technische universität dortmund

itpl

Masterarbeit

Anwendung von Data Mining auf produktionslogistische Massendaten mit Schwerpunkt Verifikation und Validierung

Jerrit Klein
Matrikelnummer: 131252
Studiengang Wirtschaftsingenieurwesen
jerrit.klein@tu-dortmund.de

Ausgegeben am: 04.07.2016
Eingereicht am: 16.12.2016

Gutachter: Prof. Dr.-Ing. Markus Rabe
Gutachter: Dipl.-Inf. Anne Antonia Scheidler

Technische Universität Dortmund, Fakultät Maschinenbau
Lehrstuhl IT in Produktion und Logistik, http://www.itpl.mb.tu-dortmund.de/

Inhaltsverzeichnis

1 Einleitung

„We are drowning in information, but we are starved for knowledge"
(Naisbitt 1982)

Obwohl fast 35 Jahre alt, erscheint das obige Zitat des Zukunftsforschers John Naisbitt in Anbetracht der heutigen Gegebenheiten aktueller als je zuvor. Immer mehr Daten werden generiert und müssen erfasst und verarbeitet werden. Der rapide Anstieg anfallender Datenmengen in vielen Lebensbereichen stellt dabei sowohl eine der großen Möglichkeiten als auch Herausforderungen der heutigen Zeit dar. Die Digitalisierung schreitet unaufhaltsam voran und damit einhergehend auch die technische Notwendigkeit der Sammlung, Verarbeitung und Auswertung extremer Datenmengen. Daten lassen sich dabei definieren als Anhäufungen von Nummern, Zeichen oder Bildern, die maschinell durch Sensoren, Barcodescanner oder auch die Tastatur registriert werden. Damit stellen Daten das niedrigste Abstraktionslevel dar, aus dem Informationen und Wissen abgeleitet werden können. Allein für das Jahr 2008 fielen geschätzt 9,57 Zettabyte an verarbeiteten Informationen durch Unternehmensserver an, was heruntergerechnet etwa 63 Terabyte pro Unternehmen und drei Terabyte pro Mitarbeiter pro Jahr bedeutet (vgl. Short et al. 2011). Für den Aufbau der hierfür benötigten Infrastruktur wurden allein 2011 ca. vier Billiarden US-Dollar in den Bereich der Informationstechnologien (IT) investiert. Zu diesem Zeitpunkt machte dies ca. sechs Prozent des weltweiten Bruttoinlandprodukts in Höhe von 65,6 Billiarden US-Dollar aus (vgl. Cortada 2012). Auch wenn solche Berechnungen und Annahmen aufgrund der Komplexität nie auf die letzte Stelle exakt sein können, vermitteln sie doch ein recht gutes Gefühl für die aktuellen und zukünftigen Herausforderungen an die Datensammlung und -analyse.

Um Daten für Unternehmen verwertbar zu machen, existiert eine Reihe von Analysemöglichkeiten. Ein Verfahren, das nach Große Böckmann et al (2013, S. 921) „im Zeitalter von Datensammlern wie Google, Facebook und Co. vermehrt Eingang in öffentliche Diskussionen [findet]", ist das Data Mining, dessen eigentliche Ursprünge bereits bis in die 1980er Jahre zurückreichen. Zusammengefasst geht es beim Data Mining „um das Herausarbeiten von Abhängigkeiten innerhalb der Datenmenge" (Lämmel 2003, S. 6). Eine bis heute allgemein anerkannte Definition lieferten Fayyad et al. (1996). Sie beschreiben das Data Mining als die Entdeckung unbekannter Muster in bekannten Daten, bei der die entdeckten Muster die Kriterien Neuheit, Allgemeingültigkeit, Nichttrivialität, Nützlichkeit sowie Verständlichkeit aufweisen. Diese lassen sich unter Zuhilfenahme verschiedener Verfahren – wie etwa *Cluster-* oder *Assoziationsanalysen* – erkennen. Die Qualität der zugrundeliegenden Daten aus den operativen

Systemen ist häufig unzureichend – etwa aufgrund von Eingabefehlern oder Redundanzen. Deshalb erscheint es ratsam, das eigentliche Data-Mining-Verfahren um vor- und nachbereitende Schritte der Datenbereinigung und der Bewertung der Ergebnisse zu erweitern. Dieser Gesamtprozess der Wissensentdeckung ist in der Literatur auch als *Knowledge Discovery in Databases* (kurz: *KDD*) bekannt.

Das Hauptziel dieser Arbeit besteht in der Durchführung eines KDD-Prozesses auf produktionslogistische Massendaten eines produzierenden Unternehmens aus Deutschland. Grundlage hierfür ist ein durch den Lehrstuhl IT in Produktion und Logistik (ITPL) der TU Dortmund entwickeltes *Vorgehensmodell zur Musterextraktion in Supply Chains* (kurz: *MESC*). Ein Schwerpunkt des MESC ist eine iterative, phasenweise *Verifikation und Validierung* (kurz: *V&V*). Durch den Einsatz der V&V wird das Ziel verfolgt, jede der Phasen des KDD sowohl gegen sich selbst als auch gegen die vorherigen Phasen zu überprüfen. Da die V&V-Techniken im KDD in der Wissenschaft bisher wenig Beachtung finden, besteht ein weiteres Ziel dieser Arbeit in der Ableitung geeigneter V&V-Techniken anderer Bereiche. Durch die Durchführung des MESC mit phasenweiser V&V soll die praktische Anwendbarkeit des gesamten Vorgehensmodells aufgezeigt werden.

Um die generellen Möglichkeiten von Datenanalysen darzustellen, erfolgt zu Beginn der Arbeit eine Einführung in den Bereich der Wissensgewinnung in Datenbanken. Dabei werden hauptsächlich die für die Arbeit relevanten Themenfelder des KDD und des Data Minings vertieft. Daraufhin wird das im Anwendungsteil verwendete MESC weiteren bekannten Vorgehensmodellen gegenübergestellt und die verschiedenen Phasen der Modelle benannt. Ein Kernelement eines jeden KDD-Vorgehensmodells stellt der eigentliche Data-Mining-Vorgang dar, weswegen im nächsten Schritt verschiedene Data-Mining-Techniken vorgestellt werden.

Da die phasenweise V&V einen bedeutsamen Teil des MESC ausmacht und im späteren Verlauf der Arbeit den einzelnen Phasen des MESC jeweils passende V&V-Techniken zugeordnet werden müssen, erfolgt im Folgenden ein Überblick über das Themenfeld der V&V. Neben einer inhaltlichen Abgrenzung der Verifikation und Validierung und einer Vorstellung beispielhafter Einsatzmöglichkeiten wird dabei die V&V im MESC vorgestellt, bevor die Behandlung der V&V mit der Einführung relevanter Techniken abschließt. Darauf aufbauend erfolgt eine Untersuchung der vorgestellten Techniken auf ihre generellen Einsatzmöglichkeiten im KDD der Produktionslogistik, bevor im Speziellen auf die Anwendung im MESC eingegangen wird. Hierbei werden die zuvor erläuterten V&V-Elemente ebenso wie V&V-Kriterien einbezogen, um geeignete Techniken für die einzelnen Phasen zu bestimmen. Die Arbeit endet mit der Durchführung des MESC auf produktionslogistischen Massendaten unter Einsatz der zuvor bestimmten V&V-Techniken.

2 Data Mining und Knowledge Discovery in Databases

Um auf das weitere Vorgehen dieser Arbeit vorzubereiten, wird in diesem Kapitel eine Einführung in die Grundlagen des Data Minings und des KDD gegeben. Dazu ist eine Einordnung des Data Minings in den Kontext der Wissensgewinnung sinnvoll, um die Frage nach dem Mehrwert einer solch zeitintensiven Form der Datenauswertung zu ergründen. Darüber hinaus stellt dieses Kapitel ausgewählte Vorgehensmodelle für die Durchführung der Wissensgewinnung sowie geeignete Data-Mining-Verfahren vor. Dabei erfolgt eine ausführlichere Vorstellung des Vorgehensmodells *MESC* sowie des Data-Mining-Verfahrens *Assoziationsanalyse*, da diese bei der späteren Durchführung des Fallbeispiels angewendet werden sollen.

2.1 Grundlagen und Einordnung

Durch die schnell voranschreitende Digitalisierung und Einbindung von Informationstechniken in die Produktion (Stichwort: Industrie 4.0) fallen täglich riesige Datenmengen an. Die Notwendigkeit diese zu speichern führt zwangsläufig zu einem immensen Wachstum der Datenbanken. Dies geschieht in zwei Dimensionen – einerseits steigt die Anzahl der Einträge in den Datenbanken, andererseits steigt auch die Anzahl Attribute, die zu diesen Einträgen erfasst werden (vgl. Fayyad et al. 1996). Gezielte Analysen helfen dabei, unbekannte Informationen und Verbindungen zwischen den Daten herauszustellen, die beispielsweise zu Wettbewerbsvorteilen oder Erkennung neuer Marktpotentiale genutzt werden können (vgl. Knobloch und Weidner 2000).

Um eine Gewinnung von Informationen aus Datenbanken zu ermöglichen, ist eine Überführung der darin enthaltenen unstrukturierten Daten in eine kompaktere, abstraktere oder nützlichere Form notwendig (vgl. Fayyad et al. 1996). Für diese Überführung existieren je nach Art des Datenanalyseproblems verschiedene Methoden. Wie in Abbildung 2.1 dargestellt, lassen sich in der Theorie grundsätzlich zwei Analysearten unterscheiden – auf der einen Seite die hypothesengetriebene und auf der anderen Seite die datengetriebene Analyse. Bei der hypothesengetriebenen Analyse – auch Top-Down-Ansatz – liegen den Untersuchungen Annahmen zugrunde, die entweder bestätigt oder widerlegt werden sollen. Dies umfasst „klassische“ Analyseverfahren wie die manuelle Analyse und Interpretation. Bei der datengetriebenen Analyse hingegen fehlt im Idealfall eine solche Annahme. Vielmehr besteht das Ziel dieser Analyse in der Ableitung von Hypothesen aus den Ergebnissen. Deswegen wird hier auch von einer hypothesenfreien Analyse gesprochen (*Bottom-Up*) (vgl. Knobloch und Weidner 2000). Durch das Fehlen einschränkender Annahmen lassen sich auch unbekannte Muster finden, nach denen aus Aufwands-

oder Voreingenommenheitsgründen ansonsten nicht gesucht worden wäre. Die Vorgehensweisen der daten- und hypothesengetriebenen Analyse werden im Idealfall kombiniert, so dass sich ein sich wiederholender Datenanalysezyklus ergibt (Abbildung 2.2).

An dieser Stelle sei angemerkt, dass in der Praxis keine scharfe Trennung in die Bereiche der daten- und hypothesengetriebene Analyse möglich ist. Eine datengetriebene Analyse kann allein aus Komplexitätsgründen nie gänzlich hypothesenfrei sein. Darüber hinaus ist eine vollkommen freie Suche nach Mustern in den Daten auch nicht zielführend (vgl. Neckel und Knobloch 2015; Prescha 2009).

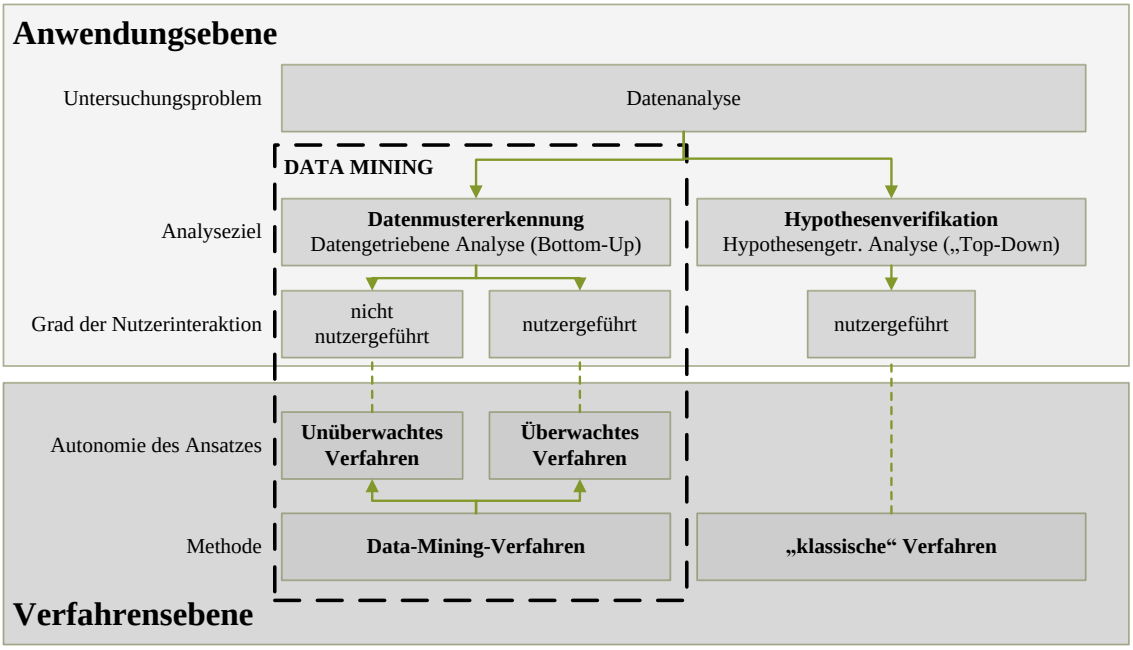


Abbildung 2.1: Methoden der Datenanalyse nach Knobloch (2000)

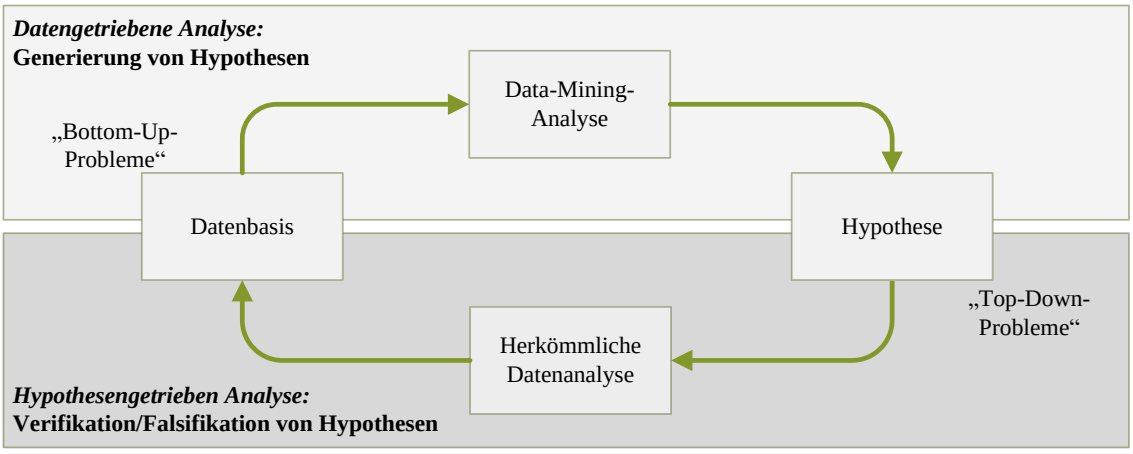


Abbildung 2.2: Datenanalysezyklus nach Knobloch (2000)

Eine Methode der datengetriebenen Analyse stellt das Data Mining dar. Dieses dient der Entdeckung unbekannter Abhängigkeiten in Datenbeständen, um so in den Daten impliziertes Wissen explizit zu machen (vgl. Lämmel 2003). Das Data Mining kann in verschiedenen Bereichen zur Wissensgewinnung eingesetzt werden. Grundsätzlich kann dabei zwischen zwei Hauptanwendungsfällen unterschieden werden – dem *deskriptiven* auf der einen und dem *prädiktiven* Problemfall auf der anderen Seite. Bei dem deskriptiven Problemfall – auch *Deskription* – liegt der Fokus auf dem Finden und Beschreiben von in den Daten vorliegenden und durch Experten auswertbaren Mustern. Bei dem prädiktiven Problemfall geht es dagegen um die Prognose von unbekannten oder zukünftigen Mustern auf Grundlage von Variablen oder Feldern in der Datenbank. Im Zusammenhang mit dem KDD hat die Deskription dabei nach Fayyad et al. (1996) die größere Relevanz. Zur Durchführung des Data Minings existieren verschieden Techniken, die in Abschnitt 2.4 genauer behandelt werden sollen.

Die Erkenntnis, dass das reine Anwenden eines Data-Mining-Verfahrens in vielen Fällen – beispielsweise aufgrund mangelhafter Datenqualität – nicht zielführend ist, macht in komplexen, nichttrivialen Systemen – wie sie etwa in der Produktion vorherrschen – weitere Schritte zur Vor- und Nachbereitung des Data-Mining-Verfahrens notwendig. Dabei ist eine vollkommen automatisierte Durchführung des Verfahrens durch die gegebene Nichttrivialität der Systeme nicht möglich. Es ist vielmehr eine Integration von Fachexperten in einen Gesamtprozess der Wissensgewinnung notwendig, der als Knowledge Discovery in Databases (KDD) bezeichnet wird (vgl. Walter 2004). Fayyad et al. (1996, S. 39) definieren den Unterschied zwischen KDD und Data Mining wie folgt: „In our view, KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data". Das Data Mining ist hierbei also nur als einer von mehreren aufeinander aufbauenden Schritten eines Gesamtprozesses anzusehen. Für das Vorgehen beim KDD existieren verschiedene Vorgehensmodelle, auf die in der Folge eingegangen werden soll. An dieser Stelle sei angemerkt, dass die Begrifflichkeiten Data Mining und KDD in der Literatur nicht eindeutig definiert sind, sondern teils unterschiedlich interpretiert oder auch synonym verwendet werden. Daneben finden sich weitere Begriffe für das selbe Vorgehen – wie etwa Wissensgewinnung oder Wissensextraktion – die im Kern dieselbe Thematik beschreiben. Deshalb sei hier festgelegt, dass in dieser Arbeit das Data Mining im Sinne Fayyads lediglich als ein Schritt der Wissensgewinnung definiert wird und im Folgenden die Begriffe KDD, KDD-Prozess und Data-Mining-Prozess synonym verwendet werden.

2.2 Vorgehensmodelle des KDD

Zur Einführung des MESC soll an dieser Stelle eine Vorstellung des generellen Vorgehensmodellbegriffs und seiner Anwendungsgebiete erfolgen, bevor im Speziellen auf die Vorgehensmodelle des KDD eingegangen wird.

2.2.1 Einführung des Vorgehensmodellbegriffs

Ein Vorgehensmodell legt den Rahmen für die Durchführung eines Prozesses fest. Ein Prozess ist dabei definiert als „eine Reihe von Aktivitäten, die untereinander in Verbindung stehen und aus einer Reihe von Eingaben ein Ergebnis für den Prozesskunden erzeugen“ (Füermann 2014, S. 1). Vorgehensmodelle können überall dort eingesetzt werden, wo Prozesse beschrieben werden müssen – wie etwa in der Softwareentwicklung, der Simulation oder im Projektmanagement. Einige dieser Modelle sollen an dieser Stelle kurz vorgestellt werden. Eine ausführlichere Übersicht existierender Vorgehensmodelle findet sich darüber hinaus etwa bei Rabe et al. (2008).

Softwareentwicklung

Als eines der ersten Vorgehensmodelle der Softwareentwicklung wurde das *Wasserfallmodell* entwickelt. In seiner ursprünglichen Form nach Benington (1956) sieht dieses ein sequentielles Durchlaufen der sechs Phasen *Planung, Anforderungsanalyse, Entwurf, Implementierung, Test* und *Betrieb* vor. Dabei ist das Verlassen einer Phase nicht vorgesehen, solange diese nicht abgeschlossen wurde. Dazu ist eine vollständige Begutachtung und Verifizierung nötig. Auf Grundlage dieses Ursprungsmodells entstanden weitere abgewandelte Modelle. Eines dieser Modelle ist das Modell von Royce (1970), das im Vergleich zu Beningtons Modell um Rücksprünge zur Vorphase ergänzt wurde. Boehm (1979) erweiterte das Wasserfallmodell zum sogenannten *V-Modell*, indem er den Phasen V&V-Aktivitäten zuordnete. Dieses Modell inklusive seiner Weiterentwicklungen – wie etwa dem *V-Modell XT* – hat insbesondere in Deutschland verstärkt Anwendung gefunden. Dies ist hauptsächlich darauf zurückzuführen, dass bei Bundeswehr und Bundesbehörden vorgeschrieben wurde, bei Softwareentwicklungen für öffentliche Auftraggeber grundsätzlich das V-Modell oder seine Weiterentwicklungen anzuwenden (vgl. Bröhl 1995; Rabe et al. 2008). Ein neuerer Ansatz zur Durchführung eines Softwareentwicklungsprozesses ist das sogenannte SCRUM-Vorgehensmodell, das sich vor allem für die Durchführung der Softwareentwicklung in Projektteams eignet und durch den regelmäßigen Austausch der Teammitglieder in kurzen Besprechungen (*Stand-Up Meetings*) sowie das Arbeiten in *Sprints* gekennzeichnet ist. Als Sprints werden ein- bis vierwöchige Abschnitte bezeichnet, in denen die Teammitglieder Arbeitssaufgaben abarbeiten müssen. Diese werden nach Abschluss des Sprints in *Sprintreviews* überprüft (vgl. Gloger 2011; Schwaber und Sutherland 2016).

Simulation

Auch wenn das Wasserfallmodell hauptsächlich für den Einsatz in der Softwareentwicklung entwickelt wurde, kann es auch für die Simulation eingesetzt werden. Neben diesem existieren noch diverse weitere Vorgehensmodelle für die Simulation. Diese können sich zwar durchaus in Komplexität und Umfang unterscheiden, bestehen aber größtenteils aus den fünf Kernelementen *Aufgabenanalyse, Modellformulierung, Modellimplementierung, Modellüberprüfung* und *Modellanwendung* (vgl. Banks et al. 1988). Als bedeutsamstes Vorgehensmodell der Simulation führen Rabe et al. (2008) das in der VDI-Richtlinie 3633 Blatt 1 beschriebene Vorgehensmodell des VDI an (vgl. VDI 2008). Dieses Modell ist insbesondere für die Produktion und Logistik im deutschsprachigen Raum von großer Bedeutung und enthält in der angeführten Version einen erhöhten Stellenwert der V&V im Vergleich zu vorherigen Versionen. Eine Darstellung des VDI-Vorgehensmodells findet sich in Abbildung A.1 des Anhangs. Bernhard et al. (2007) entwickelten darüber hinaus ein Vorgehensmodell zur Informationsgewinnung, das in verschiedene Simulationsmodelle – wie das VDI-Vorgehensmodell – integriert werden kann.

KDD

Neben den hier aufgeführten Vorgehensmodellen existieren auch verschiedene weitere, speziell zur Durchführung des KDD entwickelte Modelle. Eine Vorstellung relevanter Vorgehensmodelle des KDD findet in Abschnitt 2.2.2 statt.

Weitere Vorgehensmodelle

Die zuvor beschriebenen Vorgehensmodelle finden neben ihren ursprünglichen Bestimmungen auch in anderen Disziplinen Anwendung. So können die „klassischen“ Vorgehensmodelle der Softwareentwicklung – Wasserfall- und V-Modell – auch in der Projektdurchführung eingesetzt werden. Ebenfalls vermehrt genutzt wird in diesem Bereich das *SCRUM-Modell*. Dieses wird – neben seiner eigentlichen Nutzung bei der Softwareentwicklung – mittlerweile auch zunehmend für andere Aufgaben verwendet, die in Form eines Projektes durchgeführt werden – wie etwa die Produktentwicklung. Eine nützliche Übersicht weiterer Vorgehensmodelle der Software- sowie der generellen Produktentwicklung und eine Einordnung ihrer Anwendbarkeit findet sich bei Sharafi (2013).

2.2.2 Bedeutsame Vorgehensmodelle des KDD

Zur Durchführung des KDD existiert in der Literatur eine Vielzahl von Ansätzen, die sich in Umfang und Inhalt der Hauptelemente von Modell zu Modell teils gravierend unterscheiden. Kurgan und Musilek (2006) betrachten verschiedene Schätzungen zum maximalen zeitlichen Aufwand der einzelnen Elemente (siehe Abbildung 2.3). Dabei ist festzustellen, dass mindestens die Hälfte – einigen Schätzungen zufolge sogar 80% – des zeitlichen Aufwandes für die vorbereitenden Schritte aufzuwenden ist. Das

eigentliche Data-Mining-Verfahren hingegen ist weit weniger aufwändig und benötigt normalerweise nur ca. 10% der Gesamtzeit.

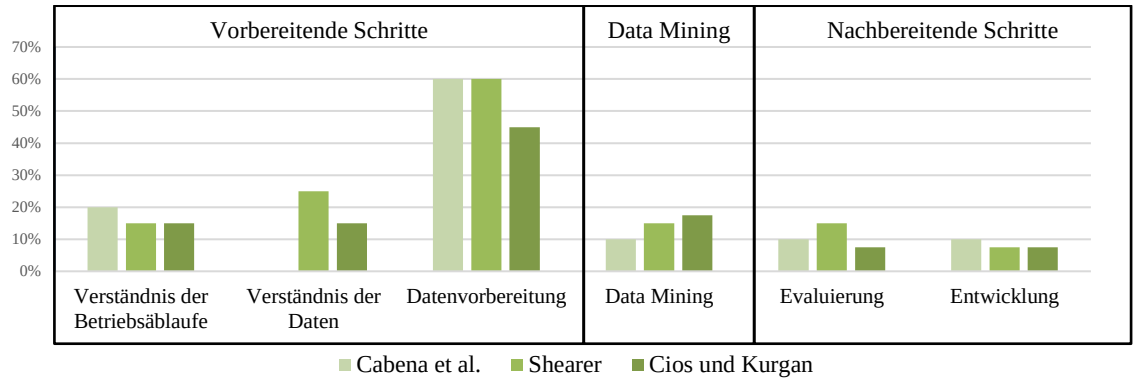


Abbildung 2.3: Anteiliger Aufwand der KDD-Elemente, eigene Darstellung nach Kurgan und Musilek (2006)

Da eine Vorstellung aller bedeutsamen Vorgehensmodelle den Umfang dieser Arbeit übersteigen würde, erfolgt an dieser Stelle lediglich eine Vorstellung einiger ausgewählter Modelle. Fayyad et al. (1996) legten mit ihrer Arbeit den Grundstein für die Forschung im Bereich des KDD, indem sie ein Modell entwickelten, das die Grundlage vieler weiterer Arbeiten bildete. Aus diesem Grund soll das KDD nach Fayyad genauer vorgestellt werden. Darüber hinaus wird auf das CRISP-DM eingegangen, da dieses Modell laut einer Umfrage zum Einsatz von Verfahren bei der Durchführung von Analyse-, Data-Mining- oder Data-Science-Projekten in 43% der 200 abgefragten Fälle eingesetzt wurde (Piatetsky-Shapiro 2014). Als drittes und letztes Vorgehensmodell wird in diesem Abschnitt das durch den Lehrstuhl ITPL der TU Dortmund entwickelte MESC betrachtet.

KDD nach Fayyad

Das Modell von Fayyad et al. (1996) basiert in seinen Grundzügen auf Beschreibungen von Brachman und Anand (1996) und umfasst insgesamt fünf Schritte (siehe Abbildung 2.4), die im Folgenden genauer erläutert werden sollen:

Auswahl (1): Diese Phase bezeichnet die Sichtung der Daten und das Entwickeln eines Verständnisses für ihren Inhalt. Darüber hinaus ist hier das Ziel des Prozesses aus Sicht des Auftraggebers festzulegen, um so die geeigneten Daten für das Data Mining ermitteln zu können *(Target Data*).

Vorverarbeitung (2): Dieser Schritt dient der Bereinigung und Vorverarbeitung der Daten für die weiteren Schritte. Dabei muss festgelegt werden, wie mit falschen oder fehlenden Daten sowie eventuellen Redundanzen umgegangen werden soll.

Transformation (3): Abhängig von der Zielstellung umfasst dieser Schritt sowohl Verfahren zur Dimensionsreduktion als auch die Anwendung von Transformationsmethoden mit dem Ziel der Reduktion der Variablenanzahl.

Data-Mining (4): In diesem Schritt gilt es zuerst ein für die Zielformulierung passendes Data-Mining-Verfahren und einen geeigneten Algorithmus zur Durchführung zu ermitteln. Anschließend findet das Data Mining im eigentlichen Sinne statt – also die Suche nach unbekannten Mustern in bekannten Daten und deren Darstellung in repräsentativer Form (z.B. Entscheidungsbaum).

Interpretation und Evaluation (5): An dieser Stelle erfolgt eine Interpretation der Data-Mining-Ergebnisse. Nach Abschluss der Interpretation kann das gewonnene Wissen genutzt werden, um weitere Schritte abzuleiten. Dies umfasst neben der direkten Nutzung des Wissens zur Anpassung der Abläufe auch eine Übertragung auf andere Prozesse oder eine Dokumentation der Ergebnisse zur Weiterleitung an Entscheidungsträger. Je nach Resultat kann es an dieser Stelle auch zu Rückschleifen in vorherige Schritte des Modells kommen. Es handelt sich beim KDD-Vorgehensmodell nach Fayyad also um ein iteratives Modell.

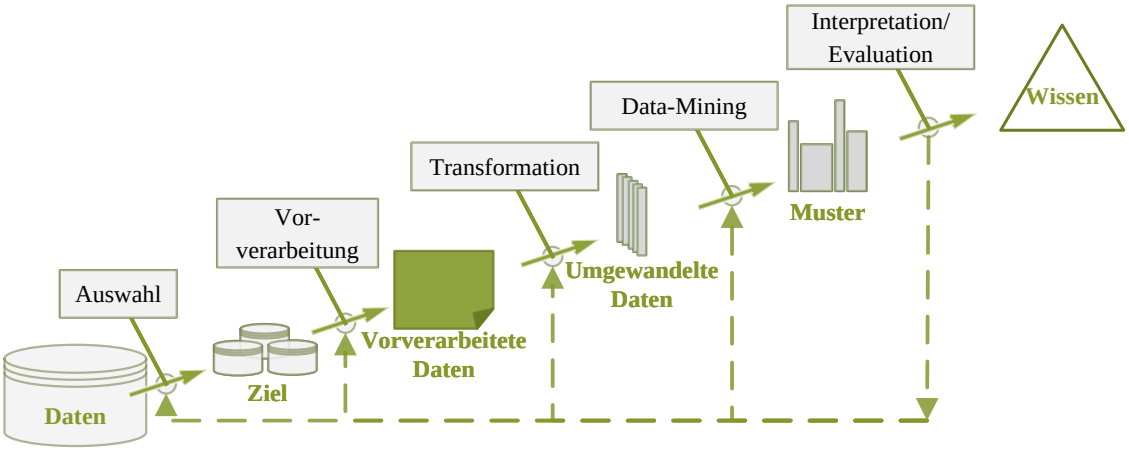


Abbildung 2.4: KDD-Prozess nach Fayyad et al. (1996)

CRISP-DM

Der ***C**ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining* (kurz: *CRISP-DM*) wurde durch ein Konsortium mit mehr als 200 Mitgliedern aus Anbietern, Unternehmensberatern und Anwendern entwickelt und ist aus diesem Grund durch seinen hohen Praxisbezug gekennzeichnet. Das CRISP-DM soll als Leitlinie verstanden werden und besteht insgesamt aus den sechs miteinander in Verbindung stehenden Phasen *Business Understanding* (Verständnis der Betriebsabläufe), *Data Understanding* (Verständnis der Daten), *Data Preparation* (Vorverarbeitung der Daten), *Modelling* (Modellierung), *Evaluation* (Evaluierung) und *Deployment* (Entwicklung), die in Abbildung 2.5 dargestellt sind (vgl. Chapman et al. 1999). Bei dem KDD nach CRISP-DM handelt es sich um einen kontinuierlichen Prozess. Dies wird in Abbildung 2.5 durch den äußeren Kreis kenntlich gemacht. Im Gegensatz zu dem Modell von Fayyad werden die Phasen des Modells nicht in einer starren Reihenfolge durchlaufen, sondern es muss teilweise zwingend zu

Rückführungen auf vorherige Phasen kommen. Dabei entscheidet das Ergebnis der Phasen darüber, welche Phase und welche spezielle Aufgabe in dieser Phase als nächstes ausgeführt werden muss. Aufgrund dieser hohen Komplexität bilden die Pfeile in Abbildung 2.5 deshalb nur die bedeutendsten und häufigsten Beziehungen ab. Im Folgenden sollen die Phasen des CRISP-DM kurz vorgestellt werden (vgl. Chapman et al. 1999):

Business Understanding (1) & Data Understanding (2): Zweck der elementaren Phasen Business Understanding und Data Understanding ist es, ein Verständnis für die Geschäftsprozesse und den Datenbestand zu erlangen und so Fehlinterpretationen zu vermeiden. Dementsprechend sind diese beiden Phasen inhaltlich nahezu deckungsgleich zu der Selektionsphase des KDD-Prozesses nach Fayyad.

Data Preparation (3): Im Rahmen der Phase der Data Preparation erfolgt die Datenvorbereitung für den eigentlichen Data-Mining-Vorgang. Hierzu zählt unter anderem das Erkennen von Dubletten oder fehlerhafter und unvollständiger Datensätze.

Modelling (4): Die Phase der Entwicklung des eigentlichen Data-Mining-Modells durch Anwendung von Data-Mining-Methoden wird als Modelling bezeichnet.

Evaluation (5) & Deployment (6): Bei der Evaluation (Evaluierung) erfolgt eine Bewertung des Modells und seiner Eignung zur Erfüllung der vorliegenden Anforderungen. Die abschließende Phase (Deployment) beschreibt die Anwendung der gewonnenen Informationen. Damit ähneln die fünfte und sechste Phase des CRISP-DM der finalen Phase des KDD nach Fayyad.

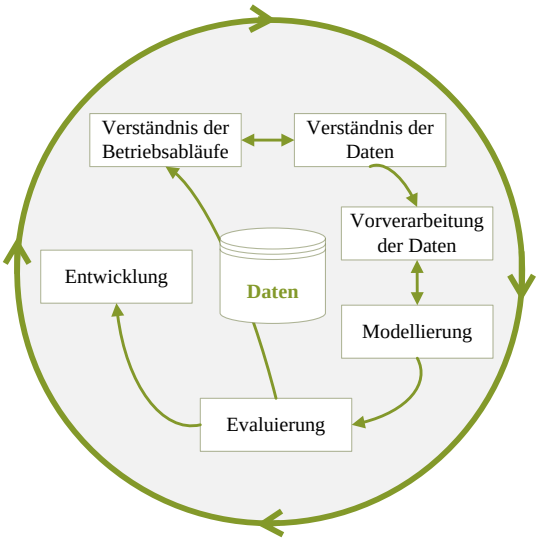


Abbildung 2.5: Phasen des CRISP-DM nach Chapman et al. (1999)

MESC

Ein weiteres Vorgehensmodell des KDD ist das durch das ITPL der TU Dortmund entwickelte *Vorgehensmodell zur Musterextraktion in Supply Chains* (kurz: *MESC*). Auch dieses umfasst über den reinen Data-Mining-Vorgang hinaus vor- und nachgelagerte

Schritte. Insgesamt besteht das Vorgehensmodell aus sieben Phasen. Da das MESC in dieser Arbeit im Rahmen eines Praxisfalls Anwendung findet, erfolgt im kommenden Abschnitt eine detailliertere Beschreibung des Vorgehensmodells sowie der darin enthaltenen Phasen.

2.2.3 MESC

Das MESC basiert nach Scheidler (2016) in seinem Aufbau auf einem für den Marketingbereich entwickelten und auf das Supply Chain übertragenen Vorgehensmodell von Hippner und Wilde (2001). Als Teil des Supply Chain Managements ist auch die *Logistik* ein mögliches Anwendungsfeld des MESC. Diese bezeichnet dabei allgemein „[die] Summe aller Tätigkeiten, die sich mit Planung, Steuerung und Kontrolle des gesamten Flusses innerhalb und zwischen Wirtschaftseinheiten befasst, der sich auf Materialien, Personen, Energie und Informationen bezieht“ (Oeldorf und Olfert 2013, S. 18). Dabei umfasst die Logistik neben den Bereichen der Beschaffung-, Distributions- und Entsorgung auch die Produktionslogistik. Ziel der *Produktionslogistik* ist die wirtschaftliche Durchführung der Produktion und damit die optimale Nutzung der Produktionskapazitäten. Zu den Aufgaben der Produktionslogistik zählen beispielsweise der innerbetriebliche Transport, die Zwischenlagerung von Teilen, die materialflussgerechte Anordnung der Maschinen oder auch die Produktionsplanung und -steuerung.

Im Folgenden soll eine Vorstellung der sieben Phasen des MESC und der darin enthaltenen Schritte erfolgen. Dabei basieren die Ausführungen auf Arbeiten des ITPL (vgl. Scheidler 2016) sowie für tiefere Beschreibungen auf dem Ursprungsmodell von Hippner und Wilde (2001):

Aufgabendefinition (1): Im Gegensatz zur Idealvorstellung der komplett datengetriebenen Analyse erfordert die Durchführung des KDD in der Praxis eine vorherige Vorgabe einer konkreten Fragestellung zur Wissensentdeckung. Hierzu wird in dieser Phase die *Aufgabenstellung* (Schritt 1.1) des KDD-Prozesses bestimmt. Dafür gilt es eine Problemstellung aus dem SCM unter Berücksichtigung gegebener Randbedingungen zu formulieren, die sowohl zeitlicher, technischer oder fachlicher Natur sein können. Im Rahmen der Produktionslogistik können fachliche Randbedingungen etwa die Art der Produktion, die Art und Anordnung der Maschinen oder die eingesetzten Transportmittel darstellen. Aus diesen gegebenen Randbedingungen sind Ziele für das Data Mining abzuleiten, die festgehalten und dokumentiert werden müssen. Das Hauptziel des Data-Minings in der Produktionslogistik ist dabei immer die Optimierung der Prozesse zur Schaffung von Wettbewerbsvorteilen. Die Ziele und Anforderungen müssen im Verlauf des KDD berücksichtigt werden und fließen zum Beispiel in die Auswahl relevanter Datenbestände oder eines geeigneten Data-Mining-Verfahrens ein.

Auswahl der relevanten Datenbestände (2): Diese Phase besteht aus den Schritten der *Datenbeschaffung* (2.1) sowie anschließender *Datenauswahl* (2.2). Bei der *Daten-*

beschaffung werden in Abhängigkeit der Aufgabendefinition Datenquellen ausgewählt. Supply-Chain-Informationen liegen selten in nur einem System, sondern vielmehr in komplexen Systemkonstrukten vor. Die Aufgabe besteht hier darin, aus den teilweise redundanten oder irrelevanten Datenbeständen mittels Kontextwissen die für die Zielsetzung relevanten Daten zu identifizieren. Kontextwissen bezeichnet hierbei Kenntnisse über Produktionsabläufe, die nicht direkt aus dem Datenbestand ableitbar sind, sondern zum Beispiel erst durch Austausch mit Experten des Unternehmens gewonnen werden können. Das Kontextwissen muss an dieser Stelle von mehreren Projektbeteiligten eingebracht werden, da eine Einzelperson nicht das komplette Konstrukt teilweise unabhängiger Systeme überblicken kann. Nach Festlegung der relevanten Datenbestände erfolgt im nächsten Schritt die *Datenauswahl* (2.2). Dabei kann die Datenmenge unter erneuter Einbeziehung von Kontextwissen reduziert werden, so dass nur noch die, für die Aufgabenstellung relevanten Informationen aus den Datenbeständen extrahiert werden müssen.

Datenaufbereitung (3): Die Phase der Datenaufbereitung ist die zeitintensivste Phase des KDD-Prozesses und dient der Erhöhung der Datenqualität. Je nach Beschaffenheit der Ausgangsdaten sind verschiedene Aktionen nötig, die sich in insgesamt vier Schritte unterteilen lassen: *Formatstandardisierung*, *Gruppierung*, *Datenanreicherung* sowie abschließende *Transformation* der Daten. Die selektierten Datenbestände stammen häufig aus relationalen Datenbanken mit einer großen Anzahl an Tabellen. Das Data-Mining-Verfahren erfordert als Standardformat aber exakt eine Tabelle, bei denen die Spalten die Attribute und die Zeilen die Datensätze darstellen. Aus diesem Grund werden in dem Schritt der *Formatstandardisierung* (3.1) die zugrundeliegenden Datenbestände in ein für das Data Mining geeignetes Format überführt. Mithilfe der *Gruppierung* (3.2) lassen sich die Datenbestände unter Berücksichtigung der Aufgabenstellung in fachliche Gruppen einteilen. Falls die zugrundeliegenden Datenbestände Lücken aufweisen, kann Kontextwissen zu *Datenanreicherung* (3.3) genutzt werden. Der Schritt der *Transformation* (3.4) dient der abschließenden Bearbeitung der Datenbestände zur Behandlung fehlerhafter Attribute, zur Reduzierung überflüssiger Attribute oder zur Reduzierung von Ausreißern.

