[383] Quelle: ebda.

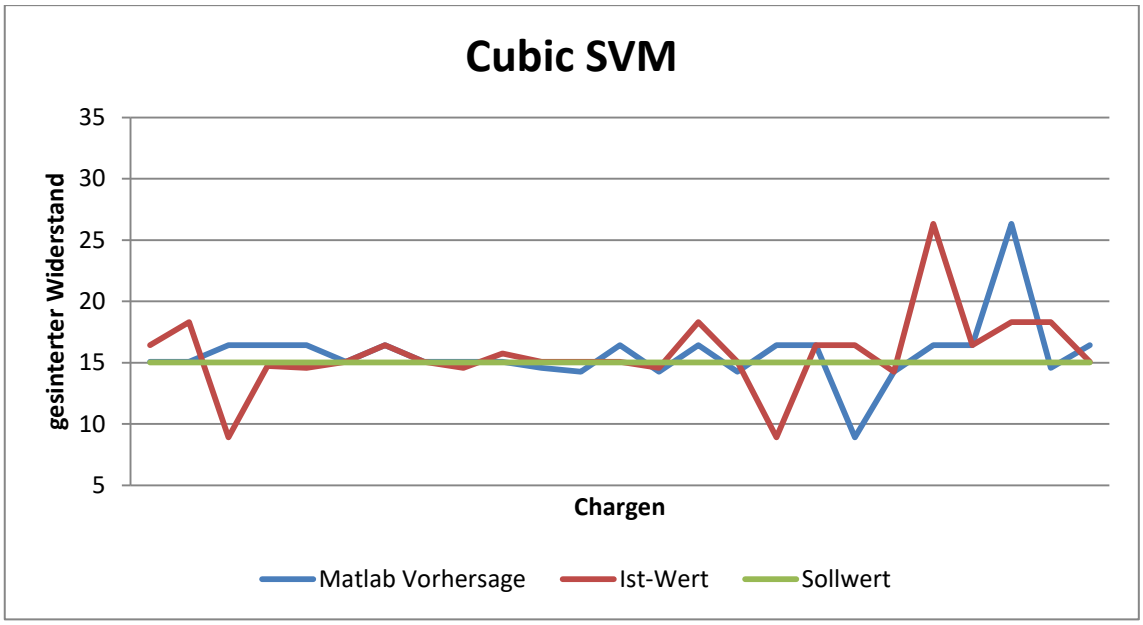


Abbildung 103: CSVM: Vergleich Vorhersage zu Ist-Wert (Šumperk ohne Ausreißer)[384]

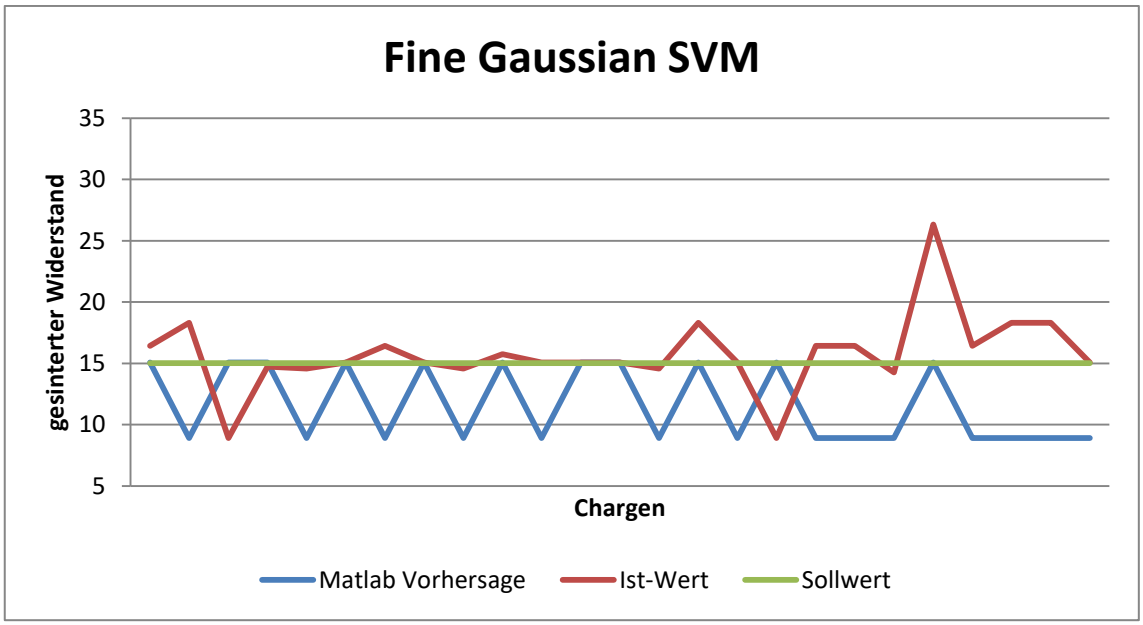


Abbildung 104: FGSVM: Vergleich Vorhersage zu Ist-Wert (Šumperk ohne Ausreißer)[385]

[384] Quelle: eigene Darstellung
[385] Quelle: ebda.

