technische universität dortmund

IT in Produktion und Logistik

Bachelorarbeit

Kategorisierung der Ausreißerbehandlung in der Produktion mittels umfassender Literaturrecherche

Zur Erlangung des akademischen Grades

Bachelor of Science (B. Sc.)

Acar Abdulsamed, 202074

Bachelor Maschinenbau

ausgegeben am:

16.05.2022

eingereicht am:

08.08.2022

Betreuer:

Dr.-Ing. Dipl.-Inform. Anne Antonia Scheidler

M. Sc. Florian Hochkamp

Technische Universität Dortmund

Fakultät Maschinenbau

Fachgebiet IT in Produktion und Logistik

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

BBA	Bagplot-Based-Adjustment
BMW	Bayerische Motoren Werke
CBLOF	Cluster-Based-Local-Outlier-Factor
IKT	Informations- und Kommunikationstechnologie
IQR	Interquartilsabstand
LMS	Least-Median-of-Squares
LOF	Local-Outlier-Factor
LTS	Least-Trimmed-Square
MDA	Mahalanobis-Distance-Approach
NNM	Nearest-Neighbor-Methode
OEE	Overall-Equipment-Effectiveness
Q_1	Unteres Quartil
Q_2	Median
Q_3	Oberes Quartil
SPS	Speicherprogrammierbare Steuerung

1 Einleitung

Jedes im Wettbewerb konkurrierende Unternehmen hat eine Strategie, die durch Planung entwickelt wird (Porter 2013). In der Automobilindustrie ist dieser Wettbewerb ebenfalls vorhanden. Der Umsatz der Automobilindustrie allein in Deutschland im Jahr 2021 betrug 411 Milliarden Euro und ist deshalb der größte Industriezweig im Land (Bundesministerium für Wirtschaft und Klimaschutz 2022). Nach Winkelhake (2021) werden durch den Konkurrenzkampf neue digitale Geschäftsmodelle frühzeitig erkannt und in das eigene Unternehmen eingebaut. Der Autor führt weiter aus, dass die Digitalisierung für viele Unternehmen eine sehr hohe Priorität hat, da sie die Aussicht auf mehr Profit und Umsatz verschafft. Es entstehen nicht immer neue Geschäftsmodelle, dennoch führt die Digitalisierung zu einer Effizienzsteigerung in der Prozessabwicklung und zu mehr Verkäufen (Winkelhake 2021). Ein großes Thema in der Digitalisierung ist die Industrie 4.0, die in der Automobilbranche ebenfalls relevant ist und die Nutzung von heutigen sowie zukünftigen Informations- und Kommunikationstechnologien (IKT) im Produktionsbereich als wesentliches Ziel hat (Hung Vo 2016). Nach Freitag et al. (2015) werden durch die Industrie 4.0-Technologien mehr Daten generiert, was dazu führt, dass Data-Science immer bedeutsamer wird. In der Produktion ermöglichen diese Daten die Optimierung von Prozessen und Systemen (Freitag et al. 2015). Bei der Datenanalyse wird geprüft, ob in dem jeweiligen Datensatz Fehler und Ausreißer vorhanden sind (Freitag et al. 2015). Walfish (2006) behauptet, dass diese Ausreißer durch bestimmte Detektionsmethoden erkannt werden und je nach Fall entschieden wird, wie diese Ausreißer zu behandeln sind. Es können durch Ausreißer nutzbare Informationen über Prozesse zur Verfügung gestellt werden, dabei liefert die Untersuchung der Entstehungsursache erste Informationen über die Ausreißer (Walfish 2006). Die Unternehmen müssen untersuchen, ob die Ausreißer als Fehler bezeichnet werden können, die durch falsche Einstellungen des Prozesses entstehen oder als unerwartete Werte, die durch einen optimalen Prozess auch auftreten können (Walfish 2006). Nach Van Stein et al. (2016) werden in der Automobilindustrie Detektionsalgorithmen von Ausreißern z. B. für die automatische Prüfung von Metallteilen benutzt. Der Anwendungspunkt beinhaltet die Erkennung von Abweichungen der Oberfläche, wie z. B. rissige Strukturen (Van Stein et al. 2016). Diese Ausreißer können je nach Fall einen signifikanten Einfluss auf die Ergebnisse einer Produktion haben und sind in technischen Bereichen vorhanden (Domanski et al. 2022). Aufgrund dieser Tatsache müssen diese Ausreißer detektiert und behandelt werden (Domanski et al. 2022).

Im Rahmen dieser Arbeit werden die verschiedenen Möglichkeiten der Ausreißerbehandlung im Hinblick auf die Produktion der Automobilindustrie vorgestellt und daraus Kriterien abgeleitet. Das Hauptziel ist eine Kategorisierung der Ausreißerbehandlung in der Automobilproduktion, die mithilfe von Kriterien erfolgt, die aus der allgemeinen Ausreißerbehandlung aus der Literatur abgeleitet werden. Zur Erreichung des Hauptziels werden Teilziele definiert. Das erste Teilziel ist die Vorstellung der Produktion in der Automobilbranche mit dem Fokus auf Daten und Datensätzen in der Produktion, da diese den Anwendungsbereich der Ausreißerbehandlung eingrenzen. Das zweite Teilziel ist die Vorstellung der verschiedene Behandlungsmöglichkeiten von Ausreißern, damit danach die Kriterien zur Kategorisierung abgeleitet werden können. Aus dieser

Absicht folgt das dritte Teilziel, dass Anforderungen für die Kriterien gestellt werden müssen. Dabei werden die Anforderungen an die Kriterien abhängig von der Automobilindustrie, der verschiedenen Detektionsmethoden und Behandlungsmöglichkeiten erstellt.
Um diese Ziele zu erreichen, wird in dieser Arbeit zuerst der Stand der Technik in Kapitel 2 bezüglich Daten in der Produktion der Automobilindustrie thematisiert. Damit wird Grundlegendes über die Anwendungsdomäne vorgestellt und der Anwendungsbereich eingegrenzt. Daraufhin wird in Kapitel 3 der Stand der Technik bezüglich der Ausreißer im Allgemeinen und deren Anwendung bearbeitet. Am Ende dieses Kapitels erfolgt die Vorstellung der Ausreißerbehandlungen, die das wesentliche dieser Arbeit einleitet. Durch die Informationen aus der Literatur bezüglich der Behandlungsmöglichkeiten folgt in Kapitel 4 die Kategorisierung der Ausreißerbehandlung in der Produktion der Automobilindustrie. Zuerst werden Anforderungen an die Kriterien zur Kategorisierung gestellt, damit relevante und passende Kriterien erfolgen. Diese Kriterien müssen im Hinblick zur Automobilindustrie eine Kategorisierung der Ausreißerbehandlung ermöglichen. Im Anschluss werden diese Kriterien exemplarisch an einer Methode angewendet. Dann werden mithilfe der Kriterien die Behandlungsmöglichkeiten von Ausreißern bezüglich ihrer Methodik verglichen und bewertet. Am Ende des Kapitels im Fazit erfolgt die Kategorisierung der Ausreißerbehandlungen, die begründet und diskutiert wird. In Kapitel 5 werden die Erkenntnisse dieser Arbeit zusammengefasst und ein Ausblick für zukünftige Themen und Forschungen gelegt.

2 Daten in der Produktion der Automobilindustrie

Dieses Kapitel thematisiert Daten und deren Relevanz in der Produktion der Automobilindustrie. Zuerst werden grundlegende Informationen zu Daten in der Produktion der Automobilbranche vorgestellt. Dann werden die Ebenen der Daten eingegrenzt und erläutert. Im Anschluss werden die Fehler und Probleme der ausgewählten Ebene präsentiert, die das nachfolgende Kapitel einleiten.

2.1 Daten und Prozesse in der Automobilindustrie

Daten dienen als Informationen, die einem Unternehmen übermittelt werden, wenn sie in einem Kontext stehen (Piro & Gebauer 2021). Laut Mertens et al. (2017) werden Daten als eine Folge von maschinell verarbeitbarer Zeichen verstanden, die durch ihre Merkmale, Produkte der realen Welt beschreiben und repräsentieren. Diese Abfolge von Zeichen können nach Datentyp unterschiedlich sein (Mertens 2017). Die Datentypen können numerisch, alphabetisch und alphanummerisch auftreten (Mertens 2017). In der Produktion der Automobilindustrie werden viele Daten erzeugt, die einen signifikanten Einfluss haben (Alvarez-Coello et al. 2020). Nach Alvarez-Coello et al. (2020) steigt das Volumen der Daten weiter exponentiell an und muss mithilfe von Software bearbeitet werden. Die Autoren behaupten weiterhin, dass heutzutage Fahrzeuge mit besonderen Rechenressourcen und hunderten sensorischen Elementen ausgestattet sind, die zu einem enorm komplex Aufbau führen. Die Komplexität resultiert nicht nur aus den erfassten Daten der Fahrzeuge, sondern auch aus den Prozessen bei der Produktion (Kropik 2021). Nach Kropik (2021) werden unterschiedliche Fertigungstechnologien für die Produktion eingesetzt. Der Einsatz dieser Fertigungstechnologien resultiert aus den verschiedenen Anforderungen der Kunden (Marschner 2004). In der nachfolgenden Tabelle 1 in Anlehnung an Kropik (2021) werden einige der Prozesse veranschaulicht, die bei der Herstellung vorhanden sind. Es existieren Unterschiede zwischen diesen Fertigungstechnologien, dennoch führt die zunehmende Digitalisierung zu der Gemeinsamkeit, dass IT-Mittel bzw. der Einsatz von Daten weiter zunimmt (Kropik 2021). Zum Beginn der Digitalisierung wurden fünf Exabyte Daten pro Jahr erschaffen, heutzutage wird diese Datenmenge innerhalb von zwei Tagen erzeugt (Steven & Klünder 2021). Laut dem Automobilhersteller Bayerische Motoren Werke (BMW) werden weitere zukünftige Themenbereiche wie das autonome Fahren und das vernetzte Fahren einen Speicherplatz von 573 Petabytes pro Jahr in Anspruch nehmen (Nolting 2021).

Tabelle 1: Fertigungstechnologien im Automobilbau in Anlehnung an Kropik (2021), S. 14

Fertigungstechnologie	Beschreibung
Presswerk	Herstellen von geformten Blechteilen durch Umformung und mechanische Bearbeitung (Pressen, Stanzen etc.)
CFK-Fertigung	Herstellung von Karosserieteilen und Interieurteilen aus Karbonfasern
Karosserierohbau	Herstellen der Rohkarosse
Lackiererei	Lackierung der Karosse
Teilefertigung	Fertigung von Teilen in mechanischer Fertigung, wobei hauptsächlich spanende Fertigungstechniken oder Spritzguss zum Einsatz kommen. Neuerdings beginnt auch 3D Druck Einzug in die Teilefertigung zu halten
Motormontage	Montage von Motoren und Getrieben
Batteriemontage	Montage des Batteriemoduls für Elektrofahrzeuge
Vormontagen	Vormontagen von Einbauteilen, wie Sitzen, Spiegel, Cockpits, Kühler, Fahrwerkskomponenten, Kabelsträngen etc.
Fahrzeugendmontage	Endmontage des gesamten Fahrzeugs
Zertifizierung	Abschließende Prüfung und Qualitätssicherung
Auslieferung	Übergabe des fertigen Fahrzeugs an den Vertrieb

Dies führt dazu, dass in der Produktion bzw. in der gesamten Automobilindustrie große Informationsmengen vorhanden sind und das diese je nach Prozess ausgewertet werden müssen (Kropik 2021; Pietsch 2021). Nach Winkelhake (2021) sind die Analyse und Auswertung dieser Informationsmengen wichtige Aspekte der Informationstechnologie in der Automobilindustrie. Ein Beispielprozess mit einer Auswertung und Analyse der Daten in der Produktion der Automobilindustrie ist das Einfügen des Navigationssystems in das Fahrzeug (Müller et al. 2006). Laut Müller et al. (2006) wird die neue Komponente in das System adaptiert und synchronisiert. Die vorhandenen Datenstrukturen unterstützen die Adaption und die Verknüpfung mit den anderen Prozessen in dem System (Müller et al. 2006). Im Hinblick auf das Auftreten der Daten und deren Strukturen werden im folgenden Kapitel die unterschiedlichen Ebenen der Daten in der Produktion vorgestellt.

2.2 Die Ebenen der Daten in der Produktion

In dieser Arbeit wird zwischen den Ebenen der einzelnen Datenpunkten, Datenbanken und Big-Data unterschieden. Die Betrachtung einzelner Datenpunkte ist aufgrund der Thematik nicht sinnvoll. Die Beobachtungen werden interessanter, wenn mehrere einzelne Datenpunkte zusammen erscheinen. Da die Produktion der Automobilindustrie thematisiert wird, werden Big-Data, Datenbanken und Datensätze im Folgenden erklärt. Große Informationsmengen von Daten werden als Big-Data bezeichnet und haben ein signifikantes Potenzial bei der Verbesserung von Prozessabläufen und bei dem Aufbau von neuen Geschäftsmodellen (Winkelhake 2021). Damit die-

ses Potenzial ausgeschöpft werden kann bei der Auswertung und Analyse, müssen zur Verarbeitung dieser Daten entsprechende Softwarewerkzeuge benutzt werden (Winkelhake 2021; Nolting 2021). Nach Fasel & Meier (2016) hat Big-Data hat folgende Kriterien:

1. Volume
2. Variety
3. Velocity
4. Veracity
5. Value

Die ersten beiden Kriterien beziehen sich auf die Menge der Daten und auf die Vielfältigkeit der verschiedenen Datenstrukturen und Datenquellen (Nolting 2021). Die enthaltenen Daten können aus strukturierten, semi-strukturierten und unstrukturierten Daten sein, d. h. das Format der Daten kann voneinander verschieden sein (Düsing 2020). Nach Düsing (2020) kennzeichnet das dritte Kriterium die Geschwindigkeit, mit der die Daten verarbeitet und erzeugt werden. Außerdem behauptet der Autor, dass es in größeren Datenbanken einen enormen Einfluss hat, da dies betriebliche Abläufe und Entscheidungsvorgänge verschnellern kann. Als viertes Kriterium gilt die Zuverlässigkeit, bei der die Genauigkeit bzw. die Richtigkeit der Daten geprüft wird (Düsing 2020; Schroeck et al. 2012). Das letzte Kriterium kennzeichnet den wirtschaftlichen Wert der Daten für das Unternehmen (Düsing 2012). Diese Kriterien kennzeichnen zusammengefasst Big-Data aus. Kropik (2021) führt weiteraus, dass ein riesiger Strom an Daten in der Produktion erzeugt wird und deshalb Big-Data einige Anwendungen findet. Eine korrekte Behandlung der Daten ist selbst mit den modernsten Big-Data-Technologien aufwändig, aufgrund der enormen Datenmenge und Volatilität der Daten, wenn die Änderungen der Bitinformationen an den wesentlichen Sensoren und Aktoren in den speicherprogrammierbaren Steuerungssystemen (SPS-Systeme) untersucht werden (Kropik 2021). Big-Data ist daher nur sinnvoll in der Produktion, wenn stationäre Geräte benutzt werden, wie im Edge-Computing (Kropik 2021). Dies wird laut dem Autor empfohlen, da die großen Datenmengen dann innerhalb eines Netzwerks bearbeitet werden (Kropik 2021). Nach Winkelhake (2021) gibt es eine Vielzahl von Technologien, die anspruchsvoller als die klassischen Datenwerkzeuge sind. Die folgende Übersicht (s. Abbildung 1) ist eine Zusammenfassung von mehreren Werkzeugen, die durch das Mitwirken unterschiedlicher Unternehmen und Technologieanbieter entstanden sind (Winkelhake 2021). In dieser Übersicht wurden sechs Anwendungsbereiche erzeugt, die als gesamtes ein Baukastensystem bilden, mit dem anforderungsgerechte Lösungen zusammengestellt werden sollen (Winkelhake 2021).

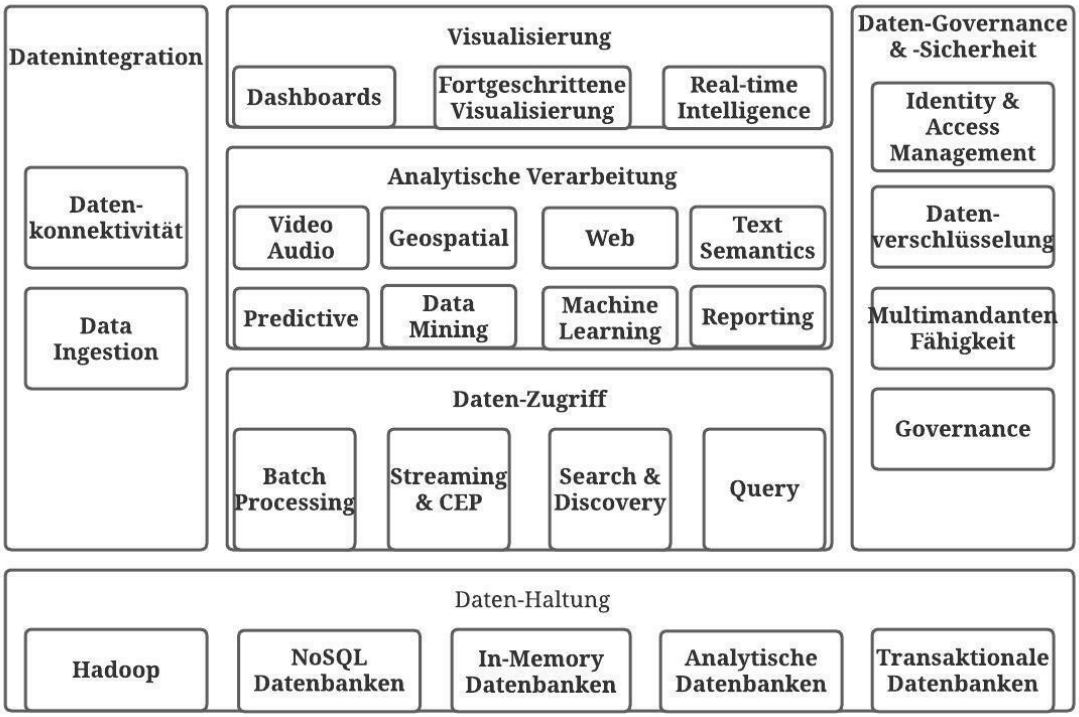

Abbildung 1: Big-Data-Werkzeuge in Anlehnung an Winkelhake (2021), S. 62

Laut Kropik (2021) können in der Automobilindustrie diese Big-Data-Werkzeuge für Projekte benutzt werden. Einer der Anwendungsbereiche ist die Aufdeckung von Verschwendungen innerhalb des Produktionssystems, bei der Redundanzen und Qualitätsmangel erkannt werden (Kropik 2021). Ein weiterer Anwendungsfall ist im Qualitätsmanagement, bei dem Serienfehler oder Prüfprozesse unterstützt werden (Kropik 2021). Außerdem kann Big-Data auch im Bereich der vorbeugenden Instandhaltung von Nutzen sein, da Ausfallzeiten durch vorbeugende Wartungen minimiert werden können (Kropik 2021). Noch ein weiteres Einsatzfeld von Big-Data wird von Winkelhake (2021) erwähnt und thematisiert das Erhöhen der Wiederverwendung von Gleichteilen. Auch wenn viele Anwendungsbereiche und Werkzeuge von Big-Data existieren, sind die Analysen in diesen Bereichen sehr abhängig vom Menschen, der dabei vom Computer und von den Programmen unterstützt wird (Kropik 2021). Der wirtschaftliche Erfolg eines Unternehmens kann durch den Einsatz von Big-Data gesteigert werden, dennoch ist die unüberlegte Nutzung davon riskant, da Fehler und Probleme entstehen können (Huber 2016). Da viele Unternehmensbereiche miteinander vernetzt arbeiten, ist der Einsatz von Big-Data noch mühsam in der Automobilindustrie (Winkelhake 2021). Nach Winkelhake (2021) ist das Problem eher organisatorisch als technisch. Neben der Ebene des Big-Data werden jetzt Datensätze und Datenbaken thematisiert, die im Hinblick auf die nächsten Kapitel eher im Fokus stehen. Als ein Datensatz werden sowohl logisch als auch physisch inhaltlich zusammenhängende Datenfelder bezeichnet (Mertens et al. 2017). Laut Vieweg et al. (2012) wird ein einzelner Eintrag von zusammengehörigen Daten in einer Datenbank als Datensatz verstanden. Datenbanken sind die Übergruppe von Datensätzen und werden als eine selbstständige, auf Dauer ausgelegte Datenorganisation, welche einen Datenbestand sicher und flexibel verwalten kann, bezeichnet (Steiner 2014). Nach Steiner (2014) hat eine Datenbank folgende Aufgaben:

- Der Zugriff auf gespeicherte Daten soll ermöglicht werden, ohne Vorkenntnisse der Organisation der Daten im System.
- Die Zugriffsberechtigung der Daten muss gesichert und vor Manipulation geschützt werden.
- Die Veränderung der Organisation der Daten soll möglich sein, ohne dabei die Anwendungsprogramme anzupassen.

