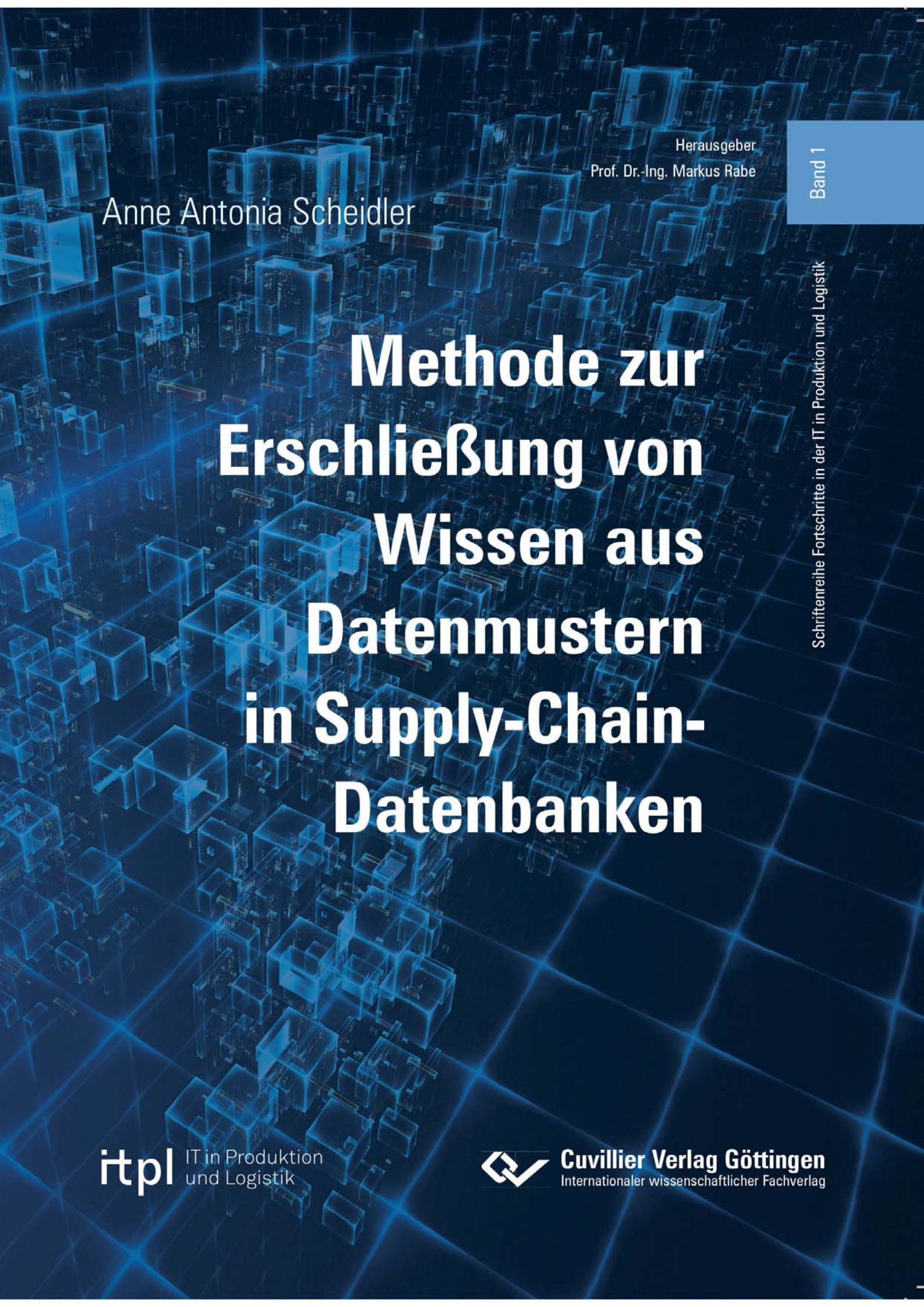
Anne Antonia Scheidler

# Methode zur Erschließung von Wissen aus Datenmustern in Supply-Chain-Datenbanken

Herausgeber
Prof. Dr.-Ing. Markus Rabe

Band 1

Schriftenreihe Fortschritte in der IT in Produktion und Logistik



itpl IT in Produktion und Logistik

Cuvillier Verlag Göttingen
Internationaler wissenschaftlicher Fachverlag

Markus Rabe (Hrsg.)

# Schriftenreihe Fortschritte in der IT in Produktion und Logistik

Band 1

Weitere Bände:
www.itpl.mb.tu-dortmund.de

# Methode zur Erschließung von Wissen aus Datenmustern in Supply-Chain-Datenbanken

Zur Erlangung des akademischen Grades eines
**Dr.-Ing.**
von der Fakultät Maschinenbau
der Technischen Universität Dortmund
genehmigte Dissertation

**Dipl.-Inf. Anne Antonia Scheidler**
aus
Bochum

Tag der mündlichen Prüfung: 11.07.2017
1. Gutachter: Prof. Dr.-Ing. Markus Rabe
2. Gutachterin: Prof. Dr.-Ing. Sigrid Wenzel

Dortmund, 2017

Prof. Dr.-Ing. Markus Rabe
Technische Universität Dortmund
Fakultät Maschinenbau
Fachgebiet IT in Produktion und Logistik
Leonhard-Euler-Str. 5
44227 Dortmund
markus.rabe@tu-dortmund.de

**Bibliografische Information der Deutschen Nationalbibliothek**
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

1. Aufl. - Göttingen: Cuvillier, 2017
Zugl.: (TU) Dortmund, Univ., Diss., 2017


Nonnenstieg 8, 37075 Göttingen
Telefon: 0551-54724-0
Telefax: 0551-54724-21
www.cuvillier.de


1. Auflage, 2017
Gedruckt auf umweltfreundlichem, säurefreiem Papier aus nachhaltiger Forstwirtschaft

ISBN 978-3-7369-9614-4
eISBN 978-3-7369-8614-5

# Vorwort

Diese Arbeit entstand während meiner Tätigkeit als wissenschaftliche Mitarbeiterin am Fachgebiet IT in Produktion und Logistik (ITPL) der Fakultät Maschinenbau an der Technischen Universität Dortmund. Ein besonderer Dank gilt zunächst meinem Doktorvater, Herrn Professor Dr.-Ing. Markus Rabe, für seine fortwährende Unterstützung und sein außerordentliches Engagement bei der Betreuung dieser Arbeit. Besonders möchte ich mich für die gewährte Forschungsfreiheit, die anregenden Diskussionen und wertvolle Ratschläge von seiner Seite bedanken. Diese haben wesentlich zum Gelingen der vorliegenden Arbeit beigetragen. Mein weiterer Dank gilt zudem Frau Professor Dr.-Ing. Sigrid Wenzel für das Interesse an meinem Thema und die Übernahme des Zweitgutachtens.

Ferner möchte ich dem gesamten Team des Fachgebiets ITPL für die stetige Diskussionsbereitschaft sowie die zahlreichen wertvollen Anregungen während der Erstellung dieser Arbeit danken. Mein besonderer Dank gilt Maik Deininger, der durch seine unermüdliche Unterstützung sowohl zum gestalterischen als auch fachlichen Gelingen der Arbeit beigetragen hat. Ferner möchte ich mich bei Sonja Drenkelforth und Astrid Klüter für die stets aufmunternden Worte und die persönliche Unterstützung während meiner Tätigkeit am ITPL bedanken. Auch die studentischen Hilfskräfte haben durch ihre Literaturrecherchen und Versuchsdurchführungen einen wertvollen Beitrag geleistet, für den ich mich bei allen bedanken möchte. Besonders möchte ich an dieser Stelle Robin Stasch erwähnen, der in der praktischen Umsetzung eine wertvolle Unterstützung war.

Mein herzlicher Dank gebührt weiterhin den zahlreichen Unterstützern aus meinem Kollegen- und Freundeskreis. Insbesondere die intensive Auseinandersetzung mit meiner Dissertation in zahlreichen Fachgesprächen und Reviews war eine große Stütze für mich. Besonderen Dank möchte ich an dieser Stelle Andreas Broede, Michael Homuth, Stefan Herwig, Ralf Junker und Katrin Sinha für ihre investierte Zeit aussprechen.

Großen Dank schulde ich zudem meiner Familie, meinem Partner Patrick Michels und meinen Freunden für ihre fortwährende Unterstützung und ihr Interesse an meiner Forschung. Insbesondere meine Mutter Ute Scheidler hat durch ihre bedingungslose Unterstützung und fortwährende Motivation zum Gelingen dieser Arbeit beigetragen.

Dortmund, August 2017
Anne Antonia Scheidler

# Inhaltsverzeichnis

# 1 Einleitung

Infolge der Globalisierung ist die Logistik zu einem der größten Wirtschaftssegmente in Deutschland herangewachsen (vgl. Steglich et al. 2016; Wannenwetsch 2010). Zu den Aufgaben der Logistik zählen nach Jünemann (1989, S. 18), „die richtige Menge, der richtigen Objekte als Gegenstände der Logistik (Güter, Personen, Energie, Informationen), am richtigen Ort (Quelle, Senke) im System, zum richtigen Zeitpunkt, in der richtigen Qualität, zu den richtigen Kosten zur Verfügung zu stellen“ . Ein wesentlicher Aspekt in der Logistik ist die Wertschöpfung mittels globaler Netzwerke, an denen unterschiedliche Teilnehmer partizipieren. Diese Netzwerke sind unter dem Begriff Supply Chain (SC, Plural: SCs) zusammengefasst und ein elementarer Bestandteil der heutigen Prozesslandschaft. Zur SC gehören alle Unternehmen bzw. eigenständig handelnden Unternehmensteile, die zur Entwicklung, Erstellung und Lieferung von Produkten oder Dienstleistungen beitragen (Beckmann 2012). In den letzten Jahren führte die permanente Zunahme der globalen Vernetzung zu einem Komplexitätsanstieg in der SC (Pfeiffer et al. 2013). Ein Ende dieser Entwicklung ist nicht abzusehen. Aufgrund des Komplexitätsanstieges entstehen beispielsweise höhere Kosten oder zeitliche Verzögerungen in der logistischen Kette, wodurch Unsicherheiten in der SC zunehmen. Dadurch ist die Notwendigkeit gegeben, die Unsicherheiten zu reduzieren, um die zunehmende Komplexität in der SC zu beherrschen (Wilding 1998). Um die Komplexität zu kontrollieren, ist konkretes Wissen über die SC notwendig. Die Fähigkeit, dieses Wissen zu nutzen sowie in geeigneter Weise zu integrieren, ersetzt mittlerweile bisherige Optimierungsziele für Unternehmen (Kemppainen und Vepsäläinen 2003) und stellt eine wichtige Unternehmensressource dar (Reiber 2013; Wenzel, Abel et al. 2011).