Vorbereitung des Data-Mining-Verfahrens (4): Nachdem die Aufbereitung der Daten abgeschlossen ist, geht es in der nächsten Phase um die Vorbereitung des Data-Mining-Verfahrens. Hierzu werden wiederum vier Schritte durchgeführt, die Verfahrens- und Werkzeugauswahl sowie die fachliche und technische Kodierung. Der erste Schritt gilt dabei der *Verfahrensauswahl* (4.1). Hierbei muss aus der Masse an verfügbaren Data-Mining-Verfahren – wie Assoziationsanalyse, Klassifikation oder Clusterverfahren – das ideale Verfahren zu der in der ersten Phase definierten Aufgabe und dazugehörigen Randbedingungen ausgewählt werden. Die Wahl des anzuwendenden Data-Mining-Verfahrens sollte unter Berücksichtigung des Problemtyps sowie verschiedener Auswahlkriterien wie etwa der Generalisierbarkeit, Interpretierbarkeit oder auch der Verfügbarkeit von geeigneteren Data-Mining-Werkzeugen. Nachdem ein geeignetes Verfahren ausgewählt

wurde, gilt der zweite Schritt der finalen *Werkzeugauswahl* (4.2) zur Durchführung des Data Minings. Der Begriff Werkzeug bezeichnet hierbei beispielsweise die Software, die zur Unterstützung der Data Minings eingesetzt werden (z.B. RapidMiner, SPSS, SAP BI). Dabei ist nach Weskamp et al. (2014) der RapidMiner mit ca. 20% die am häufigsten genutzte Software.

Oftmals ist die ursprüngliche Kodierung der Daten für das ausgewählte Data-Mining-Verfahren oder das Werkzeug nicht geeignet, da diese Verfahren besondere Anforderungen an die Attribute stellen. Aus diesem Grund findet sowohl eine fachliche als auch technische Kodierung statt. Als Grundlage der *fachlichen Kodierung* (4.3) dient das zuvor erwähnte Kontextwissen. Bei der *technischen Kodierung* (4.4) geht es darum, die Ausprägungen der Attribute in ein für das Verfahren oder Werkzeug geeignetes Format zu überführen.

Anwendung des Data-Mining-Verfahrens (5): In der fünften Phase des MESC findet nun das eigentliche Data-Mining-Verfahren Anwendung. Dabei können auch mehrere Verfahren sequentiell oder parallel angewendet werden, um der unterschiedlichen Eignung der Verfahren gerecht zu werden und die Qualität der Analyse zu erhöhen. Dazu gilt es die Schritte der *Entwicklung* (5.1) sowie des *Trainings des Data-Mining-Modells* (5.2) durchzuführen. Im ersten Schritts werden dazu die Datenbestände in Trainings-, Validierungs- und/oder Testdaten unterteilt. Die Trainingsdaten sind in der Folge Grundlage der Modellentwicklung und helfen bei der Festlegung der Modellparameter (z.B. Regressionskoeffizienten bei einer Regressionsanalyse). Im zweiten Schritt dienen die Validierungsdaten der Überprüfung der Ergebnisse der Verwendung der Trainingsdaten. Bei negativer Validierung muss die Modellentwicklung fortgesetzt werden (vgl. Gottermeier 2003; Hippner und Wilde 2001).

Weiterverarbeitung der Data-Mining-Ergebnisse (6): Die Phase der Weiterverarbeitung ist in die beiden Schritte der *Extraktion handlungsrelevanter Data-Mining-Ergebnisse* (6.1) und *Darstellungstransformation der Data-Mining-Ergebnisse* (6.2) unterteilt. Dabei gilt es im ersten Schritt interessante Ergebnisse herauszustellen, die auf den Faktoren der Handlungsrelevanz sowie technischer Maßzahlen beruhen. Auch wenn die Bewertung der Interessantheit der Ergebnisse immer hauptsächlich ein subjektiver Vorgang bleibt (z.B. in Abhängigkeit des Anwenders), existieren doch einige Kriterien, die beim Herausfiltern interessanter Ergebnisse helfen können (*Validität*, *Neuheit*, *Nützlichkeit*, *Kompaktheit*, *Verständlichkeit*). Nach der Extraktion relevanter Ergebnisse gilt es im zweiten Schritt die Ergebnisse in eine für das eingesetzte Data-Mining-Verfahren sowie die Aufgabenstellung geeignete Darstellungsform – wie etwa Klassifikations- oder Assoziationsregeln – zu überführen.

Bewertung der Data-Mining-Prozesse (7): Um den Data-Mining-Prozess bewerten zu können, muss einerseits eine *Qualitätskontrolle des Data-Mining-Prozesses* (7.1) im

Hinblick auf betriebswirtschaftliche Ziele mittels geeigneter (V&V-) Maßnahmen durchgeführt werden (siehe Abschnitt 3.2). Andererseits muss eine *Rückführung von Data-Mining-Ergebnissen* (7.2) stattfinden, die dann Grundlage weiterer Data-Mining-Prozesse bilden können. Dies führt idealerweise zu dem – in Abschnitt 2.1 beschriebenen – sich stetig wiederholenden Datenanalysezyklus.

2.3 Datenvorverarbeitung

Durch die voranschreitende Digitalisierung fallen – wie in Kapitel 1 beschrieben – immer mehr Daten an. Einhergehend mit der hohen Menge und Dimensionalität der Daten sinkt tendenziell die Datenqualität, so dass ein erhöhter Aufwand nötig ist, um die nötigen Vorarbeiten durchzuführen (vgl. Gottermeier 2003). Da der Datenvorverarbeitung in jedem der zuvor aufgeführten Vorgehensmodelle eine gewichtige Rolle zukommt, sollen die verschiedenen Verfahren der Datenvorverarbeitung hier näher erläutert werden.

Säuberlich (2000) beschreibt mögliche Problematiken in den Daten, die vor einem Data-Mining-Verfahren beseitigt werden müssen. Als solche sieht er vor allem verschmutzte, fehlende oder redundante Daten sowie ein zu großes Datenvolumen an. Aus diesem Grund ist die Datenvorverarbeitung ein bedeutsamer und zeitintensiver Teil eines jeden Vorgehensmodells des KDD. Generell lassen sich die Methoden der Datenvorverarbeitung dabei in *verfahrensunabhängige* sowie *verfahrensabhängige Methoden* unterteilen. Eine ausführliche Betrachtung der im Folgenden vorgestellten und weiterer Datenvorverarbeitungsmethoden finden sich etwa bei Bramer (2013) oder Petersohn (2005).

2.3.1 Verfahrensunabhängige Methoden

Für verfahrensunabhängige Methoden ist das später eingesetzte Data-Mining-Verfahren nicht von Bedeutung. Aus diesem Grund können diese vorverarbeitenden Schritte bereits unabhängig von der späteren Auswahl des Verfahrens durchgeführt werden (vgl. Petersohn 2005). In diesem Abschnitt sollen ausgewählte Vorverarbeitungsmethoden dargestellt werden.

Datenanreicherung

Eine Anreicherung der Daten ist vor allem dann nötig, wenn es sich um ein Analyseproblem aus der Markt- oder Absatzforschung handelt. Die Anreicherung kann dabei beispielsweise durch Heranziehen externer Daten erfolgen. Darüber hinaus können aber auch interne Informationen durch das Einbringen von Kontextwissen genutzt werden (vgl. Petersohn 2005).

Datenreduktion

Data-Mining-Prozessen liegen häufig extreme Datenmengen zugrunde. Dies kann dazu führen, dass ein performante Ausführung des Data Minings nicht mehr möglich ist. Um dieser Problematik vorzubeugen, können die Daten zweckmäßig reduziert werden. Dies ist je nach Anwendungsfall in zwei Richtungen möglich. Auf der einen Seite kann die Anzahl der Attribute (*Aggregation* und *Dimensionsreduktion*), auf der anderen Seite die Anzahl der Datensätze (*Stichprobenziehung*) verringert werden (vgl. Hippner und Wilde 2001; Petersohn 2005; Weiss und Indurkhya 1998).

Aggregation: Die Aggregation bezeichnet das Zusammenschließen mehrerer Datensätze zu einem Datensatz höherer Aggregationsebene. Beispielsweise können so Brot und Butter zu der höheren Aggregationsebene Lebensmittel zusammengefasst werden. Zu beachten ist, dass das Aggregationsverfahren zwar das Datenvolumen verringert, dies allerdings auf Kosten von Informationsverlusten geschieht (vgl. Hippner und Wilde 2001; Petersohn 2005).

Stichproben: Das Ziehen von Stichproben ist notwendig, da eine Durchführung des Data-Mining-Verfahrens auf den kompletten Datenbestand aus unterschiedlichen Gründen nicht immer möglich ist. Für das Ziehen von Stichproben existieren verschiedene Verfahren. Bei all diesen Verfahren ist es von entscheidender Bedeutung, dass die Stichproben die realen Zusammenhänge in der Grundgesamtheit widerspiegeln (vgl. Petersohn 2005):

1. **Repräsentative Stichprobe:** Bei der repräsentativen Stichprobe erfolgt eine zufällige Ziehung von Stichproben. Aus diesem Grund wird diese Art der Stichprobe auch als *Zufallsstichprobe* bezeichnet. Nach Peterson (2005) ist eine Datenreduktion durch Ziehung einer Zufallsstichprobe nicht immer sinnvoll. Ausschlaggebende Kriterien sind dabei die Anzahl der gesamten Datensätze sowie das Vorkommen der als bedeutsam bewerteten Ausprägungen eines Attributs. So kann es zum Beispiel sein, dass eine solche Ausprägung nur in einem geringen Anteil der Objekte der Grundgesamtheit enthalten ist. Dadurch wird diese Ausprägung mit hoher Wahrscheinlichkeit in der Stichprobe unterrepräsentiert sein (vgl. Hippner und Wilde 2001).
2. **Geschichtete Stichprobe:** Dieses Verfahren hilft dabei, die bei der repräsentativen Stichprobe auftretenden Probleme der Unterrepräsentation zu verhindern. Dazu enthält die gezogene Stichprobe hierbei verschiedene Teilmengen der Objekte, so dass nahezu eine Gleichverteilung der Objekte hinsichtlich ihrer Ausprägungen bedeutsamer Attribute erreicht werden kann (vgl. Petersohn 2005).
3. **Inkrementelle Stichprobe:** Hierbei wird die Stichprobe im Verlauf des Data-Mining-Verfahrens schrittweise erweitert, wodurch sehr große Stichproben entstehen, die einen hohen Analyseaufwand nach sich ziehen. Dieses Verfahren ist

hauptsächlich für Analysen mit zyklischer Durchführungswiederholung geeignet (vgl. Hippner und Wilde 2001).

4. **Selektive Stichprobe:** Hierbei werden durch einen Analysten Kriterien bestimmt, die festlegen, welche Eigenschaften die Datensätze der Stichprobe aufweisen müssen, damit sie für die Auswertung relevant sind (vgl. Petersohn 2005).

Als weitere Stichprobenverfahren sind zum Beispiel das *Average Sampling* oder das *Windowing* zu nennen, die an dieser Stelle aus Relevanzgründen für diese Arbeit allerdings nicht weiter ausgeführt werden sollen. Eine ausführliche Beschreibung dieser und weiterer Verfahren findet sich etwa bei Hippner und Wilde (2001).

Dimensionsreduktion: Die Dimensionsreduktion beschreibt das Entfernen irrelevanter oder redundanter Attribute zur Reduzierung der Datenmenge. Als irrelevant können beispielsweise Attribute angenommen werden, die eine geringe Korrelation mit Klassifikationsattributen aufweisen. Redundante Attribute sind dagegen etwa durch eine hohe Korrelation mit anderen Attributen erkennbar. Beispielsweise liefern die Attribute Geburtsjahr und Alter für das Data Mining dieselben Informationen. Die Dimensionsreduktion kann auf verschiedenen Wegen erfolgen. Einerseits können Attribute aufgrund von Kontextwissen manuell entfernt werden, andererseits kann die Reduktion etwa durch die Betrachtung der Korrelationen zwischen den Attributen automatisch erfolgen (vgl. Gottermeier 2003; Petersohn 2005).

Fehlende Werte und Ausreißer

Ein Problem der Data-Mining-Verfahren sind fehlende Werte oder Ausreißer im Datenbestand. Ausreißer sind dabei gekennzeichnet durch seltenes Auftreten in der Gesamtmenge, durch Auftreten am Rand des Wertebereichs oder durch Auftreten abseits der Mehrheit der anderen Ausprägungen (vgl. Gottermeier 2003). Zwei gängige Lösungen für dieses Problem sind das Löschen des Attributs sowie das Ersetzen der leeren Werte durch den häufigsten Wert bzw. Mittelwert (vgl. Bramer 2013). Als Grenze zur Löschung eines Attributs werden bei Arndt et al. (2001) beispielsweise 80% fehlende Werte gewählt.

Normalisierung

Die heute gebräuchlichen relationalen Datenbanken weisen im Gegensatz zu anderen, „veralteten“ Datenbankmodellen – etwa Netzwerkmodellen oder hierarchische Modellen – flache Tabellen auf, die untereinander verknüpft sind (vgl. Kemper und Eickler 2015). Werden in diesem relationalen Datenbankschema nicht zusammengehörige Informationen gemeinsam gespeichert, so ist eine *Normalisierung* notwendig. Darunter wird die Aufteilung von Attributen in mehrere Relationen verstanden, um vermeidbare Redundanzen zu beseitigen. Dabei müssen die Korrektheitskriterien der Verlustlosigkeit sowie der

Abhängigkeitserhaltung beachtet werden. Die Normalisierung wird unter Anwendung sogenannter Normalisierungsregeln durchgeführt:

Erste Normalform (1NF): Durch die *erste Normalform* wird vorausgesetzt, dass alle Attribute atomare Wertebereiche aufweisen. Dies bedeutet, dass die Attribute nicht weiter zerlegt werden können.

Zweite Normalform (2NF): Die *zweite Normalform* ist gegeben, wenn sich eine Tabelle in der ersten Normalform befindet und darüber hinaus jedes Nichtschlüsselattribut von den Schlüsselattributen abhängig ist.

Dritte Normalform (3NF): Eine Tabelle befindet sich in der *dritten Normalform*, wenn sie sich in der zweiten Normalform befindet und darüber hinaus kein Nichtschlüsselattribut von einem anderen funktional abhängig ist. Eine funktionale Abhängigkeit besteht, wenn die Werte von Attributen eindeutig durch andere Attribute dieser Tabelle bestimmt werden.

Boyce-Codd-Normalform (BCNF): Die *Boyde-Codd-Normalform* stellt eine nochmalige Verschärfung der vorherigen Normalformen dar. Sie verfolgt das Ziel Informationseinheiten exakt einmal zu speichern. Falls dies zu Abhängigkeitsverlusten führt, wird nur die dritte Normalform angewendet.

Eine detailliertere Erläuterung des relationalen Datenbankmodells sowie der verschiedenen Normalformen findet sich bei Kemper (2015).

2.3.2 Verfahrensabhängige Methoden

Verfahrensabhängige Methoden können erst nach Auswahl der Data-Mining-Technik durchgeführt werden und sind deshalb genaugenommen keine Bestandteile des Datenvorverarbeitungsschritts (vgl. Petersohn 2005). Im MESC entspricht die verfahrensabhängige Datenvorverarbeitung beispielsweise eher der fachlichen (Schritt 4.3) und technischen Kodierung (4.4) in der Phase *Vorbereitung des Data-Mining-Verfahrens* (Phase 4). Aus Gründen der Übersichtlichkeit sollen diese Methoden trotzdem an dieser Stelle vorgestellt werden.

Datenmodifikation/-transformation

Für die Überführung der Attribute in die erforderliche Form für das ausgewählte Data-Mining-Verfahren existieren verschiedene Transformationsmethoden zur Änderung, Aufteilung, Zusammenführung sowie Einteilung von Attributen. Diese Methoden sollen hier erläutert werden:

Kombination oder Separierung: Durch Zusammenfügen mehrerer Attribute zu einem neuen Attribut oder durch Zerlegung eines Attributs in seine Bestandteile lassen sich neue Informationen für die Analyse gewinnen (vgl. Cleve und Lämmel 2016). So lassen sich beispielsweise Attribute – die in ID-ähnlicher Form vorliegen und so für die Assoziationsanalyse nicht genutzt werden könnten – durch Zerlegung nutzbar machen.

Diskretisierung (Binning): Dieses Verfahren bezeichnet die Einteilung von Wertebereichen in sogenannte *Bins* zur Reduzierung der Granularität. Dies bedeutet, dass feine Daten zu gröberen Intervallen zusammengefasst werden. Neben der manuellen Einteilung in Bins kann die Diskretisierung beispielsweise auch auf Grundlage der Größe oder relativen Häufigkeit der Ausprägungen erfolgen (vgl. Pyle 1999). Die einsortierten Werte können dann zur Weiterverarbeitung beispielsweise durch die Mittel- oder Grenzwerte ersetzt werden (vgl. Cleve und Lämmel 2016). Ein Beispiel könnte etwa das Zusammenfassen von Altersangaben zu Altersgruppen sein. Die Diskretisierung wird in der Literatur auch als Teil der Datenreduktion angesehen. Da sich eine sinnvolle Einteilung meist jedoch erst nach Auswahl des Verfahrens ergibt, wird die Diskretisierung in dieser Arbeit den verfahrensabhängigen Methoden zugeordnet.

Änderung des Datentyps: Die Durchführung der verschiedenen Data-Mining-Verfahren erfordert die Daten in unterschiedlicher Form. So ist etwa für die Assoziationsanalyse das Vorliegen der Daten in sogenannter binärcodierter Form notwendig (vgl. Cleve und Lämmel 2016). Es existiert eine Vielzahl unterschiedlicher Datentypen, von denen sich drei Haupttypen identifizieren lassen: nominale, ordinale sowie metrische Daten (vgl. Bramer 2013). *Nominale Daten* liegen in qualitativen Kategorien vor (z.B. Farbe eines Objekts, Geschlecht oder Beruf. Eine Sonderform nominaler Daten sind *binäre Daten*, die lediglich zwei Werte annehmen können (z.B. 1 oder 0, ja oder nein). *Ordinale Daten* sind den nominalen Daten sehr ähnlich, mit dem Unterschied, dass sich ihre Ausprägungen in eine sinnvolle Ordnung bringen lassen (z.B. klein, mittel, groß). *Metrische Daten (Integer)* bestehen aus Zahlenwerten, mit denen im Gegensatz zu nominalen Daten arithmetische Operationen durchgeführt werden können (z.B. Einkommen, Anzahl Kinder). Eine ausführliche Erläuterung der Begrifflichkeiten der Datentypen und ihrer Skalierung findet sich bei etwa Petersohn (2005).

2.4 Data-Mining-Verfahren

Zur Durchführung des eigentlichen Data-Mining-Schritts des KDD findet sich in der Literatur eine Vielzahl unterschiedlichster Verfahren, die je nach Anwendungsfall ausgewählt werden müssen. Mögliche Anwendungsfälle können die Prüfung auf Kreditwürdigkeit (Klassifikation), die Einteilung in Kundengruppen (Segmentierung) oder auch das Erkennen von Wirkzusammenhängen zwischen Items (Assoziation) sein. Passend zu den jeweiligen Anwendungsfällen finden sich diverse Data-Mining-Verfahren wie Assoziations- oder Korrelationsanalysen bei der Aufdeckung von Zusammenhängen (vgl. Runkler 2015).

Wie in Abschnitt 2.1 ausgeführt lassen sich die Anwendungsfälle des Data Minings in der Theorie generell in *beschreibende* und *vorhersagende Problemfälle (Deskriptions-* und *Prädiktionsfälle)* einteilen. Abbildung 2.6 liefert hierzu eine Übersicht ausgewählter

Problemtypen sowie darauf anwendbarer Verfahren. An dieser Stelle sei darauf hingewiesen, dass in der Literatur unterschiedliche Ansichten über die Einteilung der Verfahren in Beschreibung und Prognose existieren. So existiert neben der hier gewählten Einteilung des Assoziationsverfahrens als Beschreibungsproblem auch eine Einordnung des Verfahrens als vorhersagendes Data-Mining-Verfahren (vgl. Cleve und Lämmel 2016). Allen Verfahren gemein ist die Nutzung von Algorithmen zur Erkennung von Mustern in den Daten.

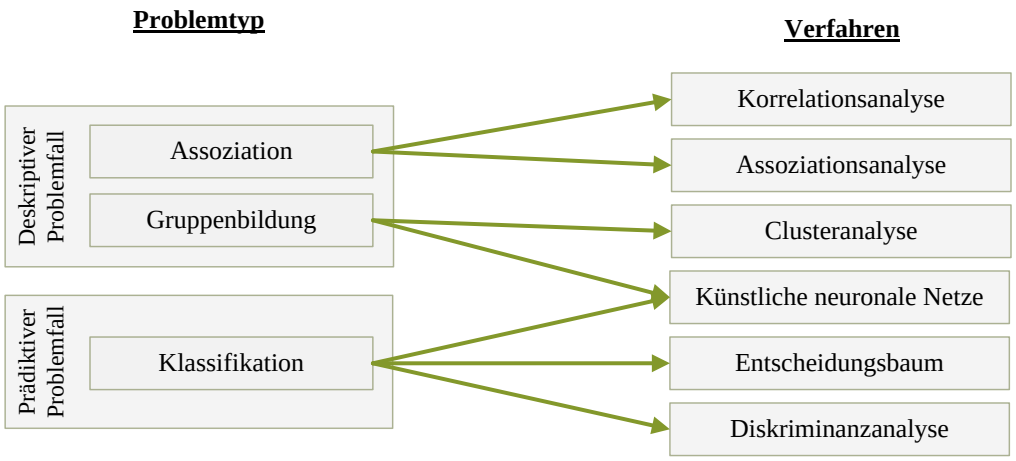


Abbildung 2.6: Problemfälle des Data Minings, eigene Darstellung nach Hippner und Wilde (2001)

2.4.1 Übersicht ausgewählter Data-Mining-Verfahren

Dieser Abschnitt dient der Einführung relevanter Verfahren des Data Minings. Dazu werden an dieser Stelle die Verfahren *Klassifikation*, *Clusteranalyse* sowie *Assoziationsanalyse* vorgestellt, da diesen in der Data-Mining-Praxis ein besonderer Stellenwert zukommt.

Klassifikation: Die Klassifikation – auch Diskriminanzanalyse – ist die geläufigste Methode des Data Minings. Ziel der Anwendung ist das Aufstellen von Klassifikationsregeln. Dabei werden mithilfe von Trainingsdaten Beispielklassen gebildet, auf deren Grundlage zukünftig Items eingeteilt werden können. Der Begriff *Items* bezeichnet hierbei beliebige, in einem Datenbestand enthaltene Objekte. *Trainingsdaten* stellen eine Stichprobe der in dem Datenbestand enthaltenen Items dar. Bekannte Beispiele der Klassifikation sind etwa das *Entscheidungsbaumverfahren*, *künstliche neuronale Netze* oder statistische Auswertungen wie die *Maximum Likelihood-Schätzung* (vgl. Cleve und Lämmel 2016).

Clusteranalyse: Auch bei der Clusteranalyse erfolgt eine Einteilung von Items in Klassen respektive Gruppen (sog. *Cluster*) aufgrund von Ähnlichkeiten. Objekte innerhalb eines Clusters sollen dabei möglichst ähnlich, Objekte unterschiedlicher Cluster möglichst unähnlich zueinander sein. Allerdings erfolgt hierbei im Unterschied zum Klassifizierungsverfahren keine vorherige Festlegung der Klassen. Vielmehr entstehen

diese erst bei Durchführung der Analyse. Anwendungsfälle wären etwa eine auf Kundengruppen zugeschnittene Werbemaßnahme oder ein optimiertes Rüsten durch Gruppierung ähnlicher Produkte, die geringe Rüstzeiten untereinander haben und so ideal zusammen gefertigt werden können (vgl. Cleve und Lämmel 2016; Weskamp et al. 2014).

Assoziationsanalyse: Ähnlich wie die Klassifikation verfolgt auch die Assoziationsanalyse das Ziel Regeln zu Korrelationen zwischen beliebigen Items zu ermitteln. Im Unterschied zur Klassifikation ist die Assoziationsanalyse dabei allerdings nicht nur auf ein Zielattribut beschränkt, sondern legt Beziehungen zwischen beliebigen Items offen und stellt diese in Form von „Wenn-Dann"-Regeln dar.

2.4.2 Assoziationsanalyse

Die Entwicklung neuer Technologien – wie Barcodescannern – und die Möglichkeit der Speicherung großer Datenmengen erleichtert die Sammlung und Analyse von Kundendaten. So lässt sich etwa mittels *Warenkorbanalyse* untersuchen, welche Produkte gemeinsam im Warenkorb eines Kunden enthalten sind. Aus diesen Informationen lassen sich im nächsten Schritt wiederum Regeln ableiten. Eine simple Regel könnte wie folgt aussehen: „Wenn ein Kunde Produkt A kauft, dann kauft er auch Produkt B". Diese Informationen können zur Verbesserung der Produktplatzierung im Supermarkt, im Katalog oder auch im Onlineshop genutzt werden. Neben dem Einsatz zur Warenkorbanalyse wird die Assoziationsanalyse heute in vielen weiteren Bereichen verwendet, etwa zur Aufdeckung (und Vorhersage) von Betrugsversuchen in der Finanzwirtschaft oder zur Optimierung von Produktionsparametern. Aufgrund dieser gesteigerten Relevanz abseits der Warenkorbanalyse hat sich die Assoziationsanalyse mittlerweile zu einem eigenen Teilgebiet des Data Minings entwickelt. Dieses ist auch unter *Association rule mining* (kurz: *ARM*) bekannt (vgl. Cleve und Lämmel 2016). Ein Einsatz der Assoziationsanalyse in der Produktion ist interessant, da hierbei unentdeckte Zusammenhänge zwischen Attributen der Produktion aufgezeigt werden können. Das Aufdecken dieser Wirkzusammenhänge kann im Idealfall zu einer Optimierung der Abläufe beitragen.

Wie in Abschnitt 2.4 dargestellt, existieren in der Literatur verschiedene Ansätze zur Einordnung der Assoziationsanalyse in die Kategorien der beschreibenden und vorhersagenden Data-Mining-Verfahren. Da die Assoziationsanalyse bereits vorhandene, aber noch nicht erkannte Zusammenhänge in Datenmengen analysiert, folgt der Autor hier der Ansicht von Hippner und Wilde (2001), die dieses Verfahren dem beschreibenden Data Mining zuordnen. Erst durch weiterführende, auf der Assoziationsanalyse aufbauende Datenanalysen, ist es möglich Hypothesen aufzustellen. Generell kann die Assoziationsanalyse als zweistufiges Verfahren angesehen werden (vgl. Han et al. 2012):

1. **Finden häufiger Itemsets:** Im ersten Schritt gilt es aus der Menge der Itemsets diejenigen herauszufinden, die mit einer gewissen Häufigkeit im Vergleich zur Gesamtmenge auftreten (vgl. Cleve und Lämmel 2016). In der Literatur werden

diese häufigen Itemsets auch als große (*large*) Itemsets bezeichnet. Itemsets ohne ausreichenden Support werden dementsprechend als kleine (*small)* Itemsets bezeichnet (vgl. Petersohn 2005).

2. **Generieren starker Assoziationsregeln:** Auf Grundlage der gefundenen häufigen Itemsets gilt es im zweiten Schritt sogenannte *starke* Assoziationsregeln zu generieren. Als *stark* werden jene Regeln bezeichnet, die sowohl das Kriterium des *minimalen Supports* als auch der *minimalen Konfidenz* erfüllen (vgl. Han et al. 2012).

Die genaueren Bedeutungen der Kriterien Support und Konfidenz soll an dieser Stelle erläutert werden. Dazu wird die vorher beispielhaft angeführte Regel etwas erweitert: „Wenn ein Kunde Produkt A kauft, dann kauft er – mit gewisser Wahrscheinlichkeit (X%) – auch Produkt B. Diese Regel ist bei Y% der Kunden zutreffend, die Produkt A kaufen“. Hierbei stellt Produkt A die sogenannte Prämisse dar, während Produkt B als Konklusion bezeichnet wird. Die Wahrscheinlichkeit X ist hier der Support, Y die Confidence.

Support: Der Support beschreibt die relative Häufigkeit eines Items in Bezug auf die Gesamtmenge und ist damit ein Maß für den Anteil der Transaktionen, die die Regel erfüllen. Ein dreiprozentiger Support sagt beispielsweise aus, dass in drei Prozent aller Transaktionen Brot und Butter zusammen gekauft wurden.

$$sup(A \rightarrow B) = \frac{|\{t \in D | (A \cup B) \subseteq t\}|}{|D|}$$

Confidence: Die Konfidenz beschreibt die Wahrscheinlichkeit, dass eine Regel der Form $A \rightarrow B$ zutrifft und damit die Stärke des Zusammenhangs zwischen den Items A und B. Dazu wird der Anteil der A und B enthaltenden Transaktionen durch die Menge aller A enthaltenden Transaktionen geteilt. Eine Konfidenz von 80% bedeutet beispielsweise, dass in 80% der Brotkäufe auch Butter dazu gekauft wurde.

$$conf(A \rightarrow B) = \frac{|\{t \in D | (A \cup B) \subseteq t\}|}{|t \in D | A \subseteq t|} = \frac{sup(A \rightarrow B)}{sup(A)}$$

Zur Durchführung der Schritte der Assoziationsanalyse existiert eine Vielzahl möglicher Algorithmen, die in Tabelle 2.1 gegenübergestellt und verglichen werden. Als eines der Standardverfahren gilt der aus dem 1993 veröffentlichten AIS-Algorithmus hervorgegangene Apriori-Algorithmus (vgl. Agrawal und Ramakrishnan 1994).

Apriori-Algorithmus

Das Ziel des Apriori-Algorithmus ist – wie in Schritt 1 beschrieben – das Finden von Itemsets aus der Menge aller Items eines Datenbestands, die einen festgelegten, minimalen Schwellwert überschreiten (minimaler Support). Diese werden beim Apriori-Algorithmus als *Frequent Itemsets* bezeichnet. Beim Apriori-Algorithmus wird eine Eigenschaft von Frequent Itemsets genutzt, die besagt, dass alle nicht-leeren Untermengen

(*Subsets*) eines häufigen Itemsets selbst auch häufig sein müssen (vgl. Han et al. 2012). Dies bedeutet im Umkehrschluss, dass ein Itemset dann kein häufiges Itemset sein kann, wenn eines der in ihm enthaltenen Subsets den minimalen Support unterschreitet. Mithilfe dieser Eigenschaft kann die Suche nach *Frequent Itemsets* als iterativer Bottom-Up-Ansatz in zwei Schritten durchgeführt werden (vgl. Hettich und Hippner 2001). Dabei werden im ersten Schritt *(Join Step)* häufige Subsets bei jedem Durchlauf um ein Item erhöht (Generierung von Kandidaten) und geprüft, ob alle Kandidaten den minimalen Support überschreiten. Kandidaten, die das Kriterium des minimalen Supports nicht erfüllen, werden im zweiten Schritt ausgeschlossen (*Prune Step)*. Der Algorithmus startet mit 1-elementigen Mengen und bricht an der Stelle ab, an der keine erfolgreiche Erweiterung der Subsets mehr möglich ist. Detailliertere Erläuterung des Apriori-Algorithmus finden sich in der Literatur etwa bei Agrawal und Ramakrishnan (1994), Cleve und Lämmel (2016), Han et al. (2012) oder Hettich und Hippner (2001).

Der Apriori-Algorithmus hat den Vorteil, dass er leicht zu implementieren ist, da es sich um simple Mengenoperationen handelt. Darüber hinaus hilft er dabei, die Anzahl zu testender Itemsets stark zu reduzieren. Da aber für jedes Itemset erneut der Support berechnet werden muss, ist die Generierung von Kandidaten durch die hohe Anzahl an notwendigen Iterationsläufen sehr zeitaufwändig (vgl. Cleve und Lämmel 2016). Aus diesem Grund existiert eine Vielzahl von Varianten und Weiterentwicklungen zum Apriori-Algorithmus. Eine dieser Weiterentwicklung ist der sogenannte *Frequent Pattern Growth*-Algorithmus (kurz: *FP-Growth*).

FP-Growth

Der große Vorteil des FP-Growth gegenüber dem Apriori-Algorithmus ist, dass dieser ohne die aufwändige Generierung von Frequent Itemsets auskommt und so einen deutlichen Geschwindigkeitsvorteil aufweist (siehe Tabelle 2.1). Dies funktioniert durch die Anwendung eines sogenannten *Divide-and-Conquer*-Ansatzes (vgl. Han et al. 2012). Die Frequent Itemsets, die bei diesem Algorithmus auch als *Frequent Pattern* bezeichnet werden, sind in einem ersten Schritt in einen *Frequent Pattern Tree* (kurz: *FP-Tree*) zu überführen. Dazu muss zuerst derselbe Scan der Datenbanken wie beim Apriori-Algorithmus durchlaufen und der Support für die einzelnen Items berechnet werden. Die Items, die das Kriterium des minimalen Supports nicht erfüllen, werden aussortiert und damit auch die Patterns, die diese Items beinhalten. Die verbleibenden werden bezüglich ihres Supports absteigend sortiert und in einen FP-Tree integriert, der die Informationen über Zusammenhänge in den Patterns darstellt (vgl. Cleve und Lämmel 2016). Im zweiten Schritt wird die jetzt vereinfachte Datenbank in eine Reihe bedingter Datenbanken aufgeteilt, die jeweils mit einem Frequent Item *(Pattern Fragment)* zusammenhängen. Nun können die bedingten Datenbanken separat mit dem Algorithmus durchlaufen werden. Dabei müssen nur noch die Datenbanken berücksichtigt werden, die mit dem Pattern Fragment in

Beziehung stehen. So kann die Anzahl der zu durchsuchenden Datensets und damit die Durchlaufzeit des FP-Growth im Vergleich zum Apriori-Algorithmus gerade bei steigendem Wachstum der Patterns deutlich reduziert werden. Für eine detailliertere Beschreibung aller in Tabelle 2.1 angeführten sowie weiterer Verfahren sei an dieser Stelle beispielhaft auf Petersohn (2005) oder Kumbhare und Chobe (2014) verwiesen. Darüber hinaus liefert Hunyadi (2011) einen ausführlichen Vergleich zwischen den hier behandelten Apriori- und FP Growth-Algorithmen zur Generierung von Assoziationsregeln.

Für die Assoziationsanalyse wird eine *Datenmenge D* mit diversen *Transaktionen t* (z.B. Einkäufen) betrachtet, die aus mehreren Items (z.B. Produkten) bestehen. Bedeutsam bei der Assoziationsanalyse ist das Kriterium der *Neuheit* der aus diesen Transaktionen abgeleiteten Regeln. Dies bedeutet, dass die Regeln bisher unbekannte Beziehungen zwischen Items aufzeigen sollen. Aus diesem Grund existieren verschiedene Maßzahlen, auch *Interessantheitsmaße* genannt, um die Relevanz einer Regel zu bestimmen. Interessantheitsmaße helfen dabei, aus der Fülle aller Regeln die interessantesten herauszufiltern (vgl. Hettich und Hippner 2001).

Tabelle 2.1: Vergleich der Algorithmen der Assoziationsanalyse nach Kumbhare und Chobe (2014)

	Ausprägung je Algorithmus				
Eigenschaft	**AIS**	**Apriori**	**AprioriTID**	**Apriori-Hybrid**	**FP-Growth**
Data Support	gering	limitiert	erscheint häufig groß	sehr groß	sehr groß
Geschwindigkeit in initialer Phase	langsam	schnell	langsam	hoch	hoch
Geschwindigkeit in späterer Phase	langsam	langsam	hoch	hoch	hoch
Genauigkeit	sehr gering	gering	genauer als Apriori	genauer als Apriori	genauer

Interessantheitsmaße

Zielsetzung der Assoziationsanalyse ist die Bestimmung interessanter Regeln zum Aufzeigen von Wirkzusammenhängen. Interessant sind meist die Regeln, die von vielen Transaktionen erfüllt werden. Deshalb hilft es für Support und Confidence untere Grenzwerte (s_{min} bzw. c_{min}) zu definieren, um so den Umfang an möglichen Regeln zu reduzieren. Hierauf basierend kann das folgende *Minimierungsproblem* für Assoziationsregeln nach Bollinger (1996) wie folgt formuliert werden: *Gegeben sei eine Menge von Transaktionen D, ein Wert für den minimalen Support* s_{min} *und ein Wert für die minimale*

Konfidenz c_{min}. Finde alle Assoziationsregeln $(A \rightarrow B)$ mit $sup(A \rightarrow B) \geq s_{min}$ und $conf(A \rightarrow B) \geq c_{min}$.

Regeln, die die Kriterien des minimalen Supports und der minimalen Konfidenz erfüllen, werden – wie im Schritt der Regelgenerierung beschrieben– als stark (strong) oder groß (large) bezeichnet. Starke Regeln müssen allerdings nicht auch zwingend interessant sein, da ein Nachteil der Konfidenz in der Nichtberücksichtigung der Wahrscheinlichkeit von B besteht. Dadurch werden Regeln erstellt, die eine Interessantheit vermuten lassen, die eigentlich nicht gegeben ist. In Bezug auf das „Brot und Butter"-Beispiel bedeutet eine 80-prozentige Konfidenz in diesem Falle eben nur, dass in 80% der Fälle, in denen Brot gekauft wurde, auch Butter gekauft wurde. Eine umgekehrte Schlussfolgerung ist auf dieser Grundlage nicht möglich. Aus diesem Grund finden sich neben den beiden bedeutsamen Maßen des Supports und der Konfidenz weitere Maße, die bei der Messung der Interessantheit und der Reduzierung der Regelanzahl helfen und hier vorgestellt werden sollen:

Lift: Die Maßzahl *Lift* (auch *Interest* oder *Strength)* ist eine Korrelationsmaß und bezeichnet die Abweichung der Konfidenz einer Regel von der erwarteten Wahrscheinlichkeit für die Konklusion dieser Regel (vgl. Hettich und Hippner 2001). Damit gibt der Lift einer Regel an, um wie viel Mal häufiger die Items in A und B in Transaktionen vorkommen als bei angenommener statistischer Unabhängigkeit der Itemsets. Ein Wert kleiner 1 (bzw. größer 1) deutet auf eine negative (bzw. positive) Korrelation. Bei einem Wert genau 1 ist die Korrelation gleich Null, d.h. es liegt eine *totale Unabhängigkeit* zwischen den Items vor. *Total abhängig* und damit *„interessant"* sind beim Lift Werte mit positiver Korrelation, also Werte größer 1. Dies weist darauf hin, dass Transaktionen, die Item A beinhalten, dazu neigen häufiger auch Item B zu enthalten, als Transaktionen ohne Item A (vgl. Bramer 2013). Allerdings muss ein hoher Wert des Lifts nicht zwangsläufig auch eine sehr interessante Regel beschreiben. So kann es beispielsweise sein, dass eine Regel mit niedrigem Lift und hohem Support interessanter als eine mit hohem Lift und niedrigem Support ist.

$$\text{lift}(A \rightarrow B) = \frac{conf(A \rightarrow B)}{sup(B)} = \frac{sup(A \cup B)}{sup(A)\ \text{sup}(B)}$$

Laplace: Die Funktion des Laplace-Maßes ähnelt in ihrem Aufbau der Konfidenz, erweitert um eine Addition mit 1 im Zähler und um eine Addition mit einer Konstanten *k*, die stets größer 1 zu wählen ist. Dadurch nimmt das Laplace-Maß einen Wert von *1/(1+k)* und *2/(1+k)* und ist dementsprechend stets kleiner als die Konfidenz ($laplace(A \rightarrow B) < conf(A \rightarrow B)$). Häufiger Einsatzbereich des Laplace-Maßes ist die Klassifikation. Idealerweise wird dazu die Konstante k mit der Anzahl der Klassen gleichgesetzt (vgl. Bayardo und Agrawal 1999; Hettich und Hippner 2001).

$$\text{laplace}(A \rightarrow B) = \frac{sup(A \rightarrow B) + 1}{sup(A) + k}$$

Gain: Der Konfidenz ebenfalls sehr ähnlich ist die Gain-Funktion nach Fukuda et al. (1996). Das Bilden einer Differenz statt eines Quotienten verhindert ein hohes Interessantheitsmaß für Regeln mit niedrigem Support. Dabei sind für den Parameter „θ“ Werte im Bereich von 0 bis 1 möglich, woraus sich für die Gain-Funktion ein Wertebereich von -θ bis *1-*θ ergibt (Hettich und Hippner 2001).

$$\text{gain}(A \rightarrow B) = \sup(A \rightarrow B) - \theta \; x \; \sup(A)$$

p-s: Die p-s-Funktion wurde durch Gregory Piatetsky-Shapiro (1991) entwickelt und stellt einen Spezialfall der Gain-Funktion dar, bei dem für den Parameter θ der Support von Itemset B ($\sup(B)$) eingesetzt wird. Dabei wird angenommen, dass der Support einer Regel bei totaler Abhängigkeit höher ist, als der Support bei angenommener totaler Unabhängigkeit. Dadurch geben Werte von 0 bis 1 eine positive Korrelation, Werte unter 0 eine negative Korrelation zwischen A und B an (vgl. Hettich und Hippner 2001).

$$\text{p-s}(A \rightarrow B) = \sup(A \rightarrow B) - \sup(A)\sup(B)$$

Conviction *(Überzeugung)*: Die Conviction ist ein Korrelationsmaß, mit einem ähnlichen Aufbau im Vergleich zum Lift. Indem es totale Abhängigkeiten besser hervorhebt, stellt es allerdings eine Verbesserung zu diesem dar. Beim Lift werden auch Werte knapp über 1 (bspw. $lift(A \rightarrow B) = 1{,}05$) trotz der Nähe zu dieser totalen Unabhängigkeit zweier Itemsets ($lift(A \rightarrow B) = 1$) als *total abhängig* bezeichnet. Bei der Conviction wird hingegen die totale Abhängigkeit durch den Wert „∞“ gekennzeichnet und somit eindeutig von der totalen Unabhängigkeit abgegrenzt. Ein weiterer Vorteil der Conviction ist, dass diese auch Veränderungen von B berücksichtigt und somit im Gegensatz zur Konfidenz keine ungewollte Resistenz gegenüber eventueller prozentualer Verschiebungen bezogen auf die Grundgesamtheit aufweist.