Laut Steiner (2014) können Datenbanken hierarchisch, relational und objektorientiert sein. Die relationalen Datenbanken sind im Hinblick auf die Produktion und der Thematik dieser Arbeit wesentlicher. In diesen relationalen Datenbanken können Datenstrukturen ergänzt werden, ohne dass die schon bestehenden Datenstrukturen beeinflusst werden (Steiner 2014). Die Datensätze in den Datenbanken besitzen Schlüssel, die aus mehreren Datenfeldern bestehen und den Datensatz eindeutig identifizieren (Mertens et al. 2017). Weiterhin behaupten die Autoren, dass die Identifizierung der Datensätze unabhängig von den restlichen Datensätzen in der Datenbank sein muss. Daten, Datensätze und Datenbanken sind in der Produktion der Automobilindustrie von Bedeutung (Dietrich 2021). Laut Dietrich (2021) werden durch die Bestimmung des Overall-Equipment-Effectiveness (OEE) Daten von Maschinen und Anlagen erfasst. Kropik (2021) behauptet, dass die Effektivität der Maschinen in der Produktion mit bestimmten Kennzahlen ausgewertet wird. Die Auswertung des OEE zeigt, ob der Prozess der Maschinen ideal abläuft oder ob währenddessen Verluste auftreten (Dietrich 2021). Im OEE werden unterschiedliche Faktoren wie z. B. die Fertigungszeit und die Planbelegungszeit in Betracht gezogen (Kropik 2021). Nach Dietrich (2021) werden für die Berechnung des OEE folgende Daten gemessen:

- Maschinenstatus
- Ausschuss
- Output
- Personalzeiten

Nach Kropik (2021) werden durch die Auswertung des OEE Fehler an Maschinen erkannt und untersucht. Der Autor führt weiter aus, dass einzelne Parameterangaben an der Maschine eine Fehlproduktion erzeugen und die gesamte Effizienz der Produktionslinie verschlechtern können (Kropik 2021). Diese Fehler und Probleme sind in der Produktion der heutigen Zeit ein essenzieller Bestandteil, die aus weiteren Faktoren entstehen können. Im Folgenden Abschnitt werden Maschinen- und Datenfehler thematisiert, da diese im Hinblick auf Kapitel 4 von Bedeutung sind.

2.3 Fehler und Probleme von Daten in der Produktion

Innerhalb der Produktion der Automobilindustrie existieren Fehler und Probleme. Nach Kropik (2021) können diese Fehler in zwei Gruppen klassifiziert werden. Die erste Gruppe beinhaltet die Fehler, die aus automatischen Produktionsmaschinen entstehen und die zweite Gruppe thematisiert Fehler, die durch Fehlhandlungen von Menschen eintreten (Kropik 2021). Laut Kropik (2021) gehört zu den Zielen der Produktion beide Arten von Fehlern zu minimieren, die Ursachen

zu erkennen und Lösungen zu finden. Die Kontrolle der industriellen Prozesse ist aufgrund des Automatisierungsgrads ein komplexes Thema (Nemeth & Peterkova 2018). In der Automobilproduktion in Europa werden menschliche Fehler häufig durch eine Automatisierung mit einer Maschine verdrängt (Kropik 2021). Diese Verdrängung hat jedoch nicht alle Fehler in der Produktion entfernt (Kropik 2021). Der Fokus in dieser Arbeit wird auf die maschinenbedingten Fehler und die daraus resultierenden Daten gesetzt. Nach Kropik (2021) verursachen die Produktionsmaschinen Fehler, da sie wie Fahrzeuge komplex aufgebaut sind und aufgrund der hohen Anforderungen an den Automatisierungsgrad beansprucht werden. Weitere Fehlerursachen nach Kropik (2021) sind in der Tabelle 2 dargestellt. Dabei fällt auf, dass äußere Faktoren, wie z. B. Teile, die bearbeitet werden sollen, Fehler verursachen (Kropik 2021). Diese Teile können nach dem Autor defekt sein und wurden vor dem Prozess von den Maschinen nicht erkannt. Der Autor führt weiter aus, dass neben den externen Teilen auch interne Maschinenteile Fehler erzeugen. Diese Art von Fehler ist durch Instandhaltungen und Prüfungen zu vermeiden (Kropik 2021).

Tabelle 2: Maschinenbedingte Fehlerursachen in Anlehnung an Kropik (2021), S. 137

Fehlerursache	Beschreibung
Teilfehler	Fehler, die durch defekte Teile verursacht werden. Typisch sind Geometrien außerhalb der Toleranz, Verschmutzung, Beschädigungen etc.
Falsche Teile	Fehler durch falsche Teile, die von den Teilekontrollen an der Maschine nicht erkannt werden.
Falsche Parameter	Die Maschine wird mit fehlerhaften oder falschen Parametersätzen betrieben.
Ablauffehler	Fehlerhafte Fertigungsabläufe im Automatikbetrieb. Dies führt z. B. zu Kollisionen.
Defekte Maschinenteile	Ausfall von Komponenten in der Maschine (Antriebe, Netzteile, Netzwerkkomponenten oder Maschinenteile).
Sicherheitsprobleme	Die Sicherheitsausrüstung an der Maschine ist fehlerhaft (Not-Aus Schalter, Lichtvorhänge etc.).
Prozessvariationen	Variationen im Fertigungsprozess, die unvermeidbar zu Ausschuss führen.
Diagnosefehler	Fehler, die von der Maschinendiagnose gemeldet werden, aber keinem realen Problem entsprechen

Nach Hallebach und Täufer (2020) müssen diese Fehler erkannt und behandelt werden, um die Produktionsqualität zu erhöhen und Kosten zu sparen. Dabei werden in der Automobilindustrie beispielsweise Sensordaten erfasst, welche die Informationen unmittelbar dem Qualitätsprüfer überliefern (Hallebach & Täufer 2020). Laut Dietrich (2021) werden mit der Datenerfassung komplexe Zusammenhänge analysiert. Der Autor führt weiter aus, dass die Erfassung in zwei Gruppen unterteilt werden kann, und zwar in die manuelle und automatische Erfassung. In dieser Arbeit ist die automatische Datenerfassung relevanter, da Daten aus Sensoren von Maschinen thematisiert werden (Dietrich 2021). Es werden verschiedene Daten während der Produktion erfasst, beispielsweise können Sensoren Maße, Drücke und Temperaturen an Produktionsmaschinen erkennen und wiedergeben (Dietrich 2021). Wenn Anomalien detektiert werden, können in der Produktion die Parameter der jeweiligen Maschinen angepasst werden, um die defekten Teile zu reduzieren (Van Stein et al. 2016). Laut Kropik (2021) sind die erfassten Daten während dieses

Prozesses signifikant, da diese die Analyse ermöglichen. Im Hinblick auf Big-Data und Datenbanken werden solche Analysen von Daten durch die Kooperation vom Menschen mit Maschinen durchgeführt, da maschinelle Algorithmen selbstständig mangelhafte Korrelationen finden können (Kropik 2021). In Abbildung 2 ist auf der rechten Seite eine Analyse bzw. eine Korrelation, die von einem Algorithmus einer Maschine erfasst wurde, zu erkennen. Nach Kropik (2021) wurden einige Daten nicht in Betracht gezogen und führen deshalb zu einer fehlerhaften Analyse. Im linken Teil der Abbildung 2 wurde die Korrelation richtig ausgeführt, da der Mensch mitgewirkt hat und den maschinellen Algorithmus prüfen konnte (Kropik 2021).

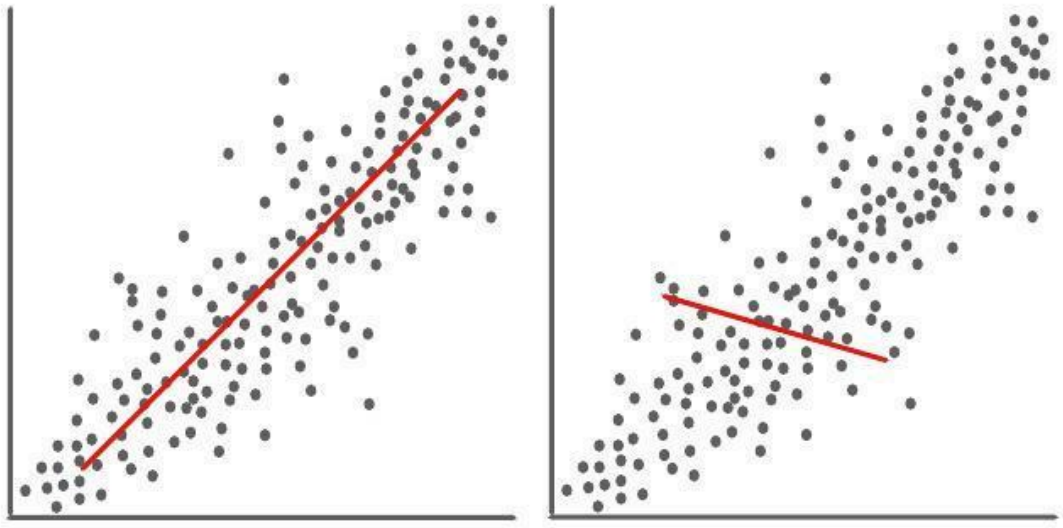

Abbildung 2: Korrekte und fehlerhafte Analyse von Daten in Anlehnung an Kropik (2021), S. 439

Nicht nur fehlerhafte Korrelationen führen zu Problemen bei der Auswertung von Daten. Die Qualität der Daten ist ebenfalls essenziell für die produzierenden Unternehmen (Mertens et al. 2017). Laut Mertens et al. (2017) bezieht sich die Datenqualität auf die Relevanz und Korrektheit von Informationen. Die Autoren führen weiter aus, dass Daten von schlechter Qualität z. B. Datenfehler, Dubletten, fehlende Werte oder falsche Formatierungen enthalten und potenzielle Fehlerquellen darstellen (Mertens et al. 2017). Laut den Autoren führen Daten von schlechter Qualität zu fehlerhaften oder wenig hilfreichen Rückmeldeergebnissen und damit zu möglichen Fehlentscheidungen (Mertens et al. 2017). Diese Qualitätsmängel existieren ebenfalls in Produktionsdaten. Bei falscher Vorgehensweise mit Qualitätsmängeln von Daten können in der Produktion Materialverluste und Fehlproduktionen entstehen (Schmid 2001). Deshalb müssen Daten aus einem optimalen Prozess unterschiedlich betrachtet werden. Der optimale Prozess wird in dieser Arbeit als Prozess verstanden, bei dem keine Fehler in der Ausführung bzw. in der Vorgehensweise zu erkennen sind. Weiterhin sind in dieser Arbeit mehrere einzelne Datenpunkte aus Datensätzen der Produktion interessant, die durch diesen optimalen Prozess entstehen und nicht dem restlichen Datensatz entsprechen.

3 Ausreißer und deren Anwendungen

In diesem Kapitel werden die Ausreißer und deren Anwendung thematisiert. Zuerst wird der allgemeine Ausreißerbegriff erklärt und dann die verschiedenen Ausreißertypen vorgestellt. Im Anschluss werden die Ausreißererkennungsmethoden vorgestellt, in denen die Detektion von den Ausreißern erläutert wird. Schließlich werden die Behandlungsmöglichkeiten von Ausreißern vorgestellt, die danach das Wesentliche dieser Arbeit einleiten.

3.1 Ausreißerbegriff und -typen

Als Ausreißer werden Datenpunkte bezeichnet, die im Vergleich zu den restlichen Datenpunkten einen signifikanten Unterschied bzw. eine Abweichung vorweisen (Hawkins 1980; Aggarwal 2017). Nicht übereinstimmende Muster in Datensätzen werden oft mit Ausreißern assoziiert und auch als Anomalien, widersprüchliche Beobachtungen, Abweichungen, Überraschungen oder Besonderheiten definiert (Aggarwal 2017; Chandola et al. 2009). Dabei ist es wichtig in diesen Datensätzen die Entstehung bzw. die Ursache der Ausreißer zu untersuchen, da die Ausreißer nutzbare Informationen über die Anomalien eines Prozesses oder eines Systems darlegen können (Aggarwal 2017; Chandola et al. 2009). Jedoch sollte das ganze System untersucht werden, da Fehler im Prozess wie z. B. Messfehler oder auch falsche Messsysteme zu unerwarteten Werten führen kann (Walfish 2006). Das Auftreten von Ausreißern kann je nach Prozess und System variieren und wird im folgenden Abschnitt thematisiert.

Nach Chandola et al. (2009) können Ausreißer in verschiedenen Typen auftreten und müssen deshalb auch unterschiedlich untersucht werden. Die grundlegendste Klassifizierung von Ausreißern ist die Unterscheidung von einzelnen, kontextualen und kollektiven Ausreißern (Divya & Sasidhar 2016; Chandola et al. 2009). Laut Divya & Sasidhar (2016) werden einzelne Ausreißer als Punktausreißer definiert und als individuelle Dateninstanz verstanden, die im Vergleich zu den erwarteten Werten eine Abweichung nachweisen. Die Punktausreißer sind die einfachsten sowie häufigsten Ausreißer und sind deshalb im Fokus der meisten Untersuchungen (Divya & Sasidhar 2016; Aggarwal 2017). Diese Punktausreißer werden nicht nur in einfachen Systemen untersucht, sondern auch in mehrdimensionalen Datenstrukturen (Gupta et al. 2014). Die Abbildung 3 veranschaulicht einen Punktausreißer in einem zweidimensionalen System. Es ist zu erkennen, dass dieser Punktausreißer im Vergleich zu den zwei Clustergruppen eine Abweichung aufweist und deshalb als Ausreißer untersucht werden kann (Ranga Suri et al. 2019).

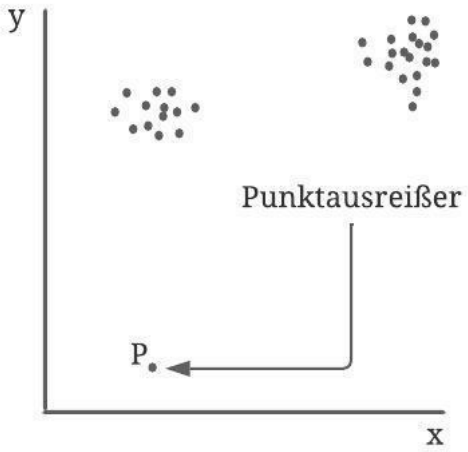

Abbildung 3: Ein Punktausreißer in einem Datensatz in Anlehnung an Ranga Suri et al. (2019), S. 4

Nach Chandola et al. (2009) weisen im Unterschied zu Punktausreißern kontextualen Ausreißer mehrere Datenpunkte auf und sind Ausreißer deren Wert innerhalb eines Kontextes von anderen Werten abweichen. Der Kontext muss als Begriff durch die Struktur des Datensatzes erkannt und als Teil der Formulierung des Problems angegeben werden (Chandola et al. 2009). Die Autoren führen weiter aus, dass jede Dateninstanz mit dem kontextualen- und Verhaltensattribut definiert wird, wobei das Verhaltensattribut die nicht kontextualen Aspekte charakterisiert. Kontextuale Ausreißer werden vorwiegend in abhängigkeitsorientierten Datentypen, wie Zeitreihen untersucht (Gupta et al. 2014).

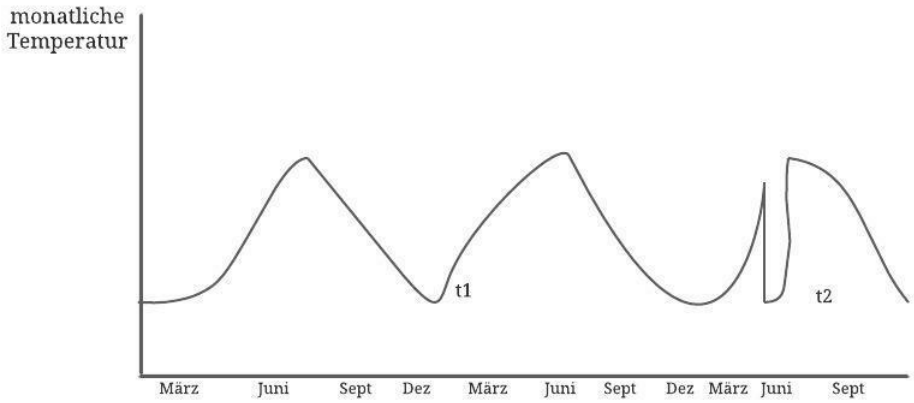

Abbildung 4: Kontextuale Ausreißer t_2 im Zeit- und Temperaturdiagramm in Anlehnung an Chandola et al. (2009), S. 8

In Abbildung 4 wird ein kontextualer Ausreißer dargestellt, der hier mit t_2 gekennzeichnet wurde. Gleichzeitig wurde eine weitere Stelle mit t_1 gekennzeichnet und hat den gleichen Wert wie t_2. Obwohl beide Temperaturen identisch sind, ist nur t_2 ein kontextualer Ausreißer, da zu der Zeit die monatliche Temperatur eine Anomalie bzw. eine Abweichung nachweist. Die kollektiven Ausreißer oder auch Cluster-Ausreißer sind mehrere Datenpunkte als Gruppe zusammengefasst

(Chandola et al. 2009). Nach Chandola et al. (2009) sind die einzelnen Ausreißer nicht nur aufgrund der Variation zu den erwarteten Werten Ausreißer, sondern auch durch das Auftreten in der Gruppe. Des Weiteren können diese kollektiven Ausreißer nur als Gruppe betrachtet werden, wenn die einzelnen Dateninstanzen in einem Zusammenhang sind (Divya & Sasidhar 2016; Chandola et al. 2009). In Abbildung 5 sind kollektive Ausreißer zu erkennen. Dabei sind die einzelnen Dateninstanzen nur Ausreißer durch das kollektive Auftreten (Chandola et al. 2009; Goldberger et al. 2000).

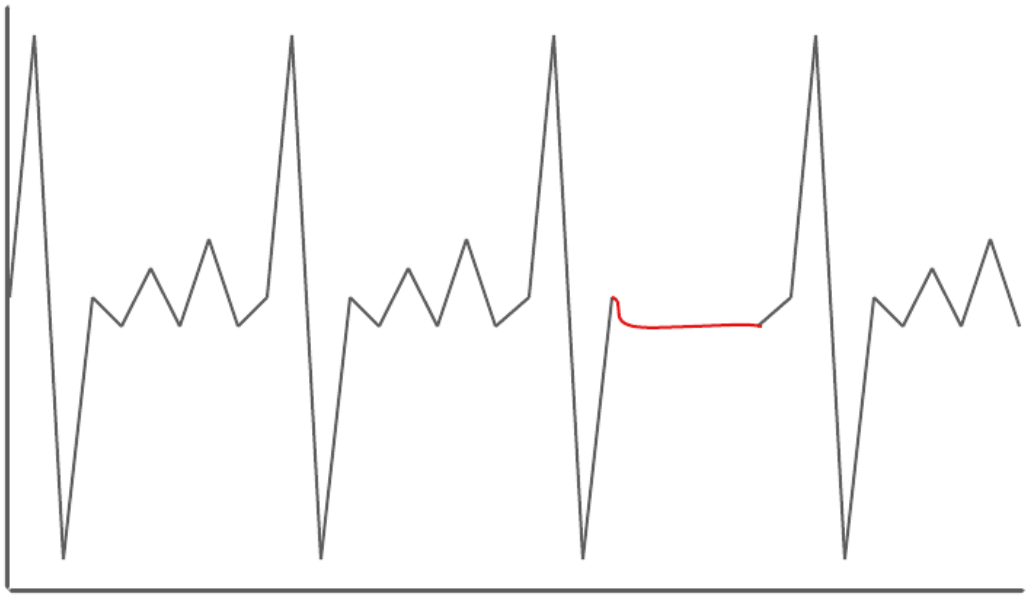

Abbildung 5: Kollektive Ausreißer in einem Elektrokardiogramm in Anlehnung an Chandola et al. (2009), S. 9 & Goldberger et al. (2000), S. 2

Unabhängig von dem Typ des Ausreißers ist die Identifizierung der Ausreißer eine der fundamentalsten Aspekte und wird im nächsten Kapitel thematisiert.