Eine spezielle Form von Wissen, die sowohl für die SCs als auch für das Supply Chain Management (SCM) von größter Wichtigkeit ist, sind Wirkzusammenhänge (Harland 1996). Wirkzusammenhänge sind immer dann von Relevanz, wenn es um die Abbildung von Systemverhalten geht. Das Systemverhalten, das zur SC-Kontrolle in der Intralogistik dient, wird z. B. mittels Simulation (Rabe und Scheidler 2014) oder logistischer Assistenzsysteme modelliert (Kuhn et al. 2008). Fachlich können diese Wirkzusammenhänge beispielsweise beschreiben, unter welchen Konstellationen Abweichungen von der durchschnittlichen Lieferzeit besonders häufig auftreten. Die Lieferung von Waren von und zu unterschiedlichen Standorten innerhalb eines logistischen Systems wird allgemein als Transaktion bezeichnet (Corsten und Gössinger 2008). Sowohl innerhalb einzelner als auch zwischen unterschiedlichen Transaktionen treten jeweils verschiedene Wirkzusammenhänge auf. Ein elementarer Schritt zur Beherrschbarkeit von SCs ist die Identifikation von Wirkzusammenhängen der logistischen Transaktionen und deren Modellierung. Dadurch besteht die Möglichkeit, Kenntnisse über das Systemverhalten bereitzustellen.

In der SC werden traditionell Materialflüsse, Finanzflüsse und Informationsflüsse unterschieden. Die Betrachtung der Flüsse erfolgt sowohl in Richtung des Lieferanten als auch des Kunden. Die Flüsse sind durch verschiedenartige Prozesse gekennzeichnet. Dementsprechend werden Materialflussprozesse in Transport-, Lager- , Umschlag- und Sortierprozesse unterteilt (Arnold und Furmans 2009). Grundlage dieser Prozesse ist eine Datenbasis, mit der die Prozessausführung ermöglicht wird. Während dieser Ausführung generiert der Prozess Daten. Ein Großteil der erzeugten Daten basiert auf Transaktionen, die Zustandsänderungen beschreiben, die mittels Mengenangaben und Zeitstempeln erfasst werden. Die gestiegene Komplexität der Netzwerke in Verbindung mit der Verbesserung der Informations- und Kommunikationstechnologien führt zu einem Anstieg der zu speichernden Datensätze und deren Detaillierungsgrad (Harrison und van Hoek 2008). Aufgrund der unüberschaubaren Datenmenge kann die Entdeckung von komplexen Wirkzusammenhängen nicht mehr manuell erfolgen.

Eine Herausforderung ist die Wissensentdeckung im Bereich der SC, da bereits bestehende Lösungen aus dem Feld der Wissensentdeckung nicht einfach übertragbar sind (Wyatt et al. 2014). Ein allgemeiner Ansatz, um Wissen wie Wirkzusammenhänge zu identifizieren, ist das Knowledge Discovery in Databases (KDD). Im KDD existieren etablierte Vorgehensmodelle zur Wissensentdeckung wie beispielsweise das Cross Industry Standard Process for Data Mining (Crisp-DM) (Gabriel et al. 2009) oder das Modell von Fayyad (Fayyad et al. 1996b). Der zentrale Aspekt des KDD ist die Wissensentdeckung mittels Data Mining. Hierbei bezeichnet Data Mining eine Sammlung von Verfahren zur Extraktion von Wissen. Klassische Problemstellungen im Bereich der Logistik nutzen bereits erfolgreich diese Art der Wissensentdeckung. Hierzu gehören Kundensegmentierung zu Marketingzwecken, Prognosen im Bereich des Controllings und Unterstützung der Produktion im Bereich der Materialbedarfsplanung (Gabriel et al. 2009). Spezifische Data-Mining-Methoden konnten ebenfalls für die SC adaptiert werden. Exemplarisch sei der Beitrag von Kamble et al. (2015) erwähnt, der Vorhersageverfahren des Data Minings für die SC genutzt hat, um Aspekte wie den Bullwhip-Effekt zu berücksichtigen. Allerdings zeigt die jüngste Studie des Fraunhofer IPA (Weskamp et al. 2014), dass hauptsächlich Unternehmen aus Branchen mit starkem Endkundenbezug KDD-Techniken nutzen. In dieser Studie empfehlen zudem 34 % der befragten Unternehmen die Logistik-Branche als großen Potentialträger für das KDD. KDD hat zum Ziel, mittels eines geeigneten zyklischen Vorgehensmodells Muster, wie beispielsweise Wirkzusammenhänge, in Daten zu entdecken. Im Vorgehensmodell des KDD gibt es unterschiedliche Musterbegriffe, die aus den verschiedenen Teilschritten resultieren. Diese Muster können auf unterschiedliche Art dargestellt werden (Klösgen und Zytkow 1996). Verifikation und Validierung (V&V) sind aufgrund der induktiven Prozess-Vorgehensweise keine Eigenschaft des KDD (Düsing 2010). Aus diesem Grund muss das KDD-Vorgehensmodell um die V&V erweitert werden. Der wachsende Datenbestand erweist sich im Zusammenhang mit dem KDD sowie der V&V als Vorteil, denn sowohl für die Methoden der Wissensent-

deckung als auch die Validierung des Wissens ist eine ausreichende Datenbasis notwendig.

Das Ziel der Arbeit ist die Entwicklung einer Methode zur Wissensentdeckung in SC-Datenbeständen, die als Element ein domänenspezifisches KDD-Vorgehensmodell beinhaltet. Für die Modellentwicklung müssen als Teilziele die bereits bestehenden Vorgehensmodelle wie Crisp-DM oder das Fayyad-Modell bezüglich der spezifischen Anforderungen und Ziele der SC analysiert und gegebenenfalls entsprechend angepasst werden. Insbesondere ist zu beachten, dass eine bedarfsgerechte Vorverarbeitung der Daten und eine Anpassung der Algorithmen zur Wissensentdeckung zu berücksichtigen sind. Ein weiteres Teilziel besteht in der Aufschlüsselung der Bedeutung des dispositiven Kontextwissens im SC-Umfeld sowie dessen anschließender Integration in die Problemkodierung der Algorithmen. Hierfür werden spezifische Charakteristika der SC wie u. a. die Berücksichtigung unterschiedlicher Transportmittel in der Vorverarbeitung entwickelt. Ein zusätzliches Ziel liegt in der Wissensrepräsentation durch Muster, wobei der Musterbegriff für die spezifischen Bedürfnisse der SC angepasst und erweitert werden muss. Dies bedeutet insbesondere, dass das Kontextwissen in den Musterbegriff des KDD-Vorgehensmodells integriert wird. Ein weiteres Ziel ist die Ermöglichung der Verwendung der entwickelten Methode bei einer unzureichenden Datengrundlage. Zu diesem Zweck wird eine Datengenerierung durch Simulation in der Methode eingeführt. Ferner ist ein weiteres Ziel dieser Arbeit die Integration einer begleitenden V&V in das zu entwickelnde Vorgehensmodell der Methode. Hierbei wird die V&V durch einen neuartigen Einsatz der Simulation unterstützt. Das Erreichen der aufgeführten Forschungsziele ist im Hinblick auf die anstehende digitale Transformation und zunehmende Automatisierung in den Unternehmen erforderlich, da dadurch eine Grundlage zur Beherrschung der Komplexität im SC-Umfeld geschaffen wird.

Zur Erreichung der zuvor genannten Ziele werden in dieser Arbeit Datenbestände von verschiedenen SCs untersucht und mögliche logistische Fragestellungen zu diesen Datenbeständen identifiziert. Aus der Datengrundlage in Verbindung mit möglichen SC-Fragestellungen werden Anforderungen abgeleitet. Basierend auf diesen Anforderungen wird ein KDD-Vorgehensmodell aus dem Bestand der existierenden Vorgehensmodelle ausgewählt. Auf Grundlage dieses Vorgehensmodells wird im Anschluss das domänenspezifische KDD-Vorgehensmodell für die SC entwickelt. Entsprechend muss hierfür insbesondere das Verhältnis von bedarfsgerechter Vorverarbeitung zu anderen Prozessphasen bestimmt werden. Die bedarfsgerechte Vorverarbeitung im SC-Kontext umfasst beispielsweise spezifische Transformationen von Daten, da Transaktionsdaten sowohl numerische Daten (z. B. Zeitstempel) als auch eine Vielzahl von nominalen Daten (z. B. Standorte) enthalten. Dies bedeutet eine zusätzliche, zeitintensive Vorverarbeitung der Daten für die Algorithmen. Aus der fachlichen Perspektive ist insbesondere der Schritt der Partitionierung relevant, der oftmals bei großen Datenbeständen für die effiziente Suche notwendig ist. Die Partitionierung ermöglicht die effiziente Suche über grö-

ßere Zeiträume, indem diese den Suchraum nach geeigneten Kriterien aufteilt. Je nach Transaktionsfrequenz und Transaktionsgranularität des Unternehmens erfordert die Größe des Datenbestandes diesen Vorgang. Im Rahmen der Vorgehensmodellerstellung sind geeignete Musterbegriffe zu definieren. Diese sind an entsprechender Stelle in das Vorgehensmodell zu integrieren. Hierbei ist insbesondere die Wechselwirkung zwischen Mustern und Kontextwissen zu beachten und für das Vorgehensmodell zu adaptieren. Unter diesem Aspekt ist mittels V&V zu prüfen, ob die entdeckten Muster relevante Zusammenhänge beschreiben und diese im realen System zu beobachten sind. Für die modellbegleitende V&V wird in der vorliegenden Arbeit ein neues Konzept in das zu entwickelnde KDD-Vorgehensmodell integriert. Zudem werden die bestehenden V&V-Techniken um die ereignisdiskrete Simulation (Discrete Event Simulation, DES) erweitert, die im Kontext der SC bereits vielfältig Anwendung findet. DES kann neben der Simulation der kompletten SC auch die konkreten Auswirkungen einzelner Zusammenhänge innerhalb des zu betrachtenden Systems isoliert untersuchen sowie die Analyse von Abhängigkeiten unterstützen. Die DES ist im Rahmen der KDD-Vorgehensmodelle ein neuer Ansatz, um inhärente Verfahrensschwächen im SC-Kontext auszugleichen. Da insbesondere in der Planungsphase von SCs oftmals keine ausreichende Datenbasis vorhanden ist, muss zusätzlich eine Möglichkeit zu einer künstlichen Generierung von Daten entwickelt werden. Dieser Vorgang ist im Kontext des DES als Data Farming bekannt und wird hier für den Einsatz in Vorgehensmodellen adaptiert.