Das Symbol „¬“ in der Conviction-Funktion drückt eine logische Verneinung (Negation) aus, weshalb $(A \rightarrow B)$ auch als $\neg(A \wedge \neg B)$ dargestellt werden können. Wird nun das Maß für die Unabhängigkeit von $(A \rightarrow \neg B)$ bestimmt, so kann unter Berücksichtigung dieses Zusammenhangs aus dem Kehrwert des Quotienten ein Maß der Abhängigkeit von $(A \rightarrow B)$ abgelesen werden (vgl. Hettich und Hippner 2001).

$$\text{conviction}(A \rightarrow B) = \frac{sup(A)\sup(\neg B)}{sup(A \wedge \neg B)}$$

Weitere Maßzahlen der Interessantheit, auf die aus Gründen der Relevanz für diese Arbeit an dieser Stelle nicht weiter eingegangen werden soll, sind beispielsweise die Maße *Leverage* oder *Improvement*. Diese werden etwa bei Bayardo und Agrawal (1999); Bramer (2013); Fukuda et al. (1996) oder Hettich und Hippner (2001) ausführlich behandelt.

3 Verifikation und Validierung

Wie sich bereits dem Titel entnehmen lässt, stellt die V&V einen wichtigen Schwerpunkt dieser Arbeit dar. Da der V&V des KDD in der Literatur bisher wenig Beachtung geschenkt wird, soll es in dieser Arbeit auch darum gehen, geeignete Maßnahmen der V&V bei der Durchführung des KDD und speziell des MESC zu ermitteln. Einleitend hierzu soll dieses Kapitel dazu dienen, die Begrifflichkeiten der V&V einzuführen. Des Weiteren werden die Anwendung der V&V im MESC aufgezeigt und V&V-Techniken aus verschiedenen Fachbereichen dargestellt, da diese anschließend in Kapitel 4 gegenübergestellt und bewertet werden sollen.

3.1 V&V-Grundlagen und beispielhafte Einsatzmöglichkeiten in der Produktionslogistik

Dieser Abschnitt dient der Beschreibung der V&V-Begrifflichkeiten und ihrer Abgrenzung voneinander. Da dieser Abschnitt als Hinleitung zur Thematik der V&V im Data-Mining-Prozess und insbesondere im MESC dienen soll, wird hier vor allem auf die V&V bei der Erstellung von Modellen und den damit einhergehenden Aufgaben eingegangen.

Die *Verifikation* (lat. *veritas* = Wahrheit, *facere* = machen) dient generell dem Nachweis, dass ein Prozess richtig repräsentiert und mit der notwendigen Korrektheit von einem Zustand in einen anderen transformiert wird. Die Korrektheitsprüfung besteht dabei nach Brade (2003) aus einer Prüfung der *Widerspruchsfreiheit* sowie der *Vollständigkeit* des Modells. Dabei ist zu prüfen, ob alle anfänglich aufgestellten Anforderungen an das Ergebnis dieses Prozesses erfüllt wurden. Die Verifikation überprüft also die formale Korrektheit eines Modells, allerdings werden hierbei die angestrebte Anwendung des Modells und äußere Faktoren außer Acht gelassen. Denn selbst eine erfolgreich durchgeführte Verifikation bedeutet lediglich, dass dieses in sich korrekt ist, aber nicht, dass es auch das richtige Modell für den vorgesehenen Anwendungsfall ist. Aus diesem Grund sind Verfahren notwendig, die der Überprüfung der „ziel- und sachgerechten Konzeption und Durchführung der Verifikation“ dienen (Felkai und Beiderwieden 2011, S. 165). Die *Validierung* (lat. *validus* = stark) soll die Eignung eines Modells garantieren, also seine *Tauglichkeit* für den geplanten Anwendungsfall, seine *Genauigkeit* sowie seine *Fehlerfreiheit* (vgl. Balci 2013; Brade 2003). Die Zusammenhänge zwischen den angeführten Begrifflichkeiten sind in Abbildung 3.1 noch einmal übersichtlich dargestellt. Zusammengefasst lässt sich durch die Verifikation also die Frage beantworten, ob ein Modell formal richtig ist. Die Validierung beantwortet darüber hinaus, ob es sich bei dem verifizierten Modell um das richtige Modell handelt, bei dem die gegebene Ausgangslage hinreichend genau abgebildet wurde (vgl. Rabe et al. 2008).

Die folgenden Abschnitte sollen einen Überblick über verschiedene Einsatzmöglichkeiten der V&V geben. Die V&V hat für die Bereiche der Softwareentwicklung und Simulation eine immense Bedeutung. Dies lässt sich durch eine Analyse der Literatur deutlich erkennen. Dort werden sowohl die Möglichkeiten der V&V der Simulation (vgl. Balci 1998; Banks 1988; Brade 2003; Kleijnen 1995; Law 2007; Rabe et al. 2008; Sargent 2011 oder Shannon 1975) als auch der Softwareentwicklung (vgl. Balzert 2008; IEEE 2012; Liggesmeyer 2009; Sommerville 2016) zahlreich behandelt. Neben diesen beiden Hauptanwendungsgebieten wird auch auf die V&V für das eigentliche Data Mining eingegangen, da auch für diesen Schritt des KDDs bereits Verfahren existieren.

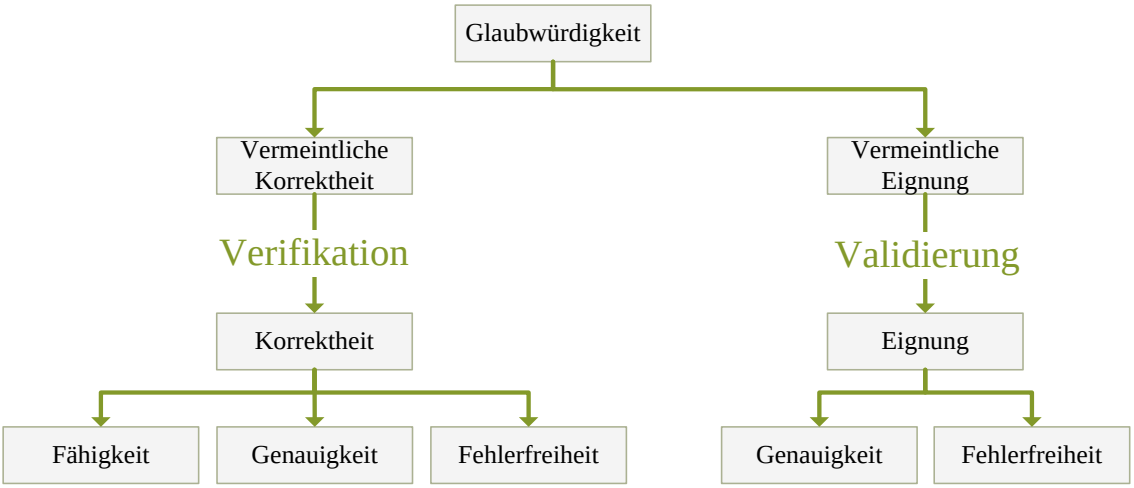


Abbildung 3.1: Zusammenhänge zwischen Begrifflichkeiten der V&V nach Brade (2003)

3.1.1 V&V in der Softwareentwicklung

Die *Softwareentwicklung* bezeichnet ein Teilgebiet der *Softwaretechnik*, die wiederum ein Teilgebiet der *Informatik* darstellt. Softwaretechnik wird durch Balzert (2009, S. 17) definiert als „zielorientierte Bereitstellung und systematische Verwendung von Prinzipien, Methoden und Werkzeugen für die arbeitsteilige, ingenieurmäßige Entwicklung und Anwendung von umfangreichen Softwaresystemen“. Hierdurch ist allerdings noch keine eindeutige Definition des Softwarebegriffs erfolgt. Im Gegensatz zu Hardware, die die materielle Teile eines Computersystems bezeichnet, umfasst Software dabei Programme, Daten und Dokumentationen, die zur Durchführung von Aufgaben mit einem Computer nötig sind (Balzert 2008). Der Begriff *Software* beschreibt also die immateriellen Teile eines Computersystems. V&V-Maßnahmen dienen in der Entwicklung von Software einerseits der Prüfung der richtigen Funktionsweise der Modelle (Verifikation), andererseits der Prüfung der Abbildungsgenauigkeit der Realität (Validierung). Da zur Überprüfung entwickelter Software meist nur einige ausgewählte Testdaten verwendet werden, kann das Ergebnis keine Gewissheit über die Korrektheit der Software für die Gesamtheit der Daten liefern. Aus diesem Grund wurden theoretische Analysemethoden zur

Überprüfung der Korrektheit der Software entwickelt, unter anderem die Softwareverifikation. Balzert (2008, S. 474) beschreibt die Verifikation als „formal exakte Methode, um die Konsistenz zwischen der Programmspezifikation und der Programmimplementierung für *alle* in Frage kommenden Eingabedaten zu beweisen“. Dazu lässt sich die Softwarevalidierung nach Sommerville (2016) in drei Phasen einteilen: *Komponententest, Systemtest* sowie *Test durch Kunden* (siehe Abbildung 3.2). Der *Komponententest* wird durch die Entwickler der Software durchgeführt. Dabei wird jede Komponente einzeln und unabhängig von den anderen Softwarekomponenten getestet. Der anschließende *Systemtest* überprüft das reibungsfreie Zusammenspiel zwischen den Komponenten und somit die Funktionsfähigkeit des Gesamtsystems. Beim abschließenden *Test durch Kunden* wird ermittelt, ob die Software den Erwartungen des Kunden entsprechen. Für die Durchführung der verschiedenen Testphasen existieren diverse Techniken, die in Abschnitt 3.3.1 vorgestellt werden.

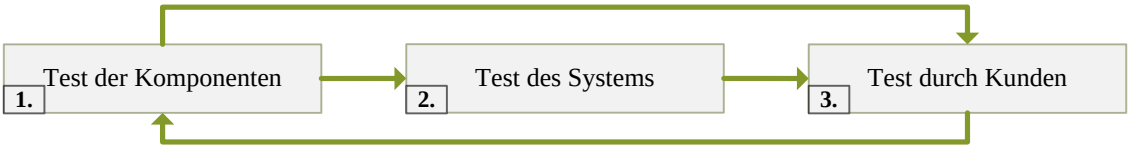


Abbildung 3.2: Testphasen der Verifikation von Software nach Sommerville (2016)

3.1.2 V&V in der Simulation

Der Begriff der Simulation bezeichnet nach VDI-Richtlinie 3633 (2008) das Überführen von dynamischen Prozessen eines Systems in experimentierfähige Modelle. Dabei sind hauptsächlich Sicherheits- und Kostengründe ausschlaggebend dafür, Probleme aus der Realität in eine abstrakte Form eines Modells zu überführen. An diesem Modell lassen sich Experimente durchführen, die in der Realität zu aufwändig, zu gefährlich oder zu teuer wären. Die Ergebnisse der Experimente können anschließend wieder auf das reale Problem übertragen werden. Simulationsmodelle können etwa in der Fabrikplanung zur Simulation von Fertigungssystemen eingesetzt werden. Um ein erstelltes Modell zu überprüfen, gibt es verschiedene Techniken der V&V. Mit diesen lässt sich zwar keine vollständige Korrektheit des Simulationsmodells garantieren, aber sie können zur Fehlererkennung in den Modellen beitragen. Aus diesem Grund beschreiben Rabe et al. (2008) das Verhindern fehlerhafter Aussagen als das übergeordnete Ziel der V&V von Simulationen in Produktion und Logistik, da diese zu falschen Schlussfolgerungen führen könnten. Zur Überprüfung des Modells sollten je nach Verwendungszweck V&V-Kriterien benannt werden (vgl. Pohl et al. 2005). Rabe et al. (2008) führen dazu verschieden Kriterien der V&V in der Simulation ein, die in Tabelle 3.1 dargestellt werden. So kann beispielsweise der Nachweis der *Korrektheit* von Inhalt und Struktur eines Modells, aber auch von Dokumenten sowie Informationen und Daten hauptsächlich

durch die Kriterien der *Vollständigkeit* und *Konsistenz* nachgewiesen werden. Während die Vollständigkeit dabei beispielsweise den Grad der Übereinstimmung zwischen gegebenen Anforderungen und Modell angibt, ist die Konsistenz u.a. ein Maß für die durchgehende Nutzung der Terminologie. Um die Angemessenheit der Ergebnisse beurteilen zu können, sind vor allem die Kriterien der *Eignung*, *Plausibilität* sowie Verständlichkeit geeignet. Die Durchführbarkeit eines Modells bezieht sich auf die organisatorische, technische sowie modelltheoretische Ebene und kann durch die Kriterien der Machbarkeit und Verfügbarkeit kontrolliert werden.

Tabelle 3.1: Bedeutsame Kriterien der Verifikation und Validierung, eigene Darstellung nach Rabe et al. (2008)

Untersuchungs-gegenstand	**V&V-Kriterium**	**Beispielhafte Fragestellung**
Korrektheit (Inhalt & Struktur)	Vollständigkeit Genauigkeit Aktualität Konsistenz	• Modell den festgelegten Anforderungen entsprechend? • Durchgehende Nutzung der Terminologie? • Fehlerhafte Modellierung? • Angemessener Detailierungsgrad?
Angemessenheit des Ergebnisses für die Anwendung	Eignung Plausibilität Verständlichkeit	• Modell für die Aufgabenstellung gültig? • Modell im Hinblick auf die Aufgabenstellung geeignet? • Ergebnisse schlüssig?
Durchführbarkeit	Machbarkeit Verfügbarkeit	• Modell lesbar und nachvollziehbar für den Anwender? • Modell technisch umsetzbar?

3.1.3 V&V im Data Mining

Das Ziel der V&V bei Data-Mining-Verfahren besteht hauptsächlich darin, die Genauigkeit eines Modells in der praktischen Anwendung aufzuzeigen. Dafür existieren verschiedene statistische V&V Techniken wie etwa die Kreuzvalidierung (siehe Abschnitt 3.3.3). Darüber hinaus können weitere statistische Techniken zur Überprüfung der Eingangs- und Ausgangswerte von Modellen herangezogen werden.

3.2 V&V im MESC

Die V&V spielt im KDD eine gewichtige Rolle und muss dementsprechend auch im MESC Berücksichtigung finden. Die Durchführung der V&V für das MESC wird durch

die Anwendung einer auf den SC-Bereich übertragenen Variante des V&V-Modells der Simulation nach Brade (2003) realisiert. Dabei werden die enthaltenen Phasen dieses Modells durch die Phasen des MESC ersetzt (vgl. Scheidler 2016). Durch die Transformation des Modells wird dieses an die im MESC vorkommenden Phasen angepasst.

V&V-Prüfungen im MESC

Ein Schwerpunkt des V&V-Modells sind Prüfverfahren, bei denen wie in Abschnitt 3.1 erläutert, phasenbegleitend sowohl Prüfungen der Phase gegen sich selbst als auch gegen die Vorphasen durchgeführt werden. Grundlage der Prüfungen gegen die Vorphase bildet die Dokumentation der entsprechenden Phase des MESC. Abbildung 3.3 liefert einen Überblick über die in den jeweiligen Phasen durchzuführenden Prüfungen (Elemente). Diese werden als zwei Elementtypen dargestellt. Die intrinsischen Prüfungen sind mit einem Kreissymbol gekennzeichnet, die Prüfungen gegen Vorphasen mit einem Pfeil in die Richtung der vorherigen Phase. Darüber hinaus sind die Elemente mit jeweils zwei Indizes in der Form *(x,y)* gekennzeichnet. Dabei kennzeichnet der erste Index das Phasenergebnis, auf das die V&V-Maßnahme durchzuführen ist. Der zweite Index gibt darüber hinaus an, gegen welches andere Phasenergebnis getestet werden soll. So zeigt beispielsweise der Index (4,3) an, dass das Ergebnis der vierten Phase gegen das Ergebnis der dritten Phase überprüft werden muss. Intrinsische Prüfungen weisen logischerweise im Index zweimal dieselbe Phase auf, da dabei die Phase gegen sich selbst geprüft wird (z.B. Element (1,1)) (vgl. Rabe et al. 2008). Durch die Prüfung der Phasen gegen die vorherigen Phasen steigt die Anzahl der durchzuführenden Prüfungen mit fortschreitender Prozessdauer kontinuierlich an. Es kommt pro Phase eine weitere Prüfung hinzu, sodass bei der letzten Phase (Phase 7) insgesamt sieben Prüfungen durchzuführen sind (siehe Abbildung 3.3).

Wichtig hierbei ist, dass die V&V-Prüfungen nicht in einer bestimmten Reihenfolge durchgeführt werden müssen. Um Irritationen vorzubeugen, werden die Phasen im Index bewusst durch Kommas statt Punkte voneinander getrennt. Auch zu beachten ist, dass die Prüfungen nicht nur am Ende der jeweiligen Phase durchzuführen sind. Zur Erkennung von Fehlern sollte validiert werden, sobald ein abgeschlossener Zwischenstand vorliegt. Des Weiteren können negative Validierungsergebnisse einer Phase auf vorhergehende Phasen zurückzuführen sein. Deshalb müssen in diesem Fall alle V&V-Elemente erneut ausgeführt werden, falls diese auf den Ergebnissen der betroffenen Phase aufbauen (vgl. Rabe et al. 2008).

Im Folgenden sollen die Prüfungen in den einzelnen Phasen des Modells erläutert werden. Da der V&V im MESC sowie in der Simulation dasselbe Ursprungsmodell von Brade (2003) zugrunde liegt, fließen hierbei teilweise auch Beschreibungen von Rabe et al. (2008) für die V&V-Elemente mit ein:

Aufgabendefinition (1): In der Phase der Aufgabendefinition ist noch keine Prüfung gegen eine Vorphase möglich. Es kann mittels intrinsischer Prüfung lediglich eine Kontrolle der Vollständigkeit und Plausibilität der bestimmten Aufgabenstellung unter Beachtung der zugrundeliegenden Randbedingungen vollzogen werden (Element (1,1)).

Auswahl der relevanten Datenbestände (2): Nachdem die Bestimmung der zu nutzenden Daten durch die Schritte der Datenbeschaffung sowie Datenauswahl erfolgt ist, muss eine intrinsische Prüfung der Daten auf ihre Korrektheit und Relevanz (2,2) durchgeführt werden. Darüber hinaus ist die Eignung der Daten im Hinblick auf die Zielbedingung (2,1) zu prüfen. Hier gilt es zum Beispiel zu erörtern, ob nach vollzogener Reduzierung der Datenmenge noch irrelevante Informationen vorhanden sind oder fälschlicherweise relevante Informationen aussortiert wurden.

Datenaufbereitung (3): Bei der Datenaufbereitung kann es durch komplexe Datenstrukturen gerade bei dem Schritt der Datentransformation leicht zu Fehlern kommen. Aus diesem Grund ist diese auf korrekte technische Ausführung zu kontrollieren (3,3). Darüber hinaus müssen die transformierten Daten auf ihre Korrektheit gegenüber den Ausgangsdaten untersucht werden (3,2). Als letztes ist in dieser Phase eine Prüfung der aufbereiteten Daten gegen die definierte Aufgabe der ersten Phase durchzuführen (3,1).

Vorbereitung des Data-Mining-Verfahrens (4): Die vierte Phase des MESC dient der Vorbereitung der Daten auf das Data-Mining-Verfahren. Dabei muss geprüft werden, ob ein geeignetes Data-Mining-Verfahren ausgewählt (4,4) und ob die Datenvorverarbeitung für dieses Verfahren korrekt angewendet wurde (4,3). Darüber hinaus gilt es erneut zu kontrollieren, ob die zugrundeliegenden Daten für das gewählte Verfahren noch immer geeignet erscheinen (4,2) und ob das Verfahren wirklich zur Beantwortung der Ausgangsfrage dienen kann (4,1) (siehe Abbildung 2.6).

Anwendung des Data-Mining-Verfahrens (5): Bei der Phase der Anwendung des Data-Mining-Verfahrens gilt es ein Modell zu entwickeln und zu trainieren. Dabei muss das Verfahren auf seine korrekte Anwendung geprüft werden (5,5). Ebenfalls Teil dieser Phase ist die Untersuchung der fehlerfreien Vorbereitung. Dabei ist beispielsweise die Auswahl des Verfahrens oder Werkzeugs für die Durchführung zu kontrollieren (5,4). Da die Datenvorverarbeitung auch von dem gewählten Verfahren abhängig ist, muss auch hier eine erneute Überprüfung hinsichtlich ihrer Eignung in Bezug auf das Verfahren erfolgen (5,3). Darüber hinaus ist die Datenauswahl auf ihre Zulässigkeit für eine fachgerechte Anwendung des Data-Mining-Verfahrens zu untersuchen (5,2). Zuletzt muss eine Prüfung der Erfüllung der Zielbedingung durch die Anwendung des ausgewählten Data-Mining-Verfahrens stattfinden (5,1).

Weiterverarbeitung der Data-Mining-Ergebnisse (6): Zur Weiterverarbeitung der Ergebnisse müssen handlungsrelevante Data-Mining-Ergebnisse gefunden, extrahiert und in eine geeignete Darstellungsform transformiert werden. In dieser Phase gilt es intrinsisch zu prüfen, ob die Ergebnisse korrekt weiterverarbeitet wurden (6,6). Daneben

ist zu kontrollieren, ob die Anwendung des Data-Mining-Verfahren eventuell keine interpretierbaren Ergebnisse liefert (6,5). Außerdem gilt es ist zu erörtern, ob das ausgewählte Data-Mining-Verfahren fachlich korrekt ist und dementsprechend theoretisch interpretierbare Ergebnisse liefern könnte (6,4). Es besteht die Möglichkeit, dass schon die Datenaufbereitung fachlich nicht richtig durchgeführt wurde, dementsprechend ist auch diese zu überprüfen (6,3). Es ist weiterhin zu analysieren, ob eine ausreichende Datenselektion für die Interpretation stattgefunden hat (6,2). Als letztes gilt es auch in dieser Phase die Phasenerkenntnisse gegen die vordefinierten Ziele der Aufgabenstellung zu kontrollieren (6,1).

Bewertung der Data-Mining-Prozesse (7): Die Bewertung der Data-Mining-Prozesse bildet den Abschluss des MESC. Dabei müssen einerseits eine Qualitätskontrolle der Prozesse und andererseits eine Rückführung der Ergebnisse stattfinden. In dieser Phase sind – wie bereits in den Phasen zuvor – verschiedene Prüfungen durchzuführen, von denen die intrinsische Prüfung aus der Untersuchung der korrekten Initiierung der Qualitätskontrolle besteht (7,7). Die restlichen Prüfungen der Bewertungsphase dienen der Kontrolle der durchgeführten Dokumentationen der vorherigen Phasen. So befasst sich die Prüfung in Element (7,6) beispielsweise mit der ausreichenden Dokumentation der Data-Mining-Ergebnisse für die Qualitätskontrolle. Da sich die restlichen Prüfungen dieser Phase hinsichtlich ihrer Durchführungsweise (Kontrolle der Dokumentationen) nicht groß von der genannten Prüfung unterscheiden, werden diese im Folgenden nur stichpunktartig aufgeführt:

- Prüfung der ausreichenden Dokumentation der Anwendung des Data-Mining-Verfahrens (7,5)
- Prüfung der ausreichenden Dokumentation der Auswahl des Data-Mining-Verfahrens (7,4)
- Prüfung der ausreichenden Dokumentation der Datenaufbereitungsprozesse und der Prozessergebnisse (7,3)
- Prüfung der ausreichenden Dokumentation des Datenauswahlprozesses (7,2)
- Prüfung der ausreichenden Dokumentation der Aufgabenstellung unter Berücksichtigung der Randbedingungen sowie Zielkriterien (7,1)

Tailoring

Das in Abbildung 3.3 dargestellte Modell der V&V im MESC ist eine generalisierte Darstellung und enthält deshalb jedes mögliche V&V-Element. Allerdings unterscheiden sich Data-Mining-Prozesse sowohl in Umfang als auch Komplexität. Aus diesem Grund muss das Modell je nach gegebener Aufgabenstellung überprüft und angepasst werden. Dadurch lässt es sich im Idealfall allerdings auch deutlich vereinfachen. Dieser Vorgang der Festlegung auf tatsächlich relevante Elemente wird *Tailoring* genannt (vgl. Rabe et al. 2008).

Dokumentation der V&V

Nach Kleijnen (1995) ist die V&V ein bedeutsamer Bestandteil zur Bewertung von Modellen. Anwender, die nicht direkt am Entstehungsprozess beteiligt sind, sollen so in die Lage versetzt werden zu beurteilen, ob die Ergebnisse eines Modells als Grundlage für Entscheidungen dienen können. Deshalb ist es zur Nachvollziehbarkeit der Durchführung der V&V sowie etwaiger Anpassungen empfehlenswert, eine V&V-Dokumentation zu pflegen. Dazu müssen für jedes der V&V-Elemente verschiedene Aspekte festgehalten werden, die bei Rabe et al. (2008) für die Simulation beschrieben werden. Dazu zählen der Gegenstand der Prüfung, eine Auflistung und Begründung weggelassener Prüfungen, die eingesetzte Techniken der V&V, der Name des ausführenden Experten, der Versionsstand der genutzten Phasenergebnisse sowie die Ergebnisse der jeweiligen Prüfung.

Durch die Dokumentation für die einzelnen V&V-Elemente des Vorgehensmodells entsteht eine Reihe von *V&V-Reporten*, die zu einem V&V-Gesamtdokument zusammengefasst werden können.

An dieser Stelle ist anzumerken, dass den Arbeiten zur V&V der Simulation nach Rabe et al. (2008) sowie der V&V des MESC nach Scheidler (2016) dasselbe Ursprungsmodell zugrunde liegt. Aus diesem Grund werden die ursprünglich für die Simulation beschriebenen Ausführungen zum Tailoring und der Dokumentation an dieser Stelle auch für das MESC als gültig angenommen.

3.3 V&V-Techniken

Nach der generellen Einführung der V&V-Begrifflichkeiten (Abschnitt 3.1) sowie der Vorstellung der im MESC durchzuführenden V&V-Maßnahmen (Abschnitt 3.2), dient der folgende Abschnitt der Vorstellung der generell für den Einsatz im MESC in Frage kommenden V&V-Techniken. Wie zuvor dargelegt, existieren speziell auf das KDD ausgelegte V&V-Maßnahmen in der Literatur bis dato nicht. Aus diesem Grund müssen V&V-Techniken aus anderen Anwendungsgebieten abgeleitet werden. Da diese Techniken nicht SC-spezifisch sein müssen, lassen sich auch geeignete Techniken aus anderen Bereichen einsetzen. Eine Darstellung aller in der Literatur benannten V&V-Techniken würde den Umfang dieser Arbeit allerdings weit übersteigen. Deswegen wird an dieser Stelle im Hinblick auf den Einsatz der Techniken im KDD bereits eine Vorauswahl auf Grundlage der folgenden Ausschlusskriterien getroffen:

1. Da es sich bei der Untersuchung der V&V-Techniken (Kapitel 4) auf Anwendbarkeit im KDD um einen *allgemeingültigen* Ansatz handeln soll, werden hauptsächlich Techniken aufgeführt, die eine gewisse Relevanz in der Literatur aufweisen. Aus diesem Grund werden beispielweise innovative Techniken bei der Betrachtung außer Acht gelassen.

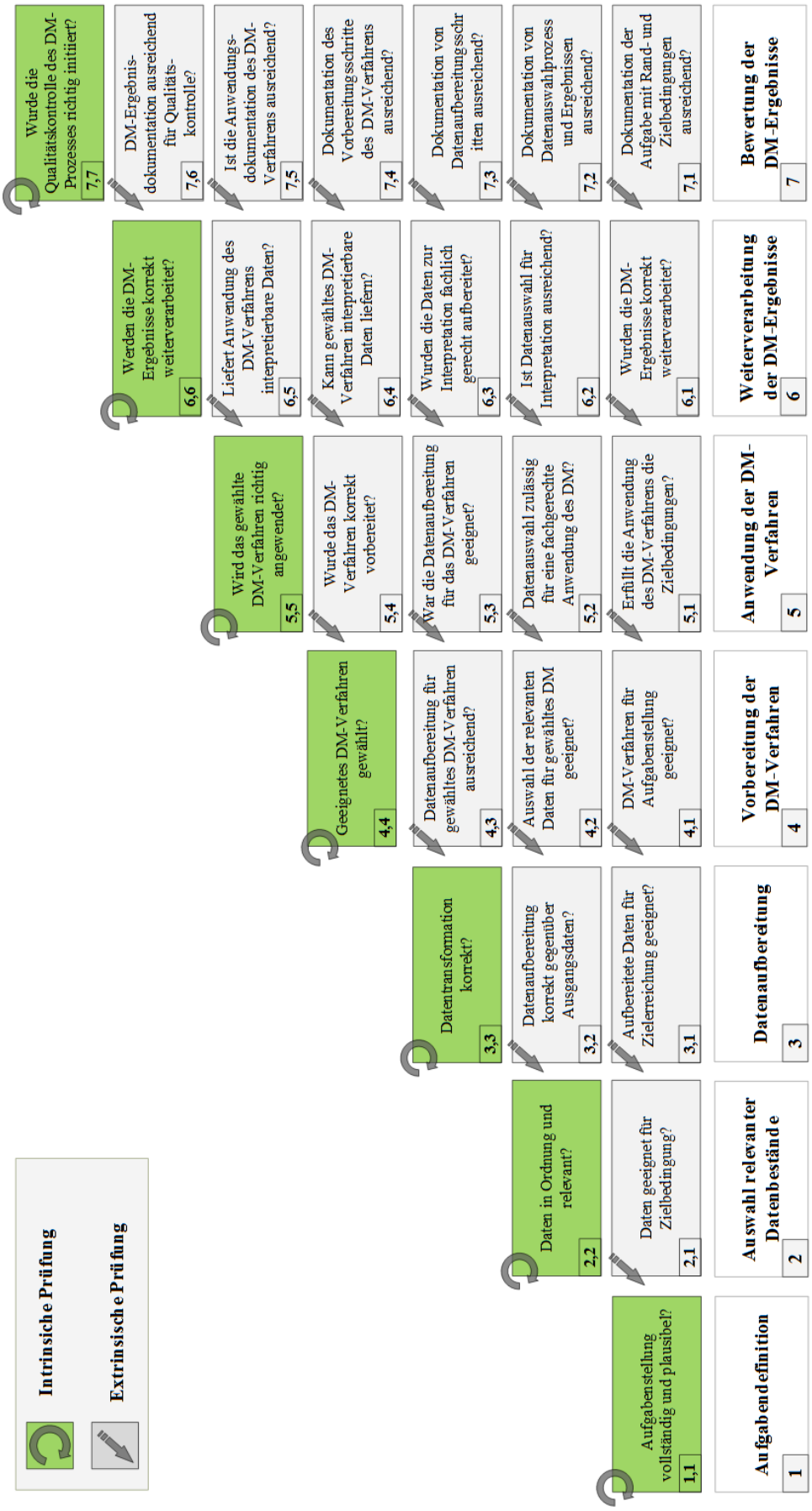


Abbildung 3.3: Prüfungsfragen der V&V-Elemente im MESC, eigene Darstellung nach Scheidler (2016)

2. V&V-Techniken, die die Untersuchung von Codes zum Ziel haben, werden an dieser Stelle vernachlässigt. Dies lässt sich wiederum mit der Allgemeingültigkeit dieser Arbeit begründen, da die meisten Data-Mining-Prozesse heute in der Regel mithilfe von Standardsoftware durchgeführt werden. Eine Studie des Fraunhofer-Institut aus dem Jahr 2014 zeigt passend dazu, dass lediglich acht Prozent der durchgeführten Data-Mining-Prozesse mit selbst geschriebener Software erfolgten (vgl. Weskamp et al. 2014). Genau genommen ist selbst bei diesen Eigenentwicklungen eine Durchführung von V&V-Techniken nicht erst bei der Anwendung, sondern bereits bei der vorherigen Entwicklung der Software nötig.
3. Techniken, die zeitliche Zustände in einem dynamischen Modell darstellen sollen, eignen sich hervorragend für die Durchführung von Simulationen. Für Data-Mining-Modelle sind sie allerdings nicht geeignet, da es sich bei diesen um statische Modelle handelt, deren zugrundeliegende Daten zu einem exakten Zeitpunkt erhoben wurden. Aus diesem Grund sind etwa visuelle V&V-Techniken, die auf die grafische Darstellung des Modells abzielen – wie etwa die Animation oder das Monitoring – ebenso wenig anwendbar wie etwa der Ereignisvaliditätstest, da auch bei dieser Technik eine Betrachtung verschiedener Ereignisse in zeitlicher Abfolge erfolgt (vgl. Rabe et al. 2008).
4. Techniken, die mit Hilfe von Manipulationen der Eingangswerte durchgeführt werden wie Äquivalenzklassenbildung oder Grenzwertanalyse – sind für einen Einsatz im KDD nicht geeignet. Die Ursache hierfür ist, dass das Data Mining genau für solche Datenbestände angewendet wird, die aufgrund ihrer Komplexität nicht mit herkömmlichen, „klassischen“ Datenanalyseverfahren untersucht werden können.
5. Formale Techniken der V&V für die Simulation – wie die Induktionsbehauptung oder der formale Korrektheitsbeweis – sind nach Rabe et al. (2008) nur in Ausnahmefällen in der Praxis anwendbar. Dabei wird auf Balci (1998, S. 378) verwiesen: “Current state-of-the-art proof of correctness techniques are simply not capable of being applied even to a reasonably complex simulation model”. Dieser Logik für Simulationsmodell folgend, werden auch bei der Auswahl von V&V-Techniken des KDD formale Techniken außer Acht gelassen.

Unter Berücksichtigung der genannten Kriterien spiegelt der Großteil der im Folgenden vorgestellten Techniken die V&V in der Simulation wider. Dies ist hauptsächlich dadurch bedingt, dass Rabe et al. (2008) verschiedene Ansätze anderer Disziplinen in ihren Ansatz der V&V in der Simulation einfließen lassen. Nichtsdestotrotz sollen auch Techniken anderer Bereiche – wie der Softwareentwicklung – in diesem Abschnitt Berücksichtigung finden und separat vorgestellt werden.

3.3.1 V&V-Techniken in der Softwareentwicklung

Nach Balzert (2008) können Testverfahren der V&V von Programmen generell in zwei Hauptverfahrensarten eingeteilt werden, in *dynamische Testverfahren* sowie *statische Testverfahren*. Darüber hinaus lassen sich sogenannte *diversifizierende Testverfahren* identifizieren (Burgdorf 2010). Abbildung 3.4 liefert eine Übersicht über die Klassifikation der Testverfahren zur V&V von Software, die in diesem Abschnitt vorgestellt werden sollen.

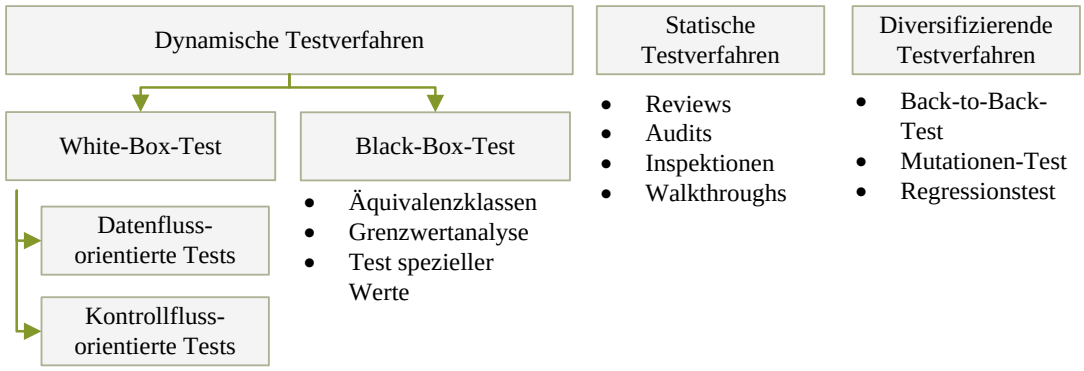


Abbildung 3.4: Testverfahren zur Softwarevalidierung, eigene Darstellung nach Balzert (2008)

Dynamische Testverfahren

Merkmaler *dynamischer Testverfahren* sind konkrete Eingabewerte, eine reale Testumgebung sowie das Durchführen einer Stichprobe. Aus diesem Grund kann durch dynamische Testverfahren die Korrektheit eines Programmes nicht vollständig bewiesen werden. Dynamischen Testverfahren lassen sich unterteilen in *White-Box-Testverfahren* (auch*: Strukturtestverfahren*) und *Black-Box-Testverfahren* (auch: *Funktionale Testverfahren*) (vgl. Balzert 2008).

White-Box-Testverfahren (selten auch: *Glass-Box-Testverfahren*) betrachten – wie der Name andeutet – die innere Struktur eines Programmes. Dies bedeutet, dass dieses Verfahren den Code eines Programms prüft. Damit sind alle White-Box-Testverfahren (wie *Anweisungs-, Pfad-, Zweig- oder Bedingungsüberdeckungstests*) – wie zuvor in Ausschlusskriterium 2 dargelegt – als V&V-Technik des KDD ungeeignet.

Black-Box-Testverfahren nutzen im Gegensatz zu *White-Box-Testverfahren* die Spezifikation des Programms zur Erstellung von Testfällen und lassen dabei die innere Struktur des Programms außer Acht. Dieses Verfahren kommt meist erst beim Systemtest (siehe Abschnitt 3.1.1) zum Einsatz, da bei vorhergehenden Phasen noch keine ausreichend genaue Spezifikation vorliegt (vgl. Liggesmeyer 2009). Da bei dieser Testweise eine Manipulation der Eingangswerte durchgeführt werden muss, sind die Black-Box-Testverfahren (z.B. *Äquivalenzklassenbildung* oder *Grenzwertanalyse*) nach Kriterium 4 für eine Betrachtung als V&V-Technik des KDD nicht geeignet. Lediglich das Verfahren

des *Testens spezieller Werte* erscheint für eine spätere Überprüfung lohnenswert und soll deshalb an dieser Stelle vorgestellt werden.