3.2 Ausreißererkennungsmethoden

Laut Chandola et al. (2009) ist die Detektion von Anomalien bzw. Ausreißern ein relevantes Thema, worüber in vielen Anwendungsdomänen geforscht wird. Es gibt viele Ausreißererkennungsmethoden, die auch als Detektionsmethoden bekannt sind und entweder für bestimmte Anwendungsdomänen spezialisiert oder generalisiert wurden (Chandola et al. 2009). Außerdem hängt die Ausreißererkennungsmethode noch von weiteren Faktoren ab, wie z. B. vom Ausreißer- und Datentyp (Gupta et al. 2014). Nach Chandola et al. (2009) ist die Datenbezeichnung signifikant, da Bezeichnungen, die einer Dateninstanz hinzugefügt wurden, wiedergeben, ob es ein normaler Wert oder eine Anomalie ist. Die Kennzeichnung der Dateninstanzen wird oft manuell von einem menschlichen Experten durchgeführt und daher ist der Prozess aufwändig (Chandola et al. 2009). Abhängig von den Kennzeichnungen der Dateninstanzen kann man zwischen den drei Verfahren wählen (Chandola et al. 2009). Neben der beaufsichtigten Ausreißererkennung existieren noch die halbbeaufsichtigten und unbeaufsichtigten Verfahren (Gupta et al. 2014). In den

beaufsichtigten Verfahren ist ein Trainingsdatensatz vorhanden, der Dateninstanzen als normal und als Anomalie kennzeichnet (Divya & Sasidhar 2016). Dabei wird ein prädikatives Modell erzeugt für normale und Anomalieklassen (Chandola et al. 2009). Dann wird jede Dateninstanz mithilfe des Modells in eine der Klassen zugeordnet (Chandola et al. 2009). Laut Chandola et al. (2009) existieren zwei Probleme, die in diesem Verfahren auftauchen. Im Trainingsdatensatz werden die Anomalien weniger mit den normalen Instanzen verglichen und die Genauigkeit der Kennzeichnung für die Anomalieklasse ist sehr anspruchsvoll (Chandola et al. 2009). Daher können die Ergebnisse von beaufsichtigten und unbeaufsichtigten Verfahren sehr unterschiedlich sein (Chandola et al. 2009; Aggarwal 2017). In den halbbeaufsichtigten Verfahren werden vom Trainingsdatensatz nur die normalen Dateninstanzen gekennzeichnet (Chandola et al. 2009). Laut Chandola et al. (2009) ist die Anwendung verbreiteter, da Anomalien nicht gekennzeichnet werden. Im Gegensatz zu den beaufsichtigten Verfahren wird das Modell dann die normalen Werte in eine Klasse hinzufügen und die restlichen Werte in den Testdaten als Anomalien identifizieren (Chandola et al.2009). Dennoch sind in diesen Verfahren auch Probleme vorhanden, die aufgrund des Trainingsdatensatzes entstehen, der nicht alle Anomalien von den normalen Werten unterscheiden kann (Chandola et al. 2009). Die letzten Verfahren sind die unbeaufsichtigten und im Vergleich zu den Verfahren davor auch die gängigsten, da keine Trainingsdaten notwendig sind (Chandola et al. 2009). Nach Chandola et al. (2009) wird hierbei die Annahme gestellt, dass normale Werte öfter als Anomalien auftreten. Des Weiteren wird eine zweite Annahme gestellt, in der behauptet wird, dass Anomalien sich qualitativ von normalen Instanzen unterscheiden (Eskin et al. 2002; Chandola et al. 2009). Durch diese Annahmen sind die Anomalien leichter zu detektieren, weil sie erstens seltener Auftreten und zweitens sich von normalen Werten unterscheiden (Eskin et al. 2002). Unabhängig vom Verfahren bzw. von der Methode geben viele Ausreißererkennungsmethoden eine Wertung für die jeweiligen Ausreißer ab, bei der die Ausreißer je nach Schwellenwert gekennzeichnet werden (Chandola et al. 2009). Dabei muss berücksichtigt werden, dass die Auswahl des Schwellenwerts einen signifikanten Einfluss hat, ob die untersuchten Werte Ausreißer sind (Aggarwal 2017; Chandola et al. 2009). Mithilfe von Rechnungen und dem festgelegten Schwellenwert können der Precision- und Recall-wert bestimmt werden (Aggarwal 2017). Nach Aggarwal (2017) gibt der Precision-wert den prozentualen Wert der Ausreißer an, die tatsächlich Ausreißer sind. Im Folgenden sind die Formeln für die Berechnung zu erkennen, dabei wird der Precision-wert mithilfe der deklarierten Ausreißer S(t) und der wahren Menge G berechnet (Aggarwal 2017). Der Recall-wert gibt schließlich den prozentualen Wert der wahren Ausreißer an, die davor als Ausreißer durch den Schwellenwert gekennzeichnet wurden (Aggarwal 2017).

$$Precision(t) = 100 * \frac{|S(t) \cap G|}{|S(t)|} \tag{1}$$

$$Recall(t) = 100 * \frac{|S(t) \cap G|}{|G|} \tag{2}$$

Nachdem Grundlegendes über die Ausreißererkennungsmethoden erläutert wurde, werden jetzt verschiede Methoden genauer vorgestellt. Dabei muss erwähnt werden, dass keine Methode universal einsetzbar ist und je nach Anwendungsbereich variiert werden kann. Die Methoden können noch weiter klassifiziert werden, und zwar in parametrische und nichtparametrische Varianten (Ranga Suri et al. 2019). Im Folgenden werden statistikbasierte, distanzbasierte, dichtebasierte und zuletzt clusterbasierte Methoden vorgestellt, die in Datensätzen benutzt werden können. Die statistikbasierten Methoden gehören zu den parametrischen Varianten, da sie ein Modell bzw. eine Verteilung benötigen für die Dateninstanzen (Ranga Suri et al. 2019). Die restlichen Methoden, die eben erwähnt wurden, gehören zu den nichtparametrischen Varianten, da sie keine Annahmen besitzen wie die statistikbasierten Methoden (Ranga Suri et al. 2019).

3.2.1 Dixon-Test

Laut Shrivastava et al. (2014) gehört der Dixon-Test zu den statistikbasierten Verfahren und ist eine Methode, die für kleine Datensätze benutzt werden sollte. Dabei ist es wichtig, dass es sich um eine Normalverteilung handelt. Außerdem soll diese Methode nur einmal pro Datensatz angewendet werden (Shrivastava et al. 2014). Die Ausführung des Dixon-Test wird mithilfe einer Beispielrechnung vorgestellt. Dafür wird in der folgenden Tabelle 3 ein sortierter Datensatz aufgeführt.

Tabelle 3: Beispiel Datensatz in Anlehnung an Vogel (2021), S. 26

Datensatz
295,1
296,9
297,5
300,8
301.5
350

In diesem Anwendungsfall des Dixon-Test wird der Abstand zwischen dem größten und zweitgrößtem Wert und zwischen dem kleinsten und dem größten Wert berechnet. Nach Vogel (2021) lautet die Formel des Dixon-Test hiermit:

$$r_{11} = \frac{(\mathrm{x}_n - x_{n-1})}{(x_n - x_1)} \tag{3}$$

Der Index n kennzeichnet die Anzahl und steht in der Formel für den höchsten Wert (Vogel 2021). In dieser Beispielrechnung resultiert ein Endergebnis von 0,883, dieses Ergebnis wird mit einem kritischen Wert verglichen, der in vorgefertigten Tabellen 4 entnommen werden kann (Vogel 2021; Walfish 2006). Laut Vogel (2021) ist der kritische Wert erstens abhängig von der Anzahl der Werte im Datensatz und zweitens von der gewünschten Irrtumswahrscheinlichkeit. Für dieses Beispiel wurde eine Irrtumswahrscheinlichkeit von 5 % ausgewählt und die Anzahl der Werte

beträgt n = 6 (Vogel 2021). Mit diesen beiden Informationen wird der kritische Wert aus der Tabelle 4 entnommen.

Tabelle 4: Kritische Werte für Dixon-Test in Anlehnung an Verma & Quiroz Ruiz (2006), S. 139

n	20 %	10 %	5 %
5	0,4508	0,5578	0,6423
6	0,3868	0,4840	0,5624
7	0,3444	0,4340	0,5077

Es ist zu erkennen, dass der Wert aus der Tabelle 0,5624 beträgt und kleiner ist als das berechnete Ergebnis. Deshalb ist der untersuchte Wert mit $x_n = 350$ statistisch signifikant als Ausreißer bestätigt (Vogel 2021). Der Dixon-Test ist durch Modifizierungen der Formel auch für die Berechnung anderer Werte nutzbar (Vogel 2021; Walfish 2006). Wenn der kleinste Einzelwert berechnet werden soll, ändert sich in der Formel der Zähler. Die neue Formel nach Walfish (2006) & Vogel (2021) lautet dann:

$$r_{11} = \frac{(x_2 - x_1)}{(x_n - x_1)} \tag{4}$$

Nach Walfish (2006) sind weitere Modifizierungen möglich, bei denen der Nenner in der Formel angepasst wird und somit z. B. eine Ausreißererkennung stattfindet, in der maximale bzw. minimale Werte ausgeschlossen werden. In der nächsten Formel wird nach Walfish (2006) die größte Beobachtung ohne Betrachtung der kleinsten Beobachtung ausgewertet.

$$r_{11} = \frac{(x_n - x_{n-1})}{(x_n - x_2)} \tag{5}$$

Neben dem Dixon-Test gibt es noch weitere statistikbasierte Verfahren, von denen zwei nachfolgend kurz erläutert werden. Ein weitere Methode ist die Maximum-Likelihood-Methode, die als Voraussetzung hat, dass die Daten mithilfe der Gauß´schen Verteilung generiert wurden (Chandola et al. 2009). Laut Chandola et al. (2009) wird dabei ein Anomaliewert erschaffen, der durch die Distanz der Dateninstanzen zum geschätzten Mittelwert entsteht. Des Weiteren hat der Anomaliewert einen gegebenen Schwellenwert, der von Dateninstanzen überschritten werden muss, um als Anomalie zu gelten (Chandola et al. 2009). Die zweite Methode ist die Außreißererkennung mithilfe des Boxplot. Der Boxplot wird durch das Minimum, das Maximum, dem unteren Quartil (Q_1), dem oberen Quartil (Q_3) und dem Median (Q_2) charakterisiert, die in der Abbildung 6 zu erkennen sind. Nach Chandola et al. (2009) wird der Abstand Q_3 -Q_1 als Interquartilsabstand IQR bezeichnet und ist wesentlich bei der Ausreißererkennung. Die Folgenden Formeln nach Chandola et al. (2009) & Laurikkala et al. (2000) zeigen an, welche Werte als Ausreißer angesehen werden.

$$1{,}5 * IQR < Q_1 \tag{6}$$

$$1{,}5 * IQR > Q_3 \tag{7}$$

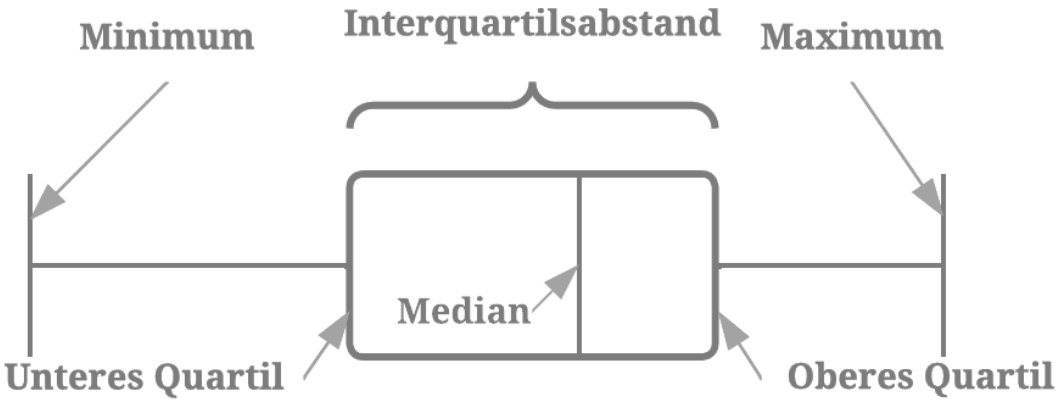

Abbildung 6: Boxplot-Darstellung in Anlehnung an Chandola et al. (2009), S. 31

3.2.2 Nearest-Neighbor-Methode

Die Nearest-Neighbor-Methode (NNM) gehört zu den distanzbasierten Verfahren und wird im Data-Mining angewendet (Ranga Suri et al. 2019). Die Methode hat die Annahme, dass normale Dateninstanzen jeweils in direkter Nähe zu den restlichen Instanzen sind und Ausreißer sich weit entfernt von den nächsten Instanzen befinden (Chandola et al. 2009). Diese Nähe wird in dieser Methode als Nachbarschaft bezeichnet (Chandola et al. 2009; Ranga Suri et al. 2019). Nach Ranga Suri et al. (2019) ist die Grundidee, dass eine Dateninstanz als Ausreißer bezeichnet wird, wenn nicht mehr als k Objekte innerhalb der Distanz d sich aufhalten. Im Vergleich zu den Methoden aus den statistikbasierten Verfahren wird hier keine Information über die Verteilung benötigt, dennoch muss die Distanz d und der k Wert festgelegt werden (Ranga Suri et al. 2019; Chandola et al. 2009; Aggarwal 2017). Diese Grundidee wurde weiterentwickelt und spezifiziert. Drei mögliche Methoden nach Ranga Suri et al. (2019) sind:

1. Distanz zu dem k-Nearest-Neighbor
2. Durchschnittliche Distanz zu den k-Nearest-Neighbor
3. Summe der Distanzen zu den k-Nearest-Neighbor.

In der ersten Methode ist es möglich durch die Distanz der Instanzen eine Ausreißerwertung zu ermitteln (Ranga Suri et al. 2019). Darüber hinaus können die ausgewerteten Ausreißer miteinander verglichen und eingestuft werden (Ranga Suri et al. 2019). Nach Ranga Suri et al. (2019) wären in diesem Fall Ausreißer mit den größten Distanzen in der Einstufung relevanter und würden deshalb auch behandelt werden. Die zwei anderen Methoden bauen auf die erste Methode auf und arbeiten ebenfalls mit Einstufungen der Dateninstanzen bei der Distanzmessung (Ranga Suri et al. 2019).

3.2.3 Local-Outlier-Factor

Der Local-Outlier-Factor (LOF) ist ein dichtebasiertes Verfahren, der Dateninstanzen durch Dichtemessungen überprüft (Ranga Suri et al. 2019). Laut Ranga Suri et al. (2019) wird der LOF durch zwei Parameter gekennzeichnet. Der erste Parameter wird als MinPts bezeichnet und steht für die minimale Anzahl der Objekte (Ranga Suri et al. 2019). Der zweite Parameter gibt das untersuchte Volumen an (Ranga Suri et al. 2019). Wie bei der Nearest-Neighbor-Methode wird eine Dateninstanz anhand ihrer Nachbarschaft ausgewertet, das bedeutet, dass eine geringe Dichte ein Kennzeichen für einen Ausreißer sein kann (Chandola et al. 2009; Ranga Suri et al. 2019). In Abbildung 7 ist ein Beispiel dargestellt, in welchem der LOF effektiver als die Nearest-Neighbor-Methode ist (Ranga Suri et al. 2019). Beide Verfahren werden im Hinblick auf P_1 und P_2 angewendet. Dabei fällt auf, dass der LOF beide Punkte als Ausreißer erkennt und die Nearest-Neighbor-Methode nur P_2 als Ausreißer akzeptiert (Ranga Suri et al. 2019). Nach Ranga Suri et al. (2019) ist die Ursache dafür, dass die Distanz von P_1 zu den anderen Dateninstanzen zu gering ist und somit den Schwellenwert bzw. den Wert für die Distanz nicht überschreitet.

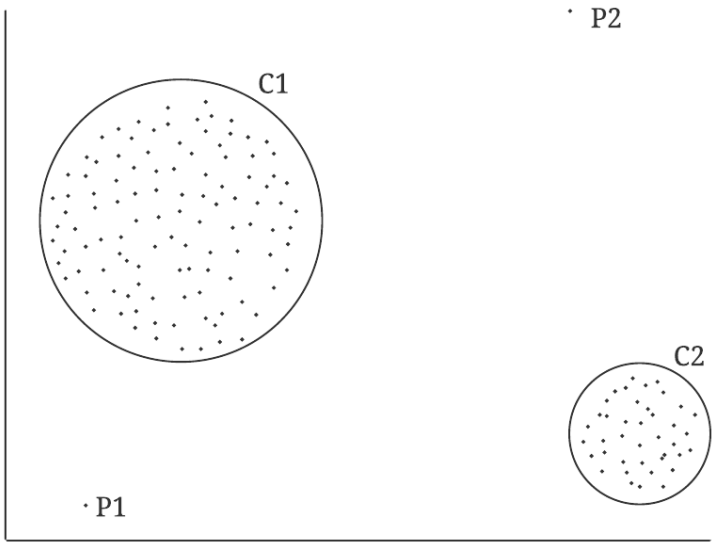

Abbildung 7: Darstellung von dichtebasierten Ausreißern in Anlehnung an Ranga Suri et al. (2019), S. 35

Die Berechnungen bei dem LOF können sehr beanspruchend werden, deshalb wurden verschiedene Modifizierungen erschaffen (Ranga Suri et al. 2019). Eine der Modifizierungen ist es Micro-Cluster innerhalb des gleichen Clusters zu erschaffen, die die Gruppe repräsentiert (Ranga Suri et al. 2019).

3.2.4 Cluster-Based-Local-Outlier-Factor

Der Cluster-Based-Local-Outlier-Factor (CBLOF) gehört zu den clusterbasierten Verfahren, in denen angenommen wird, dass normale Dateninstanzen in großen und dichten Clustern auftreten, und Ausreißer bzw. Anomalien in kleinen und spärlichen Clustern vorzufinden sind (Chandola et

al. 2009; Ranga Suri et al. 2019). Der Grundgedanke bei dieser Methode ist die Bewertung der Dateninstanzen bezüglich der Clustergröße und der Distanz zum nächstgrößeren Cluster (Chandola et al. 2009). Die clusterbasierten Verfahren sind ähnlich zu den Nearest-Neighbor-Verfahren, da die Distanz der Dateninstanzen einen signifikanten Einfluss haben (Chandola et al. 2009). Der Hauptunterschied ist, dass in den clusterbasierten Verfahren jede Dateninstanz bezüglich ihres Cluster ausgewertet wird und nicht bezüglich ihrer lokalen Nachbarschaft (Chandola et al. 2009). Nach Chandola et al. (2009) sind die Vorteile dieses Verfahrens, dass die unbeaufsichtigt angewendet werden können. Außerdem ist die Adaptierung auf andere komplexe Datenstrukturen durch simples Einfügen des Cluster-Algorithmus möglich (Chandola et al. 2009). Die Autoren führen weiter aus, dass neben den Vorteilen auch Nachteile vorhanden sind. Ein Nachteil ist, dass die Effizienz der Verfahren abhängig von den Algorithmen ist, die die Cluster-Struktur der normalen Instanzen erkennen (Chandola et al. 2009). Des Weiteren werden Anomalien von mehreren Methoden als Nebenprodukt bei der Cluster-Bildung detektiert (Chandola et al. 2009).