Um die praktische Anwendbarkeit der in dieser Arbeit entwickelten Methode zu demonstrieren, wird diese in zwei Anwendungsfeldern angewandt. Das erste Anwendungsfeld zeigt die Ausführung des entwickelten Vorgehensmodells auf Transaktionsdaten in einem konkreten Praxisfall. Hier dient das Vorgehensmodell sowie die gewonnenen Erkenntnisse bei der Ausführung des Modells als Basis für die Konzeptionierung eines unternehmensweiten Datencockpits. Das zweite Anwendungsfeld stellt ein Data-Farming-Modell vor, mit dem die Datengenerierung innerhalb des entwickelten Methode demonstriert wird. Zusätzlich wird exemplarisch dargestellt, wie die Wirkzusammenhänge der generierten Daten mittels der in dieser Arbeit entwickelten simulationsunterstützten Validierung geprüft werden können.

# 2 Wissensgewinn im Kontext der Supply Chain

In diesem Kapitel werden die Grundlagen für die in dieser Arbeit entwickelten Methode vorgestellt. Zentrale Themen sind der Wissensbegriff, die SC und ihre Datenlage sowie Verfahren, die im Kontext der Wissensgewinnung Anwendung finden. Aus den hier diskutierten Problemstellungen ergibt sich sowohl eine thematische Eingrenzung als auch der Handlungsbedarf, der am Ende des Kapitels zu Forschungsfragen zusammengefasst wird. Dafür werden zunächst die zentralen Begriffe, die sich aus dem Thema dieser Arbeit ergeben, erläutert.

Allgemein beschreibt eine Methode ein planmäßiges Vorgehen mit überprüfbaren Ergebnissen und bildet somit den Ausgangspunkt ingenieurmäßigen Vorgehens. Der Begriff Methode als „an approach to perform a systems development consisting of directions and rules, structured in a systematic way in development activities with corresponding development products“ (Brinkkemper 1996, S. 275) stammt aus dem Method Engineering. Methoden lassen sich nach Prozeduren, Notationen und deren Mischform, Konzepte, unterteilen (Goldkuhl et al. 1998), wobei diesen wiederum einzelne Elemente zugeordnet werden können. Das zentrale Methodenelement ist das Vorgehensmodell (Braun et al. 2004), dessen Entwicklung ein Ziel dieser Arbeit ist. Ein Vorgehensmodell beschreibt nach Sharafi (2013) die einzelnen Phasen eines Prozesses und strukturiert die damit verbundenen Maßnahmen. Das Vorgehensmodell, in Funktion eines Rahmenwerkes, muss deutlich vom Prozessmodell, in Funktion eines Regelwerkes, unterschieden werden. In der Literatur findet sich oftmals eine synonyme Verwendung beider Begriffe. In dieser Arbeit wird jedoch der korrekte Begriff Vorgehensmodell genutzt und nur bei direkten Zitaten oder Eigennamen der teilweise unsaubere Begriff der Literatur übernommen. Bei der praktischen Evaluierung des Vorgehensmodells wird die Durchführung einzelner Phasen im Modell diskutiert. In diesem Kontext wird der Begriff Prozess genutzt, wenn die sequentielle Ausführung von einzelnen Phasen verdeutlicht werden soll.

Der Term „Erschließen von Wissen“ beschreibt das Ziel der Methode, die Extraktion und anschließende Interpretation von Informationen, die unter gewissen Gegebenheiten, hauptsächlich Vernetzung (Dengel 2012), Wissen darstellen. Der Begriff Wissen wird in dieser Arbeit im ingenieurwissenschaftlichen Kontext untersucht. Grundlegende Betrachtungen wie die Erkenntnistheorie aus der Philosophie, die Frage, unter welchen Umständen der Begriff Wissen angewandt wird oder die Erforschung, aus welchen Grundlagen sich der Wissensbegriff historisch zusammensetzt, werden nicht diskutiert. Grundlage für die Wissensextraktion sind Datenmuster, eine Informationsrepräsentation, die aus der Verfahrenswelt des KDD, einem Vorgehensmodell zur Wissensentdeckung, stammt. Somit stellen die Muster den Betrachtungsgegenstand der zu entwickelnden Wissensentdeckung dar. Dar-

über hinaus stellt der Datenmusteraspekt in Verbindung mit dem Term „Erschließung von Wissen“ den Bezug zum KDD bezüglich des zu entwickelten Vorgehensmodells her. Viele der im KDD eingesetzten Verfahren zur Wissensentdeckung stammen ursprünglich aus der Mathematik oder der Statistik oder sind aus anderen Informatik-Disziplinen bekannt (Fockel 2009). Folglich kann konstatiert werden, dass der Begriff Wissensentdeckung in untreschiedlichen Disziplinen angesiedelt ist und praktische Lösungen in der Industrie oftmals einen interdisziplinären Ansatz erzwingen. Diese Arbeit startet in der Begriffswelt des KDD; eine darüberhinausgehende Betrachtung anderer Disziplinen sowie verwandter Gebiete der Informatik wird nur bei komplexen Begriffsklärungen berücksichtigt. Der letzte Teil des Titels „Supply-Chain-Datenbanken“ beschreibt den Kontext der zu entwickelnden Methode, deren Gültigkeit für den Einsatz in SCs gezeigt wird, jedoch nicht notwendigerweise auf diese beschränkt ist.

## 2.1 Wissen

Im Folgenden wird eine Einführung der Begriffe Wissen und Wissensmanagement gegeben. Im Anschluss werden verschiedene Kategorisierungsmöglichkeiten des Wissensbegriffs erörtert und aufgezeigt, welche Kategorisierung für den Kontext der SC Anwendung finden kann.

### 2.1.1 Definition von Wissen

In der Literatur finden sich unterschiedliche Definitionsansätze des Begriffs Wissen. So gibt es im Allgemeinen ein Verständnis des Begriffs Wissen, doch unterschiedliche Fachdisziplinen verwenden eine Vielzahl eigener Definitionen. Vorherrschend ist eine Abgrenzung der Begriffe Daten, Informationen und Wissen (Probst et al. 2006). Einige Autoren erwähnen noch zusätzlich den Begriff der Zeichen (Krcmar 2011). Diese Grundbegriffe Daten, Informationen und Wissen finden sich in der Wissenstreppe von North (2011) wieder. Abbildung 2.1 zeigt den Zusammenhang der Grundbegriffe nach diesem Abbildungsmodell. Hierbei stellen die Zeichen die unterste Ebene dar, die durch Ordnungsregeln (Syntax) in Daten überführt werden. North gibt an, dass die Anreicherung von Daten mit Bedeutung zu Information führt und verweist auf die hierarchische Beziehung zwischen den Begriffen, die von Krcmar (2011) dargelegt wurde.

> **Definition 2.1 Information:** Zeichen werden durch Syntaxregeln zu Daten, welche interpretierbar sind und damit für den Empfänger Information darstellen (vgl. North 2011; Probst et al. 2006).

In der angeführten Definition wurde der Begriff Kontext vermieden, obwohl Krcmar (2011) und Probst et al. (2006) von Kontextbezug sprechen. Die Begründung hier-