Test spezieller Werte (Error-Guessing): Bei diesem Verfahren wird das Ziel verfolgt, Testfälle zusammenzustellen, die möglichst viele Fehler aus der Vergangenheit beinhalten. Da hierfür meist über die Spezifikation hinausgehende Informationen benötigt werden, ist dieses Verfahren eigentlich nicht den Black-Box-Testverfahren zuzuordnen. Allerdings folgt der Autor hierbei Balzert (2008), der diese Einteilung damit begründet, dass die drei genannten Verfahren häufig in Kombination eingesetzt werden.

Statische Testverfahren

Mithilfe *statischer Testverfahren* können Algorithmen oder Programme, aber auch Dokumente überprüft werden. Eine Ausführung des Programms findet hierbei nicht statt, sondern nur eine statische, manuelle Analyse des Quellprogramms (vgl. Balzert 2008). Häufig eingesetzte statische Testverfahren sind das *Debugging sowie Inspektionen, Reviews* und *Walkthroughs*. Da das Debugging einen reinen Korrekturvorgang von Programmiercodes darstellt, ist es für das KDD nicht anwendbar (Ausschlusskriterium 2). Die weiteren Techniken können neben der Kodierung auch zur Prüfung von z.B. Dokumenten genutzt werden. Darum erfolgt hier eine kurze Vorstellung dieser Techniken:

Reviews: Als Review werden nach dem IEEE Standard for Software Reviews and Audits 1028 Prozesse oder Meetings bezeichnet, in denen ein bzw. mehrere Softwareprodukte oder ein Softwareprozess durch Projektbeteiligte und/oder andere interessierte Personen ausgeführt, kommentiert und abgenommen werden (IEE 2008).

Audits: Als Audits wird die unabhängige Überprüfung von Softwareprodukten oder -entwicklungsprozessen durch externe Prüfer verstanden. Dadurch soll die Erfüllung der Anforderungen (z.B. Spezifikation, Standards) erreicht werden (vgl. IEEE 2008).

Inspektionen (Inspections): Das Ziel der Inspektionen ist die Identifikation schwerwiegender Fehler im betrachteten Programm. Dazu werden das Produkt sowie seine Teilprodukte einschließlich des Erstellungsprozesses überprüft. Inspektionen werden hauptsächlich durchgeführt, um im Entwicklungsprozess entstandene Teilprodukte für die nächste Entwicklungstätigkeit freizugeben. Die Prüfung wird durch ein kleines Team von drei bis sieben Mitgliedern durchgeführt, in dem die Rollen eines Moderators, eines Autors (Entwickler), Protokollführers und eines Inspektors vorkommen. Dabei kann eine Person auch mehrere Rollen besetzen. Als Ergebnisse einer Inspektion sollen ein formalisiertes Inspektionsprotokoll mit einer Einteilung der entdeckten Fehler in leichte und schwere Fehler sowie Verbesserungsvorschläge für den Prozess und ein überarbeitetes Prüfobjekt entstehen (vgl. Balzert 2008).

Strukturiertes Durchgehen (Structured Walkthroughs): Ein Walkthrough dient ebenso wie die Inspektion der Identifizierung von Fehlern und Problemen des zu prüfenden Objekts. Allerdings kann ein Walkthrough auch das Ziel haben, das Produkt einem

Benutzer oder Mitarbeiter vorzuführen und ihn auszubilden. Im Gegensatz zur *Inspektion* ist hierbei eine Überarbeitung des Prüfobjekts ausdrücklich nicht das Ziel. Ein Walkthrough kann bereits ab einer Gruppengröße von zwei Personen durchgeführt werden. Dabei stellt der Autor, der gleichzeitig auch als Moderator fungiert, das Objekt Schritt für Schritt vor. Die Gutachter stellen dazu Fragen, um etwaige Probleme zu erkennen. Ein Vorteil des Walkthroughs gegenüber der Inspektion ist der geringere Aufwand und die Möglichkeit, das Wissen über den Entwicklungsprozess weiteren Personen zugänglich zu machen. Nachteilig ist allerdings, dass der Autor durch die gleichzeitige Ausführung der Rolle des Moderators die Sitzung dominieren und die Gutachter täuschen kann. Ein weiterer Nachteil ist, dass die Überarbeitung des Objekts kein Ziel des Walkthroughs darstellt. Dadurch bestimmt einzig und alleine der Autor, ob und auf welche Weise die aufgedeckten Fehler behoben werden sollen (vgl. Balzert 2008; IEEE 2008).

Diversifizierende Testverfahren

Von diversifizierenden Tests spricht man, wenn bei einem Test unterschiedliche Objekte miteinander verglichen werden. Es erfolgt dabei kein Vergleich der Testergebnisse zu den in der Spezifikation gegebenen Randbedingungen, sondern ein reiner Vergleich der Ergebnisse zweier (oder mehrerer) Tests. Häufig eingesetzte diversifizierende Testverfahren sind der sogenannte Back-to-Back-, der Mutationen- sowie der Regressionstest.

Back-to-Back-Test: Beim Back-to-Back-Test wird ein Objekt mit gegebener Spezifikation durch mehrere, voneinander unabhängig arbeitende Teams entwickelt. Die dabei entstandenen Ergebnisse werden bei diesem Verfahren miteinander verglichen und gegeneinander getestet. Nachteilig an diesem Verfahren ist, dass die Parallelentwicklung desselben Objekts ohne jegliche Schnittstellen zwischen den Entwicklerteams mit enormen Kosten verbunden ist, weshalb sich dieses Testverfahren im Regelfall nicht rentiert (vgl. Burgdorf 2010).

Mutationen-Test: Dieses Testverfahren dient nicht der Überprüfung eines Testobjekts auf seine Korrektheit, sondern der Überprüfung der Eignung eines bestimmten Testverfahrens. Durch das Einfügen von Fehlern in ein als korrekt erachtetes Objekt entstehen sogenannte Mutationen. Diese helfen dabei zu überprüfen, ob verwendete Testverfahren die implementierten Fehler entdecken (vgl. Liggesmeyer 2009).

Regressionstest: Beim Regressionstest werden wiederholt dieselben Testfälle eingesetzt, um bereits überprüfte Teile eines Testobjekts nach einer Veränderung (z.B. Korrektur) erneut zu testen und so vergleichbare Ergebnisse vor und nach der Modifikation zu erhalten. Ein Problem dieses Verfahrens ist die Festlegung einer maximal tolerierbaren Abweichung zwischen den Ergebnissen (vgl. Burgdorf 2010).

3.3.2 V&V-Techniken in der Simulation

Dieser Abschnitt dient der Vorstellung von V&V-Techniken in der Simulation, von denen ein Teil seinen Ursprung in der Softwareentwicklung hat. Balci (1998) unterteilt die Techniken in vier Bereiche. Neben einer Einteilung in statische und dynamische Verfahren nimmt Balci eine weitere Unterteilung in informale und formale Techniken vor. Damit ergibt sich die in Abbildung 3.5 dargestellte Einteilung in insgesamt vier Bereiche. Diese sind nach ansteigender Komplexität von links (informale Techniken) nach rechts (formale Techniken) angeordnet. Der Anstieg der Komplexität ist dabei unter anderem darauf zurückzuführen, dass formale Techniken einen hohen Anteil an mathematischen und logischen Formalismen aufweisen (vgl. Balci 1998). Mit dem steigenden Grad der Komplexität steigt allerdings auch der Grad der Objektivität der Techniken von links nach rechts.

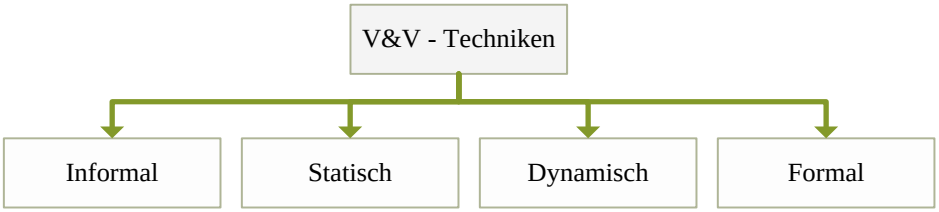


Abbildung 3.5: Einteilung der V&V-Techniken der Simulation nach Balci (1998)

In der Literatur finden sich zahlreiche Autoren, die sich mit dieser Thematik der Simulation auseinandergesetzt haben. Dabei wird immer wieder auch auf die Arbeiten von Balci (1998), Rabe et al. (2008) und Sargent (2011) verwiesen. Aus diesem Grund werden sich die Ausführungen in diesem Abschnitt hauptsächlich auf diese Arbeiten beziehen. Bei der Strukturierung folgt der Autor der gewählten Einteilung von Balci nach aufsteigender Komplexität der Techniken in informale, statische, dynamische sowie formale Techniken. Nach Ausschlusskriterium 5 sind formale Techniken für diese Arbeit allerdings nicht relevant und werden deshalb an dieser Stelle nicht weiter betrachtet. Darüber hinaus werden bei Balci V&V-Techniken, die bei der Softwareentwicklung als statische Techniken eingestuft werden (Review, Inspektion etc.), als informale Techniken eingeordnet. Dieser Einordnung folgend werden an dieser Stelle diese statischen Techniken in der Gruppe der informalen Techniken erläutert.

Informale V&V-Techniken

Nach Balci (1998) sind informale Techniken durch ihre Subjektivität und den Bedarf an menschlichem Denken gekennzeichnet. Deshalb werden hier keine formalen mathematischen Operationen eingesetzt. Allerdings können auch informale Techniken strukturierten, formalen Abläufen folgen. Richtig eingesetzt sind diese Techniken sehr effektiv und gehören deswegen zu den meistgenutzten.

Audit: Neben der Softwareentwicklung kann ein *Audit* auch bei der Simulation durchgeführt werden. Damit kann geprüft werden, wie adäquat die Simulationsstudie in Bezug auf entwickelte Pläne, Abläufe oder Richtlinien ist. Bei Auftreten eines Fehlers sollte dieser durch den Auditpfad bis zu seiner Quelle zurückverfolgt werden können. Auditierungen finden meist in periodischem Abstand oder bei Erreichung eines Meilensteins statt und werden als Mischung aus Meetings, Beobachtungen und Prüfungen durchgeführt. Audits dienen dem Management hauptsächlich als Möglichkeit der Kontrolle ihrer Mitarbeiter (vgl. Balci 1998).

Begutachtung (Review): Die *Begutachtung* dient der Überprüfung des Qualitätsgrades der Durchführung und des Ergebnisses der Simulation. Dies geschieht immer im Austausch mit dem Management sowohl auf Auftraggeber- als auch Auftragnehmerseite. Dadurch soll gewährleistet werden, dass die Simulation in Einklang mit den Anforderungen verläuft. Bedeutsam ist bei der Review-Technik, dass sich die einzelnen Teilnehmer beispielsweise unter Nutzung des *Schreibtischtests* auf das Review vorbereiten. Darüber hinaus gilt dies auch für das gesamte Team – etwa durch Anwendung des *strukturierten Durchgehens*. Hilfreich für die Bewertung der begutachteten Phasenergebnisse können die in Abschnitt 3.1.2 eingeführten V&V-Kriterien sein (vgl. Balci 1998; Rabe et al. 2008).

Dokumentenüberprüfung (Documentation Checking): Die Prüfung der Dokumentation dient der Absicherung der Korrektheit, Konsistenz, Vollständigkeit und Eindeutigkeit aller vorliegenden Dokumente. So kann es beispielsweise vorkommen, dass die Dokumentation veraltet ist oder die Modelllogik falsch festgehalten wurde (vgl. Balci 1998).

Schreibtischtest (Desk Checking): Beim *Schreibtischtest* geht es um die Kontrolle der eigenen Arbeit in Bezug auf die Kriterien der Vollständigkeit, Konsistenz und Eindeutigkeit, weshalb auch von einer Selbstinspektion gesprochen werden kann (vgl. Balci 1998). Der Schreibtischtest ist gerade in den ersten Schritten der Entwicklung sinnvoll. Problematisch bei dieser Technik ist, dass man seine eigenen Fehler selbst beim nochmaligen Durchgehen nicht immer erkennt. Deshalb ist es ratsam eine weitere Person hinzuzuziehen. Einerseits wird durch das Erklären des Handelns die Wahrscheinlichkeit erhöht den Fehler zu finden und andererseits kann so die Kompetenz der anderen Person mit einfließen. Bestandteile des Schreibtischtests sollten z.B. das Überprüfen des geschriebenen Codes, der Syntax oder der Querverweise sein.

Inspektionen (Inspections): Ursprünglich handelt es sich bei den Techniken der Inspektion um eine Technik der Softwareentwicklung (vgl. Abschnitt 3.3.1). Daneben können Inspektionen auch auf Modellierungen angewendet werden (vgl. Balci 1998).

Strukturiertes Durchgehen (Structured Walkthrough): Dieses Verfahren stellt ebenso wie die Inspektion ursprünglich eine Technik der Softwareentwicklung dar, um

die Anweisungen eines Programms in einem Team gemeinsam durchzugehen (vgl. Abschnitt 3.3.1). Sie kann aber auch für den Einsatz in der Simulation zu einer Managementtechnik erweitert werden und so für den Einsatz zur Kontrolle der Phasenergebnisse genutzt werden. Bedeutsam ist hierbei, dass die Prüfung von einem heterogenen Team aus Simulations- und Fachexperten durchgeführt wird. Dabei sollte jedoch kein direkter Projektbeteiligter involviert sein, um ein unvoreingenommenes Überprüfen der Ergebnisse zu ermöglichen. Diese Technik bietet sich hauptsächlich für in Dokumenten festgehaltene Annahmen an. Dabei sind die Dokumente solange zu diskutieren, bis ein gemeinsamer Konsens über die Sachverhalte erzielt wurde. Die Interaktion während der Treffen ist ein wesentlicher Bestandteil dieser Technik, weswegen sie nicht durch eine unabhängige Prüfung durch Fachexperten ersetzt werden kann (vgl. Balci 1998; Rabe et al. 2008; Sargent 2011).

Turing-Test: Für den Turing-Test werden Experten zwei Outputs eines Prozesses vorgelegt. Auf der einen Seite das Output des Modells, auf der anderen das des Systems. Dabei werden beide Verfahren unter Beachtung derselben Eingangsbedingungen durchgeführt. Das Expertenfeedback entscheidet über das weitere Vorgehen. Ist es dem Experten nicht möglich zwischen den Verfahren zu unterscheiden, dann bildet das Modell das System mit ausreichender Genauigkeit ab. Ist es dem Experten jedoch möglich Unterschiede zwischen den Ergebnissen festzustellen, kann dieses Feedback zur Verbesserung des Modells und so zur Annäherung der Ergebnisse genutzt werden (vgl. Balci 1998; Landry et al. 1983).

Validierung im Dialog (Face Validity): Die *Validierung im Dialog* beschreibt den Vergleich von Modell- und Systemverhalten unter Nutzung derselben Randbedingungen. Dabei wird unter Einbeziehung verschiedener Beteiligter – wie etwa Projektteammitglieder, spätere Anwender oder Experten – beurteilt, ob das Modell nachvollziehbare Ergebnisse liefert (vgl. Balci 1998). Vorteilhaft ist bei dieser Technik das schnelle Erkennen von Fehlern im Modell durch die Beteiligten oder auch schon zuvor durch den Simulationsexperten bei der Erläuterung des Modells gegenüber anderen Beteiligten. Andererseits gibt es auch Gefahren, die bei Rabe et al. (2008) mit Verweis auf Hermann (1967) angeführt werden. So kann es beispielsweise sein, dass die Beteiligten nicht ausreichend mit dem realen System vertraut sind, das Systemverhalten nicht korrekt auf das Modell übertragen können oder aus persönlichen Motiven gar nicht daran interessiert sind, fehlerhafte Zusammenhänge als solche zu identifizieren.

Dynamische V&V-Techniken

Der Einsatz dynamischer Techniken setzt die Entwicklung von Modellen voraus. Diese Modelle können dann auf ihr Verhalten bei der Ausführung untersucht werden (vgl. Balci 1998).

Akzeptanztest (Acceptance Test): Der *Akzeptanztest* wird entweder durch den Kunden selbst, durch die Entwickler im Beisein der Kunden oder durch einen unabhängigen Dienstleister durchgeführt. Ein Dienstleister wird nach Fertigstellung des Modells engagiert – aber vor Akzeptierung des Ergebnisses durch den Kunden (vgl. Balci 1998).

Alpha-Testing: Das Alpha-Testing beschreibt das Testen des fertiggestellten Modells in einem internen Bereich, der nicht in die Entwicklung des Modells involviert war (vgl. Balci 1998).

Beta-Testing: Das Beta-Testing beschreibt das Testen der sogenannten Beta-Version des Modells durch einen kleinen Kreis von Testusern und unter Einbeziehung realistischer Bedingungen. Im Gegensatz zum *Alpha-Testing* wird beim Beta-Testing eine nicht fertige Version vorab getestet (vgl. Balci 1998).

Statistische Techniken (Statistical Techniques): Statistische Techniken können dazu genutzt werden, ein Modell sowie seine Eingangs- und Ausgangsdaten zu validieren. Bei der Modellvalidierung helfen statistische Techniken bei der Bewertung des Modellverhaltens im Vergleich zum realen System (vgl. Rabe et al. 2008). Darüber hinaus ist es möglich statistische Techniken zur Überprüfung der zugrundeliegenden Daten zu nutzen – etwa durch die Analyse der Korrelation zwischen Attributen eines Datensatzes. Weiterhin kann beispielsweise untersucht werden, ob Daten verschiedener Quellsysteme richtig zusammengefasst wurden. Darauf basierend wird entschieden, ob Daten gemeinsam oder getrennt voneinander betrachtet werden müssen (vgl. Law 2007). Bei der Überprüfung von Ergebnissen der Modellbildung können statistische Techniken in Kombination mit Aufzeichnungen eines realen Systems als Vergleichsgrundlage (siehe *Vergleich mit aufgezeichneten Daten*) eingesetzt werden (vgl. Rabe et al. 2008). Da statistische Techniken im Rahmen dieser Arbeit nicht ausgeführt werden sollen, erfolgt hier nur eine kurze Beschreibung einiger Techniken. Eine Übersicht verschiedener Verfahren sowie Verweisen zu detaillierten Ausführungen zu den Techniken finden sich etwa bei Balci (1998).

- **Korrelationsanalyse:** Wie in Abschnitt 2.3.1 beschrieben, kann die *Korrelationsanalyse* genutzt werden, um den linearen Zusammenhang zwischen zwei Merkmalen zu analysieren. Dabei gelten Attribute als positiv korrelierend, wenn sie voll voneinander abhängig sind (vgl. Abschnitt 2.4.2). Allerdings ist die Korrelationsanalyse nicht geeignet, um Aussagen zu den Ursachen dieser Abhängigkeit zu treffen.
- **Regressionsanalyse:** Um den genannten Nachteil der Korrelationsanalyse auszugleichen, bietet sich die Anwendung einer *Regressionsanalyse* an. Diese kann dabei helfen, Ursache und Wirkung des Zusammenhangs festzustellen (vgl. Bosch 2015).
- **Chi-Quadrat-Test (Chi-Square-Test):** Ein weiterer Nachteil der Korrelationsanalyse besteht darin, dass mit dieser Technik lediglich lineare Zusammenhänge

zwischen Merkmalen überprüft werden. Dadurch können starke, nichtlineare Zusammenhänge fälschlicherweise als schwache Korrelation identifiziert werden könnten. Zur Quantifizierung nichtlinearer Zusammenhänge zwischen Merkmalen hilft der sogenannte *Chi-Quadrat-Test*, der die Merkmale auf stochastische Unabhängigkeit überprüft (vgl. Runkler 2015).

Test von Teilmodellen (Submodel Testing): Dieses Verfahren erfordert eine hierarchische Unterteilung des Modells in Teilmodelle. Bei Ausführung des Gesamtmodells können Ein- und Ausgangsparameter der Teilmodelle aufgezeichnet und anschließend zu Validierungszwecken mit den Parametern des Gesamtmodells abgeglichen werden. Die Validierung der einzelnen Teilmodelle kann allerdings nie direkt die V&V des Gesamtmodells ersetzen, sondern eignet sich nur zu Ergänzung zu diesen (vgl. Balci 1998; Rabe et al. 2008).

Validierung von Vorhersagen (Predictive Validation): Hierbei wird das Modell genutzt, um eine Vorhersage über das zukünftige Systemverhalten zu treffen. Voraussetzung hierfür ist ein real existierendes System, das im nächsten Schritt mit dem erstellten Voraussagemodell abgeglichen wird. Dieser Abgleich erfolgt entweder auf Basis von vorliegenden Daten aus IT-Systemen oder Beobachtungen bzw. von Messungen im realen System (vgl. Sargent 2013). Vorteil dieser Technik ist, dass die Beschreibung des Systemverhaltens erst nach der Erstellung des Modells erfolgt. Dadurch wird eine bewusste oder unbewusste Manipulation des Modells verhindert (vgl. Rabe et al. 2008).

Vergleich mit anderen Modellen (Comparing Testing): Ergebnisse des betrachteten Modells werden mit Ergebnissen anderer (valider) Modelle verglichen. Bedingung dafür ist, dass mehr als ein Modell desselben Systems zum Testen zur Verfügung steht. Dies können entweder Ergebnisse andere Simulationsmodelle oder aber auch analytische Modelle sein (vgl. Balci 1998; Sargent 2011).

Vergleich mit aufgezeichneten Daten (Historical Data Validation): Frühere aufgezeichnete Daten werden genutzt, um mit einem Teil der Daten ein Modell zu bauen, das dann zur Validierung mit dem restlichen Teil der Daten gegen das System getestet werden kann. Für diese Technik müssen logischerweise historische Daten vorhanden sein, für die ein reales System existieren muss (vgl. Sargent 2011). Ist dies nicht der Fall, kann nur die zuvor beschriebene *Validierung von Vorhersagen* durchgeführt werden. Problematisch kann bei dieser Technik sein, dass die Datenbestände aus der gleichen Quelle stammen und damit enthaltene systematische Fehler sowohl in den zur Modellbildung als auch in den zum Modelltest genutzten Daten enthalten sind. Dadurch kann dieser Vergleich fälschlicherweise eine nicht vorhandene Gültigkeit des Modells suggerieren. Um die Glaubwürdigkeit der Modellbildung zu erhöhen, ist über den Vergleich mit aufgezeichneten Daten hinaus ein Einsatz *statistischer Techniken* notwendig (vgl. Rabe et al. 2008).

3.3.3 V&V-Techniken im Data Mining

Auch zur Durchführung des Data Minings existieren bereits Techniken der V&V, die in den drei Hauptelementen des KDD eingesetzt werden können. Dabei eignen sich vor allem **statistische Techniken** (vgl. Abschnitt 3.3.2) zur Untersuchung des Datenmodells sowie seiner Eingangs- und Ausgangsdaten. Dabei kommen hier neben den in Abschnitt 3.3.2 genannten Techniken (Korrelationsanalyse, Regressionsanalyse, Chi-Quadrat-Test) weitere Techniken – wie etwa Hypothesentests, Signifikanztests oder Ziehungen von Stichproben – zum Einsatz. Auch die bei der Assoziationsanalyse berechneten Interessantheitsmaße – etwa Support und Konfidenz – stellen beispielsweise eine Form der V&V dar, da hierbei bereits irrelevante Regeln herausgefiltert werden. Ausführliche Beschreibung des Einsatzes statistischer Techniken im Rahmen des Data Minings finden sich etwa bei Runkler (2015) oder Walter (2004).

An dieser Stelle soll genauer auf die Ziehung von Stichproben eingegangen werden, da diese für den im Verlauf erläuterten Praxisfall eine erhöhte Relevanz besitzt. In Abschnitt 2.3.1 wurde die Stichprobenziehung bereits als Verfahren der Datenvorverarbeitung zur Reduzierung der Datenmenge betrachtet. Auch zur V&V der Data-Mining-Verfahren sind stichprobenbasierte Verfahren notwendig. Hauptsächlich werden Methoden des sogenannten *Resamplings* durchgeführt. Resampling bezeichnet das Bestimmen statistischer Eigenschaften von Stichprobenfunktionen durch wiederholtes Ziehen von Stichproben aus einer einzigen Ausgangsstichprobe (vgl. Rönz und Strohe 1994). Dadurch lässt sich die Vorhersagegüte eines Modells überprüfen und damit auch Generalisierbarkeit für neue Daten (vgl. Steinlein 2004). So kann auch gezeigt werden, ob sogenanntes *Over-* bzw. *Underfitting* vorliegt. Werden etwa für eine Regressionsanalyse 80% der Fälle in den Trainingsdaten korrekt klassifiziert, in den Testdaten allerdings deutlich weniger, spricht man von *Overfitting*. Dabei wird das Modell überschätzt – es hat eine höhere Trefferrate als die Grundgesamtheit. *Underfitting* bezeichnet dementsprechend das exakte Gegenteil (vgl. Schendera 2014).

Ein Nachweis der Generalisierbarkeit anhand der Trainingsdaten ist nicht möglich, da diese bereits zur Modellbildung eingesetzt wurden. Dadurch kann lediglich die Korrektheit des Modells für die Trainingsdaten nachgewiesen werden, nicht aber, wie gut ein Modell für die Anwendung auf unbekannte Daten geeignet ist (vgl. Steiner 2009). Für die Durchführung der Generalisierbarkeit existieren mehrere Resampling-Methoden, von denen Clarke et al. (2009) Kreuz- und die Bootstrapping-Validierung als bedeutsamste Methoden ansehen. Steiner führt darüber hinaus noch die *Holdoutmethode* als etablierte Methode an.

Holdoutmethode (auch Split-Validierung): Bei der Holdoutmethode werden die Daten in Stichproben zu Trainings-, Validierungs- und/oder Testzwecken aufgeteilt. Die Trainingsdaten dienen dabei zur Bildung des Data-Mining-Modells, das dann mittels Validierungsdaten überprüft und verbessert werden kann – etwa durch Anpassung der

Parameter. Testdaten werden eingesetzt, um die zuvor beschriebene Generalisierbarkeit – also die Güte des Modells – nachzuweisen. Ist die Modellgüte nach Durchführung der Holdoutmethode akzeptabel, so ist dadurch die Generalisierbarkeit gewährleistet, da die Beurteilung anhand unterschiedlicher Daten (Trainings- und Testdaten) erfolgt ist (vgl. Steinlein 2004; Steiner 2009). In der Literatur finden sich mehrere Empfehlungen zur Aufteilung der Daten. Ein Ansatz sieht eine Unterteilung des Datenbestands in lediglich zwei Teilbestände zum Lernen (Trainingsdaten) und Testen (Testdaten) vor (vgl. Cios et al. 2007). Dies geschieht im Idealfall im Verhältnis 2:1. Breiman et al. (1998) schlagen eine ähnliche Aufteilung vor, mit dem Unterschied, dass ihr Ansatz die Einteilung in Validierungs- statt Testdaten vorsieht. Berry und Linoff (2000) empfehlen hingegen eine Aufteilung in 60% Trainings-, 30% Validierungs- und 10% Testdaten. Ein weiterer Ansatz von Urban (1998) sieht vor, dass 75-90% der Daten als Trainings- und Validierungsdaten im Verhältnis 3:1 und die restlichen Daten (10-25%) als Testdaten genutzt werden. Neben den angeführten Ansätzen existieren weitere Ansätze zur Verteilung der Stichproben (vgl. Cios et al. 2007; Malthouse und Blattberg 2005; Murthy, K., Salzberg, S. 1995; Zahavi und Levin 1997). Diese werden beispielsweise bei Steiner (2009) oder Steinlein (2004) ausführlicher vorgestellt und erläutert.

Die Durchführung der Holdoutmethode mit Einteilung in Trainings-, Validierungs- und Testdaten setzt nach Hofmann (1990) eine große Datenmenge voraus. Er benennt hierfür eine Mindestanzahl von 1000 Datensätzen, bei deren Überschreitung die Holdoutmethode den anderen Methoden überlegen ist. Auch bei geringerer Anzahl an Datensätzen ist eine Durchführung der Holdoutmethode theoretisch möglich, allerdings nur unter Verzicht auf Testdaten und Verwendung der Validierungsdaten zum Modellvergleich. Dies kann unter Umständen zu dem beschriebenen Phänomen des Overfittings bezüglich der Validierungsdaten führen (vgl. Bishop 1995).

Kreuzvalidierung (Cross-Validation oder X-Validation): Das Kreuzvalidierungsverfahren ist ein statistisches Verfahren zur Evaluierung und zum Vergleich von lernenden Algorithmen. Dazu wird der Datenbestand wie bei der Split-Validierung in zwei sich gegenseitig ausschließende Teilbestände unterteilt. Der Unterschied zur Split-Validierung besteht darin, dass bei der Kreuzvalidierung das Verfahren solange mit neu unterteilten Teilbeständen wiederholt wird, bis jedes Objekt des Datenbestands einmal in der Testmenge verwendet wurde (vgl. Gottermeier 2003; Ross et al. 2009). Die Kreuzvalidierung ist hauptsächlich für die Durchführung von Data-Mining-Verfahren mit geringerer Datenmenge kleiner 1000 geeignet. Zwar werden alle Daten durch den Einsatz als Trainings- oder Testdaten optimal genutzt, allerdings steigt durch die Unterteilung in mehrere Teilmengen und das Durchlauf des Prozesses für diese Teilmengen die Gesamtzeit der Validierung mit Anzahl der Datensätze an (vgl. Kohavi 1995a; Steinlein 2004).

Bootstrapping-Validierung (Bootstrapping Validation): Diese Technik gleicht in seinen Grundzügen den zuvor beschriebenen Techniken der Split- und Kreuzvalidierung.

Dabei kommt die Bootstrapping-Validierung zum Einsatz, wenn die theoretische Verteilung der interessanten Statistiken unbekannt ist. Eine ausführliche Beschreibung findet sich bei Efron (1979).

Neben den hier aufgeführten Techniken existiert noch eine Vielzahl weiterer Validierungsmöglichkeiten, die an die obigen Techniken angelehnt oder von diesen abgeleitet sind – etwa Inkrementelle Kreuzvalidierung, Batch-Cross-Validation, Wrapper Split Validation, Wrapper-Cross-Validation oder Cross-Prediction. Aufgrund ihrer Ähnlichkeit zu den aufgeführten Techniken und mit Verweis darauf, dass sie in der Literatur wenig Beachtung finden (Ausschlusskriterium 1), wird an dieser Stelle auf eine weitergehende Behandlung dieser Techniken verzichtet. Beschreibungen dieser und weiterer ähnlicher Techniken finden sich beispielsweise bei Hofmann und Klinkenberg (2014) oder Kohavi (1995b).

4 Erläuterung von V&V-Techniken für Data-Mining-Prozesse in der Produktionslogistik

Wie in Kapitel 2 aufgezeigt, besteht die Herausforderung eines Data-Mining-Prozesses hauptsächlich in der Datenvorverarbeitung (bis zu 80% der Gesamtprozessdauer) (vgl. Abschnitte 2.1 und 2.2.2). Dies gilt aufgrund der oftmals niedrigen Datenqualitätsstandards insbesondere für Data-Mining-Prozesse der Produktionslogistik (vgl. Abschnitt 2.1). Daraus resultierend besteht gerade bei diesen Prozessen ein erhöhter Bedarf an V&V-Maßnahmen, um die Korrektheit der Datenvorverarbeitungsschritte zu gewährleisten. Nach der Einführung der V&V und ihrer Techniken in Kapitel 3 dient dieses Kapitel nun der Analyse der generellen Anwendbarkeit der V&V-Techniken für Data-Mining-Prozesse in der Produktionslogistik. In einem ersten Schritt sollen deshalb die Techniken auf ihre generelle Verwendbarkeit beim Data Mining in der Produktionslogistik überprüft werden, bevor in einem zweiten Schritt analysiert wird, welche Verfahren im vorliegenden Fall für die Anwendung des MESC relevant sind.

4.1 Untersuchung der V&V-Techniken für Data-Mining-Prozesse in der Produktionslogistik

Aufbauend auf den vorherigen Kapiteln dient dieser Abschnitt dazu, die in Abschnitt 3.3 eingeführten V&V-Techniken miteinander zu vergleichen und auf ihre generelle Anwendbarkeit im KDD zu analysieren. Dies erscheint notwendig, da nicht jede der aufgeführten Techniken zwangsläufig auch für einen Einsatz in diesem Bereich geeignet sein muss. Zur Verbesserung der Übersichtlichkeit erfolgt auch in diesem Abschnitt eine Unterteilung in V&V-Techniken von Softwareentwicklung, Simulation und Data Mining.

4.1.1 Softwareentwicklung

Wie bei der Betrachtung der V&V-Techniken der Softwareentwicklung festgestellt (vgl. Abschnitt 3.3.1), lassen sich diese in drei Kategorien unterteilen – in dynamische, statische sowie diversifizierende Techniken. Von den dynamischen Testverfahren erscheint einzig das **Testen spezieller Werte** für einen möglichen Einsatz geeignet. In seiner ursprünglichen Form dient dieses dem Zusammenstellen von Testfällen mit bekannten Fehlern der Vergangenheit unter Zuhilfenahme von Expertenwissen. Auch beim KDD-Prozess wäre eine solche Zusammenstellung bekannter Fehler interessant. Hierbei könnten durch Ziehung einer selektiven Stichprobe (siehe Abschnitt 2.3.1) gezielt Attribute oder Attributsausprägungen untersucht werden, die in der Vergangenheit Fehler aufgewiesen haben. Auch hier wäre Expertenwissen notwendig. Dieses Verfahren könnte also

im KDD in Kombination mit anderen Verfahren zur Gewinnung von Kontextwissen – etwa **Review** oder **Inspektion –** angewendet werden.

Statische Techniken erscheinen im Gegensatz zu dynamischen Techniken größtenteils geeignet für den Einsatz im KDD zu sein. Da diese Techniken aber auch bei der Simulation eingesetzt werden, erfolgt eine Diskussion an späterer Stelle.

Diversifizierende Techniken (Back-to-Back-Test, Mutationen-Test und Regressionstest) stellen hauptsächlich die Ergebnisse der Modelle in den Mittelpunkt. So werden beim **Back-to-Back-Test** die Entwicklungsergebnisse zweier parallel und unabhängig voneinander arbeitender Teams verglichen. Dies führt dazu, dass zwischen den Teams keinerlei Erkenntnisse etwa bezüglich Datenvorverarbeitung ausgetauscht werden können. Übertragen auf das KDD in der Produktionslogistik und das dort erforderliche hohe Maß an Kontextwissen, würde diese Parallelbearbeitung neben der doppelten Arbeit der Teams auch einen erhöhten Arbeitsaufwand für die Fachexperten bedeuten, da diese zur Gewinnung des Kontextwissens von beiden Teams herangezogen werden müssten. Aus diesem Grund erscheint die Durchführung theoretisch möglich, aber praktisch nicht lohnenswert. Eine parallele Ausführung des KDD erscheint nur dann sinnvoll, wenn die verfahrensunabhängigen Methoden zu Datenvorverarbeitung von einem Team gemeinsam angewendet werden und auf dieser Grundlage mehrere unterschiedliche Verfahren parallel entwickelt und ausgeführt werden – etwa Assoziations- und Clusteranalyse. Auch dabei sollte unter den Teammitgliedern zwingend weiterhin ein Austausch von hilfreichem Kontextwissen erfolgen, um eine Doppelbelastung speziell für den Fachexperten zu verhindern.

Mithilfe des **Mutationen-Tests** lassen sich Testverfahren auf ihre Eignung hin überprüfen. Dafür werden Fehler in ein korrektes Objekt eingefügt und nachfolgend kontrolliert, ob diese Fehler durch das verwendete Testverfahren entdeckt werden. Da in dieser Arbeit beispielsweise lediglich die generelle Einsatzmöglichkeit statistischer Techniken behandelt wird, könnte der Mutationstest eventuell für weitere – über diese Arbeit hinausgehenden – Untersuchungen zur Nutzung einzelner statistischer Techniken im Rahmen des KDD eingesetzt werden.

Der **Regressionstest** wird angewendet, um Teilobjekte nach Veränderungen unter Nutzung derselben Testfälle erneut zu testen. Dadurch können die Ergebnisse vor und nach der Änderung verglichen werden. Übertragen auf das KDD könnte eine solche Veränderung eines Teilobjekts beispielsweise das Weglassen eines Attributs oder das Ändern von Methodenparametern bei Durchführung des Data Minings bedeuten (vgl. Abschnitt 2.2.3).

Abschließend lässt sich bezüglich diversifizierender Techniken festhalten, dass diese nur in Abwandlung für das KDD geeignet sind und deshalb eine genauere, über die vorliegende Arbeit hinausgehende Betrachtung der Einsatzmöglichkeiten geprüft werden sollte.

4.1.2 Simulation

Den V&V-Techniken für die Simulation wurden in Abschnitt 3.3 die größte Aufmerksamkeit gewidmet. Dies geschieht vor dem Hintergrund, dass der dort aufgeführte Ansatz nach Rabe et al. zur V&V in der Simulation – wie zuvor beschrieben – bereits Techniken aus verschiedenen anderen Disziplinen aufgreift und für die Simulation adaptiert. Betrachtet man die in Abschnitt 3.3.2 aufgeführten Techniken der V&V in der Simulation, so ist festzustellen, dass einige Techniken für das KDD grundsätzlich nicht geeignet erscheinen. Für die Simulation sind visuelle Techniken – etwa die Animation – sehr nützlich, um Prozesse grafisch darzustellen und verständlich zu machen. Dies ist allerdings für das KDD nicht möglich, da hier keine dynamische Prozessnachbildung erfolgt, sondern statische Modelle verwendet werden. Wie beschrieben besteht bei KDD-Prozessen der Produktionslogistik ein erhöhter Bedarf an V&V-Maßnahmen zur Kontrolle der Ausgangsdaten. Deswegen scheinen gerade für die V&V-Elemente der Datenvorverarbeitung (vgl. Abschnitt 2.2.2) hauptsächlich solche Techniken geeignet, die der Überprüfung der Korrektheit der Daten dienen. Die Durchführung dieser Maßnahmen erfordert größtenteils Kontextwissen, um die Ergebnisse der Phasen richtig einordnen und mit den wahren Gegebenheiten überprüfen zu können. Diese Techniken scheinen neben den vorbereitenden Schritten auch für das Data-Mining-Verfahren sowie die nachbereitenden Schritte geeignet zu sein.

Die Techniken **Audit, Begutachtung, Inspektion, Strukturiertes Durchgehen, Schreibtischtest** und **Validierung im Dialog** zeichnen sich alle dadurch aus, dass bei diesen Techniken ein Objekt oder Teile des Objekts einer Überprüfung mittels Begutachtung unterzogen werden. Diese Untersuchung erfolgt durch eine Gruppe von mindestens zwei Personen. Dabei unterscheiden sich die Verfahren hauptsächlich im Zeitpunkt der Durchführung, der Zusammensetzung der Gruppe sowie der Verwertung der Ergebnisse. Bei der **Begutachtung (Review)** findet die Überprüfung sowohl durch Kunden- als auch Dienstleisterseite statt, um zu gewährleisten, dass die Anforderungen erfüllt werden. Die Durchführung eines Reviews ist jederzeit möglich. Das **Audit** wird in regelmäßigen Abständen oder bei Erreichung eines Meilensteins durch meist externe Kontrolleure durchgeführt und hat hauptsächlich die Kontrolle der Mitarbeiter durch das Management zum Zweck. Einen Meilenstein kann beispielweise der Abschluss einer Entwicklungsphase darstellen. Ebenfalls am Ende einer solchen Phase wird die **Inspektion** durchgeführt, bei der durch ein Team hauptsächlich Teile eines Objekts überprüft und im Idealfall für die nächste Entwicklungsphase freigegeben werden. Am Ende einer Inspektion sollte ein Protokoll mit entdeckten Fehlern sowie Verbesserungsvorschlägen stehen. Hier unterscheidet sich die **Inspektion** vom **Strukturierten Durchgehen,** da bei diesem die Überarbeitung von identifizierten Fehlern ausdrücklich nicht gewünscht ist. Vielmehr eignet sich der Walkthrough das Objekt vorzuführen und andere Benutzer auszubilden. Diese eher informelle Natur des Walkthroughs erlaubt eine schnellere Durchführung im

Vergleich zur Inspektion. Der **Schreibtischtest** kann den Mitgliedern eines Teams helfen, ihre eigene Arbeit zu kontrollieren. Dabei wird der Schreibtischtest idealerweise vorbereitend zu Begutachtungen durchgeführt und kann auch unter Hinzuziehung einer weiteren Person geschehen, um das Übersehen von Fehlern zu vermeiden. Die Durchführung des Schreibtischtests gleicht dann der **Validierung im Dialog**.