3.3 Behandlungsmöglichkeiten von Ausreißern

Nachdem in den letzten Kapiteln Ausreißer im Allgemeinen und ihre Detektion erläutert wurden, werden in diesem Kapitel die Behandlungsmöglichkeiten von Ausreißern thematisiert. Dabei liegt der Fokus auf den hybriden Methoden, da diese im Hinblick auf Kapitel 2 eher von Bedeutung sind. Der Zusammenhang zwischen der Ausreißerbehandlung und der Produktion wird in Kapitel 4 erläutert. Dabei hat die Ursache der Ausreißer einen signifikanten Einfluss auf die Behandlung. Wenn der Ausreißer aufgrund eines Fehler entstanden ist, wie z. B. durch falsche Einstellungen von Parametern oder Messfehlern, können diese Werte entweder angepasst oder entfernt werden (Kutner et al. 2004; Aguinis et al. 2013). Es sollte dokumentiert werden, durch welchen Fehler diese Ausreißer entstanden sind. Viel interessanter sind Ausreißer, die durch Detektieren als Ausreißer bestätigt wurden, aber nicht direkt auf Fehler zurückzuführen sind (Aguinis et al. 2013). Für Ausreißer, die genauer untersucht werden müssen, wurden Behandlungsmöglichkeiten erschaffen, die jetzt genauer vorgestellt werden.

3.3.1 Least-Trimmed-Squares

Die erste Behandlungsmöglichkeit ist die Methode der Least-Trimmed-Squares (LTS). Bei dieser Methode wird die Summe der h kleinsten quadratischen Rückstände minimiert, wobei h als Konstante gilt (Rousseeuw & Leroy 1987; Seo & Bae 2012). Im LTS werden für Datenpunkte ein lineares Modell angepasst, das aufgrund der Robustheit gegenüber Ausreißer stabil ist (Mount et al. 2014). Dabei dürfen die Ausreißer nicht über 50 % des Datensatzes darstellen (Mount et al. 2014). Es folgt die Formel für den LTS Schätzer:

$$min \sum_{i=1}^{h} \varepsilon_i^2 \qquad (\frac{n}{2} < \mathrm{h} \leq \mathrm{n}) \tag{8}$$

Der Index i gibt die geordnete Reihenfolge jedes Datenpunktes an, n steht für die Anzahl der Datenpunkte und h ist die Teilmenge von n (Seo & Bae 2012). Der Parameter ε gibt die Differenz

des aktuellen Wertes zu dem prognostizierten Wert jeder Dateninstanz an (Seo & Bae 2012). Durch das Ordnen der Rückstände ε kann LTS die getrimmte Summe begrenzen, wobei die Rückstände zuerst quadriert und dann geordnet werden (Rousseeuw & Leroy 1978). Anhand der Formel ist zu erkennen, dass die größten quadrierten Rückstände in der Summe nicht berücksichtigt werden und dabei Ausreißer vom linearen Modell fernhalten (Rousseeuw & Leroy 1978). Laut Rousseeuw & Leroy (1978) sieht man in der folgenden Abbildung 8 die Robustheit solcher linearen Modellen anhand der Methode Least-Median-of-Squares (LMS). Der Punkt, der in der Abbildung 8 mit 1 gekennzeichnet wurde, beeinflusst die Regressionsgerade nicht und kann als Ausreißer erkannt werden.

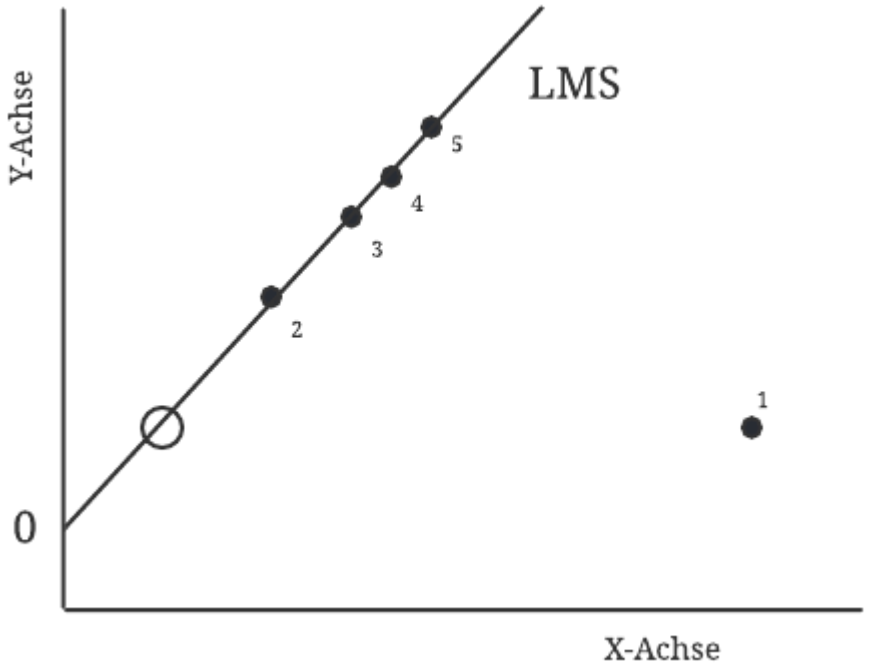

Abbildung 8: Robustheit von der LMS-Regression mit einem Ausreißer in X-Richtung in Anlehnung an Rousseeuw & Leroy (1978), S. 16

3.3.2 Bagplot-Based-Adjustment

Im Bagplot-Based-Adjustment (BBA) ist ein Bagplot vorhanden. Dieser ist im Vergleich zum Boxplot bei mehrdimensionalen Datensätzen vorhanden und besteht aus dem Bag, der 50 % der Datenpunkte hat, aus einem Fence, der Ausreißer von den Einreißern trennt, aus einem Loop, der die Datenpunkte außerhalb des Bag und innerhalb des Fence angibt (Rousseeuw et al. 1999; Avanzi et al. 2022). Nach Rousseeuw et al. (1999) entsteht der Bag durch die Auswertung der vorhandenen Datenpunkten und dem Tiefenmedian. Das Resultat ist ein konvexes Polygon, der auch als Basis für die nachfolgenden Bestandteile ist (Rousseeuw et al. 1999). Für den Fence wird meistens ein gängiger Faktor mit 3 ausgesucht, der vor dem Prozess ausgewählt wird (Rousseeuw et al. 1999). Abhängig vom Bag wird der Fence erzeugt (Rousseeuw et al. 1999). In diesem Fall wird ein zweidimensionales System thematisiert (Rousseeuw et al. 1999; Avanzi et al. 2022). Bei dieser Methode sollen Ausreißer nach der Anpassung, die außerhalb des Fence sind, an den Fence verschoben werden, jedoch erscheinen diese Datenpunkte durch die neuen Prozessberechnungen

innerhalb des Loop (Verdonck & Van Wouwe 2011). Während dieser Anpassungen ändern sich also die Formen und Größen der Bereiche, die dann zu neuen Ausreißern führen können (Avanzi et al. 2022). Nach Avanzi et al. 2022 ist in der Abbildung 9 links das System vor der Anwendung der Methode zu erkennen. Dabei ist die grüne Umrandung der Fence (Avanzi et al. 2022). Rechts in der Abbildung sieht man den Bagplot nach der Anpassung, wo die Ausreißer innerhalb des Loop sind (Avanzi et al. 2022). Außerdem ist zu sehen, dass der Loop aufgrund des Prozesses neu berechnet wurde (Avanzi et al. 2022).

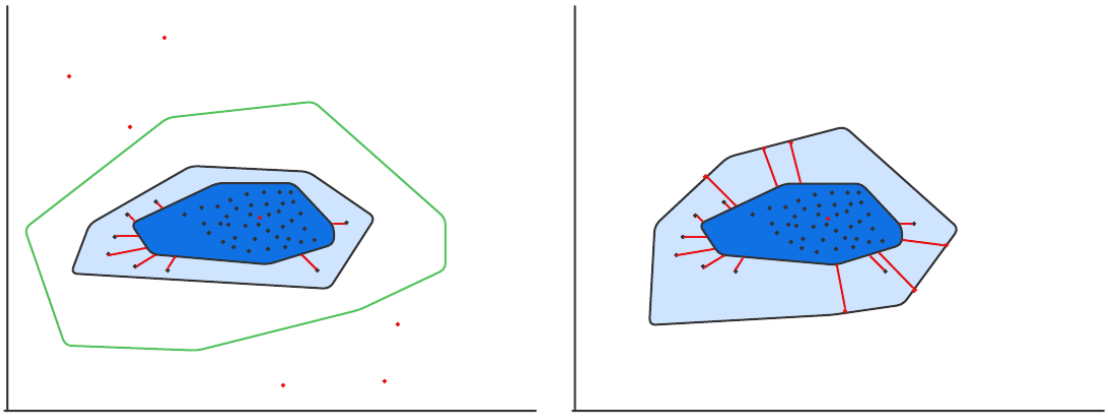

Abbildung 9: Bagplot vor der Anpassung und nach der Anpassung in Anlehnung an Avanzi et al. (2022), S.12

3.3.3 Winsorizing

Das Winsorizing schränkt den Einfluss der Ausreißer auf den Mittelwert und die Varianz des Datensatzes ein und verstärkt dabei die Abschätzung und die Variabilität des lokalen Standortes (Blaine 2018; Ghosh & Vogt 2012). Vor der Benutzung dieser Methode muss der Anwender sich für ein prozentuales Level entscheiden, welches als Grenze für die Ausreißer dienen soll (Blaine 2018). Dabei wird die Grenze für den Datensatz nach oben und unten festgelegt (Blaine 2018; Ghosh & Vogt 2012). Nachdem die Grenzen festgelegt wurden, werden Werte, die oberhalb und unterhalb der Grenze liegen, durch den Grenzwert ersetzt (Blaine 2018). Laut Blaine (2018) sind die Folgen der Ersetzung der Ausreißer, dass sich der Mittelwert und die Varianz ändern und somit auch der Einfluss von den Ausreißern. Im Folgenden wird ein Beispiel vorgezeigt mithilfe von Werten eines Datensatzes (Blaine 2018). Der gegebene Datensatz nach Blaine (2018) lautet:

(2, 2, 3, 3, 3, 4, 5, 8, 15, 25)

Der Mittelwert dieses Datensatzes lautet 7 und die Varianz 55,556. Durch die Auswahl des Winsorizing-Level von 20 %, werden bei der Perzentil Rechnung, die Zahlenwerte 2, 15 und 25 als Ausreißer erkannt und an die Werte im Bereich angepasst (Blaine 2018). Das Level von 20 % ist nach dem Autor ein gängiger Wert und der neue Datensatz nach Blaine lautet somit (2018):

(3, 3, 3, 3, 3, 4, 5, 8, 8, 8)

Durch diese Anpassung resultiert der neue Mittelwert mit 4,8 und eine Varianz von 5,3 (Blaine 2018).

3.3.4 Mahalanobis-Distance-Approach

Der Mahalanobis-Distance-Approach (MDA) wird nach Avanzi et al. (2022) für die Erkennung von Ausreißern in mehrdimensionalen Systemen benutzt. Dabei wird für jeden Datenpunkt die Distanz zum Zentrum des Datensatzes gemessen (Avanzi et al. 2022). Nach Tiwari et al. (2007) ist diese Methode auch als Behandlungsmöglichkeit zu verstehen und wird im Folgenden erklärt. Jedem Datenpunkt wird eine Gewichtung zugeordnet auf Basis ihrer Mahalanobis-Distance (Tiwari et al. 2007). Diese Gewichtung verhält sich invers bezüglich zu der Distanz, d. h. Datenpunkte mit einer größeren Distanz, bekommen eine niedrigere Gewichtung (Tiwari et al. 2007). Schließlich wird eine gewichtete Regression erzeugt, die den Einfluss der Ausreißer reduziert (Tiwari et al. 2007). Laut Tiwari et al (2007) wird die Mahalanobis-Distance nach dieser Formel berechnet:

$$D^2 = (x - \hat{x})^T * C^{-1} * (x - \hat{x}) \tag{9}$$

Wobei das D die Mahalanobis-Distance, x den originalen Datenvektor, $\hat{x}$ den geschätzten Vektor des Mittelwerts und C^{-1} die Inverse der geschätzten Kovarianz Matrix angibt (Tiwari et al. 2007).

3.3.5 Lineare Interpolation und Sigma-Approach

Als nächstes werden zwei Methoden in einem Abschnitt erklärt. Ein weiteres Verfahren zur Behandlung ist die lineare Interpolation, bei der zwei aufeinanderfolgende Datenpunkte mit einer Linie verbunden werden, um den Ausreißer zu behandeln (Wahir et al. 2018). Nach Wahir et al. (2018) wird der Ausreißer durch die Formel der linearen Interpolation abgeschätzt und lautet:

$$f(x) = b_0 + b_1(x - x_0) \tag{10}$$

In diesem Fall steht das x für die nicht abhängige Variable und das x_0 für einen bekannten Wert der nicht abgängigen Variable (Wahir et al. 2018). Das f(x) steht für den Wert der abhängigen Variable, für den Wert x der nicht abhängigen Variable (Wahir et al. 2018). Nach Wahir et al. (2018) wird die Formel der linearen Interpolation durch diese beiden Gleichungen erweitert:

$$b_0 = f(x_0) \tag{11}$$

$$b_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \tag{12}$$

Die Variablen $f(x_0)$ und x_0 sind die Ausgangspunkte und das $f(x_1)$ der Endpunkt (Wahir et al. 2018). Der Autor erwähnt eine Erweiterung der Methode, die glattere Darstellungen erstellt und Cubic-Spline-Interpolation genannt wird.

Nach Tiwari et al. (2007). ist eine weitere Methode zur Behandlung von Ausreißern der Sigma-Approach, bei dem Werte überprüft werden, die $\pm x\sigma$ vom Mittelwert abweichen. Hier ist das x

eine Zahl und Sigma (σ) steht für die Standardabweichung der Variable (Tiwari et al. 2007). Nach der Überprüfung wird der Ausreißer je nach Fall auf die Grenzwerte angepasst mithilfe der Variable y, die gleich oder größer als das x sein muss und vom Anwender abhängt (Tiwari et al. 2007).

4 Kategorisierung der Ausreißerbehandlung in der Produktion

In diesem Kapitel wird das Hauptziel dieser Arbeit thematisiert. Das Hauptziel ist die Kategorisierung der Ausreißerbehandlung. Dafür werden zuerst Ausreißer in den Daten der Produktion der Automobilindustrie vorgestellt und erklärt. Danach werden die Anforderungen an die Kriterien zur Kategorisierung gestellt, die auf Basis der Anwendungsdomäne, der Detektionsmethoden und der Behandlungsmöglichkeiten erarbeitet werden. Nach den Anforderungen werden die Kriterien zur Kategorisierung festgelegt und exemplarisch an einem Datensatz angewendet. Im Anschluss werden die Ausreißerbehandlungen bzgl. ihrer Methodik verglichen und bewertet. Schließlich folgt die Diskussion mit dem Fazit und der finalen Kategorisierung der Ausreißerbehandlungen.

4.1 Ausreißer in den Daten der Produktion

Die im vorherigen Kapitel vorgestellten Ausreißer sind in den Daten der Produktion der Automobilindustrie ebenfalls vorzufinden. In diesem Kapitel wird der Bezug zwischen den Ausreißern aus Kapitel 3 und der Daten der Produktion aus Kapitel 2 geschaffen. Wie in Abschnitt 3.1 erwähnt, können Ausreißer als Datenpunkte bezeichnet werden, die im Vergleich zum restlichen Datensatz eine Abweichung darstellen. Bei der Produktion der Automobilindustrie entstehen durch Auswertungen der Maschinen und den restlichen Fertigungsschritten, Datensätze, die Ausreißer beinhalten können. Zuerst werden die Ursachen dieser Ausreißer untersucht. Wenn die Ausreißer durch falsche Einstellungen von Parametern an den Maschinen oder falsche Erfassung entstehen, können diese meistens durch die Wiederholung des Prozesses entfernt werden. Eine weitere Methode ist das direkte Entfernen bzw. Löschen des Datenpunktes mit dem Risiko, dass die restlichen Datenpunkte nicht als Ausreißer gelten, aber dennoch von den realen Werten abweichen, die durch einen optimalen Prozess entstehen würden. Der OEE, der an den Maschinen ausgewertet wird und die Produktivität wiedergibt, kann bei der Betrachtung der Ausreißer ebenfalls relevant sein. Bei der Berechnung wird der Ausschuss von Fehlteilen in Betracht gezogen und die Produktivität der Gesamtproduktion untersucht. In dieser Untersuchung können Ausreißer auffallen, die aus Fehlern entstanden sind. Wie in Abschnitt 3.3 erwähnt, sind Ausreißer interessanter, wenn diese ohne falschen Prozess entstehen und nicht einfach entfernt werden sollten. Im Hinblick auf Datenqualität aus Abschnitt 2.3 können Ausreißer also nicht unbedingt zu den Qualitätsmängel der Produktionsdaten zugeordnet werden und zählen nicht direkt als Fehler der Produktion. Diese Ausreißer können nutzbare Informationen und nicht berücksichtigte Aspekte über den Prozess veranschaulichen. Um Datensätze dennoch nicht negativ zu beeinflussen, werden diese Ausreißer mithilfe von Behandlungsmethoden in der Produktion bearbeitet. In dieser Arbeit wurde als Datenebene in der Produktion Datenbanken und -sätze ausgewählt und darauf aufbauend hybride Detektions- und Behandlungsmethoden vorgestellt. Besonders die Behandlungsmethoden sind in der Produktion von Bedeutung, da diese den weiteren Umgang mit den Ausreißer entscheiden. Es existieren verschiedene Behandlungsmethoden, die in der Produktion

anwendbar sind, trotzdem müssen diese unterschiedlichen Methoden je nach Fall ausgewählt werden. Damit sich der Anwender zwischen den Behandlungsmöglichkeiten entscheiden kann, werden Kriterien mithilfe von Anforderungen definiert, die bei der Entscheidung der Methode unterstützen sollen.