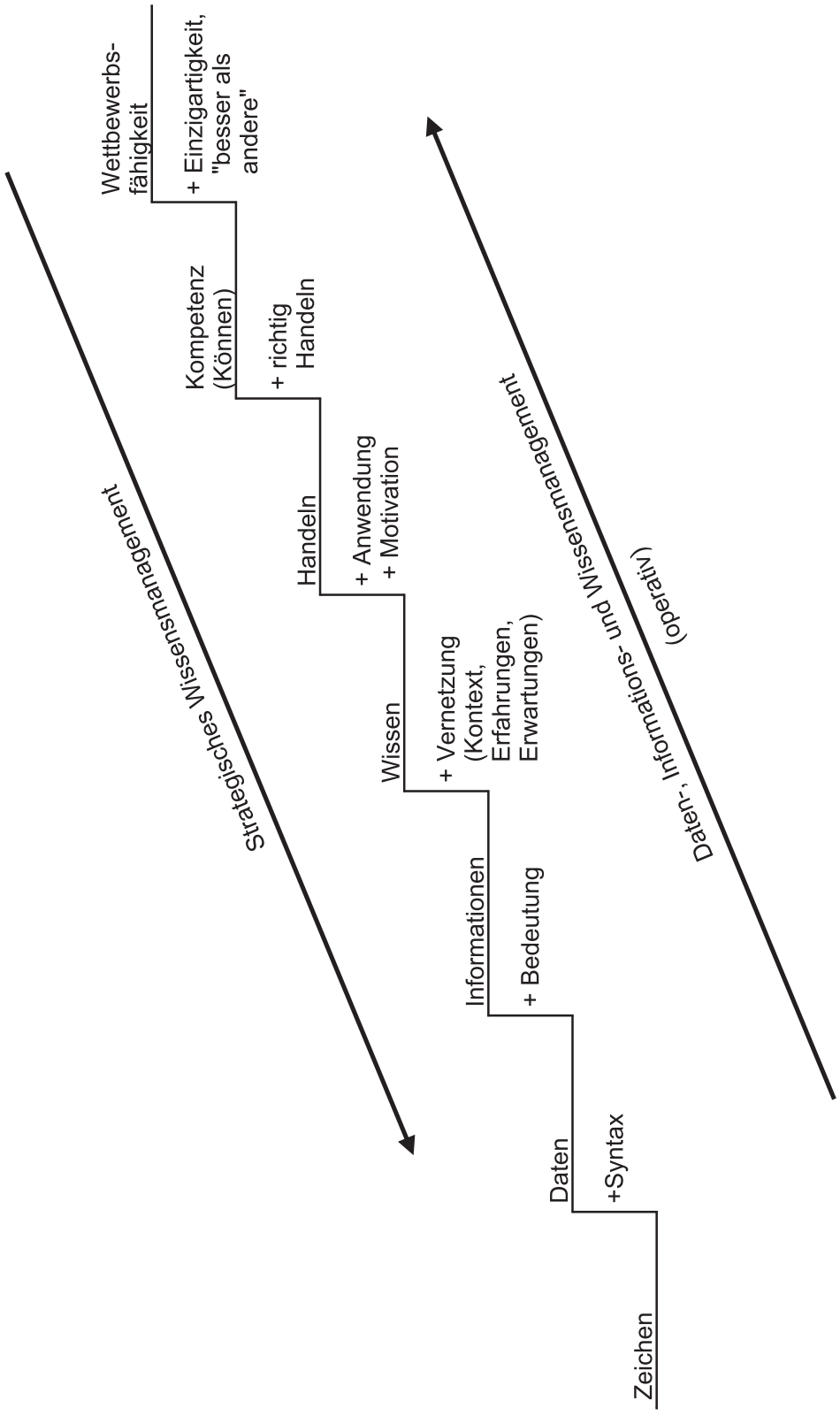



**Abbildung 2.1: Wissenstreppe nach North (2011, S. 37)**

für ist, dass der Kontextbezug in der Wissenstreppe erst bei der Überführung der Information von zentraler Bedeutung ist. Nach North (2011) entsteht Wissen, wenn Information mit weiteren Informationen, Erfahrungen oder Kontext vernetzt wird. Das wird auch in dem Zitat des Computerpioniers Feigenbaum deutlich: „Knowledge is not the same as information. Knowledge is information that has been pared, shaped, interpreted, selected, and transformed“ (Feigenbaum und McCorduck 1983, S. 121). Demgegenüber stehen spezifische Wissensdefinitionen, wie beispielsweise die von Sowa (2000), im Kontext der künstlichen Intelligenz. Hier wird Wissen als die Fähigkeit beschrieben, ein mentales Modell zu erstellen, welches die Teilaspekte der Wirklichkeit in geeigneter Weise repräsentiert. Der Begriff Wissen muss immer im Problemkontext betrachtet werden, da verschiedene Fachdisziplinen den Begriff für ihre Bedürfnisse adaptiert haben. Zusätzlich ist die Suche nach beliebigen Strukturen und somit unspezifischem Wissen basierend auf den heutigen Methoden unmöglich (Küppers 1999).

Die vorliegende Arbeit folgt in der Begriffsdefinition von „Wissen“ der weit verbreiteten Definition nach Probst, der die zentralen Begriffe Daten und Informationen aufnimmt. Er ergänzt diese Begriffe um den Begriff des „Individuums“, um die menschliche Interaktion zu verdeutlichen. Die Ergänzung um die Interaktionsmöglichkeit wird als wesentlich erachtet, da die Wissensentdeckung in Funktion eines Prozesses immer manuelle Komponenten wie die menschliche Interpretation beinhaltet. Diese Definition bildet auch eine der Grundlagen für die Wissenstreppe, die in Abbildung 2.1 dargestellt ist. Das aufgeführte Originalzitat in Definition 2.2 wurde zum späteren Zeitpunkt auch in die deutsche Spezifikation des Wissensmanagements aufgenommen (DIN SPEC 91281:2012-04), hier wird jedoch die Originalquelle zitiert.

> **Definition 2.2 Wissen:** „Wissen bezeichnet die Gesamtheit der Kenntnisse und Fähigkeiten, die Individuen zur Lösung von Problemen einsetzen. Dies umfasst sowohl theoretische Erkenntnisse als auch praktische Alltagsregeln und Handlungsanweisungen. Wissen stützt sich auf Daten und Informationen, ist im Gegensatz zu diesen jedoch immer an Personen gebunden. Es wird von Individuen konstruiert und repräsentiert deren Erwartungen über Ursache-Wirkungs-Zusammenhänge“ (Probst et al. 2006, S. 44).

Die nächsten drei Treppenstufen der Wissenstreppe in Abbildung 2.1 verdeutlichen, wie ein Unternehmen das Wissen verwerten kann, um Wettbewerbsfähigkeit zu erlangen. Dieser Treppenbereich ist dem Wissensmanagement zugeordnet. Unter Wissensmanagement ist die Nutzung des Wissens als Ressource der organisationalen Wissensbasis sowie die systematische Planung, Steuerung und Organisation der Wissensbedarfe der Organisation zu verstehen (Schaffranietz und Neumann 2009). Das Verständnis dieser Arbeit richtet sich nach der verbreiteten Spezifikation von Heisig:

> **Definition 2.3 Wissensmanagement:** „Wissensmanagement umfasst Verfahren, Methoden, Instrumente und Werkzeuge, die einen systematischen, methodengestützten Umgang mit Wissen in allen Bereichen und auf allen Ebenen der Organisation realisieren, um die organisatorische Leistungsfähigkeit der Geschäftsprozesse zu verbessern und zur Erreichung der Organisationsziele beizutragen“ (Heisig 2005, S. 18).

Die Notwendigkeit des Einsatzes von Wissensmanagementsystemen als einer Methode der Standardisierung und Systematisierung von Wissen begründet sich nach Wenzel, Abel et al. (2011) in der Komplexität der Wissensarbeit. Es gibt unterschiedliche Wissensmanagement-Modelle, die sich in Beschreibung der Aufgaben und Aktivitäten deutlich voneinander unterscheiden. Diese Modelle sollen jedoch im Kontext der vorliegenden Arbeit nicht diskutiert werden, da der Fokus auf der Wissensentdeckung liegt. In der Folge ist das Grundverständnis von der Existenz des Wissensmanagements ausreichend. Die Wissensentdeckung umfasst im Wesentlichen die Treppenstufen der Daten, Information und des Wissens in Abbildung 2.1, wobei die Unterscheidung zwischen Daten und Information im Bezug auf die Wissensentdeckung nicht immer eindeutig ist. Dies ist darin begründet, dass die Ebenen der Daten und der Informationen oftmals im Umfeld des KDDs nicht disjunkt sind, da die Zuordnung vom Betrachter abhängig ist. Folglich wird der Begriff Information nur dann verwendet, wenn eine Interpretierbarkeit der Daten durch zusätzliche Randbedingungen angezeigt ist. Dies ist oftmals in späteren Phasen des KDDs der Fall. Auf Ebene der Datenbank und den Rohdaten findet der Begriff Daten Anwendung, da hier häufig die technischen Gegebenheiten und nicht die inhaltlichen Bedeutungen der Daten im Vordergrund stehen.

### 2.1.2 Systematisierung des Wissens

In der zuvor aufgeführten Definition des Wissensmanagements wird Wissen als Ressource bezeichnet. Diese Ressource kann aus Unternehmenssicht in verschiedene Kategorien unterteilt werden. Die gebräuchlichsten Kategorien sind Explizierungsgrad (implizit oder explizit), Anwendbarkeit (kontextbezogen oder dekontextualisiert) und Zugänglichkeit (individuell oder kollektiv) (Bullinger et al. 2009). Probst et al. (2006) schlagen abweichende Kategorisierungen vor, die auf die Ebenen Daten, Information und Wissen projiziert werden. Hierbei werden beispielsweise Daten als kontextunabhängig, Wissen jedoch als kontextabhängig gesehen. Da in der Kategorisierung von Probst der Explizierungsgrad keine Berücksichtigung findet, der insbesondere für die Darstellung von Wissen von großer Bedeutung ist, dient als Grundlage die Kategorisierung von Bullinger. Abbildung 2.2 gibt eine Übersicht der wesentlichen Kategorien.

Der Begriff explizites Wissen drückt aus, dass etwas eindeutig kodierbar und somit eindeutig kommunizierbar ist. Den Gegensatz hierzu bildet das implizite (tazite) Wissen, das nicht einfach fassbar ist; beispielsweise, weil es erfahrungsbezogen oder

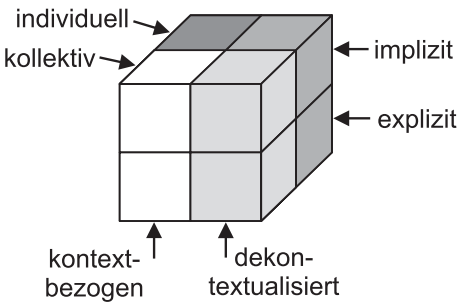



**Abbildung 2.2: Ressource Wissen nach Bullinger et al. (2009, S. 703)**

nicht formalisierbar ist. Kontextbezogenes Wissen ist spezifisch angewandtes Wissen wie Aufgaben-, Prozess- oder Organisationswissen. Eine Übertragbarkeit auf andere Objekte oder Situationen ist im Allgemeinen nicht möglich. Im Gegensatz dazu beschreibt dekontextualisiertes Wissen grundlegende Vorgehensweisen und Handlungen, die auf unterschiedliche Fragestellungen angewandt werden können. Individuelles Wissen ist nur einzelnen Personen zugänglich, kollektives Wissen immer mehreren Personen. Für eine über diese Definitionen hinausgehende Betrachtung der Kategorien und deren Kombinationsmöglichkeiten sei auf Heisig (2005) verwiesen.