Die hier aufgeführten Techniken zur Begutachtung scheinen generell alle auch für einen Einsatz im Data-Mining geeignet zu sein. Dies ist vor allem deshalb der Fall, da es beim KDD immer wieder zu Situationen kommt, in denen die bisherigen Schritte überprüft werden müssen oder Fragestellungen nur mittels Kontextwissen ausreichend bewertet werden können. Beides kann allerdings zu Rücksprüngen in vorherige Phasen führen (vgl. Abschnitt 2.2). Diese Rücksprünge machen erschweren eine Durchführung von Audits und Inspektionen bei Phasenbeendigung, da der zeitliche Aufwand dieser Methoden sehr hoch ist.

Die Dokumentation der durchgeführten Aktivitäten spielt in der V&V für die Simulation eine gewichtige Rolle, damit unbeteiligten Personen die Möglichkeit geboten wird, die Modellentstehung nachzuvollziehen und entscheiden zu können, ob die Modellergebnisse zu weiterer Verwendung geeignet sind. Dazu müssen neben der Dokumentation zu Ausgangslage, Randbedingungen etc. auch die Dokumentation der V&V-Prüfungen der einzelnen Phasen einer Überprüfung unterzogen werden (vgl. Abschnitt 3.1.2). Auch beim KDD ist eine transparente und vollständige Dokumentation des Prozessablaufs erwünscht. Besonders vor dem Hintergrund, dass dieser Prozess idealerweise in einen Datenanalysezyklus integriert ist und sich dementsprechend nach einem gewissen Zeitraum wiederholt (vgl. Abschnitt 2.1), ist es sinnvoll Datenvorbereitungsschritte oder die genutzten Modellparameter zu dokumentieren. So kann vor allem genutztes Kontextwissen für den nächsten Durchlauf erhalten werden. Aus diesem Grund ist die Technik der **Dokumentenüberprüfung** auch für das KDD anzuwenden. Idealerweise erfolgt die diese mit Beteiligten aus dem Entwicklungsteam und Fachexperten, damit beide Seiten die Dokumente gegenprüfen. Dies könnte beispielsweise im Rahmen eines Review-Termins erfolgen oder als vor- bzw. nachbereitende Tätigkeit eines solchen Termins.

Bei Durchführung des **Turing-Tests** sollen Ergebnisse des Modells mit dem realen System abgeglichen werden. Da das Ziel des KDD aber die Erkennung bisher unbekannter Muster in der Datengrundlage ist und keine Nachbildung eines realen Systems erfolgen soll, ist ein Abgleich mit einem realen System eben nicht möglich. Beispielsweise können gefundene Assoziationsregeln nicht mit existierenden verglichen werden, da es diese Regeln bisher noch nicht gibt. Aus diesem Grund erscheint der Turing-Test für die weitere Betrachtung irrelevant.

Die Testverfahren **Alpha-, Beta- oder Akzeptanztest** bedingen alle eine Ausführung des Modells zur Aufdeckung von möglichem Fehlverhalten und unterscheiden sich dabei

lediglich im Entwicklungsgrad des Modells. Während das Beta-Testing bereits in der Entwicklungsphase am nicht fertigen Modell durchgeführt wird, ist der Alpha-Test dafür vorgesehen, das fertige Simulationsmodell zu überprüfen. Idealerweise wird der Alpha-Test durch unabhängige Prüfer durchgeführt. Der Akzeptanztest wird schließlich als finale Abnahme vor Übergabe in Zusammenarbeit mit dem Kunden, durch den Kunden alleine oder durch einen Dienstleister durchgeführt. Dabei kann das Modell auf die V&V-Kriterien hin überprüft werden (vgl. Abschnitt 3.1.2). Auch beim KDD sind solche Tests des Modells während der Modellentwicklung notwendig. In Ergänzung zu den Tests des Gesamtmodells bietet sich die Anwendung des **Tests von Teilmodellen** an. Dabei wird das Modell in mehrere Teilmodelle unterteilt, die dann einzeln validiert werden können. Dies erscheint beim KDD hauptsächlich für die vorbereitenden Maßnahmen sowie für den Data-Mining-Schritt geeignet. So kann das Testen von Teilmodellen beispielsweise nach Änderungen der Attributsauswahl oder der Parameter des gewählten Data-Mining-Verfahrens notwendig sein.

Bei der Simulation dienen **statistische Techniken** der Validierung eines Modells sowie seiner Eingangs- und Ausgangsdaten. Diese Aufgaben können mithilfe statistischer Techniken auch im Rahmen des Data Minings durchgeführt werden. Dabei lassen sich beispielsweise in den Eingangsdaten miteinander stark korrelierende Attribute erkennen und von der Analyse ausschließen (Korrelationsanalyse). Auch die Ergebnisse lassen sich mit Hilfe statistischer Techniken – etwa durch die Ermittlung der Modellgüte – überprüfen. Zur Modellvalidierung können auch die im Rahmen der V&V im Data Mining eingeführten statistischen Techniken eingesetzt werden – wie etwa die Kreuzvalidierung (vgl. Abschnitt 3.1.2).

Bei Durchführung der **Validierung von Vorhersagen** wird ein Modell erstellt und anschließend die Richtigkeit der Vorhersage am realen System überprüft. Diese Beschreibung ähnelt dem Vorgehen beim Data Mining, denn auch bei diesem können – nach der Durchführung des Data Minings – die Ergebnisse durch Beobachtungen des realen Systems überprüft werden. Allerdings wird das Data Mining auf Grundlage von Daten aus der Vergangenheit durchgeführt. Deswegen scheint ein **Vergleich mit aufgezeichneten Daten** besser für das Data Mining geeignet. Diese Technik beschreibt die Nutzung von historischen Daten zum Testen des Modells. Dabei dürfen diese Daten keine Schnittmengen mit den zur Modellerstellung genutzten Daten aufweisen. Auch im KDD ist eine Modellierung unter Nutzung aufgezeichneter Daten durchzuführen. Dabei gleicht diese Technik der in Abschnitt 3.3.3 beschriebenen Split-Validierung für den Data-Mining-Schritt des KDD.

Beim **Vergleich mit anderen Modellen** geht es bei der Simulation darum, Modellergebnisse unter Verwendung derselben Eingangsdaten mit einem oder mehreren anderen Modellen zu vergleichen. Damit ähnelt diese Technik der in Abschnitt 3.3.1 beschriebenen Technik des Back-to-Back-Tests mit all seinen Vor- und Nachteilen.

4.1.3 Data Mining

Die in Abschnitt 3.3.3 eingeführten Techniken sind selbstverständlich auch für die Durchführung des KDD geeignet, da die drei Techniken der Split-, Kreuz- und Bootstrapping-Validierung dort bereits speziell für diesen Einsatz im Data-Mining-Schritt beschrieben werden. Unter Berücksichtigung der großen Datenmengen beim KDD in der Produktionslogistik scheint die Split-Validierung den anderen Validierungsmethoden vorzuziehen zu sein, da diese für Mengen mit mehr als 1000 Datensätzen am besten geeignet ist (vgl. Abschnitt 3.3.3).

4.1.4 Gesamtübersicht der generell einsetzbaren Techniken

Nach der generellen Eignungsprüfung der Anwendbarkeit der Techniken im KDD lässt sich zusammenfassend festhalten, dass von den hier diskutierten Techniken hauptsächlich solche für die Durchführung des KDD geeignet erscheinen, die auch der V&V der Datengrundlage dienen können. Dies kann durch Anwendung statistischer Methoden (Statistische Techniken), Nutzung von Kontextwissen (Begutachtungstechniken) sowie Modellvalidierung durch Aufteilen der Datenmenge (etwa Split-Validierung) erfolgen. Tabelle 4.1 stellt alle aufgeführten Techniken noch einmal übersichtlich gegenüber. Dabei werden zur Vervollständigung auch Techniken dargestellt, die in den vorherigen Abschnitten dieses Kapitels sowie in Abschnitt 3.3 bereits als ungeeignet für einen Einsatz im KDD in der Produktionslogistik erkannt wurden. Darüber hinaus werden Techniken aufgeführt, bei denen eine intensivere Betrachtung als V&V-Technik der KDD empfehlenswert sein könnte. Dies sind vor allem vergleichende Verfahren (Back-to-Back-Test oder Vergleich mit anderen Modellen), für die mehrere (parallel entstandene) Data-Mining-Modelle betrachtet werden müssen. Dies würde allerdings den Rahmen dieser Arbeit übersteigen.

4.2 Eignung der V&V-Techniken für das KDD in der Produktionslogistik

Nachdem die in Abschnitt 3.3 vorgestellten Techniken der V&V im vorherigen Abschnitt auf ihre generelle Anwendbarkeit im KDD in der Produktionslogistik überprüft wurden, sollen nun die Eignung der Techniken speziell für die einzelnen Phasen des in Abschnitt 2.2.3 eingeführten MESC und der darin enthaltenen V&V-Elemente (vgl. Abschnitt 3.2) aufgezeigt werden. Dies ist erforderlich, da nicht jede der zuvor als generell anwendbar erkannten Techniken auch in jeder Phase des MESC genutzt werden kann. Zur Festlegung der notwendigen Techniken werden hier die in Abschnitt 3.2 eingeführten Kriterien der V&V herangezogen und mögliche Fragestellungen formuliert, die bei der Durchführung der Techniken im Hinblick auf die Kriterien beantwortet werden sollten. Dieses Vorgehen soll für jedes V&V-Element des MESC durchgeführt werden, auch wenn bei der Anwendung in der Praxis meist nicht jede Prüfung notwendig ist (Tailoring) (vgl. Abschnitt 3.2).

Tabelle 4.1: Generelle Eignung von V&V-Techniken für das KDD in der Produktionslogistik

	V&V-Techniken	Eignung für KDD
Softwareentwicklung	Back-to-Back-Test	◐
	Funktionale Techniken (Äquivalenzklassenbildung etc.)	○
	Inspektionen	●
	Mutationen-Test	◐
	Regressionstest	◐
	Strukturtestverfahren (Anweisungsüberdeckungstest etc.)	○
	Test spezieller Werte	◐
Simulation	Akzeptanztest	●
	Alpha-Testing	●
	Audit	●
	Begutachtung (Review)	●
	Beta-Testing	●
	Dokumentenüberprüfung	●
	Formale Techniken (Induktionsbehauptung etc.)	○
	Schreibtischtest	●
	Statistische Techniken	●
	Strukturiertes Durchgehen	●
	Test von Teilmodellen	●
	Turing-Test	○
	Validierung im Dialog	●
	Validierung von Vorhersagen	○
	Vergleich mit anderen Modellen	◐
	Vergleich mit aufgezeichneten Daten	●
	Visuelle Techniken (Animation etc.)	○
Data Mining	Bootstrapping-Validierung	●
	Holdoutmethode (Split-Validierung)	●
	Kreuzvalidierung	●

Legende:

● generell geeignet, ◐ weitere Untersuchung empfohlen, ○ nicht geeignet

Da eine Darstellung aller Kriterien und möglichen Fragestellungen für jedes Element den Umfang dieser Arbeit übersteigen würde, erfolgt die Auswahl von Kriterien und Fragen hier lediglich beispielhaft. Es findet sich darüber hinaus zu den jeweiligen Prüfungen sicherlich eine Vielzahl weiterer möglicher Kriterien und Fragen, auf die nicht eingegangen wird. Aus diesem Grund besteht an dieser Stelle selbstverständlich kein Anspruch auf Vollständigkeit.

4.2.1 Aufgabendefinition

Die Durchführung einer Data-Mining-Analyse kann in der Praxis nur auf Grundlage einer konkreten Aufgabenstellung unter Berücksichtigung der Randbedingung und Zielkriterien erfolgen (1.1). Diese Aufgabenstellung wird in einer Zielbeschreibung in Dokumentenform festgehalten, die nach Fertigstellung überprüft werden muss (1,1). In Tabelle 4.2 werden beispielhaft Fragen für einzelne Kriterien dieser Phase aufgezeigt. Wie bereits einleitend erläutert, erfolgt die Darstellung diese Kriterien und Fragen an dieser Stelle nur beispielhaft.

Intrinsische Prüfung auf Vollständigkeit und Plausibilität der Aufgabenstellung (1,1)

Die intrinsische Prüfung dieser Phase besteht in der Überprüfung der Aufgabenstellung in Hinblick auf die V&V-Kriterien der *Vollständigkeit* und *Plausibilität*. Dazu kann etwa untersucht werden, ob alle Randbedingungen vollständig erfasst wurden und die Auswahl im Hinblick auf die Fragestellung plausibel erscheint. Dazu ist eine Überprüfung der Dokumente der Zielbeschreibung notwendig. Neben diesen beiden Kriterien scheinen auch weitere V&V-Kriterien wie *Genauigkeit* und *Aktualität* für eine Überprüfung angebracht. Die Beschreibung der Aufgabenstellung ist Grundlage jedes weiteren Schritts. Aus diesem Grund ist es von besonderer Bedeutung, dass die Dokumente so genau wie möglich sind. Eine Ungenauigkeit kann hier zu Fehlern führen, die sich in den folgenden Phasen durchziehen und so die Ergebnisse verfälschen. Auch die *Aktualität* der Zielbeschreibung ist wichtig, da beispielweise geänderte Randbedingungen große Auswirkung auf die Ergebnisse haben können. Darüber hinaus ist eine Überprüfung der *Verständlichkeit* empfehlenswert, damit auch Personen die Ausführungen nachvollziehen können, die nicht direkt an ihrer Durchführung beteiligt waren. In dieser Phase sind V&V-Techniken anzuwenden, die der Begutachtung der Zielbeschreibung dienen. Da die Ziele und Randbedingungen allerdings nur mit Kontextwissen der Fachexperten auf die genannten V&V-Kriterien überprüft werden können, sollten Begutachtungstechniken gewählt werden, die von einer Gruppe von Entwicklern und Fachexperten durchgeführt werden können. Dies ist bei den Techniken Audit, Inspektion, Review, Strukturiertes Durchgehen oder Validierung im Dialog der Fall. Die Techniken des Schreibtischtests und der Dokumentenüberprüfung sind ebenfalls geeignet, da sich auch diese im Team durchführen lassen, auch

wenn prinzipiell nur eine Person für die Durchführung nötig wäre. Aufgrund der immensen Bedeutsamkeit dieser ersten Phase für den Gesamtablauf sollte ihrer Prüfung eine besondere Aufmerksamkeit gewidmet werden. Aus diesem Grund erscheint die Anwendung einer formellen Form der Begutachtung wie beispielsweise des Audits oder der Inspektion sinnvoll. Auch weil ein späterer Rücksprung in diese Phase unwahrscheinlich ist, erscheinen diese Techniken geeignet zu sein. Ideal für diese Prüfung erscheint eine Kombination aus Schreibtischtest sowie strukturiertem Durchgehen zur Vorbereitung eines Review-Termins. Denn gerade in dieser Phase ist es enorm bedeutsam, dass jeder Aspekt der Aufgabenstellung ausreichend diskutiert und von allen Beteiligten (Data-Mining- und Fachexperten) akzeptiert wird. Hierfür ist das strukturierte Durchgehen besonders geeignet.

Tabelle 4.2: Beispielhafte Fragen zur Prüfung der Aufgabenstellung anhand der V&V-Kriterien

Untersuchungs-gegenstand	**V&V-Kriterium**	**Beispielhafte Fragestellung**
Korrektheit (Inhalt & Struktur)	Vollständigkeit	• Wurden alle Randbedingungen beachtet?
	Genauigkeit	• Ist die Beschreibung der Aufgabenstellung genau genug?
	Aktualität	• Entsprechen die Randbedingungen den aktuellen Gegebenheiten?
	Konsistenz	• Werden die Terminologien bei der Definition konsistent beibehalten?
Angemessenheit des Phasenergebnisses (Randbedingungen & Zielbedingung)	Eignung	• Ist die Zielstellung im Hinblick auf die Rahmenbedingung geeignet?
	Plausibilität	• Erscheint die Auswahl der Rahmenbedingung und die Festlegung der Zielbedingung schlüssig?
	Verständlichkeit	• Ist die Dokumentation der Zielbedingung für Unbeteiligte verständlich und nachvollziehbar?
Durchführbarkeit	Machbarkeit	• Ist die festgelegte Zielbedingung umsetzbar?
	Verfügbarkeit	• Kann auf alle nötigen Informationen zugegriffen werden?

4.2.2 Auswahl der relevanten Datenbestände

Die zweite Phase dient der Auswahl der relevanten Datenbestände und umfasst die Schritte der Datenbeschaffung (2.1) und ihrer Reduzierung (2.2). Dabei muss die Datenauswahl einer intrinsischen Prüfung hinsichtlich der Korrektheit und Relevanz (2,2) unterzogen werden, bevor erneut geprüft wird, ob die Datenauswahl für die Zielbeschreibung der ersten Phase geeignet ist (2,1).

Intrinsische Prüfung auf Korrektheit und Relevanz verwendeter Datenquellen und Datenbestände (2,2)

Bei der intrinsischen Prüfung dieser Phase ist die Datenauswahl auf ihre Korrektheit und Relevanz hinsichtlich der verwendeten Datenquellen und Datenbeständen zu untersuchen. Um den Datenabruf auszuführen, muss kontrolliert werden, ob ein Zugriff auf die gewünschten Datenquellen möglich ist (*Verfügbarkeit*) und ob eine Entnahme von Datenbeständen aus diesen Quellen technisch umsetzbar ist (*Machbarkeit*). Die Korrektheit dieser Daten kann untersucht werden, indem geprüft wird, ob die Datenbestände vollständig sind oder Lücken aufweisen (*Vollständigkeit*). Da für diese Phase wie in Abschnitt 2.2.3 bereits für die Durchführung der einzelnen Schritte Kontextwissen von mehreren Fachexperten notwendig ist, ist auch bei dieser Prüfung ihre Einbeziehung erforderlich. Deswegen sind für diese Prüfung dieselben Techniken zur Begutachtung der Daten einzusetzen wie bereits bei V&V-Element (1,1) der ersten Phase zur Begutachtung der Zielbeschreibung wie etwa Inspektion, Review oder das strukturierte Durchgehen (vgl. Abschnitt 3.3.2). Eine Darstellung möglicher Kriterien und ihrer Fragestellungen (siehe Tabelle 4.2) findet für dieses und die folgenden Elemente – mit Ausnahme von Element (5,5) – aus Platzgründen nicht statt. Anstelle dessen werden – wie bereits in Abschnitt 0 erläutert – an einigen Stellen Fragestellungen hinter den Kriterien angeführt. Auch hier sei noch einmal auf den fehlenden Anspruch auf Vollständigkeit verwiesen.

Prüfung auf Eignung der Daten für Zielbedingung (2,1)

Das zweite V&V-Element dieser Phase beinhaltet eine Überprüfung der *Eignung* der ausgewählten Datenquellen und Datenbestände für das Erreichen der in der ersten Phase festgelegten Zielbedingung. Neben dem Kriterium der Eignung können noch weitere V&V-Kriterien zur Bewertung herangezogen werden. So ist zu überprüfen, ob die ausgewählten Daten sich im Zeitraum von der Erstellung der Zielbedingung bis zur Entnahme der Daten verändert haben und damit eventuell nicht mehr für eine Analyse geeignet sind (*Aktualität*). Darüber hinaus ist zu kontrollieren, ob die Daten die benötigte Stufe der Granularität zur Erfüllung der Zielbedingung aufweisen (*Genauigkeit*). Bei der Prüfung können erneut die zuvor angesprochenen Begutachtungstechniken angewendet werden, die im Idealfall wieder in Zusammenarbeit mit Fachexperten zur Einbringung von Expertenwissen dienen.

4.2.3 Datenaufbereitung

Die dritte Phase des MESC dient der Aufbereitung der zuvor ausgewählten Daten für das Data Mining. Dabei sind die in dieser Phase durchgeführten Schritte der Formatstandardisierung (3.1.), Gruppierung (3.2), Datenanreicherung (3.3) sowie Transformation (3.4) einer Prüfung der formalen Korrektheit zu unterziehen (3,3). Darüber hinaus müssen transformierten Daten hinsichtlich der Ausgangsdaten (3,2) sowie der Zielerreichung (3,1) überprüft werden.

Intrinsische Prüfung der Datentransformation auf Korrektheit (3,3)

Die intrinsische Prüfung der dritten Phase besteht aus einer Kontrolle der durchgeführten Datentransformation auf ihre formale Korrektheit. Dazu kann beispielsweise bewertet werden, ob fehlerhafte Attribute oder vorhandene Ausreißer vollständig erkannt und behandelt wurden (*Vollständigkeit*). Auch die fehlerfreie Durchführung dieser Behandlung kann kontrolliert werden (*Genauigkeit*). Die Überprüfung, der für die Datenaufbereitung genutzten mathematischen Operatoren, kann durch einen Schreibtischtest bzw. eine Validierung im Dialog durchgeführt werden (vgl. Abschnitt 3.3.2). Darüber hinaus lassen sich auch statistische Techniken zur Kontrolle der Datenauswahl einsetzen (z.B. Korrelationsanalyse, Chi-Quadrat-Test oder Regressionsanalyse).

Prüfung der Korrektheit der Datentransformation gegenüber Ausgangsdaten (3,2)

Die zweite Prüfung dieser Phase befasst sich mit dem Nachweis der Korrektheit der Datentransformation gegenüber der Ausgangsdaten. Hier können wieder die Kriterien der *Vollständigkeit* und *Genauigkeit* herangezogen werden. Dabei kann beispielsweise geprüft werden, ob bei einer vorgenommenen Gruppierung alle Attribute erfasst (*Vollständigkeit*) und fehlerfrei in die richtige Gruppierung eingeteilt wurden (*Genauigkeit*). Darüber hinaus können die *Eignung* der angewendeten Verfahren sowie die *Plausibilität* und *Verständlichkeit* der Ergebnisse zur Bewertung der Angemessenheit gegenüber den Ausgangsdaten herangezogen werden (Fragen: Erscheinen die Ergebnisse der Datentransformation plausibel? Liefert die Durchführung der Transformation verständliche Ergebnisse im Hinblick auf die Ausgangsdaten?). Für die Durchführung der Prüfung eignen sich das Verfahren des Schreibtischtests und – im Falle von benötigtem Fachwissen – die Validierung im Dialog bzw. auch weitere Begutachtungstechniken wie das Review-Meeting oder das strukturiertes Durchgehen.

Prüfung der Eignung der aufbereiteten Daten für Zielerreichung (3,1)

Bei dieser Prüfung gilt es die aufbereiteten Daten hinsichtlich ihrer *Eignun*g für die Zielerreichung zu überprüfen (Frage: Sind die Daten nach ihrer Aufbereitung (noch) zur Beantwortung der Fragestellung geeignet?). Neben der Eignung kann für diese Prüfung analysiert werden, ob die transformierten Daten ausreichend sind oder beispielsweise

angereicht werden müssen (*Vollständigkeit*). Ähnlich zu den Prüfungen in Element (2,2) lässt sich darüber hinaus kontrollieren, ob die Granularitätsstufe nach der Transformation (noch) zu den festgehaltenen Anforderungen der ersten Phase passt (*Genauigkeit).* Für die Durchführung der Eignungsprüfung der transformierten Daten hinsichtlich der Zielerreichung sind erneut begutachtende Techniken einzusetzen (Schreibtischtest, Dokumentenüberprüfung, Validierung im Dialog etc.).

4.2.4 Vorbereitung des Data-Mining-Verfahrens

Die Phase der Vorbereitung der Durchführung des Data-Mining-Verfahrens besteht aus insgesamt vier Schritten (siehe Abschnitt 2.2.3). Die Hauptaufgabe dieser Phase besteht in der Verfahrensauswahl (4.1), die maßgebend für die weiteren Schritte der Werkzeugauswahl (4.2) als auch der fachlichen und technischen Kodierung ist (4.3, 4.4). Neben einer intrinsischen Prüfung der Verfahrensauswahl (4,4) sind dabei ebenfalls Prüfungen gegen die Vorphasen notwendig. Dabei muss hinsichtlich des ausgewählten Verfahrens die Eignung der Datenaufbereitung (4,3) sowie der Datenauswahl (4,1) kontrolliert werden. Darüber hinaus ist die Eignung des gewählten Verfahrens gegenüber der Aufgabenstellung zu prüfen.

Intrinsische Prüfung auf geeignete Auswahl des Data-Mining-Verfahrens (4,4)

Die Prüfung dieser Phase gegen sich selbst besteht in der Untersuchung der *Eignung* des ausgewählten Verfahrens sowie des ausgewählten Werkzeugs (Frage: Ist das ausgewählte Werkzeug für die Durchführung des Data-Mining-Verfahrens geeignet?). Daneben kann die Durchführbarkeit des gewählten Verfahrens untersucht werden, indem kontrolliert wird, ob die technischen Voraussetzungen für seine Durchführung gegeben sind (*Machbarkeit*) und ob auf das benötigte Werkzeug zugegriffen werden kann (*Verfügbarkeit*). Darüber hinaus sind die Daten auf ihre fachliche und sachliche Korrektheit zu überprüfen. Auch für die Durchführung dieser Prüfung sind hauptsächlich begutachtende V&V-Techniken geeignet. So können beispielsweise die für einen Einsatz in Frage kommenden Data-Mining-Verfahren im Rahmen eines Review-Termins gegenübergestellt und diskutiert werden.

Prüfung der Eignung der Datenaufbereitung für gewähltes Data-Mining-Verfahren (4,3)

Die Schritte der Datenaufbereitung (Phase 3) können unabhängig vom späteren Data-Mining-Verfahren durchgeführt werden. Deshalb muss in diesem Element beispielsweise kontrolliert werden, ob die bisherige Aufbereitung ausreichend war oder ob nach Auswahl des Verfahrens weitere, vom Verfahren abhängige Datenvorverarbeitungsmethoden angewendet werden müssen (*Eignung*). Daneben ist etwa zu untersuchen, ob die Daten

durch verfahrensabhängige Aufbereitung überhaupt in die benötigte Form gebracht werden können (*Machbarkeit*). Um die Datenaufbereitung im Bezug zum ausgewählten Verfahren zu kontrollieren, muss die fachliche und technische Kodierung überprüft werden. Dafür ist neben statistischen Techniken wiederum die Einbeziehung von Expertenwissen dringend zu empfehlen. So lassen sich ID-ähnliche Attribute nur unter Zuhilfenahme von Fachwissen auf ihre innere Struktur und damit auf eventuell nutzbare Elemente untersuchen.

Prüfung der Datenauswahl für gewähltes Data-Mining-Verfahren (4,2)

Die Prüfung der Datenauswahl im Hinblick auf das Data-Mining-Verfahren kann beispielsweise nötig werden, wenn in der für das Element (4,3) durchgeführten Prüfung eine ungeeignete Datenaufbereitung festgestellt wird. Können die Daten nicht in die richtige Form gebracht werden, ist die *Eignung* der getroffenen Auswahl zu kontrollieren und zu hinterfragen, ob neue Daten auszuwählen sind. Dies ist auch im Hinblick auf das Kriterium der *Vollständigkeit* (Frage: Sind weitere Attribute nötig?) und *Aktualität* (Frage: Ist die Datenauswahl noch für das Data-Mining-Verfahren gültig?). Für diese Prüfung muss wie bei Element (4,3) eventuell Kontextwissen eingesetzt werden, um neue Daten auszuwählen. Hierfür bietet sich etwa eine Validierung im Dialog zur Begutachtung an, aber natürlich auch weitere Begutachtungstechniken.

Prüfung der Eignung des Data-Mining-Verfahrens für Aufgabenstellung (4,1)

Wie schon in den Elementen (2,1) und (3,1) beschränken sich die Prüfungen auch in diesem Element auf die Kontrolle der Einhaltung der Ziele aus der Zielbedingung sowie der Angemessenheit des Ergebnisses für die Fragestellung. Wie in Abschnitt 2.4 dargestellt, ist nicht jedes Verfahren auch für jeden Problemtyp geeignet. Deswegen gilt es bei dieser Prüfung das ausgewählte Data-Mining-Verfahren auf seine *Eignung* hinsichtlich der Aufgabenstellung zu kontrollieren. Darüber hinaus hilft die Berücksichtigung der V&V-Kriterien der *Vollständigkeit* (Frage: Werden alle Randbedingungen bei der Auswahl des Verfahrens berücksichtigt?) sowie *Machbarkeit* (Frage: Kann das Verfahren auf Grundlage der Randbedingungen durchgeführt werden?) zur Bewertung der Durchführbarkeit des Data-Mining-Verfahrens. Dabei bieten sich erneut vor allem Techniken zur Begutachtung an – etwa der Schreibtischtest oder die Dokumentenüberprüfung.

4.2.5 Anwendung des Data-Mining-Verfahrens

Zur Anwendung des Data-Mining-Verfahrens sind die Schritte der Entwicklung eines Data-Mining-Modells (5.1) sowie des Trainings des entwickelten Modells (5.2) notwendig. Dabei sind wie in den vorherigen Phasen unterschiedliche Prüfungen zur V&V der Durchführung notwendig. Bei der intrinsischen Prüfung ist abzuklären, ob das Data-

Tabelle 4.3: Beispielhafte Fragen zur Prüfung der Modellbildung anhand der V&V-Kriterien

Untersuchungs-gegenstand	V&V-Kriterium	Beispielhafte Fragestellung
Korrektheit (Inhalt & Struktur)	Vollständigkeit	• Ist das Data-Mining-Modell den festgelegten Anforderungen entsprechend? (5,1)
	Genauigkeit	• Ist der Detailierungsgrad des Modells angemessen?
		• Ist die Modellierung fehlerfrei? (5,5)
	Aktualität	• Ist das Data-Mining-Modell für die Aufgabenstellung gültig? (5,1)
	Konsistenz	• Wird die Terminologie durchgehend genutzt? (5,5)
Angemessenheit des Ergebnisses für die Anwendung	Eignung	• Ist das Data-Mining-Modell im Hinblick auf die Aufgabenstellung geeignet? (5,1)
	Plausibilität	• Sind die Ergebnisse des Data-Mining-Verfahrens schlüssig? (5,5)
	Verständlichkeit	• Ist das Data-Mining-Modell lesbar und nachvollziehbar für den Anwender? (5,5)
Durchführbarkeit	Machbarkeit	• Ist das Modell technisch umsetzbar? (5,4)
	Verfügbarkeit	• Sind alle zur Durchführung nötigen Daten und Dokumente verfügbar? (5,3) (5,2)

Mining-Verfahren korrekt angewendet wurde (5,5). Darüber hinaus ist die Anwendungsphase gegen ihre Vorphasen zu prüfen, d.h. es ist die Vorbereitung des Verfahrens (5,4), die Eignung der Datenvorverarbeitung (5,3) sowie die Auswahl der Daten (5,2) für das Data-Mining- Verfahren zu untersuchen. Darüber hinaus ist zu kontrollieren, ob das Verfahren die festgelegte Zielbedingung erfüllt (5,1). Tabelle 4.3 stellt eine Übersicht möglicher Fragestellungen dieser Phase dar.

Dabei gleichen die Fragestellungen der V&V-Kriterien des Data Mining-Modells denen der Simulationsmodelle (Tabelle 3.1). Ergänzend dazu werden Prüfungen aufgeführt, die zur Beantwortung der jeweiligen Frage nötig sind. Wie bereits die Auswahl der Kriterien und Fragen, ist auch diese Zuteilung beispielhafter Natur.

Intrinsische Prüfung des Data-Mining-Verfahrens auf korrekte Anwendung (5,5)

Die intrinsische Prüfung dieser Phase besteht aus einer Kontrolle der Anwendung des Data-Mining-Verfahrens auf ihre Korrektheit. Hierbei geht es in erste Linie darum, die Modellbildung zu validieren. Dabei wird die Korrektheit des Modells darauf untersucht, ob alle benötigten Elemente vorhanden sind (*Vollständigkeit*) und die Modellierung dieser Elemente fehlerfrei durchgeführt wurde (*Genauigkeit*). Darüber hinaus lässt sich

die Durchführbarkeit des Modells durch Untersuchung auf technische Umsetzbarkeit sowie Verfügbarkeit aller benötigten Dokumente und Daten kontrollieren (*Machbarkeit* & *Verfügbarkeit*). Zu guter Letzt können die Modellergebnisse auf ihre Schlüssigkeit und Nachvollziehbarkeit überprüft werden (V&V-Kriterien: *Plausibilität* & *Verständlichkeit*). Zur Begutachtung der Schritte der Anwendungsphase im Hinblick auf die V&V-Kriterien bieten sich verschiedene Techniken an. So lässt sich etwa die Modelldurchführung durch Testverfahren kontrollieren. Hier könnten zum Beispiel die in Abschnitt 3.3.2 vorgestellten Alpha- Beta- bzw. Akzeptanztests durchgeführt werden, die sich – wie in Abschnitt 4.1.2 erläutert – hauptsächlich durch den Entwicklungszustand des Modells sowie die Zusammensetzung des Teams unterscheiden. Zur Durchführung dieser Tests werden idealerweise weitere V&V-Techniken herangezogen. So lässt sich ein Beta-Test beispielsweise durch den Entwickler mittels Schreibtischtest oder in Zusammenarbeit mit einer weiteren Person als Validierung im Dialog durchführen. Beim Alpha-Test kann das fertige Modell etwa im Rahmen eines Review-Termins von einer Personengruppe begutachtet werden. Dabei sind die Teilnehmer der Gruppe idealerweise kein Bestandteil des Entwicklerteams, sondern Außenstehende (vgl. Abschnitt 4.1.2). Der Akzeptanztest wird schließlich gemeinsam mit dem Kunden bzw. nur durch den Kunden oder unabhängige Dienstleister durchgeführt. Neben einer Prüfung des Gesamtmodells bietet sich gerade im Rahmen des Beta-Tests eine Unterteilung des Modells in Teilmodelle an, die dann einer Kontrolle unterzogen werden. Da die zur Modellierung des Data-Mining-Verfahrens eingesetzten Standardprogramme (z.B. RapidMiner) häufig mit Bausteinen arbeiten, können diese als eigenständige Teilmodelle auf ihre Funktionsweise überprüft werden (V&V-Technik: Test von Teilmodellen). Dadurch lassen sich diese Teilmodelle bereits vor Abschluss der Gesamtmodellbildung validieren. Dazu können dieselben Techniken wie zur Kontrolle des Gesamtmodells genutzt werden. Allerdings kann das Testen von Teilmodellen nie die Prüfung des gesamten Modells ersetzen. Darüber hinaus bietet sich im Rahmen der V&V des Verfahrens die Möglichkeit, die Daten mittels stichprobenbasierter Methoden zu beleuchten. Aus der Vielzahl der in Betracht kommenden Methoden gelten vor allem die Split-, Kreuz- sowie Bootstrapping-Validierung als geeignet zur Anwendung (vgl. Abschnitt 3.3.2). Wie schon in Abschnitt 4.1.3 festgehalten, scheint die V&V-Technik der Split-Validierung für die Anwendung im KDD aufgrund der dort vorliegenden großen Datenmengen ideal zu sein.

Prüfung auf korrekte Vorbereitung des Data-Mining-Verfahrens (5,4)

In dieser Prüfung ist – ähnlich zu Element (4,4) – erneut zu kontrollieren, ob das Data-Mining-Verfahren korrekt vorbereitet wurde (Auswahl von Verfahren und Werkzeug). Zur Durchführung der Prüfung bietet sich hier deshalb auch die Anwendung derselben V&V-Kriterien und Techniken wie bei Element (4,4) an.

Prüfung auf Eignung der Datenvorverarbeitung für das Data-Mining-Verfahren (5,3)

Bei der Anwendung des Data-Mining-Verfahrens ist ebenfalls erneut die *Eignung* der Datenvorverarbeitung für Verfahren zu überprüfen – ähnlich zu Element (4,3). So können bei der Ausführung des Verfahrens Probleme auftreten, die zuvor nicht vorauszusehen waren und eine erneute Kontrolle der Datenvorverarbeitung erfordern. Neben dem Kriterium der Eignung ist hierbei vor allem die Durchführbarkeit des Verfahrens auf Grundlage der vorverarbeiteten Daten zu untersuchen. Dabei bieten sich dieselben V&V-Techniken wie in der Vorphase an.

Prüfung auf Zulässigkeit der Datenauswahl für eine fachgerechte Anwendung des Data Minings-Verfahrens (5,2)

Falls die in Element (5,3) beschriebenen möglichen Probleme nicht auf die Vorverarbeitung zurückzuführen sind, muss die zugrundeliegende Datenauswahl einer erneuten Kontrolle unterzogen werden. Dabei müssen diese auf ihre *Eignung* für das Verfahren überprüft werden. Darüber hinaus können die Kriterien der Vollständigkeit (Frage: Sind alle für das Verfahren benötigten Daten ausgewählt?) und *Plausibilität* (Frage: Sind die Daten für das gewählte Verfahren plausibel gewählt?) untersucht werden. Für die Durchführung dieser Prüfung eignen sich die in Element (4,2) genannten Techniken der Begutachtung, da auch in dieser Prüfung Fachwissen miteinbezogen werden sollte.

Prüfung der Erfüllung der Zielbedingung durch Anwendung des Data-Mining-Verfahrens (5,1)

Hier werden die Ergebnisse dieser Phase mit der Zielbedingung verglichen und diese auf Erfüllung untersucht. Dabei ist zu kontrollieren, ob die Ergebnisse geeignet sind, um die Fragestellung vollständig zu beantworten (*Eignung* & *Vollständigkeit*). Bedeutsam ist dabei auch, die *Genauigkeit* der Ergebnisse (Frage: Haben die Ergebnisse das richtige Detaillevel zur Beantwortung der Fragestellung?). Zur Bewertung dieser Kriterien sollten Begutachtungen durchgeführt werden. So können die Ergebnisse etwa in einem Review-Termin vorgestellt und die Zielbedingung auf den Grad der Erfüllung untersucht werden.

4.2.6 Weiterverarbeitung der Data-Mining-Ergebnisse

Nach Durchführung des Data-Mining-Verfahrens geht es in der sechsten Phase um die Nutzung der Ergebnisse. Dafür sind handlungsrelevante Ergebnisse zu extrahieren (6.1) und in eine geeignete Darstellungsform zu überführen (6.2). Treten bei der intrinsischen Prüfung dieser Phase Zweifel an der Aussagekraft der Phasenergebnisse auf, müssen in weiteren extrinsischen Prüfungen die Ergebnisse der Vorphasen erneut untersucht werden. Diese erneute Untersuchung bietet sich allerdings auch generell als Abschluss der Durchführungen an, da auch korrekt erscheinende Ergebnisse der Weiterverarbeitung auf falschen Annahmen und Berechnungen beruhen können.

Intrinsische Prüfung auf korrekte Weiterverarbeitung der Ergebnisse (6,6)

Die intrinsische Prüfung dieser Phase dient der Kontrolle der Korrektheit der Weiterverarbeitung der Daten. Die Korrektheit lässt sich etwa durch die Untersuchung der Weiterverarbeitung auf ihre *Genauigkeit* und *Vollständigkeit* prüfen. Dabei kann kontrolliert werden, ob die Weiterverarbeitung formal korrekt durchgeführt wurde und alle handlungsrelevanten Ergebnisse berücksichtigt wurden. Daneben ist die Auswahl der Darstellungsform auf ihre generelle *Eignung* und *Verständlichkeit* zu überprüfen. Die Kontrolle der Weiterverarbeitung kann einerseits durch ein einzelnes Teammitglied mittels Schreibtischtest durchgeführt werden, um etwa die Auswahl der Ergebnisse zu überprüfen, andererseits bietet sich die Einbeziehung mindestens einer weiteren Person zur Vermeidung von Fehleinschätzungen an (Review, Validierung im Dialog).

Prüfung auf Interpretierbarkeit der Ergebnisse der Anwendung Data-Mining-Verfahrens (6,5)

Ist die Weiterverarbeitung der Ergebnisse formal korrekt durchgeführt worden (6,6), können eventuelle Probleme auch auf fehlerhafte Ergebnisse des Data-Mining-Verfahrens oder auf fehlende Eignung der Ergebnisse für die gewählte Weiterverarbeitungs- und Darstellungsform zurückzuführen sein. In diesem Fall sind diese Ergebnisse auf ihre Interpretierbarkeit zu untersuchen Dabei ist beispielsweise zu kontrollieren, ob die Ergebnisse in einer ausreichenden Detailierungsstufe und fehlerfrei vorliegen (*Genauigkeit*) und ob sie für die gewählte Darstellungsform geeignet sind *(Eignung)*. Diese Prüfung kann in einem ersten Schritt mittels Schreibtischtest durchgeführt werden, zur Sicherstellung der Korrektheit der Ergebnisse sollte zur Begutachtung jedoch weitere Personen herangezogen werden – etwa im Rahmen des strukturierten Durchgehens.