4.2 Anforderungen an die Kriterien zur Kategorisierung

Die Behandlungsmethoden der Ausreißer sollen im Hinblick auf die Produktion der Automobilindustrie kategorisiert werden. Damit diese Kategorisierung erfolgen kann, werden Anforderungen an die Kriterien gestellt, damit diese eine passende Kategorisierung der Ausreißerbehandlungen ermöglichen. In diesem Kapitel werden die Anforderungen an die Kriterien vorgestellt und diskutiert. Die Anforderungen sind abhängig von drei Faktoren. Der erste Faktor ist die Anwendungsdomäne, die in diesem Fall als Produktion in der Automobilindustrie ausgewählt wurde.
Es resultieren Daten aus der Produktion, die relevant für die weiteren Prozesse sind. Dabei müssen verschiedene Eigenschaften der Daten berücksichtigt werden. Eine Anforderung an die Kriterien ist das Erkennen der resultierenden Datentypen und das Unterscheiden dieser untereinander. Der Vergleich zwischen Datentypen muss möglich sein, um Ausreißer zu erkennen. Das bedeutet, dass Datensätze, die aus Integer bestehen, nicht mit Strings verglichen werden können. Außerdem ist die Datenstruktur eine weitere Anforderung an die Kriterien. Wie in Abschnitt 2.2 bezüglich Big-Data schon erwähnt wurde, können Daten aus drei verschieden strukturierten Formaten sein. Die gewonnene Informationen aus den Datenstrukturen müssen im Hinblick auf die Datentypen vergleichbar sein. Das bedeutet, dass Daten aus Tabellen, mit Daten aus Graphen verglichen werden können, wenn der Datentyp es ermöglicht. Also ist die Datenintegration der Daten aus den verschiedenen Strukturen eine weitere Anforderung. Weiterhin sind die Kriterien abhängig von der gewählten Datenebene. Im Rahmen dieser Arbeit müssen also vor allem Datensätze erfasst und Unterschiede zu größeren oder kleineren Datenebene erkannt werden. Die Größe der Datenebene hat einen signifikanten Einfluss. Die Datenmengen können im Vergleich zu Big-Data viel kleiner sein und können deshalb mit manuellen und hybriden Verfahren bearbeitet werden. Der zweite Faktor, von dem die Anforderungen abhängen, sind die Detektionsmethoden. Damit die Kriterien an die Ausreißerbehandlungen angewendet werden können, ist eine Detektion von richtigen Datenpunkten wichtig. Mit richtigen Datenpunkten werden hier wahre Ausreißer bezeichnet, die in der Detektion erkannt werden müssen. Wenn Datenpunkte als Ausreißer erkannt werden, die nicht wahre Ausreißer sind, können diese nach der Behandlung den Datensatz im Negativen beeinflussen. Außerdem folgt bei der Anwendung der Detektionsmethoden, dass alle Datenpunkte, also auch nicht auffällige Datenpunkte, untersucht werden. Also resultiert als eine Anforderung, die Auswahl der passenden Detektionsmethode. Die passende Detektionsmethode soll auf den Datensatz anwendbar sein und somit die Bedingungen des Datensatzes erfüllen. Das bedeutet, dass diese Anforderung mit den Anforderungen aus dem ersten Faktor verbunden ist. Im Hinblick auf Abschnitt 2.1 heißt das z.B., dass sich die Detektionsmethode der Datenstruktur anpassen muss. Der dritte Faktor sind die Behandlungsmethoden. In diesem Fall existiert die Anforderung, dass Ausreißer immer noch berücksichtigt und nicht indifferent entfernt werden. Als indifferentes Vorgehen wird hier, das unberücksichtigte Entfernen der Ausreißer bezeichnet. Im

indifferenten Vorgehen werden die Ausreißer aus den Datensätzen gelöscht, ohne dass die restlichen Datenpunkte verändert werden. Somit wird jeglicher Einfluss vollständig entfernt und kein Nutzen dieser Ausreißer identifiziert. Dieses Vorgehen ist in der Produktion nicht von Vorteil, da wichtige Informationen des Prozesses übersprungen werden können. Die Ausreißer können nicht nur als Fehler angesehen werden, die dann durch die indifferente Entfernen als behandelt gelten. Sie können auch aus dem optimalen Prozess entstehen und müssen dann genauer untersucht werden. Wie in Abschnitt 3.3 von Aguinis et al. (2013) erwähnt wurde, sind Ausreißer interessanter, wenn diese nicht indifferent entfernt und mithilfe der Behandlungsmethoden bearbeitetet werden In erster Linie ist hier der richtige Umgang mit den detektierten Ausreißern relevant, damit diese im weiteren Verlauf für die Produktion einen Nutzen haben. Der richtige Umgang wird durch die Auswahl der passenden Behandlungsmethode unterstützt. Als Anforderung resultiert hieraus, dass eine passende Behandlungsmethode für die interessanten Ausreißer ausgewählt werden muss. Im Folgenden werden die Anforderungen nochmal kompakt aufgezählt:

1. Datentyp
2. Datenstruktur und deren Integration
3. Datenmenge
4. Identifizierung wahrer Ausreißer durch Auswahl der passenden Detektionsmethode
5. Behandlung interessanter Ausreißer durch Auswahl der passenden Behandlungsmethode

Zusammengefasst basieren die Anforderungen an die Kriterien auf der Einwirkung der drei vorgestellten Faktoren. Alle drei Faktoren sollen gleichzeitig berücksichtigt werden und Kriterien schaffen, die eine Kategorisierung ermöglichen.

4.3 Ableitung der Kriterien zur Kategorisierung

In diesem Abschnitt werden die Kriterien zur Kategorisierung basierend auf die Anforderungen abgeleitet und vorgestellt. Mithilfe der Kriterien sollen die unterschiedlichen Behandlungsmethoden kategorisiert werden. Damit die Ableitung der Kriterien deutlich wird, werden in der Folgenden Ableitung 10 die Kriterien mit ihren Anforderungen verknüpft.

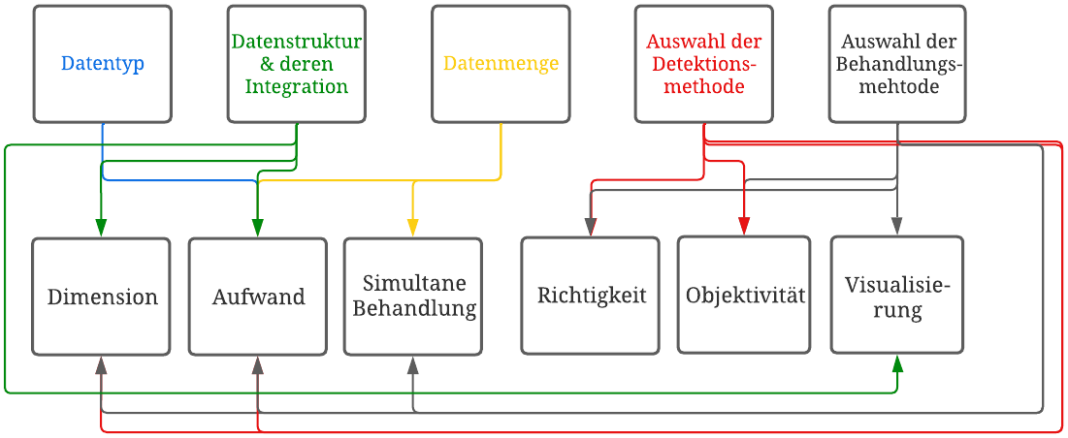

Abbildung 10: Die Ableitung der Kriterien (eigene Darstellung)

Die fünf Anforderungen aus dem vorherigen Abschnitt wurden oben in der Abbildung farblich unterteilt, damit die Unterscheidung der Einflüsse deutlicher wird. Unten in der Abbildung sind die sechs Kriterien zur Kategorisierung dargestellt. Die Kriterien werden in der Reihe von links startend erklärt. Das erste Kriterium ist die Dimension, in der sich der Datensatz in der jeweiligen Behandlungsmethode befindet. Jede vorgestellte Behandlungsmethode hat eine Dimension, in der die Ausreißer behandelt werden. In dieser Arbeit wurden für die Behandlungsmethoden ein- und zweidimensionale Datensätze thematisiert, dennoch sind durch Modifizierungen, Erweiterungen bezüglich der Dimension möglich. Diese Modifizierungen werden in Abschnitt 4.5 thematisiert. In der jeweiligen Methode muss unterschieden werden, ob die Dimension eingeschränkt ist oder ob sie erweitert werden kann. Die Dimensionen hängen stets von dem Datensatz ab, der in der Produktion erzeugt wird. Um genauer zu definieren, hängt die Dimension von der Datenstruktur und deren Integration ab. Datenstrukturen haben selbst eine Dimension, in der die Daten gespeichert sind. Die Daten bzw. Informationen aus der Struktur müssen in das gewollte Format integriert werden, damit die Ausreißer in der Behandlungs- und Detektionsmethode untersucht werden können. Neben den Anforderungen der Datenstrukturen resultiert wie schon erwähnt das Kriterium der Dimension auch aus den Anforderungen der Behandlungs- und Detektionsmethoden. Die Dimension des Datensatzes, der Detektions- und Behandlungsmethoden müssen passend sein, d. h. die Auswahl der Methoden hängt von der Dimension des Datensatzes ab. Das zweite Kriterium ist der Aufwand, der während der Behandlung entsteht. Je nach Behandlungsmethoden werden verschiedene Berechnungen ermittelt, die entweder vom Anwender oder durch die technischen Hilfsmittel durchgeführt werden. In diesem Kriterium werden erstens die Anzahl der Berechnungen und zweitens die Komplexität der jeweiligen Formel mitberücksichtigt. In Abbildung 10 ist zu erkennen, dass das Kriterium des Aufwands von allen Anforderungen aus dem vorherigen Kapitel abgeleitet wurde. Der Datentyp bestimmt im mathematischen Sinne, die Komplexität der Werte bzw. Zahlen und somit auch die Komplexität der Berechnungen. Die Datenstruktur bestimmt die Dimension, die in diesem Fall auch einen Effekt auf den Aufwand der Berechnungen hat. Je höher die Dimension des Datensatzes ist, desto mehr Rechenschritte sind notwendig. Die dritte Anforderung, die Datenmenge gibt in diesem Fall die Anzahl der Datenpunkte an. Aus dieser Anforderung in der Kombination mit den Anforderungen der Auswahl der Detektions- und

Behandlungsmethoden folgen die Hauptfaktoren des Aufwands. Je nach Methode sind die Berechnungen unterschiedlich. Unter der Berücksichtigung der Datenmenge folgt, dass mehr Datenpunkte analysiert werden und somit der Aufwand steigt. Das dritte Kriterium bezieht sich auf die Anzahl der behandelbaren Ausreißer und somit im Bezug zu Abschnitt 3.1 auf den Ausreißertyp. Das Kriterium thematisiert, ob mit der Behandlungsmethode nur Punktausreißer oder auch kollektive Ausreißer behandelt werden können. In diesem Kriterium werden kontextuale Ausreißer nicht bezüglich ihres Kontextes geprüft, deshalb sind die anderen beiden Typen im Fokus. Dieses Kriterium wird als Simultane Behandlung definiert. Wie in Abbildung 10 veranschaulicht wurde, hängt dieses Kriterium von zwei Anforderungen ab. Die erste Anforderung ist die Datenmenge, da diese durch die Anzahl der Datenpunkte, auch das potenzielle Auftreten der Ausreißer beeinflusst. Neben der Datenmenge ist die Auswahl der Behandlungsmethode ebenfalls eine Anforderung an das Kriterium. Wenn mehrere Ausreißer innerhalb eines Datensatzes auftreten, kann dies die Wahl der Behandlungsmethode beeinflussen, da je nach Fall mehrere Ausreißer behandelt werden sollen. Das vierte Kriterium setzt den Fokus auf die Detektion. Innerhalb der Behandlungsmethoden tritt eine Detektion der Ausreißer ein, in der die Ausreißer gekennzeichnet werden. Dabei ist es wichtig, dass Datenpunkte als Ausreißer behandelt werden, wenn sie durch die Detektion als wahre Ausreißer erkannt wurden. Dieses Kriterium wird als Richtigkeit definiert und soll bewerten, ob die Behandlungsmethoden wahre Ausreißer untersuchen. Die Auswahl der Detektions- und Behandlungsmethode ist die Anforderung aus denen die Richtigkeit abgeleitet wurde. Das fünfte Kriterium wird als Objektivität bezeichnet und beschreibt den Einfluss des Anwenders auf die Behandlung. Dabei muss ergänzt werden, dass bei den hybriden Behandlungsmethoden, die in dieser Arbeit im Fokus stehen, immer ein Anwender mitwirkt. Eine Folge hiervon ist, dass je nach ausgewählte Methode der Anwender einen Einfluss auf die Ausreißerbehandlung und somit auch auf den gesamten Datensatz haben kann. Der Einfluss auf den gesamten Datensatz kann Analysen der Produktion verfälschen, da in bestimmten Behandlungsmethoden der Wert der Ausreißer verändert wird. Ein Beispiel hierfür ist die Bagplot-Based-Adjustment, in der Ausreißer verschoben werden und somit der Datensatz sich ändert. In diesem Kriterium wird dann geprüft, ob das Einwirken subjektiv oder objektiv ist. Die Objektivität hängt wie das Kriterium davor von der Auswahl der Detektions- und Behandlungsmethode ab. Je nach Methode variiert der Einfluss des Anwenders und somit auch das subjektive Einwirken. Das letzte und sechste Kriterium ist die Visualisierung, in der überprüft wird, wie die Ausreißer grafisch untersucht und behandelt werden. Dieses Kriterium schließt nicht das zweite Kriterium aus, bei dem der mathematische Aufwand bewertet wird. Gleichzeitig zum mathematischen Aufwand wird hier überprüft, ob z. B. der Ausreißer mithilfe von einer Grafik angepasst oder durch Bereiche gekennzeichnet wird. Die Auswahl der Behandlungsmethode ist einer der Faktoren, die bei diesem Kriterium von Bedeutung sind. Nicht in jeder Methode wird eine Ausreißerbehandlung mit einer Grafik angepasst und als Notwendigkeit betrachtet. Dennoch existieren Methoden wie z. B. der LTS, in der eine Regressionskurve gegen die Ausreißer erstellt wird. Außerdem ist eine weitere Anforderung, die Datenstruktur. Die Datenstruktur gibt an, wie die Daten in dem jeweiligen Format vorliegen. Das Format kann z. B. ein Graph mit den Datenpunkten sein. Das bedeutet, dass die Visualisierung des Datensatzes als Grundlage schon vorhanden wäre. Dann kann je nach Be-

handlungsmethode eine grafische Analyse erfolgen. Die Datenstruktur kann aber auch eine Tabelle sein, bei der dann überprüft werden muss, ob der Datensatz visualisiert werden kann und ob es sinnvoll ist. In der Tabelle 5 werden die Kriterien nochmal kompakt dargestellt.

Tabelle 5: Kriterien zur Kategorisierung

Kriterium
Dimension
Aufwand
Simultane Behandlung
Richtigkeit
Objektivität
Visualisierung

4.4 Exemplarische Anwendung der Kriterien

Nachdem im letzten Kapitel die Kriterien zur Kategorisierung definiert und vorgestellt wurden, werden diese Kriterien jetzt an einer Behandlungsmethode angewendet. Als Behandlungsmethode wurde für diesen Fall, das Winsorizing aus Kapitel 3.3.3 ausgewählt, da ein eindimensionaler Datensatz vorliegt, der mithilfe dieser Methode bearbeitet werden soll. Die Kriterien werden in der Reihenfolge aus Tabelle 5 an dieser Methode angewendet. Das Grundprinzip der Methode ist im Eindimensionalen und ist deshalb mit weniger Aufwand verbunden. Der Datensatz für die exemplarische Anwendung hat die Bezeichnung „Parts Manufacturing – Industry Dataset“ und wurde von Kaggle entnommen (Kaggle 2022). In dem Datensatz werden Messungen von Teilen und deren Produzent angegeben. Damit das Winsorizing angewendet werden kann, wird der Datensatz nur auf die Werte der Längen untersucht. Außerdem werden für die Übersicht des Datensatzes nur die ersten zehn Werte behandelt. Der neue Datensatz wird in Tabelle 6 dargestellt:

Tabelle 6: Länge der Teile aus dem Datensatz "Parts Manufacturing - Industry Dataset

Länge der Teile
102,67
102,5
95,37
94,77
104,26
105,18
97,35
99,35
90,62
97,22

Der Mittelwert in diesem ausgewählten Datensatz beträgt 98,929 und die Varianz 22,0929. Wie eben erwähnt, liegt hier ein eindimensionaler Datensatz vor, der als Folge hat, dass die Komplexität der Berechnungen niedrig ist. Außerdem wird das Kriterium des Aufwands in diesem Fall

durch die Stichprobe verkleinert. Der originale Datensatz mit 500 Datenpunkten hat einen höheren Aufwand, der durch die technische Unterstützung im hybriden Fall zu bewältigen ist. Die Simultane Behandlung von Ausreißern ist in dieser Methode gegeben, der durch den Datensatz und dem Anwender limitiert wird. Das nächste Kriterium ist die Richtigkeit. Bei der Richtigkeit wird geprüft, ob wahre Ausreißer behandelt werden. In dieser Methode ist die Definition von wahren Ausreißern komplizierter. Ausreißer werden hier durch die Auswahl des Winsorizing-Level bestimmt. Das bedeutet, ob Datenpunkte als Ausreißer gelten, hängt indirekt vom Anwender der Methode ab. Dieser Fall leitet das nächste Kriterium ein, die Objektivität. Es wird in dieser Methode ein Level ausgewählt, welches die Berechnung der Perzentile beeinflusst und somit auch die Behandlung der Ausreißer. Dieses Level entscheidet darüber, welche Werte als Ausreißer gelten und welche nicht. Der Anwender hat bei der Auswahl des Winsorizing-Level eine freie Entscheidung, sodass Datenpunkte entfernt bzw. ersetzt werden, die subjektiv als irrelevant oder fehlerhaft angesehen werden. Um das Beispiel weiter zu bearbeiten, wird hier wie in Abschnitt 3.3.3 genannt, der gängige Wert von 20 % für das Winsorizing-Level gewählt. Diese Entscheidung wirkt zuerst als subjektiv, doch wenn in der Produktion durchgehend ein gängiger Wert für das Winsorizing benutzt wird, dann könnte diese Entscheidung als objektiv bezeichnet werden. Mit dem Winsorizing-Level von 20 % folgen die Ausreißer des Datensatzes. Die Ausreißer sind die zwei niedrigsten (90,62 & 94,77) und höchsten Werte (104,26 & 105,18). Die zwei niedrigsten Werte werden mit 95,37 und die zwei höchsten Werte mit 102,67 ersetzt. Es resultiert somit ein neuer Datensatz, der einen Mittelwert von 99,054 und eine Varianz von 10,908 nachweist. Dabei fällt auf, dass der Mittelwert sich nicht viel verändert hat und die Varianz kleiner wurde. In der vorliegenden Stichprobe des Datensatzes existierten keine Werte, die offensichtlich als Ausreißer bezeichnet werden können. Durch das Winsorizing-Level wurden Daten behandelt, die in anderen Methoden nicht als wahre Ausreißer gelten würden. Wie erwähnt wurde das Kriterium der Richtigkeit in diesem Fall durch die Wahl des Winsorizing-Level beeinflusst. Die Visualisierung ist in dieser Methode nicht gegeben. Der Datensatz wurde mit einer Tabelle dargestellt und durch die Perzentil Rechnung bearbeitet. Mithilfe eines eindimensionalen Datensatzes, kann aber ein Boxplot nachgebaut werden. Wie in Abschnitt 3.2.1 erklärt wurde, kann dieser Boxplot in dieser Behandlung auch als Detektionsmethode benutzt werden. Das Nachbauen der Bestandteile ist mithilfe der Daten aus Tabelle 6 möglich.