## 2.2 Supply Chains

Im folgenden Abschnitt werden die SC, das SCM und die aktuelle Datenlage in der SC diskutiert, um diese mit dem zuvor eingeführten Wissensbegriff zu verknüpfen. Neben dem grundlegenden Verständnis von SC und SCM liegt der Fokus auf den Datenbeständen in der SC, welche die Grundlage der Wissensentdeckung bilden. Um in den nachfolgenden Kapiteln ein gemeinsames Verständnis der Sachverhalte zu gewährleisten, ist die informationstechnische Begriffserklärung das wesentliche Ziel dieser Diskussion.

### 2.2.1 Grundlagen der Supply Chain

SCs sind ein zentraler Bestandteil der Produktionslandschaft (Stevens 1989). Zur SC gehören alle Unternehmen bzw. eigenständig handelnde Unternehmensteile, die zur Entwicklung, Erstellung und Lieferung von Produkten oder Dienstleistungen beitragen und reichen von einer Quelle (Zulieferer der Zulieferer) bis zu einer Senke (Kunden der Kunden) (Beckmann 2012). In der SC werden traditionell Materialflüsse (bzw. Dienstleistungsflüsse), Finanzflüsse und Informationsflüsse unterschieden. Die Betrachtung der Flüsse erfolgt sowohl stromaufwärts zum Lieferanten als auch stromabwärts zum Kunden. Das Ziel einer SC ist, dem Kunden die richti-

gen Güter am richtigen Ort und zum richtigen Zeitpunkt zur Verfügung zu stellen (Jünemann 1989).

Der Begriff SC bedeutet wörtlich übersetzt Versorgungs- oder Lieferkette (Wiendahl 2004). Es hat sich jedoch keine einheitliche Bedeutung durchgesetzt und die klassische SC findet sich in der deutschen Literatur am ehesten unter dem Begriff der Wertschöpfungskette, die in den 80er Jahren von Porter etabliert wurde (Arnold, Isermann et al. 2008). Jedoch mag die wörtliche Übersetzung mit dem Begriff Kette irreführend sein, wenn beachtet wird, dass die Literatur in ihren Definitionen der SC zumeist den Begriff des Netzwerks heranzieht (Arndt 2013; Arnold, Isermann et al. 2008; Günther und Tempelmeier 2011; Vahrenkamp und Kotzab 2012). So trifft Lambert (2005, S. 2) die Aussage „Strictly speaking, the supply chain is not a chain of business, but a network of business and relationships“ . Blackstone (2010, S. 213) definiert die SC als „The global network used to deliver products and services from raw materials to end customers through an engineered flow of information, physical distribution, and cash“ . Lee und Ng (1997) bezeichnen die SC als Netzwerk von Entitäten. Diese Formulierung greifen auch Fleischmann und Meyr (2001) auf. Sie bezeichnen die SC als eng verflochtenes Netz von Entitäten, in dem Produkte oder Dienstleistungen hergestellt und zum Abnehmer geliefert werden. Auch Riha (2009, S. 78) greift den Netzwerkcharakter auf, in dem die SC als „eine auf Logistikdienstleistung spezialisierte Form eines Netzwerks“ bezeichnet wird. Nach Corsten und Gössinger (2008) entsteht die SC durch den Zusammenschluss von Unternehmen zu Netzwerken, verbunden mit einer unternehmensübergreifenden Betrachtungsweise. Günther und Tempelmeier (2011) lassen als Übersetzung der logistischen Kette die Begriffe SC und Supply Network gleichbedeutend nebeneinander stehen. Christopher (2011) schlägt sogar vor, den Begriff SC um den Zusatz Network zu erweitern, um die Tatsache zu berücksichtigen, dass heute oftmals keine linearen Ketten vorliegen, sondern vielmehr Geflechte von vielen Herstellern, vielen Kunden und Kunden der Kunden. Ähnlich sehen es weitere Autoren, die versucht haben, den Begriff SC um die Komponente des Netzwerks zu ergänzen oder abzuändern (vgl. Chopra und Meindl 2014 oder Stolzle und Otto 2003).

In dieser Arbeit wird dennoch der Begriff SC verwendet, da er die Literatur bis zum heutigen Zeitpunkt dominiert, wohl wissend, dass es sich bei der Kette in der Realität zumeist um ein Netzwerk handelt. Die angeführten Überlegungen und Ansätze finden Berücksichtigung in der Definition von Christopher, die in dieser Arbeit dem Grundverständnis dient.

> **Definition 2.4 Supply Chain:** „The supply chain is the network of organizations that are involved, through upstream and downstream linkages, in the different processes and activities that produce value in the form of products and services in the hands of the ultimate consumer“ (Christopher 1998, S. 15).

Mittels Typologisierung lassen sich verschiedene Arten von SCs durch ihre Merkmale abgrenzen. Hierbei wird anhand von spezifischen Merkmalen und Merkmalsausprägungen zwischen verschiedenen Typen der SCs unterschieden, sodass darauf aufbauend beispielsweise eine Abgrenzung der Anwendungsbereiche von Methoden oder Softwaresystemen durchgeführt werden kann. Da es je nach Sichtweise eine Vielzahl von Merkmalen und Anordnungsmöglichkeiten gibt, haben sich im Verlauf der Zeit sehr unterschiedliche SC-Typologien entwickelt (Giese 2012). Im Kontext der Wissensentdeckung, die als Grundlage Datenbestände benötigt, sind viele der Typologien nicht zielführend. Diese Typologien verwenden Unterscheidungsmerkmale, die keinen Bezug zu den Datenbeständen oder Wissensentdeckungsaufgaben aufweisen. Hier fehlt generell eine Beschreibung von unterschiedlichen SC-Merkmalen und ihr Bezug zu SC-Datenbanken in der Literatur. Basierend auf dem Fokus der Arbeit wurde eine Typisierung der SC nach Meyr und Stadtler (2005) durchgeführt, die eine Reihe von funktionellen und strukturellen Merkmalen sowie Unterkategorien entwickelt haben. Während sich die funktionalen Merkmale in vier Kategorien aufgliedern (Art der Beschaffung, Art der Produktion, Art der Distribution und Art des Absatzes), werden strukturelle Merkmale in die Merkmalskategorien der Topografie der SC, Integration und Koordination eingeordnet. Eine Auswahl dieser Kategorien wurde zur Typisierung der hier behandelten SCs genutzt (vgl. Tabelle 2.1), da sich aus der Analyse von Transaktionsdaten Informationen über diese Kategorieausprägungen ableiten lassen. So können beispielsweise Distributionsmuster aus Zeitstempeln für Transporte abgeleitet werden. Kategorien, deren Ausprägungen nicht auf Datenebene zu erfassen sind, finden in der Tabelle 2.1 keine Berücksichtigung.

Tabelle 2.1 zeigt die wichtigsten Merkmalsausprägungen der verwendeten SC-Daten, die der weiteren Arbeit als Einschränkung zugrundegelegt werden. Es wird ausdrücklich darauf hingewiesen, dass branchenspezifische Supply Chains z. B. aus dem Energiesektor aufgrund ihrer anfallenden Daten und Merkmalsausprägungen nicht im Rahmen der vorliegenden Arbeit diskutiert werden.

### 2.2.2 Datenaufkommen in Supply Chains

Die Veränderungstreiber der Logistik (z. B. Globalisierung und Regionalisierung, Informations- und Kommunikationstechnologien (IuK) oder Kooperationen) wirken auf alle ihr zugeordneten Bereiche, so auch auf die SC. Bedingt durch Globalisierung und exponentielle Verbesserung der IuK steigt die Netzwerkgröße und die Komplexität der SCs signifikant (Serdarasan 2013). Der Anstieg der Netzwerkgröße und die steigende Komplexität beeinflussen unmittelbar die durch die Flüsse gekennzeichneten Prozesse der SC. Die Prozesse wiederum verwenden in der Folge eine größere Datenbasis und erzeugen selbst auch größere Datenbestände, als dies noch vor einigen Jahren der Fall war (European Commision December 2012). Die Größe der Datenbasis ist durch die zwei Dimensionen Volumen und Komplexität gekennzeichnet. Die Dimensionen und ihre Verortung in einer Tabelle können der

**Tabelle 2.1: Merkmalsausprägungen der untersuchten Supply Chains nach Meyr und Stadtler (2005)**

| Merkmalskategorie | Unterkategorien | Typische Merkmalsausprägung |
|---|---|---|
| Art der Produktion | Wiederholung der Abläufe | Losfertigung |
| | Umrüstaufwand | Hoch, abfolgeabhängige Rüstzeiten und -kosten |
| Art der Distribution | Distributionsstruktur | 3 Ebenen |
| | Distributionsmuster | Dynamisch |
| | Einsatz von Transportmitteln | Unbegrenzt |
| Art des Absatzes | Verfügbarkeit zukünftiger Nachfrage | Vorhersage |
| | Anzahl der Produktarten | Mehrere Produkte |
| | Grad der individuellen Anpassung | Standardprodukte |
| | Stückliste | Divergent |
| | Anteil an Dienstleistungen | Nicht vorhanden |
| Topographie der SC | Netzwerkstruktur | Gemischt (divergente und konvergente Elemente) |
| | Grad der Globalisierung | Mehrere Länder |
| | Position der Entkopplungspunkte | Auftragsproduktion, Produktion in Projekten |

Abbildung 2.3 entnommen werden. Die Dimension Volumen bedeutet, dass die Anzahl der einzelnen Datensätze von Relevanz ist. Die Dimension Komplexität, die auch als Varietät bezeichnet wird (Che et al. 2013), kennzeichnet, dass die Anzahl der Einzelelemente innerhalb eines Datensatzes betrachtet wird.