Prüfung ob ausgewähltes Data-Mining-Verfahren theoretisch interpretierbare Ergebnisse liefern könnte (6,4)

Sind nach Abklärung der Interpretierbarkeit der Ergebnisse (6,5) keine Fehler festzustellen, so ist das Data-Mining-Verfahren auf seine *Eignung* zur Ausgabe theoretisch interpretierbarer Ergebnisse zu kontrollieren. Dafür ist das Verfahren hinsichtlich seiner Auswahl sowie der korrekten fachlichen und technischen Kodierung der Daten zu untersuchen. Hier eignen sich dementsprechend vor allem die Techniken, die bereits in Element (4,4) genannt wurden (Begutachtungstechniken).

Prüfung auf fachlich richtige Datenaufbereitung für die Interpretation (6,3)

Erscheint das Data-Mining-Verfahren zur Ausgabe theoretisch interpretierbarer Ergebnisse geeignet, ist eine Überprüfung der in der dritten Phase durchgeführten Datenvorverarbeitungsschritte zu leisten. Dabei gilt es – ähnlich wie bei Element (3,3) – die for-

male Korrektheit der Durchführung zu überprüfen. Dies kann idealerweise von einer unbeteiligten Person oder einer Personengruppe – etwa im Rahmen des strukturierten Durchgehens – erfolgen. Allerdings ist bei einer solchen Prüfung kein Kontextwissen notwendig, da es – wie in Element (3,3) beschrieben – lediglich um die fachlich richtige Durchführung der Aufbereitung geht.

Prüfung auf ausreichende Datenselektion für die Interpretation (6,2)

Falls sich die Durchführung der Datenaufbereitung als formal korrekt erweist, so ist die der Aufbereitung zugrundeliegende Datenauswahl einer erneuten Überprüfung zu unterziehen. Dabei sind die gleichen Kriterien wie bei Element (2,2) zu beachten und es können auch dieselben Techniken verwendet werden. Bedeutsam ist hierbei die erneute Zusammenarbeit mit Fachexperten.

Prüfung ob Erkenntnisse der Interpretation den vordefinierten Zielen der Aufgabenstellung genügen (6,1)

Diese Prüfung dient dem Abgleich der Erkenntnisse der Interpretation mit den in der Aufgabenstellung definierten Zielen. Dabei ist zu hinterfragen, ob die Weiterverarbeitung der Ergebnisse sowie die Darstellungsform für die Beantwortung der in Phase 1 getroffenen Fragestellungen geeignet erscheint (*Eignung*). Ergänzend dazu kann die *Verständlichkeit* der Ergebnisse untersucht werden. Um die Korrektheit der Erkenntnisse im Hinblick auf die Aufgabenstellung zu gewährleisten, sind diese auf *Vollständigkeit* zu kontrollieren (Frage: Sind alle Fragen und Rahmenbedingungen berücksichtigt worden?). Diese Prüfung sollte unter Einbeziehung des Kunden erfolgen, da sein fachliches Feedback für die Bewertung notwendig ist. Zur Durchführung der Prüfung der Gesamtbewertung bietet sich das Strukturierte Durchgehen an. Dies kann neben einem Schreibtischtest als Vorbereitung für einen Abschlusstermin durchgeführt werden. Ein solcher Abschlusstermin kann etwa als Review-Meeting erfolgen. Aufgrund der Bedeutsamkeit dieser Prüfung kommt hier aber auch – ähnlich zu Element (1,1) – eine Begutachtung in Form eines Audits oder einer Inspektion in Frage.

4.2.7 Bewertung der Data-Mining-Prozesse

Die Phase der Bewertung des Data-Mining-Prozesses beinhaltet eine Qualitätskontrolle des Prozesses (7.1) sowie eine Rückführung der Data-Mining-Ergebnisse (7.2). Zur intrinsischen Prüfung dieser Phase muss eine Prüfung der Qualitätskontrolle auf ihre korrekte Durchführung erfolgen. Wie in Abschnitt 3.2 dargestellt, hat die Dokumentation der Phasenergebnisse sowie der durchgeführten V&V-Prüfungen eine immense Bedeutsamkeit für die V&V. Auch wenn bereits nach Abschluss jeder Phase eine Prüfung der Dokumente stattfinden sollte, muss deswegen in dieser abschließenden Phase des MESC eine erneute Überprüfung aller (V&V-) Dokumente zu den einzelnen Phasen stattfinden

(Element (7,6) - Element (7,2)). Generell sollte dazu jedes Dokument auf die in Abschnitt 3.2 geforderten Bestandteile erfolgen.

Prüfung der Qualitätskontrolle des Data-Mining-Prozesses auf richtige Initiierung (7,7)

Die intrinsische Prüfung der Bewertungsphase besteht aus einer Kontrolle der richtigen Initiierung der Qualitätskontrolle. Dazu muss ihre korrekte Durchführbarkeit überprüft werden, indem beispielsweise untersucht wird, ob vollständiger Zugriff auf alle benötigten Dokumente besteht (*Verfügbarkeit & Vollständigkeit*) und ob diese in der aktuellsten Fassung vorliegen (*Aktualität*). Für die Durchführung dieser Prüfung ist vor allem die Technik Dokumentenüberprüfung einzusetzen, aber auch weitere Begutachtungs-techniken sind unter Einbeziehung von Fachwissen geeignet, um die Qualitätskriterien zu kontrollieren, die für die Prüfung der Dokumente festgelegt wurden.

Prüfung der Dokumentation der Ergebnisse der Vorphasen (7,6) - (7,1)

Da die Prüfungen der ausreichenden Dokumentation der Vorphasenergebnisse für jede der Vorphasen wie in Abschnitt 3.2 ähnlich ist, wird auch hier auf eine Darstellung aller Phasen verzichtet. Stattdessen wird ein allgemeiner Ansatz zur Prüfung der Dokumentation vorgestellt. Zur Überprüfungen der Dokumentationen sollten diese auf ihre Korrektheit in Inhalt & Struktur untersucht werden. Für die Inhaltskontrolle kann die Aufgabenstellung beispielsweise auf Berücksichtigung aller benannten Randbedingungen untersucht werden. Darüber hinaus kann der Inhalt ebenfalls auf seine Aktualität kontrolliert werden. Bei der Prüfung der Struktur kann analysiert werden, ob der Aufbau der Dokumente einem vor Projektbeginn festgelegte Aufbau entspricht und eine einheitliche Terminologie verwendet wird (*Konsistenz*). Ganz entscheidend ist in dieser Phase die *Verständlichkeit* der Dokumente, um den KDD-Prozess für Unbeteiligte nachvollziehbar zu dokumentieren. Als Techniken für diese Prüfung bieten sich nur begutachtende Techniken an. Ideal ist hier der Einsatz der Dokumentenüberprüfung, die in Kombination mit anderen Verfahren wie etwa einem Review-Termin durchgeführt werden. Die Dokumente sollten auch in Abwesenheit des Verfassers analysiert werden, um die Verständlichkeit zu hinterfragen. Dafür bietet sich beispielsweise besonders das strukturierte Durchgehen an.

4.3 Erkenntnisse aus der theoretischen Betrachtung

Nachdem die in Kapitel 3 vorgestellten V&V-Techniken auf ihre generelle Nutzbarkeit für das KDD der Produktionslogistik (Abschnitt 4.1) und danach auf ihre Eignung speziell für die einzelnen Prüfungen je Phase des MESC (Abschnitt 4.2) untersucht wurden, sollen in diesem Abschnitt abschließend die Erkenntnisse aus der theoretischen Betrachtung dargestellt werden.

Eine vollautomatisierte Durchführung des Data-Mining-Verfahrens scheint aufgrund der gegebenen Nichttrivialität der zu suchenden Muster nicht möglich (vgl. Abschnitt 2.1). So ist die Integration von Kontextwissen in den Prozess der Wissensgewinnung zwingend erforderlich. Hierfür erscheint der Einsatz von informalen Techniken gerade in der Vorbereitung des Verfahrens unumgänglich. Das Kontextwissen lässt sich durch die Einbeziehung von Fachexperten in den KDD-Prozess gewinnen. Dies erfolgt im besten Fall im Rahmen eines Treffens mit heterogenem Teilnehmerfeld, da nicht jeder Mitarbeiter mit allen Abläufe innerhalb des Unternehmens vertraut ist. Die V&V-Prüfungen des MESC sollten nicht nur am Ende jeder Phase, sondern auch bereits zu geeignet erscheinenden Zeitpunkten während der Phasen erfolgen. Da allerdings Audits und Inspektionen meist bei Erreichen von Meilensteinen – wie etwa dem Abschluss einer Phase – durchgeführt werden, ist eine Anwendung innerhalb der Phasen problematisch (vgl. Abschnitt 3.2). Weiterhin ist der Einsatz statistischer Techniken sowohl zur Vorbereitung und Durchführung des Data-Mining-Verfahrens als auch zur Überprüfung der gefundenen Ergebnisse möglich. Insbesondere der Einsatz von Techniken zur Stichprobenziehung ist beim KDD anzuwenden.

Tabelle 4.4: Eignung der V&V-Techniken für die verschiedenen Phasen des MESC

	Phasen des MESC						
V&V-Technik	**Aufgabendefinition**	**Auswahl relevanter Datenbestände**	**Datenaufbereitung**	**Vorbereitung der DM-Verfahren**	**Anwendung der DM-Verfahren**	**Weiterverarbeitung der DM-Ergebnisse**	**Bewertung der DM -Ergebnisse**
Akzeptanztest	–	–	–	–	+	–	–
Alpha-Testing	–	–	–	–	++	–	–
Audit	++	+	O	O	+	++	++
Begutachtung (Review)	++	++	++	++	++	++	++
Beta-Testing	–	–	–	–	++	–	–
Bootstrapping-Validierung	–	–	–	–	+	–	–
Dokumentenüberprüfung	++	++	++	++	++	++	++
Holdoutmethode (Split-Val.)	–	–	–	–	++	–	–
Inspektion	++	+	O	O	+	++	++
Kreuzvalidierung	–	–	–	–	+	–	–
Schreibtischtest	++	++	++	++	++	++	++
Statistische Techniken	–	+	++	++	++	++	–
Strukturiertes Durchgehen	++	++	++	++	++	++	++
Test von Teilmodellen	–	–	–	–	++	–	–
Validierung im Dialog	++	++	++	++	++	++	++
Vergleich mit aufgezeichneten Daten	–	–	–	–	++	–	–

Legende: ++ sehr gute Eignung + gute Eignung O mittlere Eignung – keine Eignung

5 Ausführung des Data Minings auf Firmendaten aus der Branche Elektronikkleingeräte

Dieses Kapitel dient der Darstellung des KDD-Prozesses anhand eines Praxisbeispiels eines internationalen Produktionsunternehmens aus der Elektronikindustrie. Dabei soll das in Abschnitt 2.2.3 vorgestellte MESC verwendet werden. Wie dabei erläutert, umfasst dieses Vorgehensmodell auch die prozessbegleitende Anwendung von V&V-Maßnahmen (vgl. Abschnitt 3.2). Aus diesem Grund erfolgt in den jeweiligen Phasen des MESC eine Zuteilung der in Kapitel 4 ausgewählten V&V-Techniken. Um der Bedeutung der V&V im Rahmen des Vorgehensmodells gerecht zu werden, soll abschließend an einigen ausgewählten Beispielen noch einmal detaillierter auf die tatsächliche Anwendung der V&V eingegangen werden.

5.1 Aufgabenbestimmung und Datenauswahl

Dieser erste Abschnitt stellt die ersten beiden Phasen des MESC dar. Dabei findet eine Erläuterung der Problemstellung des Praxisfalls sowie der darauf basierenden Auswahl relevanter Datenbestände statt.

5.1.1 Aufgabendefinition

Die erste Phase einer jeden Ausführung des MESC besteht in der Auseinandersetzung mit den Gegebenheiten in dem Unternehmen und dem Gegenstand der Untersuchung. Dies wird in einem Dokument zur Beschreibung der *Aufgabenstellung* zusammengefasst (vgl. Abschnitt 2.2.3).

Aufgabenstellung (1.1)

Wie in Abschnitt 2.2.3 beschrieben, bedarf die Durchführung des KDD in der Praxis einer konkreten Definition der Fragestellung. Dazu müssen Randbedingungen betrachtet werden, die als Grundlage zur Ableitung der Ziele dienen. Die Aufstellung von Rand-bedingungen und Zielsetzung der Untersuchung kann an dieser Stelle nur in Zusammenarbeit von Entwicklern und Unternehmen stattfinden.

Randbedingungen: Beim betrachteten Anwendungsfall dieses Praxisbeispiels handelt es sich um ein produzierendes Unternehmen aus der Elektronikbranche. Dabei sollen produktionslogistische Vorgänge untersucht werden, für die drei verschiedene Arten von Randbedingungen festgehalten werden: fachliche, technische sowie zeitliche.

Bei den fachlichen Aspekten ist neben einer Beschreibung des Systems (Produktionslayout, Materialfluss etc.) beispielsweise auch der Untersuchungsgegenstand festzuhalten. Im vorliegenden Fall stellen dies Prüfvorgänge aus dem Qualitätsmanagement

des Unternehmens dar, die an verschiedenen Arbeitsplätzen durchgeführt werden. Dabei sind Elektrokleingeräte zu kontrollieren, die zuvor in verschiedenen Produktionslinien gefertigt wurden. Elektrokleingeräte bezeichnen dabei Elektrogeräte mit einem Gewicht kleiner fünf Kilogramm (Arnold 2008). Dies können etwa Küchengeräte wie Brotschneidemaschinen oder auch andere technische Gerätschaften wie Drucker, Bildschirme oder Smartphones sein. Bei den technischen Randbedingungen ist zu diskutieren, welche Quellsysteme für die in der zweiten Phase durchzuführende Datenauswahl zur Verfügung stehen und wie auf diese zugegriffen werden kann. Zeitlich gilt es einen geeigneten Beobachtungszeitraum zu bestimmen. Für die Auswertung der Prüfvorgänge wird an dieser Stelle die Beobachtungsgrundlage auf Datenbestände des Jahres 2015 festgelegt.

Zielbedingung: Das Unternehmen verfolgt bei der Durchführung dieser Untersuchung generell zwei wesentliche Punkte. Einerseits sollen die Daten auf Optimierungspotentiale überprüft werden, andererseits dient die Ausführung des Data-Mining-Verfahrens einer generellen Eignungsprüfung der Daten für Datenanalyseverfahren wie das Data Mining. Das Ziel der Anwendung des MESC besteht hier darin, die Datenbestände zu den Prüfverfahren der Produktion auf Ursachen von Fehlern zu untersuchen.

V&V der Aufgabendefinition

In dieser frühen Phase des MESC sind lediglich die Randbedingungen sowie die festgelegten Zielbedingungen zu überprüfen (Intrinsische Prüfung; 1,1) (vgl. Abschnitte 3.2 und 0). Diese Prüfung kann – wie bereits die Festlegung im vorliegenden Fall – nur in enger Absprache mit dem Unternehmen und unter Einbeziehung von Kontextwissen erfolgen, da von Entwicklerseite keine Aussagen über die Korrektheit der formulierten Aussagen getroffen werden können. Dies geschieht im Rahmen eines Review-Termins mit Fachleuten der Auftraggeber- und Auftragnehmerseite zur Begutachtung der festgehaltenen Bedingungen. Dazu werden die in Tabelle 4.2 dargestellten Kriterien mit Fragen zur Kontrolle der Aufgabendefinition auf Korrektheit und Angemessenheit der festgelegten Bedingungen sowie Durchführbarkeit des Verfahrens abgeprüft. Im Vorfeld des Review-Termins wird von den Beteiligten eine Auseinandersetzung mit der Thematik zur Vorbereitung des Termins erwartet (Schreibtischtest & Strukturiertes Durchgehen). Gerade für die Beteiligten auf Kundenseite ist es in Vorbereitung des Treffens bedeutsam die Bedingungen mit weiteren Mitarbeitern des Unternehmens zu diskutieren, da nicht jeder Mitarbeiter dasselbe fundierte Wissen über die Abläufe in der Produktion hat.

5.1.2 Auswahl der relevanten Datenbestände

Nach der Festlegung der Untersuchungsziele und Randbedingungen in der ersten Phase sind im nächsten Schritt des MESC relevante Daten zu beschaffen und auf eine geeignete Datenauswahl zu reduzieren.

Datenbeschaffung (2.1)

Die Beschaffung der relevanten Datenbestände für die Untersuchung beinhaltet die Auswahl benötigter Daten unter Einbeziehung der im ersten Schritt festgelegten Aufgabendefinition sowie das Auslesen der Daten aus einer Datenquelle. In dem vorliegenden Anwendungsfall geht es – wie bereits in Abschnitt 5.1.1 beschrieben – um eine Untersuchung der Abläufe bei Prüfungsverfahren der Produktion. Dafür wurden dem Lehrstuhl ITPL durch das Produktionsunternehmen Informationen in Form von Datenbankauszügen zur Verfügung gestellt. Diese sollen im Folgenden unter fachlichen und technischen Gesichtspunkten detaillierter beschrieben werden.

Fachliche Beschreibung: Bei den Ausgangsdaten handelt es sich um Datenbankauszüge aus einem Betrachtungszeitraum von zwei Jahren von Mitte 2014 bis Mitte 2016. Die Daten stammen dabei aus insgesamt drei Datenbanken. Zwei dieser Datenbankauszüge beinhalten Informationen aus dem Qualitätsmanagement des Unternehmens. Hauptsächlich unterscheiden sich diese Datenbanken durch die darin enthaltenen Ausprägungen des Attributs *Linie*. Die gelieferten Auszüge enthalten über 200 Millionen Datensätze mit zusammengenommen ca. 600 Attributen in etwa 50 Tabellen. Der Großteil der Tabellen ist in zweien der Datenbanken in identischer Struktur vorhanden. Einige Tabellen sind dabei aber auch vom Inhalt identisch oder weisen nur kleinere Unterschiede auf (z.B. *Parameterbeschreibungen* oder *Werke*). Lediglich einige wenige Tabellen sind ausschließlich in einer der Datenbanken enthalten (z.B. *Parameter* oder *Resultat*). Die dritte Datenbank beinhaltet darüber hinaus eine Tabelle mit weiteren Daten aus der Produktion (z.B. Scanner, Batch-Nr.). Die gelieferten Daten enthalten sowohl Beziehungen zwischen den Endprodukten und ihren Komponenten, als auch zwischen den Endprodukten und der Linie, auf der sie gefertigt wurden. Darüber hinaus enthalten die Ausgangsdaten Informationen über Prüfverfahren, die die Produkte durchlaufen haben sowie Ergebnisse dieser Prüfungen. Beispiele der enthaltenen Attribute sind in Tabelle 5.1 dargestellt.

Technische Beschreibung: In dem Ausgangsdatenbestand sind die Attribute in Form verschiedener Datentypen abgebildet (Erläuterung der Datentypen in Abschnitt 2.3). Hauptsächlich sind die Attribute als Zahlenwerte oder Zeichenfolge dargestellt, aber auch Datumswerte sind vorhanden. Die Datentypen werden in Tabelle 5.2 anhand der zuvor beschriebenen Beispielattribute aufgezeigt. Ergänzend zu den hier gegebenen fachlichen und technischen Beschreibungen befindet sich eine Beschreibung aller in den betrachteten Tabellen enthaltenen Attribute sowie ihrer Ausprägungen und Auffälligkeiten bei der Untersuchung im gesperrten Anhang.

Tabelle 5.1: Fachliche Beschreibung von Beispielattributen

Attribut	Beschreibung
WerkstückGuid	Bezeichner für einen Durchlauf eines Prüfvorgangs
Result_ORP	Ergebnis eines Prüfschritts
Arbeitsbeginn	Zeitstempel, der den Beginn eines Arbeitsvorgangs kennzeichnet
Linie	Bezeichner für eine Fertigungslinie
Arbeitsplatz	Bezeichner für einen Arbeitsplatz innerhalb des QM-Bereichs
Transportmittel	Bezeichner für genutztes Transportmittel

Tabelle 5.2: Vorkommende Datentypen in den Ausgangsdaten

Datentyp	Beispielhafte Attribute
Datetime	Arbeitsbeginn, Arbeitsende
Ganzzahl	Arbeitsplatz
Gleitkommazahl	Höhe, Gewicht
Zeichenkette	Resultat, Transportmittel
Eindeutiger Bezeichner (Unique Identifier)	WerkstückGuid

Datenauswahl (2.2)

Nachdem die Datenbeschaffung abgeschlossen ist, dient der zweite Schritt dieser Phase der Auswahl geeigneter Daten. Dies hilft, die Datenmenge zu reduzieren und somit den Zeitaufwand der Vorverarbeitung der Daten und des anschließenden Data-Mining-Verfahrens zu verringern. Bei der Analyse des Datenbestands ist erkennbar, dass nicht alle Tabellen relevante Informationen für die Durchführung des Data-Mining-Verfahrens liefern. So enthalten beispielsweise Tabellen mit Bestellungen keine Informationen hinsichtlich der gegebenen Aufgabenstellung. Insgesamt fünf Tabellen ausgewählt, die für die weiteren Untersuchungen verwendet werden sollen (siehe Tabelle 5.3). In Anbetracht der großen Datenmenge wird darüber hinaus – wie im ersten Schritt festgelegt – der Untersuchungszeitraum auf das Jahr 2015 eingeschränkt.

V&V der Auswahl relevanter Datenbestände

Bei der Prüfung der zweiten Phase gegen sich selbst müssen die Korrektheit und Relevanz der verwendeten Daten und Datenbestände kontrolliert werden (2,2). Wie bereits in Phase 1 ist auch dieser Vorgang nur unter Einbeziehung von Kontextwissen möglich. Dies erfolgt hier durch eine erneute Durchführung eines Review-Termins mit dem Unternehmen, bei dem etwa die Festlegung auf zu analysierende Tabellen untersucht werden. kann. Zur Vorbereitung des Treffens ist es bedeutsam, dass beide Seiten vorab die Durchführbarkeit der Datenentnahme überprüfen. Dies kann beispielsweise durch eine Kontrolle der Zugriffsverfügbarkeit der Datenquelle geschehen. Bei der intrinsischen Prüfung stellt sich eine der Tabellen – die sogenannte Rückverfolgbarkeitstabelle – als für die Analyse nicht

Tabelle 5.3: Beschreibung der ausgewählten Tabellen

Tabelle	Beschreibung	Beispielattribute
Verfahrensprotokoll	Daten ausgeführter Prüfverfahren	Arbeitsablauf, Schicht, Result-Code, Arbeitsbeginn, Arbeitsende
Protokoll der Verfahrensergebnisse (ORP)	Ergebnisse der ausgeführten Schritte der Prüfverfahren in Ergänzung der Operation Protocol-Tabelle	Parameterbeschreibung, Wert, Resultat
Parameterbeschreibung	enthält die Parameter der einzelnen Prüfverfahren	Parameter, Minimalgrenzwert, Maximalgrenzwert
Werkstück	Beschreibung der Werkstücke	WerkstückGuid, Anzahl Reparaturen
Rückverfolgbarkeit	enthält weitere produktionslogistische Informationen	Lieferanten-Chargennummer, Scannertyp, Scannerwert, Bezeichner

nutzbar dar. Die genauen Gründe sollen in Abschnitt 5.4 genauer erläutert werden. Die zweite Prüfung dieser Phase dient der Kontrolle der Datenauswahl auf ihre Eignung im Hinblick auf die in der ersten Phase definierte Zielbedingung (2,1). Dabei wird die Auswahl der Tabellen als für die Aufgabenstellung und somit für eine weitere Nutzung geeignet erkannt.

5.2 Vorverarbeitung der Daten für das Data Mining

Um die in der zweiten Phase ausgewählten Daten für die Nutzung im Data-Mining-Verfahren vorzubereiten, müssen verfahrensunabhängige Methoden der Datenvorverarbeitung angewendet werden (vgl. Abschnitt 2.3.1). Bei der vorliegenden Arbeit wurde auf das Gruppieren von Attributen (Schritt 3.2) verzichtet, da es als nicht zielführend erkannt wurde. Ebenfalls nicht durchgeführt wurde die Anreicherung der Daten mittels Kontextwissen (3.3), da auch hier kein Bedarf einer solchen gesehen wurde. Aufgrund der Größe der zugrundeliegenden Datenbestände (vgl. Abschnitt 5.1.2) sowie der Tatsache, dass nicht alle Vorverarbeitungsschritte vollautomatisiert vollzogen werden können, ist bereits vorab das Ziehen einer Stichprobe sinnvoll. Dabei wird in dieser Arbeit zuerst eine Kombination aus repräsentativer sowie inkrementeller Stichprobenziehung mit zufälligen Datensätzen gewählt (vgl. Abschnitt 2.3.1), um sich ein Bild der bestehenden Daten zu

machen und die Umsetzung der Modellentwicklung in Phase 5 zu vereinfachen. Dabei werden in der Größe steigende Stichproben mit bis zu 400.000 Datensätzen gezogen.

Formatstandardisierung (3.1)

Zur Durchführung von Data-Mining-Verfahren müssen die Daten aus den verschiedenen Quellformaten standardisiert und in eine Gesamttabelle überführt werden (Schritt 3.1). Im vorliegenden Fall liegen die Informationen bereits größtenteils in Relationen vor. Lediglich in der Tabelle *Parameterbeschreibung* existieren hierarchische Beziehung innerhalb eines Attributs, was die dritte Normalform der relationalen Datenbanken verletzt. Zur Lösung dieses Problems kann die Tabelle in zwei neue Tabellen zerlegt werden. Dabei werden nur die Attribute mit direkter Abhängigkeit vom Attribut *Parameterbeschreibung* in einer Tabelle mit diesem Attribut gespeichert. Alle anderen werden aus der Tabelle *Parent-Parameterbeschreibung* geerbt (siehe Tabellen A.1 - A.3 des Anhangs).

Zur Verknüpfung der ausgewählten Tabellen werden verschiedene Attribute in Tabellen als Schlüssel verwendet. Beispielsweise dient das Attribut *Parameter* in Tabelle *Protokoll der Verfahrensergebnisse* als verknüpfendes Attribut mit der Tabelle *Parameterbeschreibung*. Anhand dieser Schlüssel können die Tabellen miteinander verbunden und zu einer Gesamttabelle zusammengefügt werden. Durch diesen Vorgang entsteht aus den Haupttabellen des Qualitätsmanagements eine neue Tabelle mit insgesamt ca. 100 Attributen. Zusätzlich dazu soll auch die Rückverfolgbarkeitstabelle mit den anderen Tabellen kombiniert werden. Dies ist aus mehreren Gründen nicht ohne größeren Aufwand möglich. Auf die genauen Gründe soll in Abschnitt 5.4 genauer eingegangen werden.

Transformation (3.4)

Ziel der Transformation ist die Reduzierung der Daten. Dies kann wie in Abschnitt 2.3.1 geschildert mittels Aggregation, Stichprobenziehung oder Dimensionsreduktion erfolgen. Darüber hinaus ist auch die Reduzierung der Anzahl der Attribute sowie verschiedene weitere Transformationsmethoden durchzuführen. Bei der Betrachtung der ausgewählten Tabellen ist festzustellen, dass viele der darin enthaltenen Attribute für eine Analyse mittels Data-Mining-Verfahren aus verschiedenen Gründen nicht geeignet sind. Ein häufiger Grund dafür ist, dass diese Attribute nicht gefüllt sind (leer oder „NULL“) oder lediglich eine Standardausprägung aufweisen. Ein weiterer Grund sind Redundanzen. Einige Attribute sind in mehreren Tabellen mit denselben Ausprägungen enthalten (z.B. Linie). Darüber hinaus existieren auch innerhalb der Tabellen Redundanzen (z.B. Arbeitsende und Zeitstempel). Auch Ausreißer und fehlende Werte innerhalb der Ausprägungen müssen beseitigt werden. Eine Übersicht einiger ausgewählter Beispiele ist in Tabelle 5.4 gegeben. Darüber hinaus findet sich eine detailliertere Betrachtung der Datenvorverarbeitung bei Li (2016).

Tabelle 5.4: Beispiele für Auffälligkeiten in den Datenbeständen

		Beispiel	
Auffälligkeit	**Beschreibung**	**Attribute**	**Ausprägungen**
leere Attribute	Attribute weisen nur Ausprägung 0 oder „NULL“ auf		immer „NULL“, immer 0
mehrfach vorkommende Attribute (Redundanz)	Attribute weisen identische Werte mit weiteren Attributen auf	Arbeitsende & Zeitstempel, Linie	
lediglich eine Ausprägung	Attribute brauchen mehrere Ausprägungen, sonst unbrauchbar	Werk, Routing,	immer 001, immer Rout0000003
Datentypen	alle Attribute in Rückverfolgbarkeitstabelle sind als Zeichenkette gespeichert		
Asiatische Zeichen	Nicht interpretierbare Zeichen	Wert	鱸蒯濛娣عي怃策

V&V Datenvorverarbeitung

Bei der intrinsischen Prüfung dieser Phase (3,1) muss kontrolliert werden, ob die Schritte der Datenvorverarbeitung korrekt durchgeführt wurden. Im Praxisbeispiel ist hier etwa die richtige Ausführung der Stichprobenziehung oder die Auflösung der hierarchischen Beziehungen zu untersuchen. Darüber hinaus sind diese Schritte auch gegen die Ausgangsdaten zu prüfen (3,2). So ist bei der Normalisierung abzuklären, ob die in Abschnitt 2.3.1 eingeführten Kriterien der Verlustlosigkeit sowie der Abhängigkeitserhaltung nach Durchführung der Normalisierung erfüllt sind. Bei der Kontrolle der Daten im Hinblick auf die Zielerreichung (3,1) ist vor allem darauf zu achten, dass die transformierten Daten noch für die Erreichung der Ziele geeignet sind. Zur formalen Untersuchung werden hier statistische Techniken angewendet. Darüber hinaus müssen Begutachtungstechniken eingesetzt werden, die sowohl selbstständig (Schreibtischtest, Dokumentenüberprüfung) als auch in Kooperation mit dem Unternehmen zur Abklärung offener Fragen (Strukturiertes Durchgehen, Review) durchgeführt werden.

5.3 Durchführung des Data Minings auf einen Datenbestand der Produktionslogistik

Dieser Abschnitt befasst sich mit den übrigen Phasen des MESC, die nach der Festlegung der Aufgabendefinition sowie der Auswahl und der verfahrensunabhängigen Vorver-

arbeitung der Daten durchgeführt werden müssen. Dies sind neben der Vorbereitung und Anwendung des Data-Mining-Verfahrens (Phasen 4 und 5) die Weiterverarbeitung der Data-Mining-Ergebnisse (Phase 6). In der letzten Phase (Phase 7) wird sowohl eine qualitative Bewertung des durchgeführten Prozesses als auch eine auf den Ergebnissen basierende Ableitung von Handlungsanweisungen vorgenommen. Da diese Phase aus zeitlichen Gründen nicht mehr in dieser Arbeit behandelt werden kann, wird an dieser Stelle auf eine Beschreibung verzichtet und auf die allgemeinen Ausführungen im vorherigen Verlauf hingewiesen (vgl. Abschnitte 2.2.3, 3.2 und 4.2.7).

5.3.1 Vorbereitung des Data-Mining-Verfahrens

Nach Abschluss der verfahrensunabhängigen Schritte muss in dieser Phase die Anwendung des Data-Mining-Verfahrens vorbereitet werden. Dazu sind die Schritte der Verfahrens- und Werkzeugauswahl sowie der für das Verfahren notwendigen fachlichen und technischen Kodierung notwendig.

Verfahrensauswahl (4.1)

Bei der Auswahl des anzuwendenden Data-Mining-Verfahrens müssen die zugrundeliegenden Daten sowie die festgelegten Randbedingungen und Fragestellungen des KDD-Prozesses berücksichtigt werden. Im vorliegenden Fall handelt es sich um einen produktionslogistischen Datenbestand, bei dem die Informationen über die Produktion und durchzuführende Prüfverfahren in verschiedenen Systemen und Datenbanken vorliegen (vgl. Abschnitt 5.1). Von besonderem Interesse erscheint die Untersuchung möglicher Wirkzusammenhänge zwischen den verschiedenen Faktoren der Produktion. Ein Verfahren, das für die Betrachtung einer solchen Ausgangssituation als sehr geeignet erscheint, ist die Assoziationsanalyse, da dieses genau dem Aufdecken unbekannter Zusammenhänge und Muster in Datenbeständen dient. Eine detaillierte Beschreibung der Assoziationsanalyse findet sich in Abschnitt 2.4.2. Für die Durchführung der Assoziationsanalyse wird der *FP-Growth-Algorithmus* gewählt, da dieser im Vergleich zu anderen Algorithmen – etwa dem *Apriori-Algorithmus* – deutlich schneller ist (siehe Tabelle 2.1). Überdies ist der FP-Growth-Algorithmus in der gewählten Software standardmäßig vorhanden. Neben der Assoziationsanalyse scheint auch die Clusteranalyse für die Ausgangssituation geeignet. Eine ausführliche Behandlung dieser Technik und ihrer Durchführung innerhalb des MESC für dieses Fallbeispiel erfolgt bei Li (2016).

Werkzeugauswahl (4.2)

Wie in den Abschnitten 2.2.3 und 3.3 dargestellt, erfolgt die Umsetzung von Data-Mining-Prozessen größtenteils unter Einsatz von Standardsoftware. Bei der Durchführung der Assoziationsanalyse im vorliegenden Anwendungsfall wird die Software Rapid-Miner ausgewählt RapidMiner ist das am häufigsten eingesetzte Data-Mining-Werkzeug

und kommt darüber hinaus standardmäßig am Lehrstuhl ITPL zum Einsatz. Dadurch ist es für die Ausführungen leicht zugänglich. Darüber hinaus wird mit weiteren Standardprogrammen gearbeitet – etwa zur Datenabfrage (SQL) oder zur Datenverarbeitung (Excel).

Fachliche und technische Kodierung (4.3, 4.4)

Neben den verfahrensunabhängigen Datenvorverarbeitungsschritten (vgl. Abschnitt 0) sind noch weitere Schritte nach Auswahl des passenden Verfahrens notwendig (vgl. Abschnitt 2.3.2). Dies erfolgt sowohl auf Grundlage von Kontextwissen (fachliche Kodierung) als auch auf Grundlage von erforderlichen Formaten der Attributsausprägungen (technische Kodierung). An dieser Stelle sollen einige Beispiele der Kodierung angeführt werden.

Umwandlung von Datumswert: Zur Gewinnung neuer Informationen lassen sich Datumswerte aufteilen und umwandeln. Dieses Verfahren wird im vorliegenden Fall auf das Attribut *Arbeitsbeginn* angewendet, um Aussagen über den Wochentag und die Schicht des Arbeitsbeginns treffen zu können. So ist es möglich das Element Wochentag mit den Ausprägungen Montag, Dienstag, Mittwoch, Donnerstag und Freitag sowie das Element Schicht mit den Ausprägungen Früh-, Spät-, und Nachtschicht zu gewinnen. Der Umwandlungsprozess ist exemplarisch in Abbildung A.2 des Anhangs dargestellt.

Innere Struktur: Bei relationalen Datenbanken wird nach der ersten Normalform eine Atomarität der Attribute gefordert. Zusammengesetzte Attribute sind aus diesem Grund nicht zulässig (vgl. Abschnitt 2.3.2). Diese Forderung wird bei den vorliegenden Datenbeständen nicht immer erfüllt. So weisen einige Attribute eine innere Struktur auf oder sind aus anderen zusammengesetzt. Die *Seriennummer* beispielsweise kann in seiner vorliegenden, ID-ähnlichen Form keinen Mehrwert für das Assoziationsverfahren liefern und müsste dementsprechend aus der Auswahl entfallen. Bei Seriennummern handelt es sich allerdings oftmals um zusammengesetzte Attribute, die in ihrer inneren Struktur Elemente wie etwa das Baujahr enthalten können. Deshalb sind eine Zerlegung des Attributs und eine anschließende Prüfung der einzelnen Elemente auf eine weitere Verwendung hin möglich. Die Zusammensetzung der Seriennummer und eine Beschreibung der Elemente ist Abbildung A.3 zu entnehmen. Nach Rücksprache mit dem Fachexperten kommen für die Durchführungen dieser Arbeit lediglich die Attribute Gerätekennzahl (GZ) und Konstruktionsstand (KS) in Betracht.

Diskretisierung: Im vorliegenden Fall existiert das Attribut *Anzahl der Reparaturen* mit Ausprägungen von 0 bis 15. Da die exakte Anzahl von Reparaturen für die betrachtete Ausgangslage irrelevant ist, erscheint eine Einteilung der Ausprägungen in Intervalle sinnvoll (vgl. Abschnitt 2.3.2). Zur Einteilung der Klassen werden statistische Techniken eingesetzt, die aufzeigen, dass die Anzahl der Werte mit steigender Reparaturanzahl

rapide abnimmt und der Großteil der Werte von 1 bis 7 liegt. Aus diesem Grund wird hier entschieden, die in Tabelle 5.5 dargestellte Einteilung zu wählen.

Tabelle 5.5: Einteilung der Klassen bei Durchführung der Diskretisierung

Klasse (Kurzform)	Grenzen
keine Reparaturen (keine)	0
geringe Anzahl Reparaturen (gering)	1 - 2
mittlere Anzahl Reparaturen (mittel)	3 - 4
hohe Anzahl Reparaturen (hoch)	4 - ∞

V&V der Verfahrensvorbereitung

Auch in dieser Phase des MESC sind verschiedene Maßnahmen der V&V notwendig (vgl. Abschnitt 4.2.4). Bestandteil der intrinsischen Prüfung (4,4) ist dabei etwa die Kontrolle des ausgewählten Data-Mining-Verfahrens auf seine Nutzbarkeit für die vorliegende Aufgabe. Dabei zeigt sich, dass das Verfahren geeignet ist, um Wirkzusammenhänge zwischen den Attributen der Produktion zu bestimmen. Weiterhin ist das Verfahren hinsichtlich seiner Eignung für die Aufgabenstellung (4,1) zu untersuchen. Auch die Aufbereitung der Daten (4,3) sowie die grundlegende Auswahl der Daten (4,2) sind im Hinblick auf die Assoziationsanalyse erneut auf ihre Korrektheit zu kontrollieren. Die Prüfungen dieser Phase sind zwingend in enger Abstimmung mit dem Kunden durchzuführen. So ist beispielsweise die innere Struktur der Seriennummer nur durch die Kommunikation mit dem Unternehmen zu entschlüsseln. Der Austausch von Informationen mit den Experten kann hierbei auch außerhalb der Review-Termine per Telefon oder E-Mail erfolgen.

5.3.2 Anwendung des Data-Mining-Verfahrens und Weiterverarbeitung der Ergebnisse

Nach Abschluss der für das ausgewählte Verfahren notwendigen vorbereitenden Maßnahmen wird in der fünften Phase das eigentliche Data-Mining-Verfahren durchgeführt. Dies erfolgt in den zwei aufeinanderfolgenden Schritten der *Entwicklung eines Data-Mining-Modells (5.1)* sowie des *Trainings des entwickelten Modells (5.2).* Da im Folgenden mehrere Modellentwicklungs- und Modelltrainingsprozesse beschrieben werden sollen, wird auf eine Unterteilung in die genannten Schritte verzichtet. Aus diesem Grund wird in diesem Abschnitt auch bereits die Weiterverarbeitung der Data-Mining-Ergebnisse (Phase 6) dargestellt. Eine separate Darstellung dieser Ergebnisse scheint aufgrund der wiederholten Ausführung des Verfahrens wenig zielführend.

Entwicklung und Training der Modelle

Nach Auswahl der Assoziationsanalyse als anzuwendendes Data-Mining-Verfahren (vgl. Abschnitt 5.3.1), dient dieser Schritt der Entwicklung eines Modells. Dazu liefert die genutzte Software verschiedene Bausteine, etwa zur Anwendung des FP-Growth-Algorithmus und anschließender Darstellung von Assoziationsregeln oder der zuvor nötigen Schritte der Datenvorverarbeitung (Korrelationsanalyse etc.). Aufgrund der Menge der Datenbestände wurden diese bereits in Phase 3 mittels Stichprobenziehung reduziert (vgl. Abschnitt 0).