4.5 Methodenbasierter Vergleich und Bewertung der Ausreißerbehandlungen

Nachdem im letzten Kapitel Kriterien an dem Winsorizing angewendet wurden, werden jetzt die restlichen Behandlungsmethoden mithilfe der Kriterien verglichen und bewertet. Dafür wird eine Tabelle erstellt, in der die Behandlungsmethoden und die Kriterien aufgelistet werden. Die Tabelle 7 auf Seite 35 soll als Übersicht dienen und den direkten Vergleich der Behandlungsmethoden darstellen. Als erstes werden die Dimensionen der einzelnen Methoden thematisiert. In der Tabelle 7 sind die Dimensionen aus Kapitel 3.3 zu erkennen. Nicht alle Methoden sind fest an die Dimensionen gebunden und können durch Anpassungen verändert werden. Das BBA und MDA sind z. B. zweidimensionale Methoden, die ebenfalls im dreidimensionalen funktionieren würden. Außerdem ist eine Erweiterung vereinfachter als bei den anderen Methoden. Das LTS, der Sigma-

Approach und die lineare Interpolation sind auch zweidimensional, deren Erweiterung in höhere Dimensionen etwas komplizierter ist im Vergleich zu den anderen Methoden. Die Komplexität bei der Erweiterung wirkt sich auf den Aufwand aus, da höhere Dimensionen die gegebenen Formeln ausweiten und zu mehr Berechnungen führen. Der Sigma-Approach, der in der bekannten Normalverteilung benutzt werden kann, würde bei der Erweiterung der Dimension, die restlichen Kriterien beeinflussen. Das LTS müsste in der Erweiterung die Rückstände in höheren Dimensionen berechnen, und dann die Regression erzeugen. Die Lineare Interpolation hat wie Wahir et al. (2018) erwähnt hat, die Cubic-Spline-Interpolation als Erweiterung, die die Komplexität der Methode erhöht. Das Winsorizing ist eindimensional und die Erweiterung ist wie in den anderen Methoden nicht in anderen Dimensionen vorgestellt worden. Die Erweiterung des Winsorizing ist ebenfalls komplizierter und hat vom Prinzip eine Ähnlichkeit zum BBA und MDA. Das Winsorizing im zweidimensionalen würde wie im BBA Ausreißer erkennen und mit dem bestimmten Grenzwert ersetzen. Außerdem werden in allen drei Methoden Zentren erfasst bzw. im Winsorizing der Mittelwert und darauf aufbauend abweichende Werte untersucht. Deshalb wird hier angenommen, dass das Winsorizing eindimensional ist und für höhere Dimensionen die anderen beiden Methoden benutzt werden können. Durch diese Annahme folgt, dass das Winsorizing durch die Einschränkung in diesem Kriterium am schlechtesten ist. Die Methoden des BBA und MDA sind bezüglich dieses Kriteriums am besten, da die Erweiterungen in höhere Dimensionen den Aufwand am wenigsten beeinflussen. Außerdem können diese Methoden in das Eindimensionale transformiert werden, da die Methoden ihre Grundideen nicht durch die Abstufung der Dimensionen verlieren. Wenn das Kriterium der Dimension im Bezug zu der Anwendungsdomäne untersucht wird, existiert ein wesentlicher Aspekt, und zwar die Datenstruktur. Neben der Untersuchung, ob die Behandlungsmethode an ihre Dimension gebunden ist, ist das Erfassen der Dimensionen der Datensätze auch relevant. Je nach erfassten Datensatz aus der Produktion muss eine Behandlungsmethode ausgewählt werden, die der Dimension des Datensatzes entspricht. Also müssen die Anwender in der Produktion entweder eine Methode wählen, die in der Dimension des Datensatzes benutzt werden kann oder durch Anpassungen, die Dimension der Methode umwandeln. Bei der Umwandlung muss überprüft werden, ob sich der Aufwand rentiert oder ob eine andere Methode benutzt werden sollte. Diese Entscheidung hängt in der Produktion von mehreren Faktoren ab, wie z. B. vom Wettbewerbsdruck, der zeitnahe Lösungen verlangt. Als Schlussfolgerung des Vergleichs für das erste Kriterium folgt, dass die Dimension stets den Aufwand der Methode beeinflusst. Höhere Dimensionen erhöhen immer den Aufwand. Dies hat als Folge, dass das Winsorizing im Eindimensionalen einen niedrigen Aufwand hat und somit im Kriterium des Aufwands im Vergleich zu den anderen Methoden am besten ist. Das Zweite Kriterium, der Aufwand, ist im Kontext der hybriden Methoden schwierig zu vergleichen, wenn die Dimensionen der Methoden identisch sind. Bei Hybriden Methoden werden jegliche Berechnungen in der Produktion, die in diesem Fall ebenfalls den Aufwand beeinflussen, mit technischen Hilfsmitteln durchgeführt. Deshalb wird der Aufwand mit dem Kriterium der Simultanen Behandlung verknüpft und dann bewertet. Nachdem das Winsorizing aufgrund der Dimension als die Methode mit dem wenigsten Aufwand bezeichnet wurde, folgt als nächste Methode mit geringem Aufwand, die Lineare Interpolation, da in dieser Methode jeweils nur ein Ausreißer behandelt wird. Die Formeln, die Wahir et al. (2018) definiert hat, können durch Umstellungen in eine Formel umgewandelt werden und dadurch mit einer Rechnung den Ausreißer ermitteln. Als

nächstes folgt der Sigma-Approach, bei dem der Aufwand erstens von der Berechnung der Standardabweichung und zweitens von den Ausreißern abhängig ist. Die Anzahl der behandelbaren Ausreißer wird vom Anwender im Hinblick auf die Verteilung der Datenpunkte bestimmt. Die Berechnungen in dieser Methode sind ähnlich zum Winsorizing. Es werden grundlegende statistische Werte ermittelt, wie die Standardabweichung und der Mittelwert. Im LTS und BBA können bis zu 50 % des Datensatzes aus Ausreißern bestehen. Die Menge der Datenpunkte und der inhärenten Ausreißer beeinflussen ebenfalls den Aufwand. Je mehr Ausreißer behandelt werden, desto mehr Berechnungen müssen ausgeführt werden. Das bedeutet, dass im LTS und BBA ein höherer Aufwand vorhanden ist, da mehrere Punkte berechnet werden können. Dazu sind die Berechnungen und Formeln komplexer als die von den vorherigen Verfahren. Als letztes Verfahren wird der MDA überprüft. In dieser Methode basiert der Aufwand auf den Berechnungen der Distanzen von den einzelnen Punkten und der Zuordnung der Gewichtungen. Wie in den anderen Methoden hängt der Aufwand auch von der Simultanen Behandlung ab. Im MDA werden nach Avanzi et al. (2022) und Tiwari et al. (2007) alle Datenpunkte untersucht und die Distanz zum Zentrum ermittelt. Beide Quellen thematisieren ein Zentrum in dieser Methode, aber definieren das Zentrum nicht genau. Deshalb wird in dieser Arbeit für diese Methode ein mögliches Zentrum definiert und die Folgen davon erklärt. Für das Zentrum wird sich am BBA orientiert. In dieser Methode werden 50 % der Datenpunkte als Zentrum angenommen, daraus resultiert im BBA, dass der restliche Datensatz aus Ausreißern bestehen kann. Wird diese Definition des Zentrums auch im MDA angenommen, dann existieren bis zur 50 % Ausreißer, die aber nicht den Aufwand der Methode beeinflussen, da für jeden Datenpunkt die Distanz ermittelt wird. Also werden auch Datenpunkte ausgewertet, die mithilfe einer Detektionsmethode nicht als Ausreißer erkannt werden. Also ist der Aufwand in dieser Methode im Vergleich zu den anderen Methoden im Kontext mit der Simultanen Behandlung am größten, da jeder Datenpunkt gemessen wird. Dies hat zur Folge, dass in dieser Methode Ausreißer nicht gesondert behandelt werden, also ist keine strikte Trennung vorhanden. Nachdem erwähnt wurde, dass in der Produktion der Automobilindustrie Behandlungsmethoden aufgrund ihrer Dimension und des daraus entstehenden Aufwands ausgewählt werden müssen, folgt dass der Aufwand der Methode auch von der Simultanen Behandlung abhängig ist. Das bedeutet, dass mehr Ausreißer in der Produktion den Aufwand erhöhen können. Mit der Beachtung der vorhandenen Ausreißer und durch die passende Auswahl der Behandlungsmethode kann der Aufwand variieren. Anwender sollten je nach Anzahl der zu behandelnden Ausreißer und unter Berücksichtigung der Dimension des Datensatzes eine passende Behandlungsmethode auswählen. Als letztes wurde im MDA behauptet, dass keine strikte Trennung zwischen Ausreißer und erwarteten Werten existiert. Dies leitet das Kriterium der Richtigkeit ein, bei dem überprüft wird, ob wahre Ausreißer behandelt werden. Im MDA hat die Richtigkeit im Vergleich zu den anderen Methoden keine diskrete Antwort. Alle Datenpunkte werden gemessen und bekommen abhängig davon eine Gewichtung. Der Übergang von den Ausreißern zu den erwarteten Werten kann als stetig angesehen werden, der durch eine stetige Gewichtung widergespiegelt wird. Also können in der Produktion Ausreißer im Datensatz unverändert mitwirken und werden weiterhin berücksichtigt. Um aber herauszufinden, welche Werte wahre Ausreißer sind, muss eine Detektionsmethode benutzt werden. Im BBA werden Punkte, die außerhalb des Fence sind als Ausreißer erkannt. Also trennt die Methode erwartete Werte von den Ausreißern und benötigt in der Behandlung keine Detektionsmethode. In dieser Methode werden die Ausreißer

an den Fence verschoben und erhalten deshalb einen neuen Wert. Die Veränderung der Ausreißer führt zu neuen Berechnungen der Bestandteile des Bagplot. Wie in Abschnitt 3.3.2 erwähnt, können dadurch neue Ausreißer entstehen. Somit könnte diese Methode keine vollständige Ausreißerbehandlung garantieren, die dazu führt, dass die Anwender in der Produktion eine weitere Behandlung durchführen müssen. Hier entsteht die Frage, ob die Behandlung mithilfe einer anderen Methode geschehen sollte, oder ob wieder das BBA benutzt wird. Wenn eine vollständige Ausreißerbehandlung erwünscht ist, könnte das BBA eine endlose Behandlung einleiten und somit keine richtige Lösung darbieten. Schon nach der ersten Behandlung ist der Datensatz manipuliert und bei wiederholender Behandlung wird der Datensatz sich immer weiter von dem ursprünglichen Datensatz entfernen. Daraus folgt, dass diese Methode in der Produktion einen weiteren Aufwand mitbringen könnte, der schließlich nicht zu der Lösung der Probleme führt. In der linearen Interpolation wird der Ausreißer durch die Verbindung der zwei aufeinander folgenden Datenpunkte abgeschätzt. Die Benutzung einer Detektionsmethode könnte in dieser Methode sinnvoll sein und muss die Erkennung eines Punktausreißers ermöglichen. Im LTS werden Ausreißer durch das lineare Modell erkannt, da die Ausreißer nicht von der Regression beachtet werden. Im Sigma-Approach werden wahre Ausreißer mithilfe der Standardabweichung bestimmt. Der Anwender bestimmt den Wert der Variable x und somit auch die Ausreißer. Wie in 4.4 schon besprochen wurde, hängen die wahren Ausreißer im Winsorizing vom Anwender ab, der das Winsorizing-Level in der Methode festlegt. Alle Behandlungsmethoden wurden jetzt bezüglich des Kriteriums der Richtigkeit überprüft. Der MDA war die einzige Methode, die keine integrierte Detektion bzw. Festlegung von Ausreißern hat. Eine Bewertung bezüglich der Richtigkeit ist komplex, da die Bezeichnung eines Ausreißers je nach Behandlungs- und Detektionsmethode variiert. Es fällt auf, dass in den Methoden der Anwender bei der Richtigkeit einen Einfluss hat. Das LTS und die lineare Interpolation sind die einzigen Methoden, die diesbezüglich nicht vom Anwender betroffen sind. Damit die Richtigkeit bewertet werden kann, wird das Kriterium der Objektivität miteinbezogen. Bei der Untersuchung dieses Kriteriums werden die lineare Interpolation und das LTS nicht berücksichtigt, da kein subjektiver Einfluss des Anwenders zu erkennen ist. Im BBA wurden die Werte außerhalb des Fence als Ausreißer bezeichnet, dabei hat der Anwender einen besonderen Einfluss. Der Anwender bestimmt vor der Benutzung der Methode einen Faktor für den Fence, der abhängig vom Bag erzeugt wird. Dieser Faktor bestimmt die Größe des Fence und somit auch die Werte, die als Ausreißer angesehen werden können. Nach Rousseeuw et al. (1999) wird für den Fence ein gängiger Faktor ausgewählt. Mit dem gängigen Faktor kann das BBA noch objektiv gelten, wenn dieser in der Produktion fast immer gewählt wird. Wenn in der Produktion der Faktor je nach Anwender variiert, dann kann die Behandlung der Ausreißer subjektiv sein. Der Anwender im BBA wirkt nicht nur vor der Behandlung, sondern auch danach. Wie eben erwähnt, können nach der Behandlung wieder Ausreißer entstehen, da die Bestandteile des Bagplot neu kalkuliert werden. Der Anwender muss in diesem Fall entschieden, ob der Prozess der Behandlung wiederholt werden soll. Im MDA wirkt der Anwender bei der Gewichtung der Datenpunkte. Die Gewichtung muss bezüglich der Distanz invers sein. Das bedeutet, der Anwender muss eine stetige Reduzierung der Gewichtung mit steigender Distanz sichern. Wenn die Gewichtung auf Basis der Distanz erzeugt wird, dann kann der Einfluss des Anwenders objektiv bezeichnet werden. Aber vor der Zuordnung der Gewichtung benötigt der

Anwender einen Richtwert bzw. einen Startwert. In Abschnitt 4.4 wurde das Winsorizing als subjektiv beeinflussbares Verfahren bezeichnet, da der Anwender indirekt durch das Winsorizing-Level die Ausreißer bestimmt. Hier entsteht die Frage, ob der Einfluss des Anwenders als objektiv bezeichnet werden kann, wenn in der Produktion durchgehend ein gängiges Level gewählt wird. Daraus resultiert die weitere Frage, ob es Sinn ergibt, immer einen gängigen Faktoren zu benutzen. Wenn durchgehend ein gängiger Faktor benutzt wird, dann wäre die Methode als objektiv anzusehen, da die subjektive Entscheidung des Anwenders entfällt und er nicht mehr je nach Datensatz die Ausreißer bestimmen kann. Aber immer den gleichen Faktor zu benutzen, würde bei bestimmten Datensätzen nicht wirklich das Problem lösen. Das bedeutet, dass der Faktor, der gewählt wird, abhängig vom Datensatz sein kann. Zum Beispiel in Abschnitt 4.4 war ein Datensatz vorhanden, der keine auffälligen Ausreißer hatte. Durch den gängigen Faktor wurde der Datensatz behandelt, ohne dass es eine große Veränderung bei dem Mittelwert gab. Bei dem Winsorizing lohnt sich daher eine Detektion mithilfe einer separaten Detektionsmethode, mit der im Anschluss entschieden werden kann, ob eine Behandlung notwendig ist. Wahre Ausreißer können danach mithilfe des gängigen Faktors richtig behandelt werden. Die letzte Methode, die bezüglich ihrer Objektivität überprüft werden muss, ist der Sigma-Approach. Nach Tiwari et al. 2007 werden in der Methode Werte die x-mal von σ entfernt sind als Ausreißer erkannt. Dieses x kann sich je nach Datensatz unterscheiden und hängt von der Verteilung ab. Der Anwender bestimmt dann mit der Variable x welche Werte als Ausreißer gelten. Darüber hinaus beeinflusst der Anwender mit der Variable y die Behandlung der Ausreißer. Also wirkt der Anwender bei der Detektion und bei der Behandlung. Bei der Betrachtung des Kriteriums wirkt der Eingriff des Anwenders subjektiv, da er mit der Auswahl der Variablen den Prozess beeinflusst. Alle Behandlungsmethoden wurden bezüglich ihrer Objektivität überprüft. In vier dieser Methoden hat der Anwender einen Einfluss. Bis auf das MDA können die restlichen Methoden, wo der Anwender einen Einfluss hat, durch den Anwender subjektiv beeinflusst werden. Besonders das BBA wirkt im Vergleich aufgrund der Prozesswiederholungen als eine schlechte Auswahl. Der Anwender in der Produktion muss sich bewusst sein, dass in dieser Methode neben der Faktorauswahl, die die Richtigkeit bestimmt, auch durch die Prozesswiederholungen den Datensatz enorm beeinflussen kann. Durch eine geregelte Zuordnung der Gewichtung, die invers zu der Distanz bestimmt wird, kann das Mitwirken des Anwenders als objektiv betrachtet werden. Deshalb erfüllt das MDA neben den Methoden, in denen der Anwender keinen Einfluss hat, dieses Kriterium. Als letztes Kriterium wird die Visualisierung überprüft. Das Winsorizing und der Sigma-Approach wurden bei den benutzten Quellen (s. Tabelle 7) ohne eine Visualisierung erklärt. Dennoch existiert die Möglichkeit, das Winsorizing durch einen Boxplot und den Sigma-Approach durch eine Gaußsche Kurve darzustellen. Dabei muss erwähnt werden, dass die Darstellung in diesen Methoden keinen Effekt auf die Behandlung hat, sondern nur den jeweiligen Datensatz veranschaulichen würde. Außerdem muss bei der Darstellung des Sigma-Approach in der Gaußsche Kurve beachtet werden, dass dann der Datensatz normalverteilt sein muss. Im LTS ist die Visualisierung wesentlich, da die Ausreißer mit der Erzeugung der Regressionskurve behandelt werden. Diese Behandlung wurde von Rousseeuw & Leroy (1978) als Stabilität bezeichnet, da die Kurve den Ausreißer ignoriert. Im BBA ist die Visualisierung im ersten bis zum letzten Schritt vorhanden. In der Methode werden zuerst die Bestandteile erzeugt, die durch den Datensatz entstehen. Dann werden dadurch

Ausreißer identifiziert und mithilfe der Bestandteile behandelt. Am Ende kriegt man neu kalkulierte Bestandteile, die im optimal Fall alle Ausreißer behandelt haben. Im MDA kann der zweidimensionale Datensatz visualisiert werden sowie die gewichtete Regression. Dabei ist die gewichtete Regression wesentlicher, da diese den Einfluss der Ausreißer reduziert und somit den essenziellen Teil der Behandlung darstellt. Die letzte Methode ist die lineare Interpolation, in der die Visualisierung durch die Verbindung der zwei aufeinanderfolgenden Punkten geschieht. Dieser Schritt wird vor der Behandlung bzw. Berechnung getätigt. Die zwei aufeinanderfolgenden Punkte sind die Voraussetzung für die folgende Behandlung.
Nachdem alle Methoden bezüglich alle Kriterien überprüft wurden, folgt die Erkenntnis, dass die Kriterien nicht nur von den Anforderungen abhängig sind. Es existiert unter den Kriterien selbst eine Abhängigkeit, die je nach Behandlungsmethode variiert. In diesem Kapitel wurde z. B. eine Abhängigkeit zwischen den Kriterien des Aufwands, der Dimension und der Simultanen Behandlung erkannt. Diese Abhängigkeit wurde nicht in jeder Methode festgestellt. Bei der linearen Interpolation z. B. hat die Simultane Behandlung keinen Einfluss auf den Aufwand, da in der Methode nur jeweils ein Punktausreißer behandelt wird. Eine weitere Abhängigkeit existiert zwischen dem Kriterium der Richtigkeit und der Objektivität. In den Methoden, in denen der Anwender subjektiv mitwirken kann, bestimmt er indirekt die Richtigkeit. Das bedeutet, dass der Anwender selbstständig wahre Ausreißer definieren kann und dann diese behandelt werden. Im Hinblick auf die Produktion können gleiche Ausreißer je nach Anwender zu verschiedenen Ergebnissen führen. Daher müssen je nach Anwendungsfall in der Produktion, verschiedene Behandlungsmethoden mit der Sachlage geprüft werden. Die Anwender müssen außerdem auswählen, welche Kriterien für den vorhandenen Datensatz wichtig sind und darauf aufbauend ihre Methode auswählen. Dabei sind die Kriterien der Richtigkeit und Objektivität von enormer Bedeutung, da diese Kriterien nicht wie die anderen immer eindeutig sind. Innerhalb der Behandlungsmethode können die Aussagen dieser Kriterien durch die gegebenen Datensätze und der Produktion variieren. Also kann der Einfluss des Anwenders innerhalb einer Methode abhängig vom Datensatz oder der Produktion als subjektiv oder auch als objektiv bezeichnet werden. Diese Thematik wird im folgenden Kapitel genauer diskutiert und im Hinblick auf die Kategorisierung der Ausreißerbehandlung miteinbezogen.