Die SC-Datenbestände sind inhomogen bezüglich der verwendeten Datenkategorien, da Daten sowohl in numerischer als auch alphanumerischer Form enthalten sind. Die unterschiedlichen Kategorien, in denen die Daten vorliegen, stehen in Wechselwirkung zu den Methoden der Wissensentdeckung. Das ist damit begründet, dass alle Methoden spezifische Eingabedaten benötigen und somit Anforderungen an die zugrundeliegende Datenbasis stellen. Neben möglichen Datenkategorien

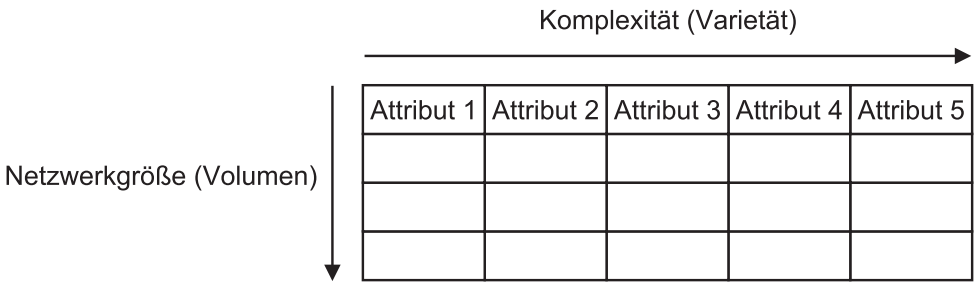



**Abbildung 2.3: Netzwerkgröße und Komplexität in Zuordnung zur Tabellendarstellung**

ist die Datenqualität ein maßgeblicher Einflussfaktor, denn ohne hinreichende Datenqualität sind die Ergebnisse der Wissensentdeckung nicht aussagekräftig. Aufgrund der angesprochenen Wechselwirkungen sollen in den folgenden Abschnitten die Daten, ihre mögliche Kategorisierung und die Datenqualität im Kontext der SC thematisiert werden.

#### 2.2.2.1 Datendefinitionen im Kontext der SC

Im Kontext der Datenbasis als Grundlage der Wissensentdeckung wird der Begriff „Datenbestand“ im Plural verwendet, sofern nicht ein spezifischer Datenbestand im Rahmen von Verfahrensbeschreibungen oder durchgeführten Experimenten adressiert ist. Die Begründung liegt darin, dass im Regelfall mehrere Datenbestände aus separaten Systemen eine Datenbasis bilden. Ein Datenbestand besteht aus einzelnen Datensätzen, die wiederum aus Daten bzw. Datenfeldern zusammengesetzt sind. Zwischen Daten und Datenfeldern lässt sich inhaltlich keine Unterscheidung treffen. So nutzt Mertens (Mertens, Bodendorf et al. 2012) den Begriff Datenfelder als kleinste Einheit, während das Lexikon für Wirtschaftsinformatik (Stickel et al. 1997) in einer äquivalenten Formulierung Daten als kleinste Einheit definiert. Hier soll als kleinste Einheit der Begriff Datenfelder gewählt werden, um eine deutliche Unterscheidung zum Begriff Daten zu ermöglichen.

Datensätze können in Dateien oder Datenbanken gespeichert werden. Im Umfeld der SC sind Datenbanken, insbesondere relationale Datenbanken, von zentraler Bedeutung. Datenbanken werden aus den Rohdaten der unterschiedlichen operativen SC-Systeme mittels Extraktion-Transformation-Laden (ETL) befüllt (Mertens, Back et al. 2001). Der Speichermechanismus, der sich lediglich ein oder mehreren Dateien als Speicherort bedient, findet in komplexen Systemen häufig nur beim Datenaustausch Verwendung. Dieser Datenaustausch kann beispielsweise zwischen zwei Systemen mittels sogenannter Schnittstellen stattfinden. Auf den Datenbanken basieren nachgelagerte Modelle, wie Business Layer, Data-Warehouse oder Data Marts (vgl. Parimala und Pahwa 2008). Diese Modelle greifen mittels ETL auf die Datenbestände der zugrundeliegenden Datenbanken zu (zumindest auf Ebene der einzelnen Zuliefersysteme) oder werden mittels Schnittstellen-Exporten

aus Datenbanken beliefert. Die genannten Modelle wie das Data Warehouse bestehen oftmals aus aggregierten Werten, d. h. sie haben eine hohe Aggregationsstufe, und lassen nur schwer einen Rückschluss auf die zugrundeliegenden Datensätze zu (Messaoud et al. 2006). Daher wird zur Vereinfachung von einer Speicherung in heute üblichen relationalen Datenbanken ausgegangen.

Eine relationale Datenbank besteht aus einer oder mehreren Tabellen, in welchen die Datensätze gespeichert werden. Dabei stellt jede Zeile einer Tabelle einen Datensatz dar – dieser Datensatz wird im Kontext der relationalen Algebra auch als Tupel bezeichnet. Jedes Tupel enthält ein oder mehrere Datenfelder, wobei gleichartige Datenfelder spaltenweise angeordnet werden und als Attribute gekennzeichnet werden. Jedes Attribut kann mit einem Attributnamen versehen werden. Die konkrete Wertebelegung der einzelnen Attribute wird als Attributsausprägung bezeichnet. Alternativ werden die Einträge in Tabellen häufig als Entitäten, ihre Spalten als Merkmal und die Ausprägungen als Merkmalsausprägungen bezeichnet. Eine Erklärung hierfür kann in den Entity-Relationship-Modellen (ERM) gefunden werden. Hier entsprechen die Entitätstypen den Tabellen und die konkreten Instanzen den Entitäten. Im Allgemeinen bezeichnet der Begriff Merkmal die zu beobachtende Eigenschaft eines Objektes und wird auch im spezifischen Kontext als statistische Variable, Untersuchungsmerkmal oder kurz Variable bezeichnet (Fahrmeir et al. 2010, S. 148). Die nachfolgende Definition ist Grundlage für die vorliegende Arbeit:

> **Definition 2.5 Datensatz:** Ein Datensatz ist eine Entität, seine Datenfelder sind Merkmale und die konkreten Belegungen der Datenfelder werden als Merkmalsausprägungen bezeichnet.

#### 2.2.2.2 Kategorisierung der Supply-Chain-Daten

Basierend auf den Veränderungen der Unternehmen (vgl. Abschnitt 2.2.2) nehmen Umfang und Komplexität der Datenbestände in SCs fortwährend zu. Hierbei ist es wichtig, die vorhandenen Daten grundlegend zu kategorisieren, da die verschiedenen Methoden der Wissensentdeckung unterschiedliche Anforderungen an die zugrundeliegende Datenbasis stellen. Dementsprechend sind Daten und die damit verbundenen Kategorisierungssysteme im Allgemeinen nicht für spezifische SC-Typen ausgelegt, sodass der Begriff der Datenkategorisierung deutlich von der SC-Typologie abgegrenzt werden muss (vgl. Abschnitt 2.2.1).

Datenbestände können im Allgemeinen nach verschiedenen Kriterien kategorisiert werden. Eine Vielzahl von Autoren hat unterschiedliche Kategorisierungssysteme vorgeschlagen, wie beispielsweise die Unterteilung in analog und digital (Dworatschek 1989). Dabei ist es heutzutage nicht zweckmäßig, auf diesen Betrachtungsebenen anzusetzen, da die Datenbasis einer globalen SC regelmässig in digitaler Form vorliegt. Für die problemorientierte Sichtweise wird eine Kategorisierung

nach Piro und Gebauer (2011) vorgenommen. Die Autoren nennen sechs Kategorien, die im Folgenden kurz erläutert werden und zur Einführung und Diskussion der grundlegenden Begrifflichkeiten dienen. Anschließend wird ein spezifisches Datenmodell eingeführt und die Daten der SC bezüglich ihrer Einordnung in dieses Modell geprüft.

Das erste Kriterium des Kategorisierungssystems nach Piro und Gebauer (2011) ist das Format. Hierbei ist jedes Datenfeld mit einem ausgezeichneten Datentyp assoziiert. Ein Datentyp gibt dabei an, welchen Wertebereich ein Datenfeld annehmen kann. Zugleich legt der Datentyp die Operationen fest, die auf dem zugeordneten Datenfeld ausgeführt werden können. Es gibt verschiedene Kategorien von Datentypen, die jedoch immer abhängig von Prozessor und Programmiersprache zu sehen sind. In erster Annäherung können Daten in alphabetisch, numerisch, alphanumerisch sowie Bildzeichen (Index und Symbolbilder) unterteilt werden (Dworatschek 1989). Hierbei werden alphabetische Daten aus alphabetischen Zeichen gebildet und numerische Daten bestehen aus Ziffern. Alphanumerische Daten bilden wiederum Elemente bestehend aus Ziffern, Buchstaben oder Sonderzeichen ab. In der darunterliegenden Abstraktionsstufe wird jedoch deutlich, dass beispielsweise der Bereich der numerischen Daten weiter gegliedert werden kann. Tabelle A.1 im Anhang gibt eine Übersicht über mögliche Datentypen und dient in der vorliegenden Arbeit als Referenz. Da es verschiedene Kategorien von Datentypen in unterschiedlichen Referenzsprachen gibt, ist es notwendig, die Datentypdefinitionen an eine formale Sprache zu koppeln. In der vorliegenden Arbeit wurde als formale Sprache die Datenbanksprache Structured Query Language (SQL) genutzt. SQL wurde ausgewählt, da in den hier betrachteten SC-Architekturen von Speichermechanismen mittels relationaler Datenbanken ausgegangen wird. Alle gängigen relationalen Datenbanken unterstützen SQL.