Zur Durchführung der Entwicklung sowie dem späteren Training ist es notwendig, diese Daten weiter zu unterteilen. Dies sichert gleichzeitig auch die Qualität der Entwicklung und entspricht somit einer Form der V&V. Bei der Aufteilung der Daten wird bei diesem Praxisfall die sogenannte Holdout-Methode verwendet. Diese ist anderen gängigen Aufteilungsarten (Kreuz- oder Bootstrapping-Validieurung) für die Anwendung bei großen Datenbeständen mit mehr als 1000 Datensätzen überlegen (vgl. Abschnitt 3.3.3 und 4.1.2). Dabei gibt es bei der Holdout-Methode verschiedene Möglichkeiten für die Splittung der Daten in Trainings-, Test- und/oder Validierungsdaten. In dieser Arbeit wird die gezogene Stichprobe in Anlehnung an die Methode von Urban (vlg. Abschnitt 3.3.3) zuerst in eine Trainings- und Validierungsdatenmenge sowie eine Testdatenmenge unterteilt. Dabei beträgt das Verhältnis von Trainings- und Validierungsdaten zu Testdaten sowie von Trainings- zu Validierungsdaten jeweils 4:1 (siehe Tabelle 5.6). Diese Aufteilung ist sinnvoll, um die Entwicklung des Modells und die Bestimmung der Modellparameter unter Verwendung der Validierungsdaten durchführen und im nächsten Schritt das Modell mit Hilfe der Trainingsdaten trainieren zu können. Daneben dienen die Testdaten der späteren Validierung der Ergebnisse auf ihre Allgemeingültigkeit. Damit wird auch verhindert, dass alle Daten zum Training des Modells verwendet werden und es zu einem Overfitting des Modells kommt (vgl. Abschnitt 3.3.3).

Tabelle 5.6: Aufteilung der Datenmenge in Trainings-, Validierungs- und Testdaten

Art der Datenmenge	Anteil an Gesamtmenge	Anteil an Untermenge
Trainingsdaten	80%	80%
Validierungsdaten		20%
Testdaten	20%	-

Für die Entwicklung des Data-Mining-Modells stellt der RapidMiner notwendige Bausteine der Vorverarbeitung sowie Anwendung der Data-Mining-Prozesse zur Verfügung. Unter Zuhilfenahme dieser Bausteine kann ein Modell der Assoziationsanalyse erstellt werden (siehe Abbildung 5.1). Die einzelnen Bausteine dieses Modells sollen im Folgenden kurz vorgestellt werden.

Datenabfrage: Theoretisch lässt das verwendetete Programm Abfragen der Daten direkt aus den Datenbanken zu. Da aber außerhalb der Programmumgebung bereits Vorverarbeitungsschritte durchzuführen sind (z.B. Stichprobenziehung, Verknüpfung der Tabellen und Transformationsschritte), kann keine direkte Einbindung der Datenbanken erfolgen. Aus diesem Grund werden die zu nutzenden Datenauszüge im CSV-Format gespeichert und abgefragt.

Datenvorverarbeitung: Das Modell enthält automatische Schritte der in Abschnitt 0 beschriebenen verfahrensunabhängigen Datenvorverarbeitung. Dies umfasst neben dem Entfernen korrelierender und unnötiger Werte auch das Ersetzen fehlender Werte. Da bei diesen automatischen Reduzierungsschritten nicht immer alle irrelevanten Attribute korrekt herausgefiltert werden, ist eine anschließende manuelle Sichtung und Selektion notwendig. So kann auch die vorselektierte Auswahl weiter variiert werden, um eine gezielte Betrachtung bestimmter Attribute zu ermöglichen. Tabelle 5.7 stellt beispielhaft die finale Auswahl der Attribute des ersten Versuchs der Assoziationsanalyse dar. Dabei sind die Elemente der Seriennummer noch nicht berücksichtigt, da eine Zerlegung in die Einzelteile erst im Verlauf der Untersuchungen erfolgte. Nach Auswahl der Vorverarbeitungsschritte verbleiben also lediglich sieben Attribute. Der letzte Schritt der Datenvorverarbeitung dient der Diskretisierung, deren Durchführung in Abschnitt 5.3.1 beschrieben wurde.

Tabelle 5.7: Auswahl der zu nutzenden Attribute für die Assoziationsanalyse (Versuch-Nr. 1)

Selektierte Attribute	Beschreibung
Linie	Produktionslinie
Anzahl Reparaturen	gibt die nötige Anzahl der Reparaturen an
Parameterbeschreibung	Identifier, unter der die Parameterbeschreibung zu finden ist
Resultat_ORP	Resultate der Prüfungen aus Tabelle *OperationResult-Protocol* (PASS, FAIL, …)
Result-Code	Code der Fehlerbeschreibung bei auftretenden Fehlern
Arbeitssequenz	Arbeitsschritt-Nummer
Arbeitsplatz	Arbeitsplatz-Nummer

Die in den vorliegenden Datenbeständen enthaltenen Attribute weisen verschiedene Datentypen auf (vgl. Abschnitt 0). Da für die Assoziationsanalyse binäre Attributsausprägungen benötigt werden, müssen hier verschieden Schritte zur eine Änderung des Datentyps vorgenommen werden (numerisch zu polynominal, numerisch zu binär, nominal zu binär) (vgl. Abschnitt 2.3.2).

Assoziationsanalyse: Die letzten beiden Bausteine stellen die eigentliche Ausführung der Assoziationsanalyse dar. Dabei lassen sich die in Abschnitt 2.4.2 vorgestellten Schritte des Findens häufiger Itemsets sowie des Generierens von

Assoziationsregeln als Bausteine nachbilden, die durch Variierung der Parameter angepasst werden können (Interessantheitsmaße, vgl. Abschnitt 2.4.2). Zu beachten ist hierbei, dass bei der Nutzung des Bausteins zur Generierung von Assoziationsregeln lediglich einer der Mindestwerte der Interessantheitsmaße (Konfidenz, Lift, Conviction, p-s, Gain und Laplace) aktiv festgelegt werden kann. Eine Regelerstellung mit einer Kombination aus Mindestwerten meherer Parameter ist dementsprechend nicht möglich. Zur Durchführung der generellen Modellbildung wird in dieser Arbeit der Parameter Konfidenz verwendet, da dieser das geläufigste Kriterium der Regelerkennung darstellt (vgl. Abschnitt 2.4.2).

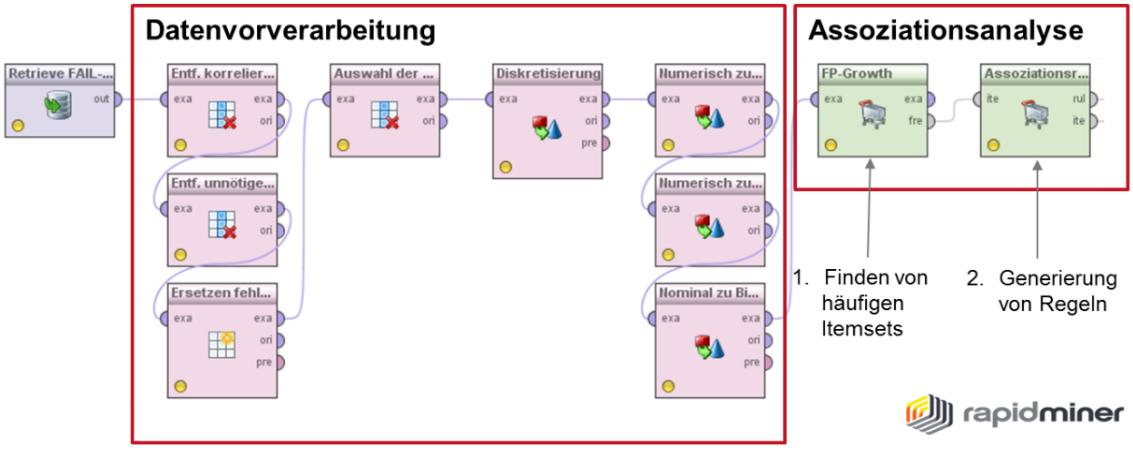


Abbildung 5.1: Beispielhafte Modellierung eines Data-Mining-Prozesses mittels RapidMiner

Versuchsdurchführung

Durch die Variierung der Modellbausteine sowie der darin enthaltenen Parameter – etwa Support, Konfidenz oder Auswahl der Attribut – kann eine Vielzahl an Versuchen zur Modellbildung durchgeführt werden. Da eine Darstellung aller Versuche den Rahmen dieser Arbeit übersteigen würde, soll im Folgenden anhand geeigneter Beispiele das Vorgehen und das Finden von Regeln dargestellt werden. Eine ausführlichere Beschreibung der Stichproben-, Attributs- sowie Parameterauswahl der Versuche findet sich im gesperrten Anhang. Darüber hinaus finden sich dort auch alle gefundenen Regeln sowie entwickelte Modelle.

Zur Reduzierung des Aufwands zur Datenentnahme werden zu Beginn zwei Vereinfachungen betreffend der Stichprobengröße und -quelle festgelegt. Wegen der goßen Gesamtdatenmenge in verschiedenen Datenbanken sollen erste Annahmen unter Ziehung einer vergleichsweise kleinen Stichprobe von 10.000 Datensätzen aus lediglich einer Datenbank getroffen werden. In dieser Datenbank sind nur drei der insgesamt sechs Ausprägungen des Attributs *Linie* enthalten (2, 3, 4). Die Datenmenge ist hier deshalb so gering gewählt, da es in diesem ersten Schritt weniger um finale Ergebnisse, als um eine

Bestimmung der Modellparameter sowie der Ableitung erster Annahmen und des weiteren Vorgehens geht. Aus diesem Grund wäre eine höhere Datenmenge zur Validierung der Teilprozesse des Modells aus zeitlichen Gründen wenig zielführend.

Ein erster Versuch erfolgt auf Grundlage der in Tabelle 5.7 festgelegten Attribute sowie Support- und Konfidenzwerten von 0,8. Durch diese Einstellung ist sichergestellt, dass nur Regeln mit einem starken Wirkzusammenhang der Attribute erzeugt werden. Bei Betrachtung der gefundenen Regeln ist zu erkennen, dass zwar eine Vielzahl an Regeln generiert werden kann, aber das Ziel der Assoziationsanalyse – das Entdecken unbekannter Wirkzusammenhänge – bei dieser Kombination aus Attributenauswahl und Parametereinstellungen nicht erfüllt wird. Es werden lediglich erwartbare Zusammenhänge – wie etwa zwischen einem bestandenen Prüfverfahrensschritt und dem Result-Code 0 – aufgezeigt. Hierbei ist allerdings auffällig, dass die Konfidenz nicht den Wert 1 aufweist. Dieser wäre zu erwarten, da es fachlich logisch erscheint, bei einer bestandenen Prüfverfahren (Resultat_ORP = „PASS“) den Result-Code 0 zu vergeben. Dies könnte auf ein Problem der Datenqualität hinweisen (vgl. Tabelle 5.8).

Tabelle 5.8: Ausgewählte Ergebnisse der Durchführung der Assoziationsanalyse (Versuch-Nr. 1)

Nr.	Prämisse	Konklusion	Support	Konfidenz
124	Resultat_ORP = PASS	Result-Code = 0	0,949	0,987
33	Result-Code = 0	Anzahl Reparaturen = keine	0,888	0,904

Unter Berücksichtigung dieser Auffälligkeit wird für die weiteren Untersuchungen die Festlegung getroffen, dass diese, der fachlichen Logik widersprechenden Regeln von der Betrachtung ausgeschlossen werden. Im späteren Verlauf dieser Arbeit findet darüber hinaus eine kritische Diskussion dieses Aspektes statt. In weiteren ähnlichen Versuchen werden die Parameter Support und der Konfidenz variiert sowie die Stichprobengröße schrittweise bis auf 400.000 Datensätze erhöht. Dabei werden allerdings keine neuen, interessanten Regeln aufgedeckt. Aus diesem Grund wird auf eine Darstellung dieser Versuche an dieser Stelle bewusst verzichtet.

Die Ausgangsfrage der Untersuchung nach den Ursachen fehlerhafter Teile (vgl. 5.1.1) lässt sich also auf Ebene aller Datensätze nicht beantworten. Dies lässt sich damit erklären, dass der Anteil dieser Teile an der Gesamtmenge zu gering ist, um in den Regeln wiedergefunden zu werden. Darum ist eine Untersuchung verschiedener Gruppierungen zu prüfen (vgl. Abschnitt 2.3.2). Hierfür erscheinen in erster Linie diejenigen Attribute geeignet, deren Ausprägungen Prüfungsergebnisse darstellen (hier: *TotalResult, Result_OP* (Operation Protocol) sowie *Result_ORP*). Dabei geben diese Attribute Auskunft über Ergebnisse auf verschiedenen Ebenen: TotalResult auf Werkstückebene, Result_OP auf Ebene der Prüfverfahren und Result_ORP auf Ebene der einzelnen

Schritte dieser Prüfverfahren. Zum weiteren Vorgehen wird in diesem Fallbeispiel das Attribut Result_ORP und insbesondere die Gruppierung der Ausprägung „FAIL“ ausgewählt. Diese Ausprägung kennzeichnet fehlgeschlagene Prüfverfahrenschritte und somit die niedrigste Ebene des Prüfprozesses. Zu Beginn wird auch hier dieselbe Stichprobenquelle und -größe wie in Versuch 1 gewählt (10.000 fehlerhafte Datensätze der Linien 2, 3 und 4). Nach der erneuten Modellierung und Anpassung des Modells mittels Validierungsdaten erfolgt die Durchführung des Trainings mit Hilfe der zuvor eingeteilten Trainingsdaten. Da mit den vorherigen Parametereinstellungen (Support = 0,8, Konfidenz = 0,8) keine Regeln gefunden werden können, werden die Parameter zur Inkludierung seltenerer Kombinationen schrittweise abgesenkt. Tabelle 5.9 stellt die interessantesten Regeln dieses Versuchs dar.

Tabelle 5.9: Ergebnisse der Durchführung der Assoziationsanalyse für die Gruppierung Resultat_ORP = „FAIL“ (Versuch-Nr. 2)

Nr.	Prämisse	Konklusion	Support	Konfidenz
2	Result-Code = 2	Anzahl Reparaturen = gering	0,298	0,703
11	Arbeitsplatz = 81, Arbeitssequenz = 2	Anzahl Reparaturen = gering	0,211	0,841
14	Linie = 4	Anzahl Reparaturen = gering	0,207	0,528

Die gefundenen Regeln dieser Durchführung lassen mehrere Hypothesen zu, die im Folgenden dargestellt werden.

Hypothese 2.1: Hauptsächlich ist eine geringe Reparaturanzahl notwendig (Attribut in den ausgewählten Regeln immer als Konklusion gegeben).

Hypothese 2.2: Produktionslinie 4 scheint überproportional häufig von Fehlern betroffen, da diese Linie als einzige in den Regeln wiederzufinden ist (Regel 14).

Hypothese 2.3: Wenn Fehler an Arbeitsplatz 81 in der zweiten Arbeitssequenz auftreten, dann ist eine geringe Reparaturanzahl notwendig (Regel 11). Ähnliche Vermutungen finden sich für Arbeitsplatz 82 und Arbeitssequenz 3.

Hypothese 2.4: Fehler mit dem Result-Code 2 tritt verstärkt auf (einziger Code in den generierten Regeln).

In Absprache mit dem Unternehmen wird auf Grundlage der Regeln und getroffenen Hypothesen das weitere Vorgehen bestimmt. Dabei erscheint eine Betrachtung der einzelnen Produktionslinien von größtem Interesse. Hierfür ist es jedoch notwendig, die Untersuchungen auf eine größere Stichprobe auszuweiten. Um alle bestehenden Linien zu enthalten, muss diese aus beiden Datenbanken gezogen werden. Da diese Gruppierung insgesamt wesentlich weniger Datensätze (ca. 350.000) umfasst als der Gesamtdatenbestand, ist die Handhabung der Datenmenge weitaus unkomplizierter. Aus diesem Grund wird hier erstmals die in Tabelle 5.6 beschriebene Aufteilung des Datenbestands in Teilmengen genutzt. Dabei verläuft die Einteilung der Daten unter

Anwendung der Methode der Zufallsstichprobe (vgl. Abschnitt 2.3.1). Auch bei diesem Versuch wird erst durch Herabsetzen der Parameterwahl eine ausreichende Menge an interessanten Regeln generiert, von denen Tabelle 5.10 eine Auswahl darstellt. Im Hinblick auf die zuvor aufgestellten Hypothesen ist zu erkennen, dass sich auch bei der Durchführung auf alle Linien eine geringe Reparaturanzahl als häufigste Ausprägung zeigt (vgl. Hypothese 2.1). Ein zuvor vermutetes, erhöhtes Auftreten einer bestimmten Linie lässt sich in der Gesamtdatenmenge hingegen selbst bei Absenkung der Parameterwerte auf Support = 0,1 und Konfidenz = 0,5 nicht mehr feststellen (vgl. Hypothese 2.2). Dagegen scheint sich ein, in Hypothese 2.3 vermutetes, erhöhtes Vorkommen von Fehlern an Arbeitsplatz 81 in Arbeitssequenz 2 auch hier zu bestätigen. Darüber hinaus ist jedoch Arbeitsplatz 82 – anders als zuvor – mit einem höheren Support enthalten. Der Result-Code 2 findet sich auch hier als häufigster Code in Kombination mit einer geringen Reparaturanzahl wieder (vgl. Hypothese 2.4), dies allerdings mit einer niedrigeren Konfidenz als im vorherigen Versuch (82% zu 70%).

Tabelle 5.10: Ergebnisse der Durchführung der Assoziationsanalyse für Gruppierung Resultat_ORP = „FAIL“ über alle Linien (Versuch-Nr. 3)

Nr.	Prämisse	Konklusion	Support	Konfidenz
60	Result-Code = 2	Anzahl Reparaturen = gering	0,285	0,694
93	Arbeitsplatz = 82	Arbeitssequenz = 3	0,243	0,917
78	Arbeitsplatz = 81	Arbeitssequenz = 2	0,233	0,844

Auf Grundlage der gefundenen Regeln des dritten Versuchs (siehe Tabelle 5.10) ergeben sich also neue, leicht veränderte Hypothesen.

Hypothese 3.1: Es treten vermehrt Fehler an Arbeitsplatz 82 (bzw. 81) in Arbeitssequenz 3 (bzw. 2) auf (Regel 93/78).

Hypothese 3.2: Es wird eine erhöhtes Auftreten von Result-Code 2 in Kombination mit einer geringen Reparaturanzahl vermutet (Regel 60).

Unter Berücksichtigung dieser Hypothesen erscheint im nächsten Schritt eine genauere Untersuchung der Arbeitsplätze sinnvoll, für die sich zwei Vorgehensweisen anbieten. Einerseits können die Gruppierungen des Attributs *Arbeitsplatz* untersucht werden, um zu schauen, welche Auffälligkeiten sich dabei zeigen. Andererseits erscheint es darüber hinaus interessant, die Arbeitsplätze innerhalb der Gruppierung der Produktionslinien zu untersuchen. So kann überprüft werden, ob sich Arbeitsplatz 81 und 82 innerhalb der Linien als am häufigsten betroffene Arbeitsplätze herausstellen. Desweiteren kann gezeigt werden, ob sich dies über alle Linien hinweg zeigt oder nur einzelne Linien betroffen sind. Deshalb wird an dieser Stelle für das weitere Vorgehen entschieden, die Linien zu vergleichen. So kann auch gleichzeitig dem zuvor erläuterten Wunsch des Unternehmens Rechnung getragen. Durch die erneute Anwendung der

Gruppierungtechnik und das Herausfallen des Attributs *Linie* aus der Auswahl, wird die Anzahl der betrachteten Attribute auf lediglich sechs reduziert. Daher bietet sich hier die Möglichkeit zu prüfen, ob die Menge der ausgewählten Attribute durch Aufteilung nichtgenutzter Attribute in ihre Elemente künstlich erhöht werden kann. Im Hinblick auf die Aufgabenstellung erscheint es vorallem interessant, an welchen Wochentagen und in welchen Schichten die häufigsten Fehler auftreten. Diese Informationen können aus dem vorhandenen, bisher nicht genutzen Attribut *Arbeitsbeginn* gewonnen werden. Darüber hinaus können auch zwei Elemente aus dem Attribut *Seriennummer* verwendet werden. So lässt sich die Menge an Attributen künstlich um die Attribute Schicht und Wochentag (aus *Arbeitsbeginn*) sowie Konstruktionsstand (KS) und Gerätekennzahl (GZ) (aus *Seriennummer*) auf insgesamt zehn erhöhen. Der Ablauf der Vorverarbeitung wurde bereits in Abschnitt 5.3.1 beschrieben.

Die generierten Regeln werden in Tabelle 5.11 präsentiert, dabei werden die interessanten Regeln aller Linien gegenübergestellt. Im Folgenden sollen diese unter Berücksichtigung der zuvor aufgestellten Hypothesen verglichen werden. Wie in Hypothese 3.1 beschrieben, scheint es linienübergreifend vermehrt zu Fehlern an den Arbeitsplätzen 81 und 82 zu kommen. Bei der Betrachtung der einzelnen Linien ist allerdings festzustellen, dass sich diese Auffälligkeit nicht in allen Linien wiederfindet. So sind die Arbeitsplätze nur in den generierten Regeln von drei der insgesamt sechs Linien gemeinsam zu identifizieren (Linien 2, 3 und 4). Bei diesen treten die beiden Arbeitsplätze im Zusammenhang mit der Arbeitssequenz mit einem ähnlichen Supportwert auf (20% bis 27%). Allerdings weist Arbeitsplatz 81 dabei stets einen höheren Supportwert auf (ca. 3% höher). Bei Betrachtung der Linien ohne Arbeitsplatz 81 in den generierten Regeln (1, 5) fallen die wesentlich höheren Supportwerte (47% und 55%) im Vergleich zu den restlichen Linien (23% bis 27%) auf. Dies bedeutet, dass Arbeitsplatz 82 bei diesen beiden Linien ca. doppelt so häufig in den Itemsets der fehlerhaften Datensätze gefunden wird. Für Linie 11 ist lediglich der Arbeitsplatz 81 in den Regeln zu finden. Auffällig bei diesen Regeln ist allerdings der geringe Supportwert von maximal 20%. So weist die Regel 5 mit Arbeitsplatz 81 und Arbeitssequenz 2 in der Prämisse und einer geringen Anzahl Reparaturen als Konklusion lediglich einen Supportwert von 18% auf. Damit ist für diese Linie eine Festlegung auf Arbeitsplatz 81 als vermeintlich problematischer Arbeitsplatz nicht möglich.

Weiterhin wurde vor der Untersuchung der Linien ein erhöhtes Auftreten von Result-Code 2 in Kombination mit einer geringen Reparaturanzahl vermutet (Hypothese 3.2). Auch hier ist dieser Result-Code nur in den Linien 2, 3 und 4 zu identifizieren. Dabei tritt dieser stets zusammen mit Arbeitsplatz 81 auf, allerdings mit relativ niedrigen Supportwerten (ca. 15%). Auch hier erscheint der Support zu gering, als dass hieraus eine Schlussfolgerung abgeleitet werden könnte. Über die vorherigen Hypothesen hinaus erscheint bei der Betrachtung der verschiedenen Linien das verstärkte Auftreten eines

Parameters in Kombination mit einem Arbeitsplatz interessant. So tritt der Parameter „PDES00000068“ häufig in Kombination mit Arbeitsplatz 82 auf (ähnlich Parameter „PDES00000048“ und Arbeitsplatz 81). Wie zuvor ist auch hier eine Schlussfolgerung aufgrund des geringen Supportwerts schwierig.

Betrachtet man die zuvor künstlich erzeugten Attribute, so ist festzustellen, dass diese einen geringen bis keinen Einfluss auf die Entstehung von Fehlern zu haben scheinen:

- Der Konstruktionsstand (*KS*) ist lediglich mit der Ausprägung „0“ in den Regeln vorhanden. Dies könnte erneut auf Probleme mit der Datenqualität zurückzuführen sein.
- Die Gerätekennzahl (*GZ*) tritt hauptsächlich in Kombination mit KS = 0 auf und dementsprechend wenig aussagekräftig.
- Das Attribut *Wochentag* findet sich nicht in den generierten Regeln wieder.
- Für das Attribut *Schicht* ist festzustellen, dass dieses lediglich bei vier der sechs Linien in den Regeln enthalten ist. (1, 2, 3, 5), dabei allerdings meist mit einem Supportwert unter 20%. Interessant scheint hier lediglich eine genauere Betrachtung der Schichten der Linie 5. Dabei kommen die Ausprägungen „Frühschicht“ und „Spätschicht“ mit fast identischem Support- und Konfidenzwert in Kombination mit Arbeitsplatz 82 und Arbeitssequenz 2 vor (Supp.: ca. 23%, Konf.: ca. 99%). Dies spiegelt sich in ähnlicher Form auch in den drei anderen Linien wieder. Dementsprechend ist zwar keine der gegebenen Schichten als auffällig zu identifizieren – es kann aber vermutet werden, dass die Nachtschicht am wenigsten Probleme verursacht, da sie nicht oder mit geringeren Support- und Konfidenzwerten in den Regeln enthalten ist.

Überprüfung der generierten Regeln

Nach Durchführung von Modellbildung und Training müssen die gefundenen Regeln unter Nutzung der Testdatenmenge – der bisher nicht genutzten 20% der Datensätze – überprüft werden. Dies ist auf zweierlei Wegen möglich. Einerseits kann ein Abgleich durch eine rechnerische Bestimmung der Modellgüte erfolgen, andererseits können erneut Regeln für die Testdatenmenge bestimmt und mit den zuvor für die Trainingsdatenmenge bestimmten Regeln abgeglichen werden. Im vorliegenden Fall erscheint dieses Vorgehen wesentlich geeigneter, als lediglich einen, in Prozent ausgedrückten Wert der Modellgüte zu ermitteln. Auch die Regeln der Testdatenmenge werden in Tabelle 5.11 dargestellt, um einen direkten Vergleich mit den Regeln der Trainingsdatenmenge zuzulassen. Wie dort zu erkennen, ist der Großteil der zuvor gefundenen Regeln auch in den Testdaten mit annähernd denselben Support- und Konfidenzwerten vertreten. Lediglich drei der als interessant eingestuften Regeln sind

Tabelle 5.11: Ergebnisse der Durchführung der Assoziationsanalyse für Gruppierung Resultat_ORP = „FAIL“ und Produktionslinien (Versuch-Nr. 4)

			Training		Test	
Nr.	**Prämisse**	**Konklusion**	**Supp.**	**Konf.**	**Supp.**	**Konf.**
Linie = 1						
1	Arbeitsplatz = 82	Arbeitssequenz = 3	0,467	0,979	0,467	0,980
6	Arbeitsplatz = 82, Parameterbeschreibung = PDES00000068	Arbeitssequenz = 3	0,335	0,981	0,336	0,983
Linie = 2						
1	Arbeitsplatz = 81	Arbeitssequenz = 2	0,274	0,890	0,289	0,891
3	Arbeitsplatz = 82	Arbeitssequenz = 3	0,242	0,912	0,242	0,919
11	Result-Code = 2, Arbeitsplatz = 81	Arbeitssequenz = 2	0,152	0,912	0,163	0,921
13	Parameterbeschreibung = PDES00000048	Arbeitsplatz = 81	0,150	1,0	0,156	1,0
Linie = 3						
1	Arbeitsplatz = 81	Arbeitssequenz = 2	0,231	0,870	0,233	0,867
3	Arbeitsplatz = 82	Arbeitssequenz = 3	0,206	0,899	0,211	0,905
5	Arbeitssequenz = 2	Anzahl Reparaturen = gering	0,196	0,845	0,202	0,867
12	Parameterbeschreibung = PDES00000048	Arbeitsplatz = 81	0,134	1,0	0,135	1,0
13	Result-Code = 2, Arbeitsplatz = 81	Arbeitssequenz = 2	0,133	0,913	0,137	0,905
Linie = 4						
1	Arbeitsplatz = 81	Arbeitssequenz = 2	0,267	0,881	0,268	0,868
3	Arbeitsplatz = 82	Arbeitssequenz = 3	0,234	0,892	0,238	0,903
9	Arbeitsplatz = 81, Arbeitssequenz = 2	Anzahl Reparaturen = gering	0,229	0,858	0,230	0,859
11	Result-Code = 2, Arbeitsplatz = 81	Arbeitssequenz = 2	0,150	0,905	/	/
Linie = 5						
1	Arbeitsplatz = 82	Arbeitssequenz = 3	0,546	0,992	/	/
6	Arbeitsplatz = 82, Parameterbeschreibung = PDES00000068	Arbeitssequenz = 3	0,437	0,995	/	/
Linie = 11						
5	Arbeitsplatz = 81, Arbeitssequenz = 2	Anzahl Reparaturen = gering	0,181	0,893	0,223	0,891
7	Parameterbeschreibung = PDES00000048	Arbeitsplatz = 81	0,151	1,0	0,189	1,0

nicht 1:1 wiederzufinden. Bei genauerer Betrachtung ist festzustellen, dass der in Regel 11 der Linie 4 gefundene Zusammenhang zwischen Result-Code 2, Arbeitsplatz 81 und Arbeitsschritt 2 in ähnlicher Form in der Testdatenmenge vorhanden ist – etwa mit vertauschter Prämisse und Konklusion. Darüber hinaus ist festzuhalten, dass Linie 5 einen Ausnahmefall bei der Betrachtung darstellt. Bei Generierung der Regeln der Testdaten werden für diese Linie insgesamt lediglich zwei Regeln gefunden. Diese entsprechen dabei nicht den zuvor bestimmten Regeln, deuten aber auf bereits zuvor erkannte Zusammenhänge hinsichtlich Result-Codes, Reparaturen, Arbeitsplätzen und Parameterbeschreibungen hin (siehe Tabelle 5.12). Ein Grund für die geringe Regelmenge in den Testdaten der Linie 5 sowie fehlende Übereinstimmungen zu den vorherigen Regeln ist in der geringen Datenmenge zu vermuten. So besteht die Testdatenmenge der Linie 5 lediglich aus ca. 500 Datensätzen und ist damit deutlich geringer als die Testdatenmengen der restlichen Gruppierungen.

Tabelle 5.12: Ergebnisse der Durchführung der Assoziationsanalyse für Gruppierung Resultat_ORP = „FAIL“ und Produktionslinie 5 (Testdatenmenge)

Nr.	Prämisse	Konklusion	Support	Konfidenz
1	Result-Code = 2	Anzahl Reparaturen = gering	0,184	0,951
2	Parameterbeschreibung = PDES00000068	Arbeitsplatz = 82	0,181	1,0

V&V der Verfahrensanwendung und der Weiterverarbeitung der Data-Mining-Ergebnisse

Auch die V&V-Maßnahmen dieser Phasen sind für das Gesamtergebnis des KDD-Prozesses von entscheidender Bedeutung. Dazu werden mithilfe von V&V-Techniken verschiedene Prüfungen durchgeführt, die in den Abschnitten 4.2.5 und 4.2.6 vorgestellt wurden. Für Phase 5 wurden dazu bereits beispielhaft verschiedene Fragestellungen zur Prüfung der V&V-Kriterien in Tabelle 4.3 aufgelistet. Hauptsächlich geht es in diesen Prüfungen darum, die Durchführung des Data-Mining-Verfahrens auf ihre Korrektheit zu untersuchen sowie diese Phase gegen die vorherigen Phasen zu kontrollieren. Hauptsächlich werden auch hier wieder Techniken zur Begutachtung, aber auch statistische Techniken, eingesetzt. Auch nach der Weiterverarbeitung und Darstellung der Regeln in geeigneter Form ist eine erneute Überprüfung notwendig. Dies wird hier vor allem durch die Aufteilung der Datenmenge mittels Holdoutmethode und Überprüfung der Regeln sichergestellt. Als Beispiel für den Einsatz der Begutachtungstechniken sei erneut auf das Attribut *Seriennummer* verwiesen. Dieses kann erst nach Rücksprache mit Fachleuten

entschlüsselt und für die Assoziationsanalyse nutzbar gemacht werden. Auch die getroffene Festlegung auf die Untersuchung der Gruppierungen erfolgt in Absprache mit dem Unternehmen. Durch diese Änderungen der zugrundeliegenden Daten ist ein Rücksprung in dieser Phase unausweichlich, um wieder in die Datenvorverarbeitung einzusteigen. Ein weiteres praktisches Anwendungsbeispiel ist die Prüfung der verwendeten Bausteine. Dabei ist festzustellen, dass mittels automatischer Bausteine nicht alle Attribute sauber entfernt werden. Deshalb muss dies im Schritt der Selektion überprüft und im Zweifelsfall manuell angepasst werden. Dies entspricht der V&V-Technik des Tests von Teilmodellen.

5.4 Tatsächliche Anwendung der V&V-Techniken

In den vorherigen Abschnitten wurden die benötigten Schritte zur Durchführung der Phasen des MESC beschrieben. Bei dieser chronologisch gehaltenen Beschreibungsweise der Abläufe ist eine ausreichende Behandlung der V&V-Maßnahmen schwierig, da dabei immer wieder auch Rücksprünge in vorherige Phasen nötig werden können. Deshalb soll dieser Abschnitt in Ergänzung zu den V&V-Beschreibungen in den Phasen dienen und interessante Beispiele ausführlicher darstellen. Diese Beispiele dienen gleichzeitig dem Aufzeigen einiger Auffälligkeiten hinsichtlich der Datenqualität.

Künstliche Erzeugung von Attributen

Bei der Anwendung des Data-Mining-Verfahrens kann es nötig sein, die Form von Attributen zu verändern, um diese so in eine geeignete Form für das Verfahren zu bringen. Dazu kann beispielsweise eine Separierung durchgeführt werden, um zusammengesetzte Attribute ihre atomaren Elemente zu zerlegen (vgl. Abschnitt 2.3.2). Im vorliegenden Fall ist das Attribut Seriennummer als ID vorhanden und in dieser Form als unbrauchbar eingestuft worden (Phase 3). Im Rahmen der Prüfung der Modellentwicklung (Phase 5) ist in V&V-Element (5,4) die Modellierung auf richtige Vorverarbeitung der Daten zu kontrollieren. Dabei sind auch die Attribute mit ID-ähnlicher Form (z.B. Seriennummer) erneut zu untersuchen. Durch Kontaktaufnahme per Email zu dem entsprechenden Fachexperten kann eine innere Struktur erkannt und das Attribut in seine Elemente zerlegt werden (vgl. Abschnitt 5.3.2). Bei dieser Zerlegung treten allerdings mehrere Auffälligkeiten zu Tage:

- Eine Vielzahl von Seriennummern bricht nach der dritten Ziffer ab und ist „genullt“ (z.B. 15300000000000000). Dadurch bedingt, kann für die genullten Teile keine Aufteilung der Seriennummern erfolgen. Davon betroffen ist auch das Element Herstelltag (T), das theoretisch den benötigten Wochentag der Herstellung liefern könnte. Erst dadurch, dass eine Auswertung dieses Elements nicht möglich ist, wird eine Zerlegung des Datumwerts überhaupt nötig (vgl. Abschnitt 5.3.2).

- Auffällig ist dabei auch, dass alle Seriennummern der Linie 5 komplett nach der dritten Ziffer „genullt“ sind.
- Bei dem Element Linie (L) fällt auf, dass für dieses lediglich eine einstellige Ziffer verwendet wird. Da aber mit Linie 11 auch eine zweistellige Ziffer als Ausprägung vorkommt, werden die Seriennummern dieser Linie falsch aufgeteilt. Das hat zur Konsequenz, dass die zweite Ziffer der Linie 11 in das nächste Element geschoben wird und damit auch alle folgenden Elemente falsch dargestellt werden. Aus diesem Grund ist eine automatisierte Zerlegung nicht möglich, da beispielsweise alle Datensätze der Linie 11 fälschlicherweise der Linie 1 zugeordnet werden würden. Für die vorliegende Arbeit ist dies jedoch weniger problematisch, da die Liniennummer noch in weiteren Attributen vorkommt und die folgenden Elemente Laufende Nummer, Prüfziffer nach Absprache mit dem Unternehmen nicht in die Betrachtung eingezogen werden soll.

Wie in Abschnitt 5.3.2 beschrieben, ist für die Ausführung der Assoziationsanalyse im Praxisbeispiel zu überprüfen, ob die verschiedenen Arbeitsschichten und Wochentage einbezogen werden können. Es ist festzustellen, dass der Wochentag theoretisch bereits in der Seriennummer enthalten ist, aber praktisch nicht genutzt werden kann. Ähnlich verhält es sich mit der Arbeitsschicht. In den zugrundeliegenden Daten existiert bereits ein Attribut mit der Bezeichnung *Schicht.* Dabei enthält dieses Attribut verschiedene Ausprägungen zwischen 0 bis 23. Da eine Entschlüsselung der Bedeutung dieser Ausprägungen trotz gemeinsamer Begutachtung mit dem Unternehmen nicht möglich ist, kann dieses Attribut nicht für die weitere Betrachtung genutzt werden. Aus diesem Grund ist es notwendig, auch die Arbeitsschicht künstlich herzuleiten. Eine Beschreibung des Vorgehens zur Gewinnung des Attributs findet sich bereits in Abschnitt 5.3.1. Darüber hinaus soll hier noch einmal auf Probleme bei der Umformung eingegangen werden, um die Wichtigkeit der V&V herauszustellen.

Die Umformung des Datumwerts wurde mit Hilfe eines VBA Makros in Excel durchführt. Dabei wurden die Uhrzeiten fälschlicherweise in das amerikanische Format umgewandelt, wodurch Einträge nach 12 Uhr mittags nicht der Früh- bzw. Spätschicht, sondern der Nachtschicht zugeordnet wurden (z.B. 00:05 Uhr statt 12:05 Uhr). Dies führt zu einer immensen Verzerrung der Ergebnisse, konnte aber bei einer erneuten Kontrolle der Vorverarbeitung im Rahmen der V&V-Maßnahmen der fünften Phase des MESC erkannt und beseitigt werden. Hierdurch wird deutlich, dass die Kontrollen der Phasen gegen die Ergebnisse der Vorphasen dringend notwendig sind, um solche Fehler aufdecken zu können.

Rückverfolgbarkeitstabelle

Bei der Datenauswahl in der zweiten Phase werden Tabellen festgelegt, die für das Data-Mining-Verfahren genutzt werden sollen (vgl. Abschnitt 5.1.2). Eine dieser Tabellen ist die *Rückverfolgbarkeitstabelle*, die Logistikdaten aus der Produktion enthält (z.B. Scanner, Scannerwert). Allerdings sind bei der Betrachtung einige Auffälligkeiten zu beobachten, die eine Einbeziehung dieser Tabelle fraglich machen:

- Die Tabelle beinhaltet insgesamt ca. 50 Attribute, von denen eine Vielzahl entweder leer oder fehlerhaft ist. Darüber hinaus existieren hierarchische Beziehungen innerhalb der Tabelle, die erneut aufwendig beseitigt werden müssten.
- Es gibt keinen eindeutigen Schlüssel in dieser Tabelle, durch den sich diese mit den anderen Tabellen verknüpfen lassen würde. Es müsste mithilfe eines zu erzeugenden künstlichen Schlüssels gearbeitet werden.
- Darüber hinaus liegt die Tabelle in einer anderen Datenbank vor, die Tabellen müssten zur Verknüpfung erst auf eine gemeinsame Datenbank transferiert werden.
- Hauptgrund für eine Nichtberücksichtigung ist, dass in der Tabelle keine neuen, für das Assoziationsverfahren relevanten Informationen enthalten sind.

Bezieht man die genannten Gründe in seine Überlegungen mit ein, steht ein sehr hoher Aufwand einem vermeintlich geringen Ertrag gegenüber. Aus diesem Grund wird nach gemeinsamer Begutachtung und in Absprache mit den Fachleuten des Unternehmens auf eine Einbeziehung der Attribute dieser Tabelle verzichtet.