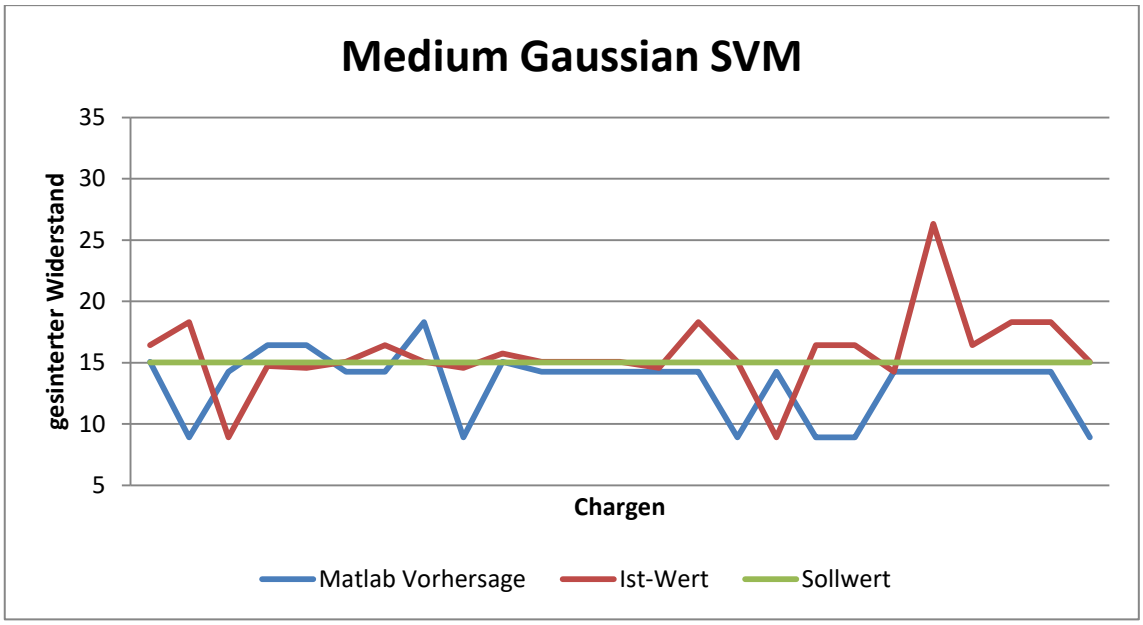


Abbildung 105: MGSVM: Vergleich Vorhersage zu Ist-Wert (Šumperk ohne Ausreißer)[386]