Das zweite Kategorisierungskriterium ist die Strukturierung der Daten. Dieses Kategorisierungskriterium mischt sich oftmals mit dem Formatkriterium und ist manchmal unter Begriffen wie formatiert und unformatiert in der Literatur aufzufinden (Mertens, Back et al. 2001). Ist von Strukturiertheit im engeren Sinne die Rede, können Datenbestände in strukturiert, semi-strukturiert (auch als halbstrukturiert bezeichnet) und unstrukturiert unterteilt werden. Im Umfeld von Big Data findet sich zusätzlich der Begriff der Polystrukturiertheit. Die polystrukturierten Daten bezeichnen hierbei eine Mischung aus strukturierten, unstrukturierten sowie maschinengenerierten Daten wie es beispielsweise bei Daten aus RFID-Systemen der Fall ist. Strukturierte Daten hingegen sind zum Beispiel Kundenstammdaten mit alphabetischen und numerischen Anteilen. Allen strukturierten Daten gemeinsam ist, dass sie sich mit klassischen Datenbanksystemen wie einer relationalen Datenbank speichern und verarbeiten lassen. Die Unterteilung von semi-strukturierten und unstrukturierten Daten hingegen ist nicht deutlich ausgeprägt. Dies liegt insbesondere daran, dass auch bei unstrukturierten Daten zumindest ein geringer Teil der Daten eine Struktur aufweist. Häufige Vertreter von unstrukturierten Daten, die zumindest teilweise Strukturen aufweisen, sind u. a.

E-Mails. Klassische unstrukturierte Daten haben hingegen kein Schema, keine Tags oder Metadaten, die Informationen über Inhalte bereitstellen. Die Gesellschaft für Informatik schlägt wegen der unsauberen Trennung der Begrifflichkeiten weitere Aufteilungen im Bereich der semi-strukturierten und unstrukturierten Daten vor. Da jedoch die SC-Daten in Unternehmensdatenbanken oder ähnlichen Konstrukten verwaltet werden (vgl. Abschnitt 2.2.2.1) und ursprüngliche Rohdaten nach Transformation üblicherweise gelöscht werden, handelt es sich um strukturierte Daten und eine weitere Aufschlüsselung ist für diese Arbeit nicht erforderlich.

Ein weiteres Kriterium ist die Stabilität, d. h. die Zeitdauer, in der die Daten unverändert bleiben. Stammdaten (z. B.) sind Daten, die sich nur geringfügig über den Zeitverlauf verändern und (u. a.) zur Identifikation, Klassifikation und Charakterisierung von Objekten dienen (Otto und Hüner 2009). Daher wird auch synonym der Begriff feste oder fixe Daten verwendet. Ein klassisches Beispiel sind Adressdaten, wie sie in der SC für Lieferanten und Kunden gehalten werden. Einige Autoren führen Daten, die im Zusammenhang mit Änderung der Stammdaten stehen, separat ein und nennen diese Änderungsdaten (Tiemeyer 2007). Auf diese Unterteilung wird jedoch im Kontext dieser Arbeit verzichtet, da diese für das zu entwickelnde Vorgehensmodell nicht zielführend ist. Alle nicht-fixen Daten zählen zur Kategorie der variablen Daten. Hierzu zählen z. B. Bestandsdaten, die Mengen- und Wertestrukturen in der Datenhaltung kennzeichnen. Bestandsdaten sind durch die betrieblichen Prozesse Änderungen unterworfen. Bewegungsdaten hingegen sind Daten, die starke Änderungen über die Zeit erfahren. Die zur SC-Planung notwendigen Informationen werden aus den Bewegungsdaten, wie z. B. Kapazitäten, Terminen und Auftragsdaten, gewonnen und sind damit Voraussetzung für ein erfolgreiches Informationsmanagement (Hellingrath, Laakmann et al. 2004). Bewegungsdaten entstehen durch den betrieblichen Leistungsprozess und werden durch diesen verändert. Die Bewegungsdaten bewirken Änderungen an den Bestandsdaten. Die Bewegungsvorgänge werden als Transaktionen bezeichnet, die Daten synonym als Transaktionsdaten. In einem SC-Datenbestand gibt es aufgrund der Vielzahl von möglichen Quellsystemen prinzipiell Daten aus verschiedenen Stabilitätsklassen. Der größte Teil des Datenbestandes wird jedoch von Transaktionsdaten gebildet. Da die Transaktionsdaten für die Daten-und Managementprozesse der SC von essentieller Bedeutung sind, bilden sie die Grundlage für die (angestrebte) Wissensentdeckung (Anane et al. 2002).

Ein weiteres Kriterium ist der Funktionsbezug der Daten. Hier wird zwischen der eigentlichen Informationsbeschreibung (auch als Nutzdaten oder Inhaltsdaten bezeichnet) und Metadaten (auch als Steuerdaten bezeichnet) unterschieden (Hansen und Neumann 2005). Je nach zugrundeliegender Struktur der SC-Systemlandschaft kommen beide Varianten in einem Datenbestand vor. Als Informationsbeschreibung werden Daten bezeichnet, die einen Sachverhalt darstellen und Informationen beinhalten. Metadaten wiederum sind strukturgebend für die Inhaltsdaten, weisen jedoch keine eigenen Inhalte auf. Alternative Kategorisierungsmodelle (z. B. Lassmann 2006) unterscheiden in diesem Zusammenhang Kriterien basierend auf

den Aufgaben der Daten im Informationsverarbeitungsprozess. Alle diese Modelle basieren auf den zentralen Begriffen der Informationsbeschreibung und der Metadaten.

Des Weiteren können Daten bezüglich ihres Verarbeitungsstandes innerhalb der Prozessketten unterschieden werden. Hierbei wird zwischen den Kategorien Eingabe-, Speicher- und Ausgabedaten unterschieden. Die SC-Daten können über den zeitlichen Verlauf der Betriebsprozesse die Kategorie des Verarbeitungsstandes wechseln. Da die Wissensentdeckung größere Zeiträume benötigt, um interessante Zusammenhänge zu entdecken, erfolgt die Suche auf historisierten Daten. Historisierte Daten sind immer Speicherdaten.

Zuletzt können alle Daten einem oder mehreren Geschäftsobjekten (Business-Objekte, BOs) zugeordnet werden. Ein BO repräsentiert in einem System ein Objekt der Geschäftswelt. BOs enthalten neben der reinen Beschreibungen von Objekten auch Beschreibungen zu deren Funktionsweise. Typische BOs im Kontext der SC sind Lieferanten, Güter, Lager, Hubs und Endkunden. Die Daten der SC umfassen je nach korrespondierenden Prozessen eine Vielzahl von BOs.

Eine zusätzliche Dimension entsteht in der Kategorisierung durch sogenannte Kontextkriterien, die neben der reinen Beschaffenheit der Daten auch Auskunft über die Art der Informationen geben, die in den Daten hinterlegt sind. Diese Kontextkriterien sind notwendig, um eine eindeutige Bedeutung und in der Folge Interpretation der Daten innerhalb von Unternehmen zu ermöglichen. Piro und Gebauer (2011) unterscheiden beispielsweise explizit zwischen dem Unternehmenszweck und dem Prozessbezug, weisen jedoch darauf hin, dass die Kontextkriterien je nach Anwendungsbereich, in dem die Datenkategorisierung genutzt werden soll, individuell gestaltbar sind und insbesondere ein konkreter Datenbestand mehreren Kontextkriterien zugeordnet werden kann. Die verschiedenartigen Kontextkriterien sind von unterschiedlichen Autoren in eigene Modelle überführt worden. Da es jedoch in der Literatur kein Kategorisierungsmodell für die spezifischen Anforderungen der SC-Daten gibt, sind Vorarbeiten am Fachgebiet IT in Produktion und Logistik (ITPL) zu diesem Bereich betreut worden. Das Modell in der Abbildung 2.4 gibt die Essenz eines Datenkategorisierungsmodells von SC-Daten basierend auf Ziegler (2015) wieder und zeigt die Aufschlüsselung der zeichenorientierten, strukturierten Daten. Die Zuordnung der einzelnen Daten sowie der Bezugsrahmen orientiert sich dabei vorwiegend an dem Modell von Oedekoven (2011).

#### 2.2.2.3 Transaktionen

Bei der Analyse der SC-Datenbestände in Abschnitt 2.2.2.2 wurde die Stabilität der Daten diskutiert. Daten, die im Kontext der Stabilität von besonderer Bedeutung sind, sind die Transaktionsdaten. Transaktionen beschreiben betriebswirtschaftlich gesehen den Wechsel eines materiellen oder immateriellen Objektes aus dem Wirkungskreis eines Akteurs in den eines anderen (Corsten und Gössinger