Fachliche Logik

Die gefundenen Regeln wurden – wie in Abschnitt 5.3 beschrieben – unter Berücksichtigung des Ausprägung „FAIL“ des Attributs *Result_ORP* bestimmt, da dies die Ergebnisse der einzelnen Prüfprozessschritte darstellt. Unter logischen Gesichtspunkten sollten die Attribute *Result_ORP*. *Result-Code* und *Anzahl Reparaturen* die folgende Beziehung aufweisen: Wenn ein Prüfungsschritt als bestanden vermerkt wird (*Result_ORP* = „PASS“), sollte auch der Result-Code gleich „0“ sein. Genauso sollten bei einem Result-Code 0 keine Reparaturen notwendig sein. Wie aber bereits in Tabelle 5.8 dargestellt, ist dies nach Auswertung der Regeln für den Gesamtdatenbestand aber nicht immer der Fall. So müssten die Regeln eine Konfidenz von 1 aufweisen, um einen totalen Zusammenhang zwischen den Attributen darzustellen. Dies kann auf Probleme in der Datenqualität hinweisen. Auf dieser Annahme basieren die dargestellten Durchführungen des Data-Mining-Verfahrens und die gefundenen Regeln. Im Verlauf der Durchführungen und der erneuten Kontrolle der Attribute wird allerdings die Frage aufgeworfen, ob die getroffene Annahme vor dem Hintergrund der fachlichen Logik haltbar erscheint. Es ist bei weiteren Ausführungen zu prüfen, ob die Attribute *Result-Code* bzw. *Anzahl Reparaturen* tatsäch-

lich mit diesem Attribut verknüpft werden können oder es sich lediglich auf die „übergeordneten“ Attribute *Result_OP* (Prüfungsebene) bzw. *TotalResult* (Werkstückebene) bezogen werden dürfen. Wäre dies der Fall, so wäre zu prüfen, ob eine ganzheitliche Betrachtung der Attribute möglich ist oder ob eine Einbeziehung der Attribute *Result-Code* und *Anzahl Reparaturen* lediglich auf den jeweiligen Ebenen möglich ist.

5.5 Erkenntnisse aus der praktischen Anwendung des MESC auf einen produktionslogistischen Datensatz

Die vorherigen Abschnitte dieses Kapitels dienten der Darstellung der prektischen Durchführung des MESC aus einen Datenbestand der Produktionslogistik. Dazu wurde nach der Auswahl und Vorverarbeitung der Daten (vgl. Abschnitte 5.1 und 0) die Assoziationsanalyse durchgeführt (vgl. Abschnitt 5.3). Darüber hinaus diente vorherige Abschnitt dazu, noch einmal genauer auf verschiedene Punkte der V&V einzugehen. Das Resultat der Assoziationsanalyse sind verschiedene Hypothesen, die aus den gefundenen Regeln abgeleitet werden. Diese Hypothesen dienen weiterhin als Grundlage zur erneuten Durchführung der Assoziationsanalyse und zur Verfeinerung der Regeln. Dies entspricht dem in Abschnitt 2.1 eingeführten Datenanalysezyklus. Bedingt durch u.a. die erneut notwendige Vorverarbeitung der Daten (z.B. zur Zerlegung von Attributen) ist diese, sich ständig wiederholende Durchführung sehr zeitaufwändig. Dadurch kann dieser Zyklus im Rahmen dieser Arbeit nur bis zu einem bestimmten Grad behandelt werden. Aus diesem Grund sollen an dieser Stelle finale Hypothesen aufgestellt werden, den Auswertung aus den vorherigen Abschnitten basieren:

Hypothese 1: Eine erhöhte Fehlerhäufigkeit einer bestimmten Linie ist nicht feststellbar.

Hypothese 2: Arbeitsplatz 81 und 82 scheinen – mit Unterschieden zwischen den Linien – die problematischen Arbeitsplätze zu sein.

Hypothese 3: Vor allem Result-Code 2 ist in den Regeln vertreten, dabei verstärkt in Verbindung mit Arbeitsplatz 81.

Hypothese 4: Der Wochentag scheint keinen großen Einfluss auf die Entstehung von Fehlern zu haben.

Hypothese 5: Es kann keine Schicht als hauptsächlich problematisch erkannt werden, lediglich die Vermutung aufgestellt werden, dass die Nachtschicht auf Grundlage der Regeln als am wenigsten betroffen anzunehmen ist.

Diese Hypothesen sollten im Rahmen des Datenanalysezylkus die Grundlage weiterer Untersuchung bilden.

Die Datenbank ist in ihrem aktuellen Aufbau mit hierarchischen Beziehungen sowie leeren, fehlerhaften, redundanten oder nicht interpretierbaren Attributen für die Durchführung des Data-Mining-Verfahrens nur bedingt geeignet. Hier sollte eine Entscheidung getroffen werden, ob die Datenbank an die Anforderungen des Data

Minings angepasst werden soll, um die Durchführung zu erleichtern bzw. das Ziel der automatisierten Durchführung des Data Minings zu verfolgen.

Bezogen auf das V&V kann durch die praktische Anwendung erkannt werden, dass es bei der Durchführung immens wichtig ist, die Ergebnisse jeder einzelnen Phase zu überprüfen sowie diese Prüfungen in den weiteren Phasen zu wiederholen. Wie in den vorherigen Abschnitten dargestellt, können so durch eine erneute Kontrolle vorher unentdeckte Fehler gefunden und beseitigt werden. Was ebenfalls deutlich wird, ist, dass aufgrund der nicht ausreichenden Datenqualität hauptsächlich jene Techniken zum Einsatz kommen müssen, mit deren Hilfe auf vorhandenes Kontextwissen zurückgegriffen werden kann. Dies kann vor allem durch informale Techniken und insbesondere durch Begutachtungstechniken wie etwa Review -Meetings erfolgen. Ein enger Austausch mit Prozessexperten ist für eine erfolgreiche Durchführung unumgänglich.

6 Zusammenfassung

Das Ziel der vorliegenden Arbeit war es, die Anwendung eines Data-Mining-Verfahrens auf einen produktionslogistischen Datensatz im Rahmen des MESC aufzuzeigen. Ein Schwerpunkt hierbei lag darauf, Verifikations- und Validierungsmöglichkeiten für die phasenbegleitende Überprüfung der Zwischenergebnisse darzustellen.

Durch die Ausführung des Verfahrens sollte die Frage beantwortet werden, ob die zugrundeliegenden Daten des betrachteten Praxisfalls generell für Data-Mining-Verfahren geeignet sind und ob dabei Wirkzusammenhänge zwischen verschiedenen Attributen der Produktion erkannt werden können. Zu diesem Zweck wurden zu Beginn der Arbeit die Begriffe des Data Minings und des übergeordneten KDD eingeführt, denen im Rahmen von Big Data, Industrie 4.0 und Digitalisierung eine immer bedeutsamere Rolle zugesprochen wird. Bei der Betrachtung wurden das *KDD nach Fayyad* und das *CRISP-DM* als entscheidende Vorgehensmodelle des KDD erkannt – ersteres aufgrund der herausragenden Bedeutsamkeit bei der Entwicklung des KDD, letzteres aufgrund seines heutigen Stellenwerts für die praktische Anwendung.

Um dem Schwerpunkt der V&V bei der Durchführung des MESC gerecht zu werden, erfolgte darüber hinaus eine Einführung der Begrifflichkeiten der Verifikation und Validierung sowie eine Darstellung von V&V-Techniken verschiedener Bereiche. Dabei wurden hauptsächlich V&V-Techniken der Softwareentwicklung und Simulation betrachtet. Dies geschah vor dem Hintergrund, dass es sich bei den Prozessen dieser Bereiche – wie beim KDD – um (Modell-) Entwicklungsprozesse handelt, die einer gewissenhaften Prüfung erfordern. Darüber hinaus finden diese Bereiche in der Literatur eine breite Berücksichtigung. Zusätzlich erfolgte eine Betrachtung relevanter V&V-Techniken des Data Minings. Um die Menge der zu analysierenden V&V-Techniken vorab zu reduzieren, wurden Ausschlusskriterien – wie etwa die Relevanz in der Literatur – definiert und der Auswahl zugrunde gelegt. Bei der darauffolgenden generellen Überprüfung ihrer Einsatzfähigkeit im KDD wurden insbesondere informale Techniken als geeignet erkannt, die in einem Austausch mit Fachexperten zur Begutachtung von Dokumenten, Daten oder Modellen sowie zur Erklärung fachlicher Wirkzusammenhänge zwischen Attributen eingesetzt werden können – etwa Reviews oder Validierung im Dialog. Bei der anschließenden Prüfung der Tauglichkeit der Techniken speziell für das MESC zeigte sich, dass sich für alle Phasen – aber insbesondere für die Auswahl und Vorverarbeitung der Daten – hauptsächlich Techniken eignen, mit denen sich Kontextwissen in die Analyse der Daten einbeziehen lässt. Diese Erkenntnis wurde in der Folge auch durch die praktische Durchführung des MESC bestätigt. So wäre die Anwendung des Data-Mining-Verfahrens aufgrund nicht geeigneter Strukturen bzw. nicht ausreichender Datenqualität ohne eine vorherige Betrachtung der Daten durch Fachexperten komplizierter, fehleranfälliger oder gar nicht

zu realisieren gewesen. Immer wieder kam es bei der Durchführung des MESC zu Situationen, in denen Zusammenhänge zwischen Attributen falsch interpretiert wurden und erst durch Rücksprache mit den Fachverantwortlichen korrekt erfasst werden konnten. Allerdings traten auch mehrmals Situationen auf, in denen die Bedeutung eines Attributs selbst nach Rücksprache mit den Prozessbeteiligten unklar blieb und diese aus der Betrachtung entfallen mussten.

Für den Praxisfall wurde die Assoziationsanalyse als Data-Mining-Verfahren ausgewählt, da diese für das Ziel der Untersuchung – die Aufdeckung von Wirkzusammenhängen – von allen Verfahren als das geeignetste erkannt wurde. Als Ergebnis wurden Assoziationsregeln generiert, die zur Ableitung von Hypothesen genutzt werden konnten. Dabei zeigte sich, dass für die Beantwortung der Ausgangsfrage nach Ursachen fehlerhafter Prüfresultate eine Unterteilung der Bestände in Gruppierungen sinnvoller erscheint als eine Betrachtung des Gesamtdatenbestands. Die hier gefundenen Regeln und die daraus abgeleiteten Hypothesen konnten durch die Bestimmung weiterer interessanter Gruppierungen und durch das erneute Aufstellen von Regeln in diesen überprüft werden. Dieses Vorgehen führt allerdings zu einem Data-Mining-Prozess, der durch Rücksprünge in vorherige Phasen – etwa zur erneuten Vorverarbeitung – zu einem, sich stets wiederholenden Vorgang wird, bei dem der Detaillierungsgrad der Betrachtung stetig zunimmt. Aufgrund des zeitlichen Rahmens dieser Arbeit musste dieser Vorgang zwangsläufig frühzeitig beendet werden. Die dabei abschließend generierten Regeln der verschiedenen Produktionslinien wurden mithilfe der Holdout-Methode getestet. Diese Methode erschien aufgrund der Größe des Datenbestands am geeignetsten für die Überprüfung. Dabei konnte gezeigt werden, dass der Hauptteil der in der Trainingsdatenmenge gefundenen Regeln auch in der Testdatenmenge nachgewiesen werden kann.

Das Ergebnis der praktischen Anwendung des Data-Mining-Verfahrens in dieser Arbeit konnte aus den zuvor angeführten Gründen nur aus Hypothesen bestehen. Auch wenn in den Untersuchungen keine der Produktionslinien als besonders betroffen zu identifizieren war, zeigten die gefundenen Regeln über alle Linien hauptsächlich Auffälligkeiten an zwei Arbeitsplätzen. Dabei waren allerdings die Linien in unterschiedlichem Maße betroffen. Ebenfalls häufig in den Regeln wiederzufinden waren spezielle Parameterbeschreibungen. Allerdings kamen dieser Regeln mit einer relativ geringen Konfidenz vor, sodass die Untersuchung dieser Zusammenhänge geringer priorisiert werden sollte.

Zusammengefasst lässt sich sagen, dass die Ziele dieser Arbeit erreicht wurden. Vorbereitend auf die praktische Anwendung des MESC konnten V&V-Techniken bestimmt werden, die für den Einsatz in der KDD geeignet sind. Hier könnte eine zukünftige Aufgabe in der weiteren Untersuchung dieser Techniken bestehen, um zu prüfen, ob sich die Erkenntnisse dieser Arbeit bei weiteren KDD-Prozessen bestätigten lassen. Auch für V&V-Techniken, bei denen eine weitere Untersuchung lohnenswert erscheint, sollte

durch eine tiefergehende Analyse überprüft werden, unter welchen Bedingungen diese für einen zukünftigen Einsatz in der praktischen Anwendung genutzt werden können.

Im Praxisfall konnten Regeln über anzunehmende Wirkzusammenhänge in der Produktion aufgestellt und somit auch die generelle Anwendbarkeit des Data Minings auf die Datenbestände des Praxisfalls aufgezeigt werden. Allerdings müssen die gefundenen Regeln und die abgeleiteten Hypothesen differenziert betrachtet werden. Es ist zu beachten, dass die vermeintlichen Auffälligkeiten in der Produktion auf Grundlage verschiedener Annahmen getroffen wurden. Deshalb ist es notwendig, die fachliche Logik der getroffenen Annahmen zu hinterfragen, bevor eine weitere Untersuchung dieser Hypothesen mittels Data Mining erfolgen kann. Darüber hinaus ist es empfehlenswert, sich genauer mit den aufgezeigten Auffälligkeiten hinsichtlich der Datenqualität und des generellen Aufbaus der genutzten Datenbanken zu beschäftigen. Bei einer Entscheidung für eine Optimierung der Datenbanken im Hinblick auf Data-Mining-Prozesse, böte sich etwa der Einsatz eines sogenannten Data-Mart an. Mit diesem Abbild der Ursprungsdatenbestände lassen sich Daten schon vor dem Data-Mining-Prozess in entsprechender Form speichern, ohne Änderungen in den aktiven Datenbeständen vornehmen zu müssen (vgl. Bauer 2013). So kann die Datenvorverarbeitung bei künftigen Data-Mining-Prozessen verkürzt sowie qualitativ verbessert werden, um so die Aussagekraft der Ergebnisse zu erhöhen. Für die Zukunft sollte zusätzlich über die Durchführung der KDD-Prozesse in Form eines Data-Mining-Projekts nachgedacht werden. Dadurch ließe sich der Prozess einfacher koordinieren und notwendiges Kontextwissen besser abrufen und bündeln. Denn auch für zukünftige Durchführungen wird es unerlässlich sein, auf das Know-how der Fachexperten aus Produktion und IT zurückzugreifen.

Die in der vorliegenden Arbeit erkannten Potentiale des Data Minings und der V&V werden auch durch im Jahr 2017 stattfindende Konferenzen bestätigt. So finden zahlreiche Veranstaltung zu dem Gebiet der KDD und des Data Minings statt – wie etwa die *22. ACM SIGKDD Conference On Knowledge Discovery and Data Mining*. Allerdings wird in Rahmen dieser Konferenzen die V&V – wenn überhaupt – nur am Rande thematisiert. Aber auch für die V&V sind diverse Konferenzen angesetzt. So findet im kommenden Mai das *Verification and Validation Symposium (VandV)* mit dem Schwerpunkt der V&V in der Simulation statt. Auch für die Softwareentwicklung ist eine ähnliche Veranstaltung mit der *10. IEEE International Conference on Software Testing, Verification and Validation* geplant. Auch wenn diese Konferenzen nicht direkt die V&V in der KDD thematisieren, sind die Ergebnisse dieser Konferenzen sehr wohl von Interesse. Denn wie in dieser Arbeit herausgestellt, sind die V&V-Techniken dieser Bereiche auch für einen Einsatz in der KDD geeignet.

Insgesamt lässt sich festhalten, dass das Data Mining im Rahmen des MESC ein geeignetes Mittel zur Durchführung von Datenanalysen ist. Allerdings ist hierfür – bedingt durch die vorherrschende Datenqualität – aktuell noch eine ausführliche manuelle

Datenvorverarbeitung notwendig, die den Einsatz von V&V-Techniken zur Begutachtung erfordert. Die Vision eines automatisierten Data-Mining-Verfahrens ist nur durch die Verbesserung der Datengrundlage zu erreichen. Da sich durch den Data-Mining-Einsatz enorme Wettbewerbsvorteile erzielen lassen, stellt dies jedoch eine lohnenswerte Investition dar

Literaturverzeichnis

Agrawal, R.; Srikant, R. (1994): Fast algorithms for mining association rules in large databases. In: Bocca, J. B. (Hrsg.): Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94), Santiago (Chile). Hove, East Sussex: Morgan Kaufmann, S. 487-499.

Arndt, D.; Gersten, W.; Wirth, R. (2001): Kundenprofile zur Prognose der Markenaffinität im Automobilsektor. In: Hippner, H.; Küsters, U.; Meyer, M. und Wilde, K. D. (Hrsg.): Handbuch Data Mining im Marketing. Knowledge discovery in marketing databases. Braunschweig [u.a.]: Vieweg (Vieweg Gabler business computing), S. 591-605.

Arnold, D.; Isermann, H., Kuhn, A.; Tempelmeier, H.; Furmans, K. (2008): Handbuch Logistik. 3. Aufl. Berlin: Springer.

Balci, O. (1998): Verification, Validation, and Testing. In: Banks, J. (Hrsg.): Handbook of simulation. Principles, methodology, advances, applications, and practice. New York: Wiley, S. 335-393.

Balci, O. (2013): Verification, validation, and testing of models. In: Gass, S. I. und Fu, M. C. (Hrsg.): Encyclopedia of operations research and management science. 3. Aufl. New York: Springer, S. 1618-1627.

Balzert, H. (2008): Lehrbuch Grundlagen der Informatik. Konzepte und Notationen in UML, Java und C++, Algorithmik und Software-Technik, Anwendungen. 2. Aufl. [Nachdruck], München: Spektrum Akademischer Verlag (Lehrbücher der Informatik).

Balzert, H. (2009): Lehrbuch der Softwaretechnik. Basiskonzepte und Requirements-Engineering. 3. Aufl. Heidelberg: Spektrum Akademischer Verlag (Lehrbücher der Informatik).

Banks, J.; Gerstein, D.; Searles, S. P. (1988): Modeling processes, validation, and verification of complex simulations. A survey. In: *Methodology and validation* 19 (1), S. 13-18.

Bauer, A.; Günzel, H. (2013): Data-Warehouse-Systeme. Architektur, Entwicklung, Anwendung, Heidelberg: dpunkt.

Bayardo, R. J.; Agrawal, R. (1999): Mining the most interesting rules. In: Fayyad, U. M. (Hrsg.): Proceedings of the fifth ACM SIGKDD International Conference on Knowledge discovery and Data Mining. San Diego (USA). New York: ACM Press, S. 145-154.

Benington, H. D. (1956): Production of large computer programs [Nachdruck]. In: *IEEE Annals of the History of Computing (1983)* 5 (4), S. 350-361.

Bernhard, J.; Hömberg, K.; Jodin, D.; Kuhnt, S: Schürmann, C.; Wenzel, S. (2007): Vorgehensmodell zur Informationsgewinnung – Prozessschritte und Methodennutzung. Technical Report – Sonderforschungsbereich 559 „Modellierung großer Netze in der Logistik“ 06008, Dortmund, ISSN 1612-1376.

Berry, M. J. A.; Linoff, G. (2000): Mastering data mining. The art and science of customer relationship management. New York: Wiley.

Bishop, C. M. (1995): Neural networks for pattern recognition. Oxford [u.a.].: Oxford University Press.

Boehm, B. W. (1979): Guidelines for verifying and validating software requirements and design specifications. In: Samet, P. A. (Hrsg.): Euro IFIP 79. Proceedings of the European Conference on Applied Information Technology of the International Federation for Information Processing. North-Holland Publishing Company, S. 711-719.

Bollinger, T. (1996): Assoziationsregeln. Analyse eines Data Mining Verfahrens. In: *Informatik-Spektrum* 19 (5), S. 257-261.

Bosch, K. (2015): Großes Lehrbuch der Statistik. [Nachdruck]. Berlin, Boston: De Gruyter.

Brachman, R.; Anand, T. (1996): The process of knowledge discovery in databases. In: Fayyad, U. M. (Hrsg.): Advances in knowledge discovery and data mining. 5. Aufl. Menlo Park: AAAI Press, S. 37-57.

Brade, D. (2003): A generalized process for the verification and validation of models and simulation results. Dissertation, Universität der Bundeswehr München, München.

Bramer, M. A. (2013): Principles of data mining. 2. Aufl. London: Springer (Undergraduate topics in computer science).

Breiman, L.; Friedman, J.; Stone, C. J.; Olsen, R. A. (1998): Classification and regression trees. Boca Raton: Chapman & Hall.

Bröhl, A. P. (1995): Das V-Modell. Der Standard für die Softwareentwicklung mit Praxisleitfaden. 2. Aufl. München: Oldenbourg (Software, Anwendungsentwicklung, Informationssysteme).

Burgdorf, F. (2010): Eine kunden- und lebenszyklusorientierte Produktfamilienabsicherung für die Automobilindustrie. Karlsruhe: KIT Scientific Publishing.

Chapman, P.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R. (1999): CRISP-DM 1.0. Step-by-step data mining guide. The CRISP-DM consortium.

Cios, K. J.; Kurgan, L. A.; Pedrycz, W.; Swiniarski, R. W. (2007): Data mining. A knowledge discovery approach. Boston: Springer.

Clarke, B.; Fokoué, E.; Zhang, H. H. (2009): Principles and theory for data mining and machine learning. New York: Springer (Springer series in statistics).

Cleve, J.; Lämmel, U. (2016): Data Mining. 2. Aufl. Berlin [u.a.]: De Gruyter Oldenbourg (De Gruyter Studium).

Cortada, J. W. (2012): The digital flood. The diffusion of information technology across the U.S., Europe, and Asia. Oxford: Oxford University Press.

Efron, B. (1979): Bootstrap methods. Another look at the jackknife. In: *The Annals of Statistics* 7 (1), S. 1-26.

Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P. (1996): From data mining to knowledge discovery in databases. In: *AI Magazine* 17 (3), S. 37-54.

Felkai, R.; Beiderwieden, A. (2011): Projektmanagement für technische Projekte. Wiesbaden: Vieweg+Teubner.

Füermann, Timo (2014): Prozessmanagement. Kompaktes Wissen, konkrete Umsetzung, praktische Arbeitshilfen München: Hanser Verlag

Fukuda, T.; Morimoto, Y.; Tokuyama, T.; Morishita, S. (1996): Data mining using two-dimensional optimized association rules. In: Jagadish, H. V. (Hrsg.): 1996 proceedings ACM SIGMOD International Conference on Management of Data. Montreal (Kanada). New York: ACM Press (SIGMOD record, 25.2), S. 13-23.

Gloger, B. (2011): Scrum. Produkte zuverlässig und schnell entwickeln. 3. Aufl. München: Hanser Verlag

Gottermeier, C. (2003): Data Mining. Modellierung und Durchführung ausgewählter Fallstudien mit dem SAS Enterprise Miner. Diplomarbeit. Online verfügbar unter http://archiv.ub.uni-heidelberg.de/volltextserver/4073/1/Diplomarbeit_Christian_Gottermeier.pdf

Große Böckmann, M.; Krappig, R.; Stolorz, M.; Schmitt, R. (2013): Data-Mining in der Produktion*. Neue Methoden für eine robuste Prozessentwicklung. In: *Werkstattstechnik online* 103 (11/12), S. 921-925.

Han, J.; Kamber, M.; Pei, J. (2012): Data mining. Concepts and techniques. 3. Aufl. Amsterdam: Elsevier (The Morgan Kaufmann series in data management systems).

Hermann, C. F. (1967): Validation problems in games and simulations with special reference to models of international politics. In: *Behavioral Science* 12 (3), S. 216-231.

Hettich, S.; Hippner, H. (2001): Assoziationsanalyse. In: Hippner, H.; Küsters, U.; Meyer, M. und Wilde, K. D. (Hrsg.): Handbuch Data Mining im Marketing. Knowledge discovery in marketing databases. Braunschweig [u.a.]: Vieweg (Vieweg Gabler business computing), S. 427-464.

Hippner, H.; Wilde, K. D. (2001): Der Prozess des Data Mining im Marketing. In: Hippner, H.; Küsters, U.; Meyer, M. und Wilde, K. D. (Hrsg.): Handbuch Data Mining im Marketing. Knowledge discovery in marketing databases. Braunschweig [u.a.]: Vieweg (Vieweg Gabler business computing), S. 21-92.

Hofmann, H. J. (1990): Die Anwendung des CART-Verfahrens zur statistischen Bonitätsanalyse von Konsumentenkrediten. In: *Zeitschrift fur Betriebswirtschaft* 60, S. 941-962.

Hofmann, M.; Klinkenberg, R. (Hrsg.) (2014): RapidMiner. Data mining use cases and business analytics applications. Boca Raton [u.a.]: CRC Press (Chapman & Hall/CRC data mining and knowledge discovery series, 33).

Hunyadi, D. (2011): Performance comparison of Apriori and FP-Growth algorithms in generating association rules. In: Leandre, R. (Hrsg.): Proceedings of the European Computing Conference (ECC ‘11). Athen: WSEAS Press (Proceedings of the WSEAS international conferences), S. 376-381.

IEE (2008): IEEE Std. 1028-2008. IEEE standard for software reviews and audits (2008). Piscataway: IEEE.

IEE (2012): IEEE Std. 1012-2012. IEEE standard for system and software verification and validation (2012). New York: IEEE.

Kemper, A.; Eickler, A. (2015): Datenbanksysteme. Eine Einführung. 10. Aufl., Berlin [u.a.]: De Gruyter Oldenbourg (De Gruyter Oldenbourg Studium).

Kleijnen, J. P.C. (1995): Verification and validation of simulation models. In: *European Journal of Operational Research* 82 (1), S. 145-162.

Knobloch, B.; Weidner, J. (2000): Eine kritische Betrachtung von Data Mining-Prozessen. Ablauf, Effizienz und Unterstützungspotenziale. In: Jung, R. und Winter, R. (Hrsg.): Data Warehousing 2000. Methoden, Anwendungen, Strategien. Heidelberg: Physica, S. 345-365.

Kohavi, R. (1995a): A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Mellish, C.S. (Hrsg.): Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Montreal (Kanada). San Francisco: Morgan Kaufmann, S. 1137–1143.

Kohavi, R. (1995b): The power of decision tables. In: Lavrač, N. und Wrobel, S. (Hrsg.): Proceedings of the Eighth European Conference on Machine Learning. Iraklio (Griechenland). Berlin: Springer (Lecture notes in computer science, 912), S. 174-189.

Kumbhare, T. A.; Chobe, Santosh V. (2014): An overview of association rule mining algorithms. In: *International Journal of Computer Science and Information Technologies* 5 (1), S. 927-930.

Kurgan, L. A.; Musilek, P. (2006): A survey of knowledge discovery and data mining process models. In: *The Knowledge Engineering Review* 21 (1), S. 1.

Lämmel, U. (2003): Data-Mining mittels künstlicher neuronaler Netze. Wismar: Hochschule Wismar, Fachbereich Wirtschaft (Wismarer Diskussionspapiere, 7).

Landry, M.; Malouin, J.-L.; Oral, M. (1983): Model validation in operations research. In: *European Journal of Operational Research* 14 (3), S. 207-220.

Law, A. M. (2007): Simulation modeling and analysis. 4. Aufl. Boston: McGraw-Hill (McGraw-Hill series in industrial engineering and management science).

Li, Y. (2016): Anwendung von Data Mining auf produktionslogistische Massendaten mit Schwerpunkt Datenvorverarbeitung. Masterarbeit.

Liggesmeyer, P. (2009): Software-Qualität. Testen, Analysieren und Verifizieren. Heidelberg: Spektrum Akademischer Verlag.

Malthouse, E. C.; Blattberg, R. C. (2005): Can we predict customer lifetime value? In: *Journal of Interactive Marketing* 19 (1), S. 2-16.

Murthy, K., Salzberg, S. (1995): Lookahead and pathology in decision tree induction. In: Mellish, C. S. (Hrsg.): Proceedings of the fourteenth International Joint Conference on Artificial Intelligence, Montreal (Kanada). San Francisco: Morgan Kaufmann, S. 1025-1031.

Naisbitt, J. (1982): Megatrends. Ten new directions transforming our lives. New York: Warner Books.

Neckel, P. R.; Knobloch, B. (2015): Customer relationship analytics. Praktische Anwendung des Data Mining im CRM. 2. Aufl. Heidelberg: dpunkt.

Oeldorf, G.; Olfert, K. (2013): Material-Logistik. 4. Aufl. Herne: Kiehl (Kompakt-Training praktische Betriebswirtschaft).

Petersohn, H. (2005): Data Mining. Verfahren, Prozesse, Anwendungsarchitektur. München [u.a.]: Oldenbourg.

Piatetsky-Shapiro, G. (2014): What main methodology are you using for your analytics, data mining, or data science projects? Online verfügbar unter http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html.

Pohl, S.; Bel Haj Saad, S.; Best, M.; Brade, D.; Hofmann, M.; Kiesling, T.; Krieger, T. (2005): Verifizierung, Validierung und Akkreditierung von Modellen, Simulationen und Föderationen. Abschlussbericht Studienkennziffer E/F11S/2A280/T5228. Neubiberg: ITIS.

Prescha, M. (2009): Data-Mining im Immobilien-Business. Hamburg: Diplomica Verlag.

Pyle, D. (1999): Data preparation for data mining. San Francisco: Morgan Kaufmann.

Rabe, M.; Spiekermann, S.; Wenzel, S. (2008): Verifikation und Validierung für die Simulation in Produktion und Logistik. Vorgehensmodelle und Techniken. Berlin [u.a.]: Springer (VDI-Buch).

Rönz, B.; Strohe, H. G. (1994): Lexikon Statistik. Wiesbaden: Gabler

Ross, K. A.; Jensen, C. S.; Snodgrass, R. T.; Skiadopoulos, S.; Sirangelo, C.; Larsgaard, M. L. et al. (2009): Cross-Validation. In: Liu, L. und Özsu, M. T. (Hrsg.): Encyclopedia of database systems. New York: Springer (Springer reference), S. 532-538.

Royce, W. W. (1970): Managing the development of large software systems. In: *Proceedings of IEEE WESCON* 26 (8), S. 328-338.

Runkler, T. A. (2015): Data Mining. Methoden und Algorithmen intelligenter Datenanalyse. Wiesbaden: Springer.

Sargent, R. G. (2011): Verification and validation of simulation models. In: Jain, S.; Creasey, R., Himmelspach, J.; White, K. P., Fu, M. C.: Proceedings of the 2011 Winter Simulation Conference, Phoenix (USA). New York: ACM, S. 183-198.

Sargent, R. G. (2013): Verification and validation of simulation models. In: *Journal of Simulation* 7 (1), S. 12-24.

Säuberlich, F. (2000): KDD und Data Mining als Hilfsmittel zur Entscheidungsunterstützung. Frankfurt am Main: Lang, Peter Frankfurt (Entscheidungsunterstützung für ökonomische Probleme, 18).

Scheidler, A. A. (2016): Vorgehensmodell zur Musterextraktion in SCs (MESC). Whitepaper

Schendera, C. F. G. (2014): Regressionsanalyse mit SPSS. 2. Aufl. München: De Gruyter Oldenbourg.

Schwaber, K.; Sutherland, J. (2016): The scrum guide. The definitive guide to scrum, the rules of the game. Online verfügbar unter http://www.scrumguides.org/docs/scrumguide/v2016/2016-Scrum-Guide-US.pdf.

Shannon, Robert E. (1975): Systems simulation. The art and science. Englewood Cliffs: Prentice-Hall.

Sharafi, A. (2013): Knowledge discovery in databases. Eine Analyse des Änderungsmanagements in der Produktentwicklung. Wiesbaden: Springer.

Short, J. E.; Bohn, R. E.; Baru, C. (2011): How much information? 2010. Report on enterprise server information. San Diego: UCSD Global Information Industry Center.

Sommerville, I. (2016): Software engineering. 10. Aufl. Boston: Pearson.

Steiner, V. (2009): Modellierung des Kundenwertes. Ein branchenübergreifender Ansatz. Wiesbaden: Gabler.

Steinlein, U. (2004): Data Mining als Instrument der Responseoptimierung im Direktmarketing. Methoden zur Bewältigung niedriger Responsequoten. Göttingen: Cuvillier.

VDI (2008): VDI-Richtlinie 3633 Blatt 1. Simulation von Logistik-, Materialfluss und Produktionssystemen. Berlin: Beuth.

Walter, J.A. (2004): Datamining. Methoden integrativer Datenpräsentation. Göttingen: Cuvillier.

Weiss, S. M.; Indurkhya, N. (1998): Predictive data mining. A practical guide. San Francisco: Morgan Kaufmann.

Weskamp, M.; Tamas, A.; Wochinger, T.; Schatz, A. (2014): Einsatz und Nutzenpotenziale von Data Mining in Produktionsunternehmen. Stuttgart: Fraunhofer IPA.

Zahavi, J.; Levin, N. (1997): Applying neural computing to target marketing. In: *Journal of Direct Marketing* 11 (1), S. 5-22.

Anhang

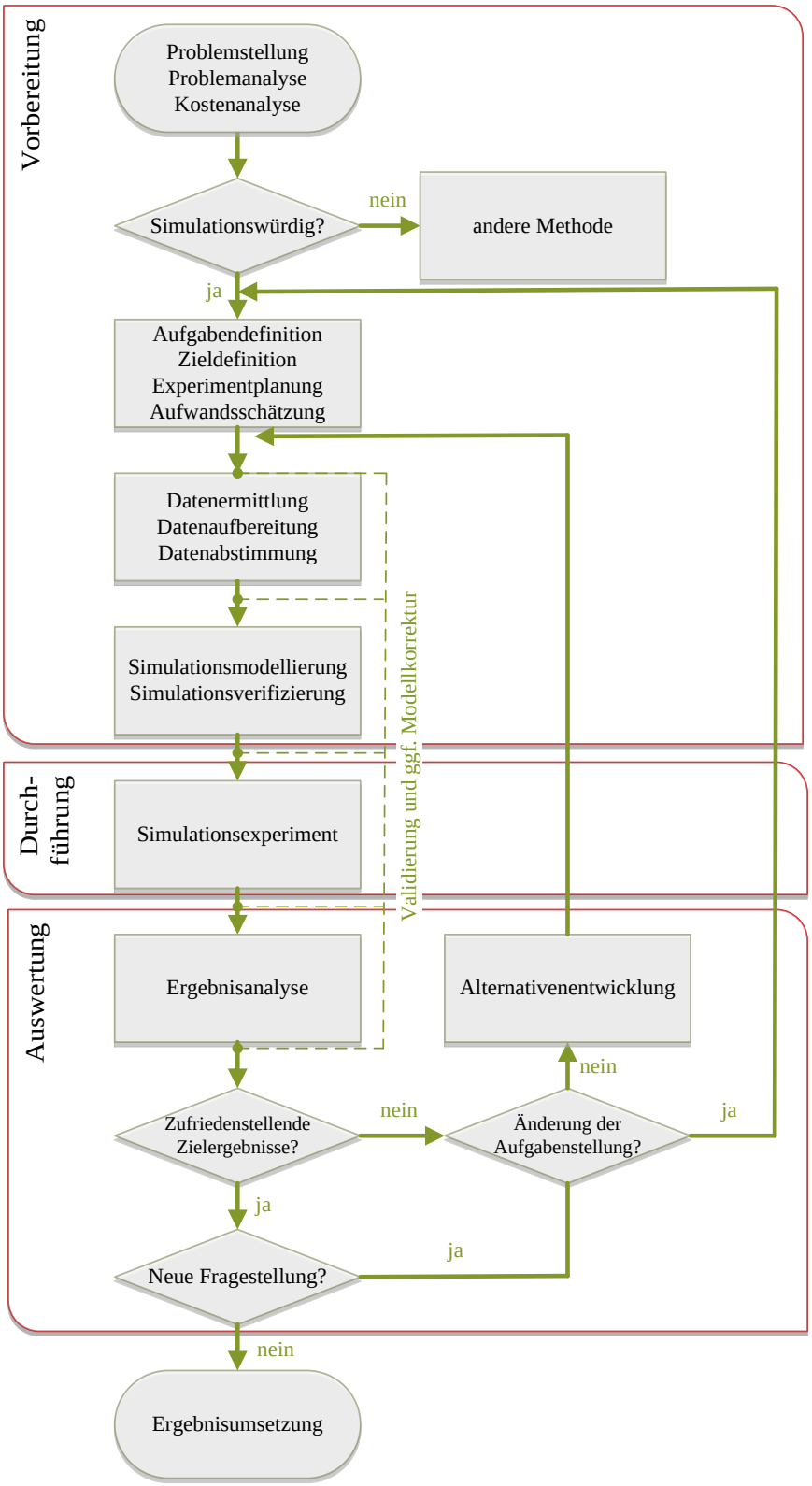
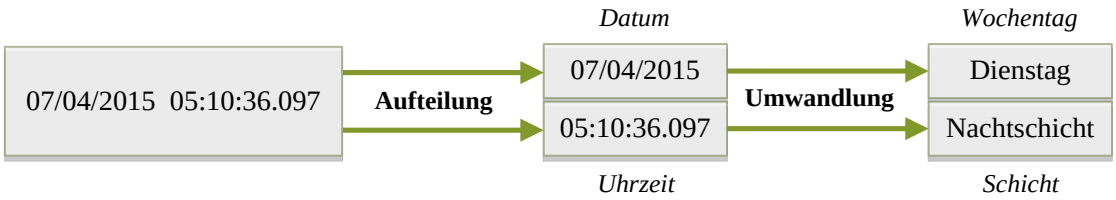
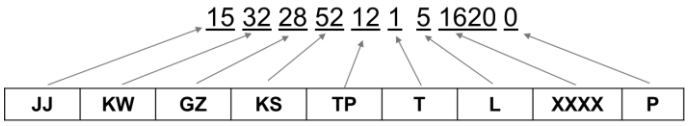


Abbildung A.1: VDI-Vorgehensmodell nach VDI (2008)



Attribut	Ausprägungen
Wochentag	Montag, Dienstag, Mittwoch, Donnerstag, Freitag
Schicht	Frühschicht, Spätschicht, Nachtschicht

Abbildung A.2: Beispielhafte Umwandlung eines Datumwerts



Abkürzung	Beschreibung	Kommentar zur Verwertbarkeit
JJ	Herstelljahr	nicht relevant, da nur 2015 betrachtet wird
KW	Herstellwoche	nicht relevant
GZ	Gerätekennzahl	kann für die Betrachtung genutzt werden
KS	Konstruktionsstand	kann für die Betrachtung genutzt werden
TP	Gerätetyp	nicht relevant
T	Herstelltag	Wochentag, sehr interessant
L	Linie	mehrfach in anderen Attributen vorhanden
XXXX	laufende Nummerierung am Herstelltag	ID-ähnlich, nicht nutzbar
P	Prüfziffer	ID-ähnlich, nicht nutzbar

Abbildung A.3: Zusammensetzung der Seriennummer

Tabelle A.1: Umwandlung hierarchischer Beziehungen in der relationalen Datenbank

Parameterbe-schreibung	Parent-Parameter-beschreibung	Kategorie	Standard-wert	…
PDES0000001	NULL	NTM	NULL	…
PDES00000002	*PDES00000001*	NTM	NULL	…
PDES00000003	*PDES00000001*	NTM	NULL	…

Tabelle A.2: Neue Tabelle für Parent-Parameterbeschreibung

Parameter-beschreibung	Kategorie
PDES00000001	NTM

Tabelle A.3: Neue Tabelle für Parameterbeschreibung

Parameter-beschreibung	Parent-Para-meterbeschrei-bung	Standard-wert
PDES00000002	PDES00000001	NULL
PDES00000003	PDES00000001	NULL

Abbildungsverzeichnis

Tabellenverzeichnis

Eidesstattliche Versicherung

_____ _____

Name, Vorname Matr.-Nr.

Ich versichere hiermit an Eides statt, dass ich die vorliegende Masterarbeit mit dem Titel **Anwendung von Data Mining auf produktionslogistische Massendaten mit Schwerpunkt Verifikation und Validierung** selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

_____ _____

Ort, Datum Unterschrift

Belehrung:
Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden. (§ 63 Abs. 5 Hochschulgesetz - HG -).

Die Abgabe einer falschen Versicherung an Eides statt wird mit Freiheitsstrafe bis zu 3 Jahren oder mit Geldstrafe bestraft.

Die Technische Universität Dortmund wird gfls. elektronische Vergleichswerkzeuge (wie z.B. die Software „turnitin“) zur Überprüfung von Ordnungswidrigkeiten in Prüfungsverfahren nutzen.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

_____ _____

Ort, Datum Unterschrift