Tabelle 7: Vergleich und Anwendung der Kriterien an den Behandlungsmethoden (eigene Darstellung in Anlehnung an 3.3 und 4.3)

Behandlungsmethode\Kriterien	Dimension	Aufwand	Simultane Behandlung	Richtigkeit	Objektivität	Visualisierung
Least Trimmed Squares (Rousseeuw & Leroy 1987; Seo & Bae 2012; Mount et al. 2014)	Zweidimensional	Berechnung der Rückstände	n/2 Ausreißer	Punkte, die nicht von der Regression beachtet werden	Anwender hat keinen Einfluss	Regressionskurve
Bagplot Based Adjustment (Rousseeuw et al. 1999; Avanzi et al. 2022; Verdonck & Van Wouwe 2011)	Zweidimensional	Berechnung der Bestandteile, Auswertung der Distanzen	n/2 Ausreißer	Punkte, die außerhalb des Fence sind	Anwender bestimmt Fence-Faktor und Prozesswiederholung	Bag, Fence & Loop
Mahalanobis Distance Approach (Avanzi et al. 2022; Tiwari et al. 2007)	Zweidimensional	Berechnungen der Distanzen, Zuordnung der Gewichtung	Es werden alle Datenpunkte miteinbezogen	Keine Definition von wahren Ausreißern	Anwender bestimmt die Gewichtung	Distanz zum Zentrum, Erzeugung einer Regression
Winsorizing (Blaine 2018; Ghosh & Vogt 2012)	Eindimensional	Berechnung der Perzentile basierend auf Winsorizing-Level	Anzahl hängt von den dem Winsorizing-Level ab	Werte, die nach dem ausgewählten Level unter/über dem bestimmten Perzentil sind	Anwender bestimmt Winsorizing-Level	Nicht in der Methode integriert (Boxplot Darstellung möglich)
Lineare Interpolation (Wahir et al. 2018)	Zweidimensional	Berechnung durch bekannte Interpolationsformel	Ein Ausreißer wird behandelt	Punkt zwischen der Verbindung der zwei Punkten	Anwender hat keinen Einfluss	Verbindung der Datenpunkte zur Behandlung
Sigma-Approach (Tiwari et al. 2007)	Zweidimensional	Berechnungen von grundlegender Statistik	Anzahl abhängig von der Verteilung und der Variable x	Werte, die x-mal vom Mittelwert abweichen	Anwender bestimmt das x und y	Nicht in der Methode integriert

4.6 Diskussion und Fazit

In diesem Kapitel werden verschiedene Aspekte und die Erkenntnisse aus den vorherigen Kapiteln diskutiert bzw. kritisch hinterfragt. Darüber hinaus wird am Ende dieses Abschnitts das Hauptziel die Kategorisierung der Ausreißerbehandlung mithilfe der Kriterien erfolgen.
In der letzten Thematik wurde diskutiert, dass innerhalb einer Methode je nach Fall die Objektivität des Anwenders variieren kann. Das bedeutet, dass z. B. das BBA nicht immer subjektiv durch den Anwender beeinflusst wird. Wenn in der Produktion Leitlinien für die Benutzung bestimmter Methoden erzeugt werden, dann kann der Einfluss des Anwenders objektiv wirken. Die Objektivität entsteht dadurch, dass verschiedene Anwender innerhalb einer Produktion anhand der Leitlinie Ergebnisse finden, die durch die gleichen Voraussetzungen entstanden sind. Im Hinblick auf das BBA wäre eine beispielshafte Leitlinie, den gängigen Faktor von drei für den Fence zu benutzen, der von Rousseeuw et al. (1999) genannt wurde. Außerdem müsste in dieser Methode eine Leitlinie für die Prozesswiederholung festgelegt werden, in der z. B. steht, wie oft ein Prozess wiederholt werden sollte, oder ob überhaupt eine Prozesswiederholung stattfinden soll. Je nach Unternehmen können diese Leitlinien variieren, dennoch ist es sinnvoll innerhalb der Produktion eines Unternehmens so eine Leitlinie zu benutzen. Wenn die Fertigungstechnologien nach Kropik (2021) aus Abschnitt 2.1 betrachtet werden, dann liefert jede Technologie, Daten, die im Hinblick auf Datenstrukturen verschiedene Datensätze hervorbringen. In diesen Datensätze können Ausreißer vorhanden sein, die behandelt werden sollten. Wenn z. B. in der Produktion ein Presswerk vorhanden ist, wo Blechteile gemessen werden, dann sollten dort verschiedene Anwender durch die Leitlinie in einer Behandlungsmethode bei demselben Datensatz gleiche Ergebnisse hervorbringen. Weitere Leitlinien können im Winsorizing und im Sigma-Approach benutzt werden. In diesen Methoden hat der Anwender ebenfalls die Möglichkeit subjektiv Entscheidungen zu treffen, die durch gängige Faktoren mit der Erzeugung von Leitlinien objektiv bezeichnet werden können. Das MDA kann aufgrund ihrer Methodik schon als Verfahren mit einer Leitlinie betrachtet werden, da dort vorgegeben ist, dass die Gewichtung anhand der Distanz invers festgelegt werden muss. Im Gegensatz dazu, existieren Verfahren, wo der Anwender keinen Einfluss hat und keine Entscheidungen bezüglich der Behandlung trifft. Hierbei entsteht die Frage, ob diese Verfahren überhaupt als hybride Methoden bezeichnet werden können. In der Arbeit wurde der Fokus auf diese hybriden Methoden gelegt, d. h. es sollten Methoden thematisiert werden, die eine Kooperation von Menschen und technischen Hilfsmitteln benötigt. Ein Grund für die Nutzung der hybriden Methoden wurde in Abschnitt 2.3 bei der Analyse von Daten vorgestellt. Automatisierte Methoden können selbstständig zur fehlerhaften Erkennung von Korrelationen führen, die durch das Wirken eines Anwenders verhindert werden. Dabei entsteht die Frage, ob Methoden erst als hybrid bezeichnet werden können, wenn der Anwender auch einen Einfluss hat. Bei dem LTS und bei der linearen Interpolation ist kein Einfluss des Anwenders ersichtlich. Dennoch sind diese Methoden für den hybriden Fall geeignet. Der Anwender im LTS kann nach der Ausreißerbehandlung wie in Abschnitt 2.2 zur Überprüfung der Regressionskurve mitwirken. Das bedeutet, dass der Einfluss auf eine Kontrolle der Methode abzielt. Diese Vorgehensweise kann in der linearen Interpolation ebenfalls benutzt werden. Dort kann der Anwender

kontrollieren, ob wirklich zwei aufeinanderfolgende Datenpunkte erfasst wurden, die zur Behandlung notwendig sind.

Ein weiterer Aspekt der hybriden Methoden ist die gewählte Datenebene. Die Detektions- und Behandlungsmethoden aus Kapitel 3 wurden im Hinblick auf Datenbanken und Datensätze ausgewählt. Neben diesen Datenebenen wurde auch Big-Data vorgestellt und von den kleineren Datenebenen abgegrenzt. Daraus resultiert die Frage, inwiefern die vorgestellten Methoden auch bei Big-Data nutzbar sind. Dafür werden die Kriterien nach Fasel und Meier (2016) mit den Voraussetzungen der Methoden verglichen und ausgewertet. Das Volumen ist das erste Kriterium und weist zu den Datensätzen einen großen Unterschied auf. Wie Nolting (2021) schon erwähnt hat, sind die Mengen an Daten bei Big-Data enorm. Dies führt bei den hybriden Behandlungs- und Detektionsmethoden dazu, dass der Aufwand steigt sowie die Wahrscheinlichkeit der Simultanen Behandlung, da mehr Datenpunkte berechnet werden. Durch diese Faktoren fällt die Wahrscheinlichkeit der Nutzung der linearen Interpolation weg, mit der nur ein Punktausreißer behandelt wird. Es wird vermutet, dass die Wahrscheinlichkeit von einem Punktausreißer bei Big-Data sehr gering ist. Im Kontext mit der Datenmenge fällt auch die Benutzung des Dixon-Test weg, da der als Voraussetzung hat, dass kleine normalverteilte Datensätze bearbeitet werden. Das nächste Kriterium von Big-Data ist die Vielfalt, die die verschiedenen Datenstrukturen thematisiert. Damit die Methoden funktionieren, existiert die Voraussetzung, dass der Datensatz innerhalb einer Datenstruktur ist oder mithilfe des gleichen Datentyps in ein Format integriert werden kann. Wenn im Big-Data verschiedene Datenstrukturen vorhanden sind, die nach einer Datenintegration nicht miteinander verglichen werden können, dann kann keine Methode benutzt werden. Diese Aspekte sind wie in den Abschnitten 4.2 und 4.3 auch erwähnt, unabhängig von Big-Data essenzielle Faktoren, die die Benutzung der Behandlungsmethoden beeinflussen. Das nächstes Kriterium von Big-Data ist die Geschwindigkeit, die sich mit der Datenverarbeitung befasst. Die Geschwindigkeit der Daten innerhalb von Big-Data haben keinen direkten Einfluss auf die Methoden. Dennoch ist das schnelle Erzeugen der jeweiligen passenden Daten innerhalb einer Methode, die durch technische Hilfsmittel unterstützt werden, bedeutsam. Außerdem wird die Geschwindigkeit der Datenverarbeitung mit dem steigenden Aufwand in Big-Data ausgeglichen. Das vierte Kriterium von Big-Data thematisiert die Zuverlässigkeit der Daten. Diese Zuverlässigkeit kann auf die Fehlererfassung vor der gesamten Thematisierung der Ausreißer bezogen werden. Es sollten in der Produktion Ausreißer untersucht werden, die durch einen optimalen Prozess (s. 2.3 & s. 3.3) entstehen. Der optimale Prozess verspricht die Richtigkeit der Daten und ermöglicht somit die Behandlung von interessanten Ausreißern. Das letzte Kriterium von Big-Data kennzeichnet den Wert der Daten. In diesem Fall hat auch dieses Kriterium keinen direkten Einfluss auf die Methoden. Durch die Detektion und Behandlung der Ausreißer in Daten soll die Produktion analysiert und verbessert werden. Im Endeffekt führt das zu der Wertsteigerung der Daten. Nachdem die Kriterien von Big-Data in Verbindung mit den Detektions- und Behandlungsmethoden gesetzt wurden, folgt die Erkenntnis, dass Methoden existieren, die auch für Big-Data nutzbar sind. Dabei ist der Aufwand in diesen Methoden der größte Faktor. Die Methoden müssen die enormen Datenmengen bearbeiten und für die Produktion schnelle Ergebnisse liefern. Die Kalkulationen bei der Bearbeitung in diesen Methoden werden auch mit den technischen Hilfsmitteln länger dauern.

Dennoch lohnt es sich bei Big-Data auch auf hybride Methoden zu setzen, da fehlerhafte Korrelationen unabhängig von der Datenebene entstehen können. Die vorgestellten hybriden Methoden in dieser Arbeit sollten einen Überblick erschaffen. Es existieren weitere Methoden, die für die Nutzung von Big-Data geeigneter wären. Außerdem wurde vermutet, dass die Wahrscheinlichkeit von einem Punktausreißer in Big-Data sehr gering ist. In solchen riesigen Datenmengen ist das Erscheinen von kollektiven und kontextualen Ausreißern wahrscheinlicher. Die kontextualen Ausreißer wurden bis jetzt nicht genauer in der Produktion thematisiert und deshalb werden sie hier untersucht.

Unabhängig von Big-Data ist die Untersuchung der kontextualen Ausreißer aus Kapitel 3.1 auch im Hinblick auf die Produktion interessant, da bei diesen Ausreißern auch der Produktionskontext analysiert werden muss. Im Grunde können diese kontextualen Ausreißer auch als kollektive Ausreißer betrachtet werden, die dann durch den Kontext abgegrenzt werden. Bei den vorgestellten Detektions- und Behandlungsmethoden wurden die kontextualen Ausreißer nicht spezifisch thematisiert. Die Erkennung dieser kontextualen Ausreißer innerhalb der Methoden muss durch einen weiteren Faktor erweitert werden. Die Abbildung der kontextualen Ausreißer aus Abschnitt 3.1 zeigt, dass kontextuale Ausreißer nicht nur durch die Abweichung des Wertes erkannt werden. Für die Produktion muss vor der Detektion eine Festlegung des Kontexts bzw. eine Abhängigkeit des Werts von der benutzen Datenstruktur erzeugt werden. Ein Beispiel für diesen Fall ist die Herstellung von unterschiedlichen Teilen zu verschiedenen Zeitintervallen in der Produktion. Die kontextuale Ausreißer wären in diesem Fall die Gruppe von Teilen, die in einem falschen Zeitintervall hergestellt werden. Also entstehen diese Ausreißer nicht nur aufgrund der Abweichung der Werte, sondern durch die Kombination der Abweichung und der zeitlichen Abhängigkeit. Wie erwähnt, wurden diese kontextualen Ausreißer bis jetzt nicht in dieser Arbeit thematisiert, dennoch existiert dieser Typ von Ausreißer auch in der Produktion und ist somit relevant für die Detektion und Behandlung. Die Abhängigkeiten können z. B. durch das Mitwirken eines Anwenders oder durch Erweiterungen der Methoden erkannt werden. Nach der Detektion der kontextualen Ausreißer muss eine Behandlungsmethode ausgewählt werden, die diese kontextualen Ausreißer im Rahmen der Bedingungen behandeln kann. Als Behandlungsmethode fällt das Winsorizing weg, da die Methode eindimensional ist und somit keine Abhängigkeiten der Datenstruktur besitzt. Die restlichen Behandlungsmethoden sind zweidimensional und können mit den richtigen Anpassungen solche Ausreißer behandeln. Zum Beispiel im BBA könnte die X-Achse als Zeit und die Y-Achse als Länge der Teile benutzt werden. Das Zentrum in dieser Methode wird von den Teilen ausgemacht, die wirklich in diesem Zeitintervall hergestellt werden sollten. Die Ausreißer sind dann die Teile, die in einem anderen Zeitintervall richtig wären, aber in diesem System aufgrund der zeitlichen Abhängigkeit abweichen.

Neben den kontextualen Ausreißern wurden die Detektionsmethoden auch zu wenig thematisiert, da die Behandlungsmethoden inhärente Detektionen verfügen. Deshalb werden jetzt Detektionsmethoden überprüft, die eine Kompatibilität mit den Behandlungsmethoden aufweisen und welche Folgen die verbundene Benutzung hat. Als erste Detektionsmethode wird der Dixon-Test untersucht. Aufgrund der Dimension ist der Dixon-Test nur mit dem Winsorizing kompatibel. Wenn das Beispiel aus Abschnitt 3.2.1 betrachtet wird, fällt auf, dass der Wert für $x_n = 350$ ein

Ausreißer ist. Durch den Einsatz des Winsorizing-Level mit der inhärenten Detektion in der Methode ist der gleiche Wert auch als Ausreißer definiert und würde behandelt werden. Außerdem fällt auf, dass beim Winsorizing der Wert $x_1 = 295{,}1$ ebenfalls als ein Ausreißer erkannt wird. Bei der Überprüfung im Dixon-Test durch die Formel (4) ist dieser Wert kein Ausreißer und müsste nicht behandelt werden. Also wird im Winsorizing ein Wert behandelt, der nach dem Dixon-Test kein Ausreißer ist. Dies zeigt auf, dass die Nutzung des Winsorizing negative Aspekte beinhaltet und mit Risiken verbunden ist. Als nächstes wird die Nearest-Neighbor-Methode ausgewertet. Diese Methode kann für die Nutzung des BBA und des MDA nützlich sein, da es vom Prinzip der inhärenten Detektion von den Behandlungsmethoden ähnelt. Wie in den Behandlungsmethoden wird hier die Ausreißererkennung durch eine Nachbarschaft und einer Distanzmessung ermittelt. Ein großer Unterschied ist, die Anzahl der Nachbarschaften bzw. der Zentren. Im BBA wird ein Zentrum erstellt, wo Werte in kurzer Distanz 50 % des Datensatzes ausmachen und somit nicht als Ausreißer gelten. Das gleiche gilt auch für das MDA, da dort auch nur ein Zentrum definiert wird und darauf aufbauend Distanzen gemessen werden. Im NNM können mehrere Zentren entstehen und somit den Behandlungsmethoden widersprechen. Ein Beispiel für diesen Fall ist ein Datensatz mit zwei Nachbarschaften, die 50 % und 40 % der Datenpunkte darstellen. Im BBA würde die Nachbarschaft mit 40 % als Ausreißer gelten, wenn die Entfernung zum Zentrum mit 50 % über die Größe des Fence hinaus geht. In Abschnitt 4.5 wurde die Definition des Zentrums vom BBA auch auf das MDA übertragen. Nach dieser Definition und der Methodik des MDA würde die Nachbarschaft mit 40 % eine niedrigere Gewichtung zugeteilt bekommen als das Zentrum. Dennoch existiert eine weitere Möglichkeit eines Zentrums im MDA. Das Zentrum kann auch durch die Auswertung des ganzen Datensatzes entstehen, indem für jede Achse bzw. Dimension eine Mitte bestimmt wird. Das bedeutet im zweidimensionalen, dass für die X- und Y-Achse einen Mittelwert bestimmt und dann aus den Mittelwerten ein Punktzentrum festgelegt wird. So wäre keiner der Nachbarschaften das Zentrum in der Methode, sondern ein Punkt zwischen den Nachbarschaften, welcher näher zu der größeren Nachbarschaft liegt. Die Detektion mit der NNM hat in beiden Methoden keinen Mehrwert, da verschiedene Ergebnisse entstehen. Im Allgemeinen erzeugt die inhärente Detektion der Behandlungsmethoden unterschiedliche Ergebnisse zu den Detektionsmethoden. Das LTS würde für das Beispiel aus der Abbildung (s. 3.3.1) wahrscheinlich das gleiche Ergebnis wie in der NNM liefern, dennoch ist dies nur für den Fall. Das bedeutet nicht, dass jeder Datensatz im LTS, die gleiche Erkenntnis wie in der NNM hat. Da die Detektionsmethoden aus Kapitel 3.2.3 und 3.2.4 auf dem Prinzip der NNM aufgebaut sind, gelten die gleichen Erkenntnisse. In der linearen Interpolation wird der Ausreißer durch die Verbindung der zwei aufeinanderfolgenden Punkten detektiert. Die Verwendung der vorgestellten Methoden ist nicht zielführend, da die Methodik aus der linearen Interpolation nicht durch die Methoden adaptiert werden kann. Der Ausreißer in dieser Methode liegt zwischen den aufeinanderfolgenden Punkten, dies würde die Methoden 3.2.2 bis 3.2.4 ausschließen, da das Nachbarschaftsprinzip hier nicht wirkt. Der Dixon-Test aus 3.2.1 ist ebenfalls keine Methode, die hier benutzt werden kann, da erstens eindimensionale Datensätze behandelt werden und zweitens der Datensatz normalverteilt sein muss. Der Sigma-Approach hat eine inhärente Detektion und hat wie die lineare Interpolation mit den vorgestellten Detektionsmethoden keine nutzbaren Ergebnisse