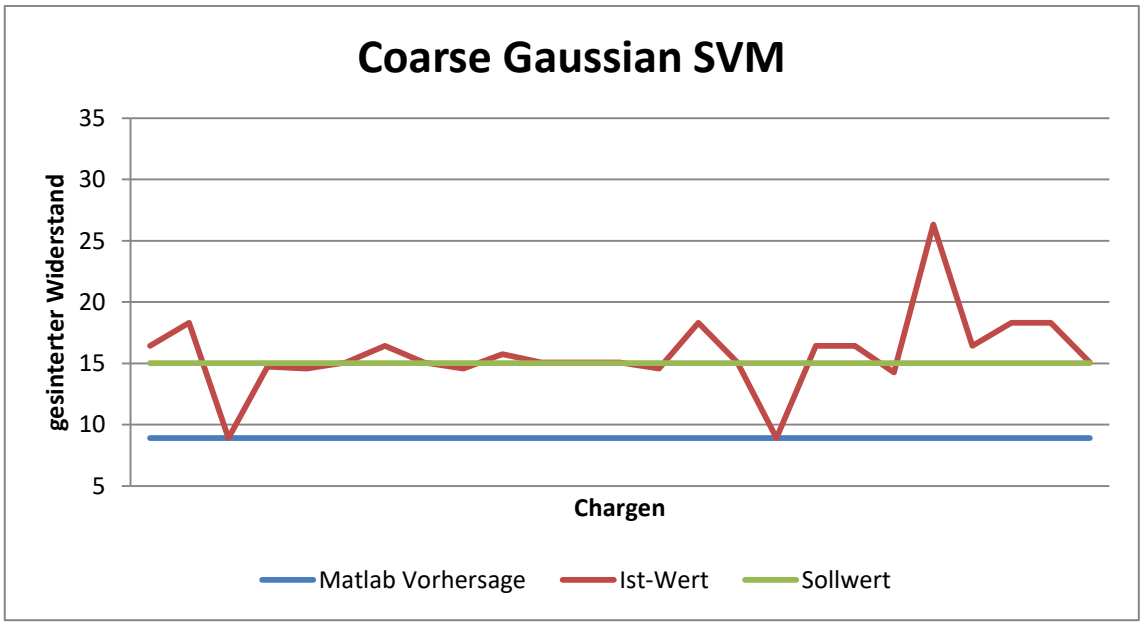


Abbildung 106: CGSVM: Vergleich Vorhersage zu Ist-Wert (Šumperk ohne Ausreißer)[387]

[386] Quelle: eigene Darstellung
[387] Quelle: ebda.

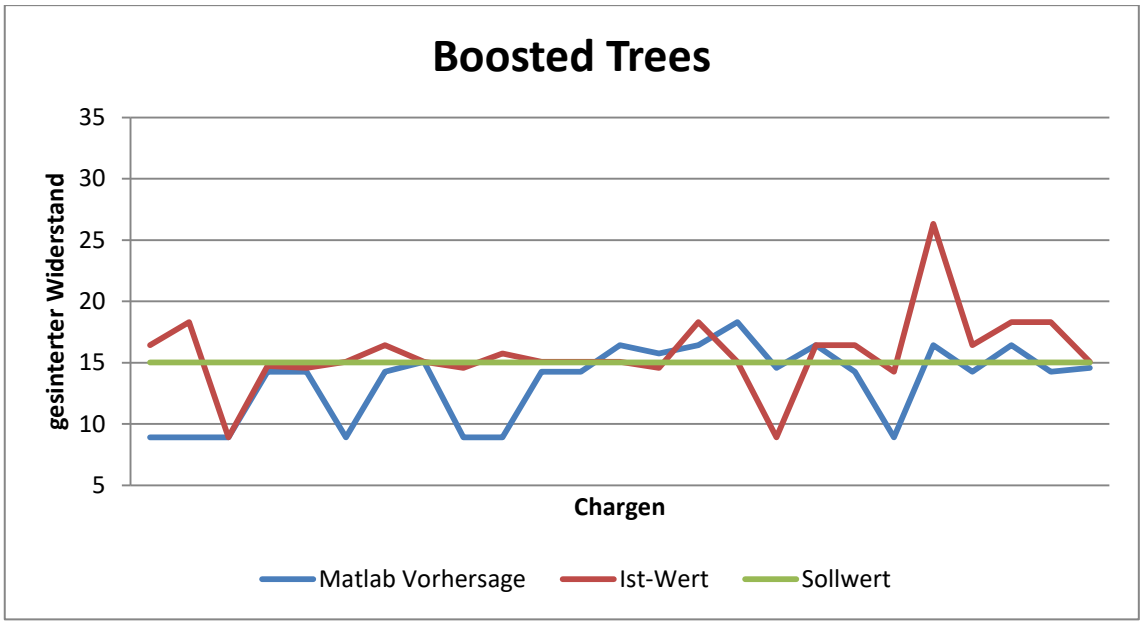


Abbildung 107: BoT: Vergleich Vorhersage zu Ist-Wert (Šumperk ohne Ausreißer)[388]

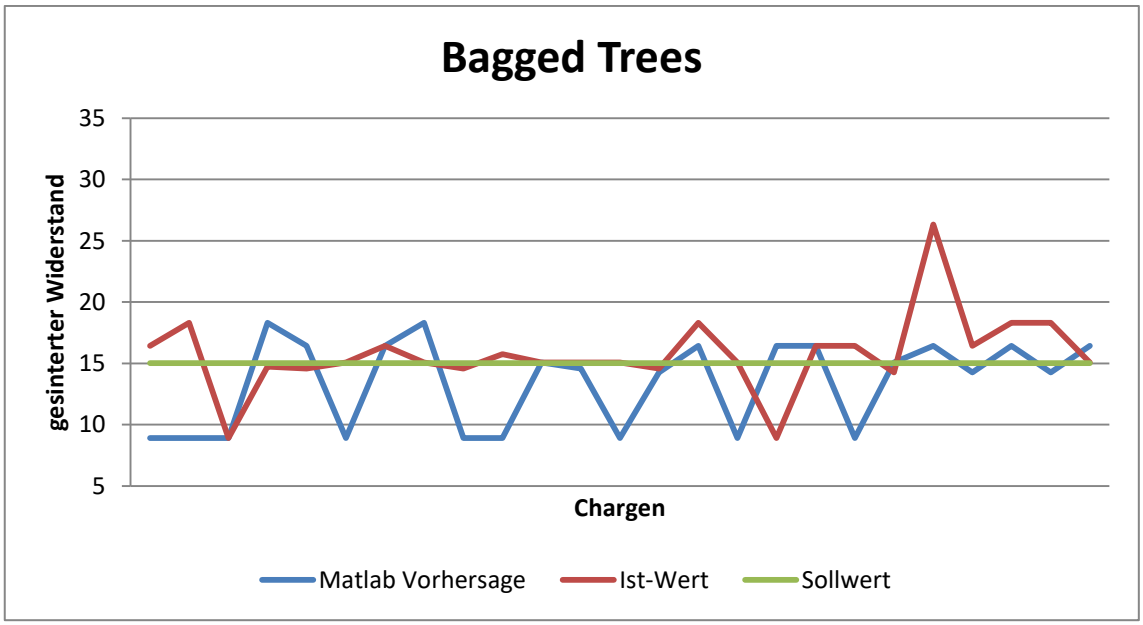


Abbildung 108: BaT: Vergleich Vorhersage zu Ist-Wert (Šumperk ohne Ausreißer)[389]

[388] Quelle: eigene Darstellung
[389] Quelle: ebda.

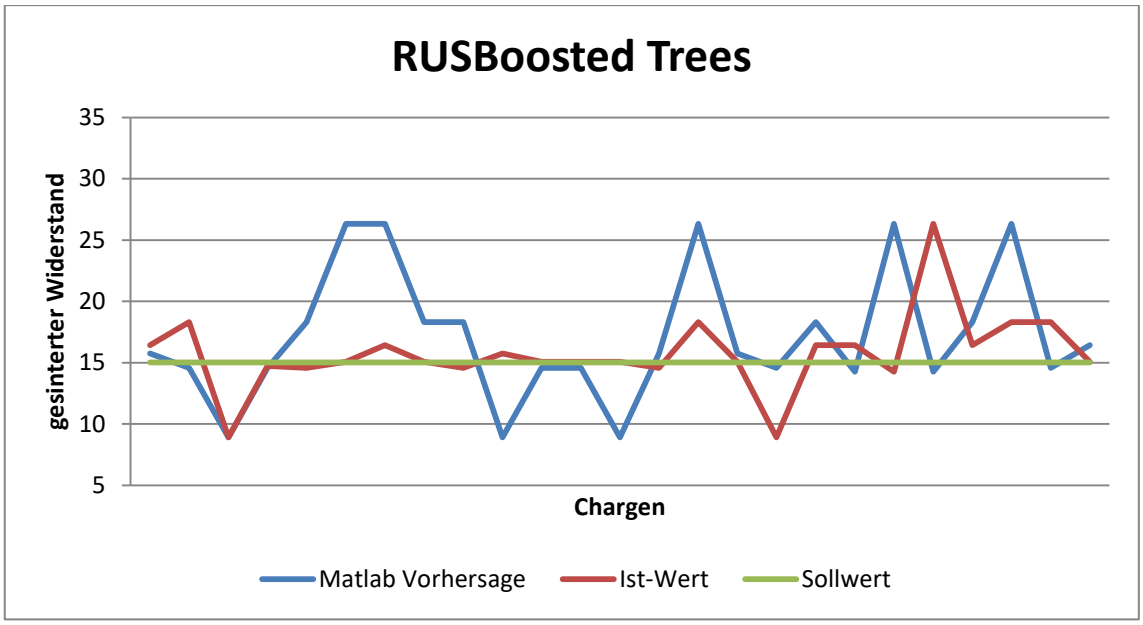


Abbildung 109: RUSBT: Vergleich Vorhersage zu Ist-Wert (Šumperk ohne Ausreißer)[390]