Die Gründe dafür sind die gleichen wie bei der linearen Interpolation. Nachdem die Behandlungsmethoden bezüglich ihrer Detektion überprüft wurden, folgt die Erkenntnis, dass vor der Behandlung die inhärente Detektion viel zielführender ist als die Benutzung einer der vorgestellten Detektionsmethoden. Im Hinblick auf die Produktion wäre die direkte Nutzung einer Behandlungsmethode effektiver als die eigenständige Detektion. Der Anwender muss also bei der Auswahl der Behandlungsmethoden auch die inhärente Detektion beachten.
Einige Aspekte aus dieser Arbeit wurden in diesem Kapitel diskutiert, dennoch wurde das Hauptziel noch nicht thematisiert. In dieser Arbeit soll eine Kategorisierung der Ausreißerbehandlung in der Produktion stattfinden, die mithilfe der Kriterien aus Abschnitt 4.3 erfolgt. Dafür wurden in der Einleitung Teilziele definiert. In Kapitel 2 wurde das erste Teilziel abgearbeitet und war die Vorstellung der Daten in der Produktion. Dort wurde die Datenebene für die Analyse der Datenpunkte eingegrenzt. Danach wurde in Kapitel 3 das zweite Teilziel bearbeitet, indem verschiedene Behandlungsmöglichkeiten erklärt und diese voneinander unterschieden wurden. Nachdem beide Teilziele erreicht wurden, konnte in Kapitel 4 der Eigenanteil folgen. Im Eigenanteil wurden die vorherigen Kapitel verknüpft, um das dritte Teilziel zu erreichen. Das dritte Teilziel mit den Anforderungen wurde aus den drei Faktoren, also von der Produktion der Automobilindustrie im Hinblick auf Daten, von den Detektions- und Behandlungsmethoden, erarbeitet. Die Anforderungen, die daraus entstanden sind, waren die Basis für die Kriterien, die zur Kategorisierung notwendig sind. Aber bevor die Kategorisierung stattfindet, muss die Relevanz für die Forschung verdeutlicht werden. In der Produktion der heutigen Zeit entstehen durch die verschiedenen Schritte Daten, die zur Analyse und zur Weiterbearbeitung notwendig sind. Diese Daten können neben den erwarteten Werten die thematisierten Ausreißer beinhalten. Damit diese Ausreißer einerseits die Aussage des Datensatzes nicht zerstören sollen, müssen diese behandelt werden. Die vorher erwähnte indifferente Behandlung würde die nutzbaren Informationen der Ausreißer vernichten. Damit die Ausreißer noch berücksichtigt werden sollen, ist die Benutzung von Behandlungsmöglichkeiten empfehlenswert. Diese Behandlungsmöglichkeiten können in der Produktion, dem Analysten bzw. dem Anwender behilflich sein, indem die Informationen der Ausreißer benutzt und das Gesamtbild des Datensatzes erhalten wird. Die Arbeit kann in diesem Prozess der Produktion eingeordnet werden. Es existieren verschiedene Behandlungsmöglichkeiten, die der Anwender auswählen kann. Damit der Anwender nicht unbedacht zwischen den verschiedenen Behandlungsmethoden wählen muss, werden diese in verschiedene Kategorien zugeordnet. Nach der Einordnung der Arbeit folgt jetzt die Vorstellung der Kategorisierung der Ausreißerbehandlung in der Produktion.
Die erste Kategorie beinhaltet die Behandlungsmöglichkeiten, die mithilfe von Leitlinien subjektive Entscheidungen in der Behandlung verhindern sollen. In diese Kategorie können alle vorgestellten Methoden bis auf das LTS und die lineare Interpolation eingeordnet werden. Die Auswahl dieser Kategorie resultiert aufgrund der bestimmten Eingriffe des Anwenders, die objektive Ergebnisse subjektiv verändern können. Diese Kategorie wird als „Standardisiert“ bezeichnet, da durch die Leitlinien Standards definiert werden. Die zweite Kategorie resultiert aus den Methoden, die keine Leitlinien benötigen und somit vom Anwender nicht subjektiv beeinflusst werden können. Diese Kategorie wird als „Kontrolliert“ bezeichnet und beinhaltet das LTS und die lineare Interpolation. Der Anwender kann, wie vorher erwähnt, als Kontrolleur in diesen Methode

mitwirken und somit fehlerhafte Korrelationen verhindern. Die beiden Kategorien gelten als Obergruppen und sind aus der Richtigkeit sowie aus der Objektivität entstanden. Aus den Obergruppen entstehen zwei Untergruppen, die jeweils aus beiden Oberkategorien folgen. In diesen Unterkategorien werden zwischen komplexen und simplen Behandlungsmöglichkeiten unterschieden. Der Ursprung, dieser Unterkategorien resultiert aus den Kriterien der Dimension, des Aufwands und der Simultanen Behandlung. In Abschnitt 4.5 wurde verdeutlicht, dass der Aufwand von der Dimension und der simultanen Behandlung abhängig ist. Mit diesem Zusammenhang wurden die Untergruppen erzeugt. Die Untergruppen „Simpel" beinhalten die Behandlungsmöglichkeiten deren Aufwand gering ist. Hier in der Kategorisierung wurden das Winsorizing und die lineare Interpolation in diese Untergruppen zugeordnet, da ersteres durch die Eindimensionalität und zweiteres durch die Behandlung nur eines Punktausreißers zu einem geringen Aufwand führen. Die restlichen Methoden sind zweidimensional und behandeln mehrere Datenpunkte bzw. Ausreißer. Deshalb wurden diese Methoden in die Untergruppen „Komplex" eingeordnet. Erwähnenswert ist der Einfluss des Kriteriums der Visualisierung. In Abschnitt 4.5 wurde keine direkte Abhängigkeit bzgl. der Visualisierung erkannt. Außerdem wurde in 4.3 erklärt, dass die Kriterien des Aufwands und der Visualisierung koexistieren und sich nicht gegenseitig ausschließen. Hier muss noch erweitert werden, dass je nach Methode, die Visualisierung vom Aufwand abhängt. Der MDA, das BBA und das LTS haben als Folge des Aufwands, dass eine Visualisierung erstellt wird. Darüber hinaus führt die Visualisierung im BBA auch wieder zu neuen Berechnungen, die im Kriterium des Aufwands berücksichtigt werden müssen. Also haben diese Kriterien im BBA eine Wechselwirkung, die die Einordnung des BBA in die Kategorie „Komplex" unterstützen. Als nächstes folgt die Abbildung 11, in der die finale Kategorisierung der Ausreißerbehandlung in der Produktion dargestellt wird.

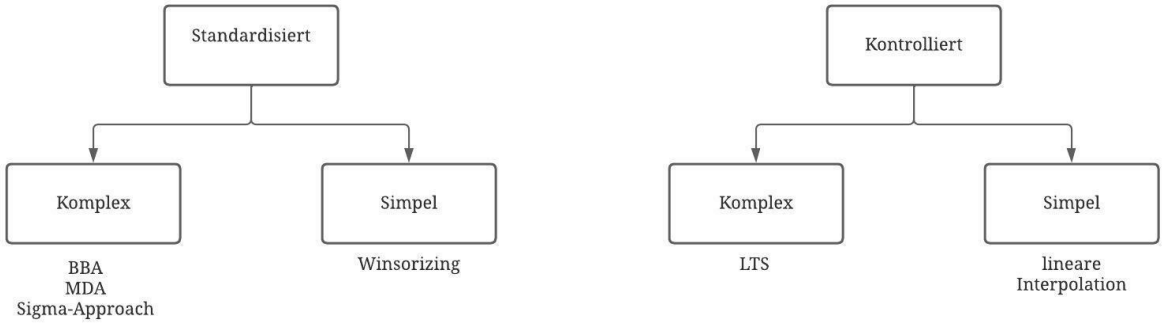

Abbildung 11: Kategorisierung der Ausreißerbehandlung in der Produktion (eigene Darstellung in Anlehnung an 4.5)

In der Industrie können diese Kategorien mit den Erklärungen bzgl. der Kriterien auf verschiedene Behandlungsmöglichkeiten angewendet werden, damit Anwender für ihre Fälle, die passende Methode auswählen. Dafür ist es sinnvoll, den Anwendern eine weitere Übersicht wie die Tabelle 7 auszuhändigen, damit vor allem Dimensionsprobleme nicht entstehen. Das benutzen von den Kategorien und Kriterien auf die Ausreißerbehandlungen soll in der Produktion als Unterstützung dienen. Aus der Kategorie „Simpel" wird z. B. die Dimension nicht ersichtlich, da die

Eindimensionalität keine Bedingung der Gruppe ist, weil auch Methoden mit trivialen Rechenschritten, wie die lineare Interpolation, in diese Kategorie eingeordnet werden können. Schließlich wurde mit diesen Kategorien und Kriterien eine Grundlage erzeugt, die durch weitere Behandlungsmöglichkeiten erweitert werden kann, aber für die erste Anwendung in der Praxis geeignet sein sollte.

5 Zusammenfassung und Ausblick

In dieser Arbeit wurden für die Ausreißerbehandlungen in der Produktion Kriterien abgeleitet, die eine Kategorisierung ermöglicht haben und in der Industrie als Unterstützung für die Ausreißer benutzt werden sollen. Dafür wurde zuerst der Datenbegriff und die Relevanz der Daten in der Produktion der Automobilindustrie thematisiert. Mit der Schlussfolgerung, dass die Prozesse aufgrund der Digitalisierung Daten ausgeben. Im Anschluss wurden verschiedene Datenebenen vorgestellt, damit die Ausreißer aus Kapitel 3 auf eine Ebene begrenzt werden. Nach der Datenebene wurden Fehler und Probleme von Daten in der Produktion erklärt. Dabei wurden die Fehler von automatischen Produktionsmaschinen fokussiert, da diese in der Industrie relevanter sind. Mithilfe dieser Klassifizierung wurden Datenfehler bzw. Qualitätsmängel von Daten aus Produktionsmaschinen eingegrenzt und somit die Überleitung zum optimalen Prozess erschaffen. Der optimale Prozess hat die Ausreißer eingeleitet und diese nicht als Fehler definiert. Daraufhin wurden die verschiedenen Ausreitertypen und deren Detektion thematisiert, die als Grundwissen für die Behandlungsmöglichkeiten notwendig sind. Diese verschiedenen Behandlungsmöglichkeiten von unterschiedlichen Autoren wurden bzgl. ihrer Methodik und ihrer Voraussetzungen vorgestellt, damit diese im Eigenanteil für die Kategorisierung benutzt werden können. Im Eigenanteil wurde die Verbindung zwischen den Daten in der Produktion und den Ausreißern erklärt und darauf aufbauend Anforderungen an die Kriterien gestellt. Von den Anforderungen wurden dann die Kriterien abgeleitet im Hinblick auf die finale Kategorisierung. Durch die Anwendung der Kriterien auf die Behandlungsmöglichkeiten wurden verschiedene Ergebnisse erkannt, die zu verschiedenen Kategorien geführt haben. Im letzten Abschnitt wurden die Kategorien erstellt und die Behandlungsmöglichkeiten in diese eingeordnet.

In der Einleitung wurde die fortlaufende Digitalisierung erklärt, die für die Entstehung von Daten sorgt, mit der Erkenntnis, dass die Digitalisierung sich immer weiter durchsetzen wird. Schließlich wurden hier die Kategorien im Hinblick auf die Produktion erzeugt, die sich mit der Zeit auch immer weiterentwickelt. Die Weiterentwicklung der Produktion wird zu einer Veränderung führen, die sich auf die drei Faktoren und somit auf die finale Kategorisierung auswirkt. Also ist die Veränderung der Kategorien abhängig vom Forschungsstand bezüglich der Produktion, der Detektions- und Behandlungsmethoden. Weitere Behandlungsmöglichkeiten, die in der Zukunft noch entstehen werden, können zu einer Erweiterung der Kriterien bzw. der Kategorien führen.

Literaturverzeichnis

Aggarwal, C. C. (2017): *Outlier Analysis: An introduction to Outlier Analysis*, 2. Aufl., Cham, Schweiz: Springer Nature.

Aguinis, H., Gottfredson, R.K., Joo, H. (2013): Best-Practice Recommendations for Defining, Identifying, and Handling Outliers, in *Organizational Research Methods*, Bd. 16, Nr. 2, S. 270–301, [online] doi:10.1177/1094428112470848.

Alvarez-Coello, D., Wilms, D., Bekan, A., Marx Gomez, J. (2021): Towards a Data-Centric Architecture in the Automotive Industry, in Procedia Computer Science, Bd. 181, S. 658-663, [online] doi:10.1016/j.procs.2021.01.215.

Avanzi, B., Lavender, M., Taylor, G., Wong, B. (2022): *Detection and treatment of outliers for multivariate robust loss reserving*. [online] verfügbar unter: https://arxiv.org/pdf/2203.03874.pdf [Zugriff am 27 Juni 2022].

Blaine, B. E. (2018): The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation: Winsorizing, in *Fisher Digital Publications*, S. 1817-1818, [online] doi:10.4135/9781506326139.

Bundesministerium für Wirtschaft und Klimaschutz (2022): Automobilindusrie, [online] verfügbar unter: https://www.bmwk.de/Redaktion/DE/Textsammlungen/Branchenfokus/Industrie/branchenfokus-automobilindustrie.html#:~:text=Die%20Automobilindustrie%20ist%20die%20gr%C3%B6%C3%9Fte,besch%C3%A4ftigten%20direkt%20knapp%20786.000%20Personen. [Zugriff am 28. Juni 2022].

Chandola, V., Banerjee, A., Kumar, V. (2009): Anomaly detection, in *ACM Computing Surveys*, Bd. 41, Nr. 3, S. 1–58, [online] doi:10.1145/1541880.1541882.

Dietrich, M. (2021): *Digitales Shopfloor Management in SAP-Systemumgebungen: Roadmap und Lösungsalternative für die Umsetzung*, Wiesbaden, Deutschland: Springer Vieweg.

Divya, D. & Sasidhar, B. (2016): Methods to detect different types of outliers, in *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, S. 23-28, [online] doi:10.1109/sapience.2016.7684114.

Domański, P., Chen, Y. Q., Ławryńczuk, M. (2022): *Outliers in control engineering: fractional calculus perspective*, 1. Aufl., Berlin, Deutschland: De Gruyter.

Düsing, R. (2020): *Big Data: Anwendung und Nutzungspotenziale in der Produktion: Big Data Analytics – Begriff, Prozess und Ausrichtungen*, Stuttgart, Deutschland: Kohlhammer Verlag.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., Stolfo, S. (2002): A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data, in *Advances in Information Security*, S. 77-101, [online] doi:10.1007/978-1-4615-0953-0_4.

Fasel, D. & Meier, A. (2016): Big Data: Grundlagen, Systeme und Nutzungspotenziale, (HMD Edition), 1. Aufl., Wiesbaden, Deutschland: Springer Vieweg.

Ghosh, D. & Vogt, A. (2012): Outliers: An Evaluation of Methodologies, in *Joint Statistical Meetings*, S. 3455-3460 [online] verfügbar unter: http://www.asasrms.org/Proceedings/y2012/Files/304068_72402.pdf.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. Ch., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K., Stanley, H. E. (2000): PhysioBank, PhysioToolkit, and PhysioNet, in *Circulation*, Bd. 101, Nr. 23, S. 1-6, [online] doi:10.1161/01.cir.101.23.e215.

Gupta, M., Gao, J., Aggarwal, C. C., Han, J. (2014): *Outlier detection for temporal data*, San Rafael, USA: Morgan & Claypool.

Hallebach, J. & Täufer, L. (2020): *Wie können Fehler in der Produktion selbstständig und zuverlässig erkannt werden?* [online] verfügbar unter: https://www.industry-of-things.de/wie-koennen-fehler-in-der-produktion-selbststaendig-und-zuverlaessig-erkannt-werden-a-978698/ [Zugriff am 14. Juni 2022].

Hawkins, D.M. (1980): *Identification of Outliers*, New York, USA: Springer Publishing.

Huber, W. (2016): *Industrie 4.0 in der Automobilproduktion: ein Praxisbuch*, Wiesbaden, Deutschland: Springer Vieweg.

Hung-Vo, P. (2015): *Automobilindustrie und die Bedeutung innovativer Industrie 4.0 Technologien*, Hamburg, Deutschland: Diplomica.

Kropik, M. (2021): *Produktionsleitsysteme für die Automobilindustrie Digitalisierung des Shop-Floors in der Automobilproduktion*, Berlin, Deutschland: Springer Vieweg.

Kutner, M. H., Nachtsheim, C. J., Neter, J., Li, W. (2004). *Applied linear statistical models*, 5. Aufl., Boston, USA: McGraw-Hill/Irwin.

Laurikkala, J., Juhola, M., Kentala, E. (2000): Informal identification of outliers in medical data, in *The Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology*

(IDAMAP-2000), S. 20-24 [online] verfügbar unter: https://researchportal.tuni.fi/en/publications/informal-identification-of-outliers-in-medical-data [Zugriff am 25. Juni 2022].

Marschner, K. (2004): *Wettbewerbsanalyse in der Automobilindustrie: ein branchenspezifischer Ansatz auf Basis strategischer Erfolgsfaktoren*, 1. Aufl., Wiesbaden, Deutschland: Deutscher Universitätsverlag.

Mertens, P., Bodendorf, F., König, W., Schumann, M., Hess, T., Buxmann, P. (2017): *Grundzüge der Wirtschaftsinformatik*, 12. Aufl., Wiesbaden, Deutschland: Springer Vieweg.

Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y. (2012): On the Least Trimmed Squares Estimator, in *Algorithmica*, Bd. 69, Nr. 1, S. 148–183, [online] doi:10.1007/s00453-012-9721-8.

Nemeth, M. & Peterkova, A. (2018): Proposal of Data Acquisition Method for Industrial Processes in Automotive Industry for Data Analysis According to Industry 4.0, in *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)*, S. 157-162, [online] doi:10.1109/ines.2018.8523853

Nolting, M. (2021): *Künstliche Intelligenz in der Automobilindustrie mit KI und Daten vom Blechbieger zum Techgiganten*, Wiesbaden, Deutschland: Springer Vieweg.

Pietsch, W. (2021): *Big data*, Cambridge, United Kingdom: Cambridge University Press.

Piro, A. & Gebauer, M. (2021): *Daten- und Informationsqualität: Die Grundlage der Digitalisierung: Definition von Datenarten zur konsistenten Kommunikation im Unternehmen*, 5. Aufl., Wiesbaden, Deutschland: Springer Vieweg.

Porter, M. E. (2013): *Wettbewerbsstrategie: Methoden zur Analyse von Branchen und Konkurrenten*, 12. Aufl., Frankfurt am Main, Deutschland: Campus-Verlag.

Ranga Suri, N. N. R., Athithan, G., Narasimha Murty, M. (2019): *Outlier detection: Techniques and Application: A Data Mining*, New York, USA: Springer Publishing.

Rousseeuw, P.J. & Leroy, A.M. (1987): *Robust regression and outlier detection*, Hoboken, USA: Wiley.

Rousseeuw, P.J., Ruts, I., Tukey, J.W. (1999): The Bagplot: A Bivariate Boxplot, in *The American Statistician*, Bd. 53, Nr. 4, S. 382–387, [online] doi:10.1080/00031305.1999.10474494.

Schmid, U. (2001): Schwachstelle Datenqualität, in *Ökologisches Wirtschaften*, Bd. 16, Nr.6, S. 16-18, [online] doi: 10.14512/oew.v16i6.133

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., Tufano, P. (2012): Analytics: The real-world use of big data – How innovative enterprises extract value from uncertain data, in *IBM Institute for Business Value*, [online] verfügbar unter https://www.bdvc.nl/images/Rapporten/GBE03519USEN.PDF. [Zugriff am 11. Juli 2022].

Santello, G. (2022): Datensatz Parts Manufacturing – Industry Dataset. [online] verfügbar unter: https://www.kaggle.com/datasets/gabrielsantello/parts-manufacturing-industry-dataset?resource=download [Zugriff am 8. August 2022]

Seo, Y.-S. & Bae, D.-H. (2012): On the value of outlier elimination on software effort estimation research, in *Empirical Software Engineering*, Bd. 18, Nr. 4, S. 659–698, [online] doi:10.1007/s10664-012-9207-y.

Shrivastava, S., Rajesh, A., Bora, P.K. (2014): Sliding window Dixon's tests for malicious users' suppression in a cooperative spectrum sensing system, in *IET Communications*, Bd. 8, Nr. 7, S. 1065–1071, [online] doi:10.1049/iet-com.2013.0609.

Steiner, R. (2014): *Grundkurs Relationale Datenbanken: Einführung in die Praxis der Datenbankentwicklung für Ausbildung, Studium und IT-Beruf*, 8. Aufl., Wiesbaden, Deutschland: Springer Vieweg.

Steven, M., & Klünder, T. (2020): *Big Data: Anwendung und Nutzungspotenziale in der Produktion*, Stuttgart, Deutschland: Kohlhammer Verlag.

Tiwari, K., Mehta, K., Jain, N., Tiwari, R., Kanda, G. (2007): Selecting the Appropriate Outlier Treatment for Common Industry Applications, in *NESUG Conference*, [online] verfügbar unter https://lexjansen.com/nesug/nesug07/sa/sa16.pdf. [Zugriff am 12. Juli 2022].

Van Stein, B., Van Leeuwen, M., Wang, H., Purr, S., Kreissl, S., Meinhardt, J., Back, T. (2016): Towards Data Driven Process Control in Manufacturing Car Body Parts, in *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, S. 459-462, [online] doi:10.1109/csci.2016.0093.

Verma, S.P. & Quiroz-Ruiz, A. (2006): Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering, in *Revista mexicana de ciencias geológicas*, Bd. 23, Nr. 2, S. 133–161, [online] verfügbar unter: http://www.scielo.org.mx/scielo.php?script=sci_abstract&pid=S1026-87742006000200003&lng=es&nrm=iso&tlng=en [Zugriff am 27. Juni 2022].

Vieweg, I., Werner, C., Wagner, K. P., Hüttl, T., Backin, D. (2012): *Einführung Wirtschaftsinformatik: IT-Grundwissen für Studium und Praxis*, 8. Aufl., Wiesbaden, Deutschland: Springer Gabler.

Vogel, P. (2021): *Laborstatistik für technische Assistenten und Studierende*, Wiesbaden, Deutschland: Springer Fachmedien.

Wahir, N. A., Nor, M. E., Rusimann, S. R., Khuneswari, G. P. (2018): Alternative Method: Outlier Treatments with Box-Jenkins and Neural Network via Interpolation Method, in *Journal of Science and Technology*, Bd. 10, Nr. 2, S. 122-127, [online] doi:10.30880/jst.2018.10.02.020.

Walfish, S. (2006): A review of statistical outlier methods, in *Pharmaceutical Technology* [online] verfügbar unter: https://www.semanticscholar.org/paper/A-review-of-statistical-outlier-methods-Walfish/0469f92dbce67ec4444d094e818eeebaaed3d8d5 [Zugriff am 27. Juni 2022].

Winkelhake, U. (2021): *Die digitale Transformation der Automobilindustrie: Treiber – Roadmap – Praxis*, 2. Aufl., Wiesbaden, Deutschland: Springer Vieweg.