[390] Quelle: eigene Darstellung

Medium Tree Šumperk ohne Ausreißer

```
1  if SchlickerdichteMIN_VM_DL<0.571801 then node 2 elseif
SchlickerdichteMIN_VM_DL>=0.571801 then node 3 else 11.664
 2  if Symm3MW_DL<-0.435851 then node 4 elseif Symm3MW_DL>=-
0.435851 then node 5 else 16.5
 3  if pHWertSchlicker_VM_DL<-0.550633 then node 6 elseif
pHWertSchlicker_VM_DL>=-0.550633 then node 7 else 11.664
 4  if d10GranulatNM_DL<-0.147212 then node 8 elseif
d10GranulatNM_DL>=-0.147212 then node 9 else 14.589
 5  if SchlickertempMAX_VM_DL<0.0256806 then node 10 elseif
SchlickertempMAX_VM_DL>=0.0256806 then node 11 else 16.5
 6  class = 15.68
 7  if SchlickertempStabw_VM_DL<-1.26292 then node 12 elseif
SchlickertempStabw_VM_DL>=-1.26292 then node 13 else 11.664
 8  class = 14.241
 9  if Heizgruppe11_VM_DL<0.660181 then node 14 elseif
Heizgruppe11_VM_DL>=0.660181 then node 15 else 14.589
10  if d90VM_VM_DL<-0.773474 then node 16 elseif d90VM_VM_DL>=-
0.773474 then node 17 else 15.089
11  if Symm3MW_DL<0.44651 then node 18 elseif
Symm3MW_DL>=0.44651 then node 19 else 16.5
12  class = 15.089
13  if RadialeAuffederung_DL<-0.793705 then node 20 elseif
RadialeAuffederung_DL>=-0.793705 then node 21 else 11.664
14  class = 14.589
15  class = 14.241
16  class = 15.089
17  if Heizgruppe13_VM_DL<0.605687 then node 22 elseif
Heizgruppe13_VM_DL>=0.605687 then node 23 else 18.336
18  class = 16.5
19  class = 18.336
20  class = 25.907
21  if Ofen in {19 20 22 25 BV} then node 24 elseif Ofen=24 then
node 25 else 11.664
22  class = 14.241
23  class = 18.336
24  class = 11.664
25  class = 18.336
```

Complex Tree Šumperk ohne Ausreißer

```
1  if SchlickerdichteMIN_VM_DL<0.571801 then node 2 elseif
SchlickerdichteMIN_VM_DL>=0.571801 then node 3 else 11.664
 2  if Symm3MW_DL<-0.435851 then node 4 elseif Symm3MW_DL>=-
0.435851 then node 5 else 16.5
 3  if pHWertSchlicker_VM_DL<-0.550633 then node 6 elseif
pHWertSchlicker_VM_DL>=-0.550633 then node 7 else 11.664
 4  if d10GranulatNM_DL<-0.147212 then node 8 elseif
d10GranulatNM_DL>=-0.147212 then node 9 else 14.589
 5  if SchlickertempMAX_VM_DL<0.0256806 then node 10 elseif
SchlickertempMAX_VM_DL>=0.0256806 then node 11 else 16.5
 6  class = 15.68
 7  if SchlickertempStabw_VM_DL<-1.26292 then node 12 elseif
SchlickertempStabw_VM_DL>=-1.26292 then node 13 else 11.664
 8  class = 14.241
 9  if Heizgruppe11_VM_DL<0.660181 then node 14 elseif
Heizgruppe11_VM_DL>=0.660181 then node 15 else 14.589
10  if d90VM_VM_DL<-0.773474 then node 16 elseif d90VM_VM_DL>=-
0.773474 then node 17 else 15.089
11  if Symm3MW_DL<0.44651 then node 18 elseif
Symm3MW_DL>=0.44651 then node 19 else 16.5
12  class = 15.089
13  if RadialeAuffederung_DL<-0.793705 then node 20 elseif
RadialeAuffederung_DL>=-0.793705 then node 21 else 11.664
14  class = 14.589
15  class = 14.241
16  class = 15.089
17  if Heizgruppe13_VM_DL<0.605687 then node 22 elseif
Heizgruppe13_VM_DL>=0.605687 then node 23 else 18.336
18  class = 16.5
19  class = 18.336
20  class = 25.907
21  if Ofen in {19 20 22 25 BV} then node 24 elseif Ofen=24 then
node 25 else 11.664
22  class = 14.241
23  class = 18.336
24  class = 11.664
25  class = 18.336
```