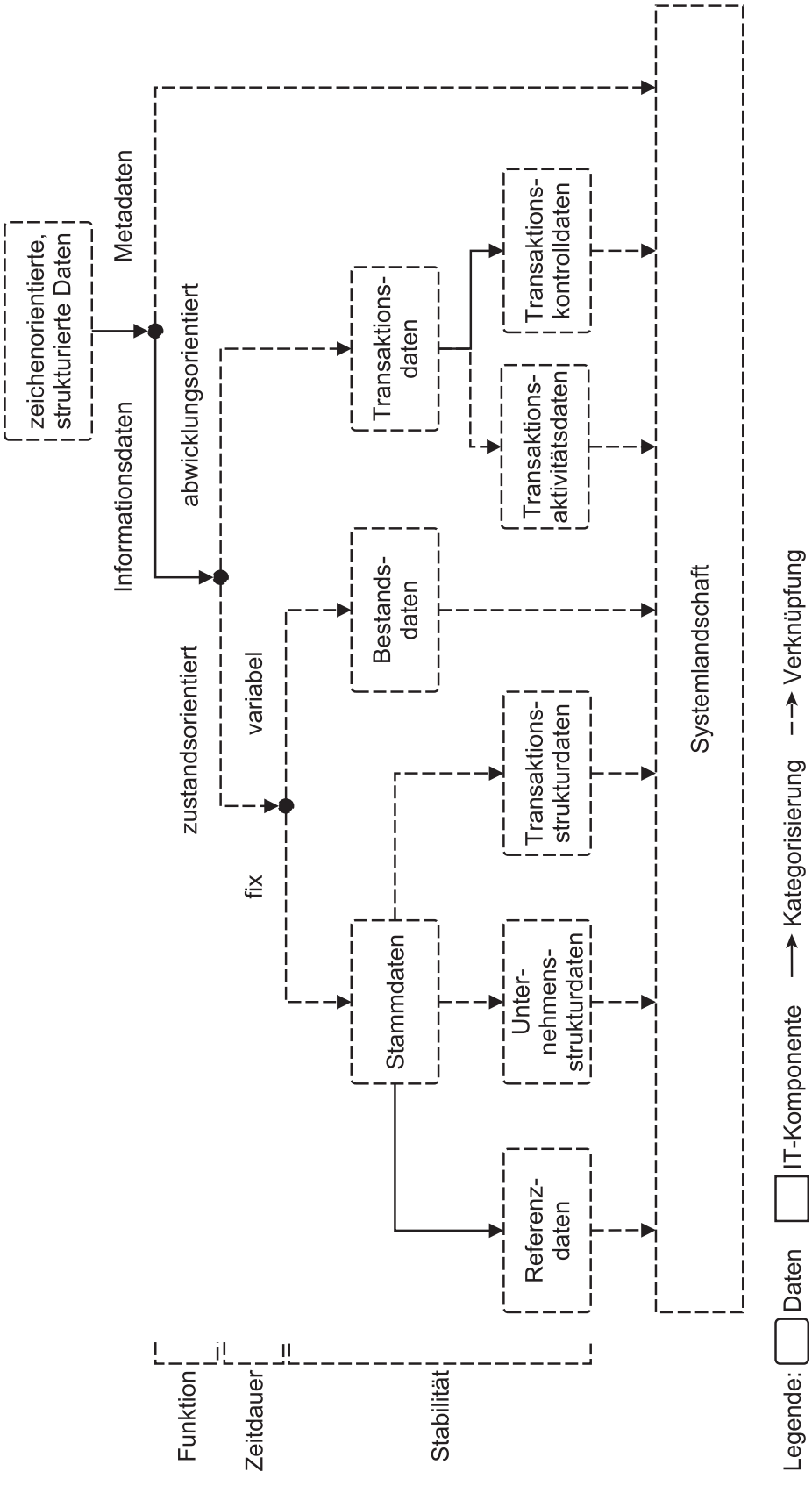



**Abbildung 2.4: Kategorisierungsmodell SC-Daten**

2008). Auch im Gabler Kompakt-Lexikon findet sich eine ähnliche Definition, in der eine Transaktion als der Austausch von Gütern und Leistungen erklärt wird. Dabei werden Güter als materielles oder immaterielles Mittel zur Befriedigung menschlicher Bedürfnisse genannt und eine Leistung aus betriebswirtschaftlicher Sicht als Ausbringungsmenge eines Produktionsprozesses definiert (Stickel et al. 1997). Eine Transaktion besteht aus kennzeichnenden Merkmalen, insbesondere Zeitstempel und Mengenangaben (Moody und Kortink 2000). Eine wesentliche Eigenschaft von Transaktionsdaten ist, dass diese im Idealfall konsistenzerhaltend sind. Dies wird insbesondere in der angeführten Definition von Corsten und Gössinger (2008) deutlich, der den Güterwechsel als zentralen Transaktionsaspekt benennt.

Aus Sicht der Datenverwaltung ist eine Transaktion ein Datensatz, der ein oder mehrere Datenfelder beinhaltet, die verschiedene Formate aufweisen können. Transaktionsdaten umfassen häufig eine Transaktions-ID mit numerischem Datentyp und Adressdaten mit alphanumerischem Datentyp. Diese Daten sind folglich inhomogen bezüglich der vorliegenden Datentypen. Abbildung 2.4 zeigt, dass die Transaktionsdaten in Aktivitäts- und Kontrolldaten unterteilt werden können. Transaktionsaktivitätsdaten beschreiben die eigentlichen Transaktionen durch ihre zugehörigen Eigenschaften wie Zeitstempel, Art der Güter oder Liefermenge. Transaktionskontrolldaten wiederum beinhalten Protokolldaten (Auditdaten) der Transaktionsprozesse, wie Datenbanklogeinträge oder separate Logdateien. Da in der Praxis die Logdateien eher technischer Natur sind und wenig Analysepotential für das SCM bieten, beschränkt sich der hier verwendete Begriff Transaktionsdaten nur die Transaktionsaktivitätsdaten. Eng mit den Transaktionsdaten sind die Transaktionsstrukturdaten verbunden (vgl. Abbildung 2.4). Diese beschreiben die BOs sowie ihre Relationen, die an Transaktionen beteiligt sind (z. B. Lieferanten, Kunden, Produkte).

Otto und Hüner (2009) sowie Schemm (2012) kategorisieren Transaktionsdaten als dynamisch. Hierbei wird insbesondere hervorgehoben, dass Transaktionsdaten Veränderungen zu beliebigen Zeitpunkten erfahren können. Dies unterscheidet sie von anderen dynamischen Daten wie Streaming-Daten, bei denen Updates kontinuierlich vonstatten gehen. Die Kategorisierung der Stammdaten als dynamisch ist aus aggregierter Sicht als zutreffend einzustufen, denn eine Transaktion wird beispielsweise in einem Monitoring-System mitunter nur in ihrer aktuellen Ausprägung angezeigt. Es muss jedoch berücksichtigt werden, dass auf Datenbankebene oftmals keine Datensätze überschrieben werden, sondern bei Veränderungen ein neuer Datensatz angelegt und der alte historisiert wird. An diesem Beispiel wird deutlich, das die Kategorisierung eng an eine betriebswirtschaftliche Sichtweise geknüpft ist und nicht losgelöst betrachtet werden darf (vgl. Abbildung 2.5).

Eine Besonderheit der SC ergibt sich aus den Wechselwirkungen der einzelnen Unternehmen und Unternehmensteile. So umfasst das zu entdeckende Wissen auf keinen Fall nur einzelne Unternehmensteile und deren Transaktionen oder lässt sich hieraus erschließen. Vielmehr müssen wesentliche, zusammenhängende Abschnitte der SC betrachtet werden, um Eigenschaften, die sich aus komplexen dynamischen

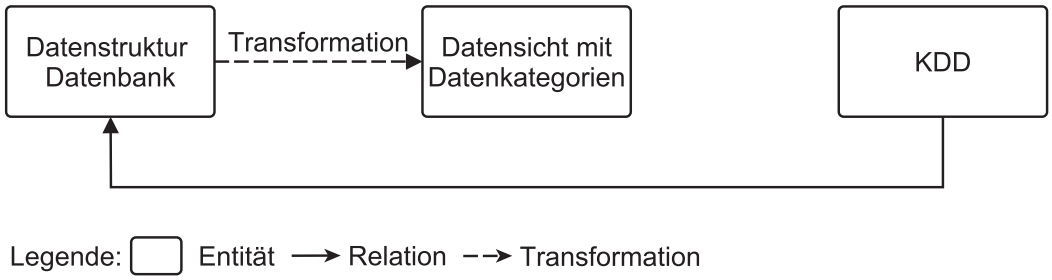



**Abbildung 2.5: KDD und benötigte Datengrundlage**

Prozessen im Gesamtsystem ergeben, zu erkennen (vgl. hierzu auch „supply chain complexity triangle“ in Wilding 1998). Durch die emergenten Eigenschaften des Systems besteht die Forderung nach möglichst großen zusammenhängenden Datenmengen. Große zusammenhängende Datenmengen ermöglichen eine Suche nach Effekten, die zwischen einzelnen, zeitlich entfernt liegenden Datensätzen auftreten können.

#### 2.2.2.4 Datenqualität

Neben dem Volumen und der Komplexität der vorhandenen Daten (vgl. Abschnitt 2.2.2) ist auch die Datenqualität ein Kennzeichen der vorhandenen SC-Datenbestände. Unzureichende Datenqualität birgt ein erhebliches Risiko für operative Entscheidungen der SC und kann zu einer ineffizienten SC führen (Christopher und Lee 2004). Hier muss zwischen der Qualität der Daten und der Qualität der Datenstruktur unterschieden werden (Küppers 1999). Datenqualität zielt hierbei auf die eigentliche Qualität auf Ebene einzelner Datensätze ab – beispielsweise fehlende oder fehlerhafte Attribute (Wang und Strong 1996). Hingegen ist die Qualität von Datenstrukturen eher von mangelnder interner und externer Interoperabilität gekennzeichnet. Dies wird beispielsweise an unternehmensweit nicht einheitlichen Begriffsdefinitionen, inkonsistenten Relationen und sich überschneidenden Strukturen deutlich (Küppers 1999). Es gibt unterschiedliche Definitionsansätze zum Thema der Datenqualität, in Abhängigkeit des Umfeldes. Sowohl Six Sigma (Töpfer 2009; Weigel 2011) wie auch die Deutsche Gesellschaft für Informations- und Datenqualität, die sich auf die Forschungsansätze von Wang und Strong (1996) beziehen, haben in zahlreichen Beiträgen Vorschläge unterbreitet. Diese Arbeit stützt sich auf die generische Definition von Würthele (2003), da sie die Unterscheidung in Qualität der Daten und Qualität der Struktur gestattet. Diese Unterscheidung ist bei stark strukturellen Daten wie den Transaktionsdaten zwingend notwendig, da insbesondere Fehler in den zugrundeliegenden Strukturen gesondert betrachtet werden müssen.

> **Definition 2.6 Datenqualität:** Datenqualität ist ein „mehrdimensionales Maß für die Eignung von Daten, den an ihre Erfassung/Generierung gebundenen Zweck zu erfüllen. Diese Eignung kann sich über die Zeit ändern, wenn sich die Bedürfnisse ändern“ (Würthele 2003, S. 21).

Hierbei ist des Weiteren zu beachten, dass Methoden zur Messung der Datenqualität oftmals einzeln betrachtet nur unzureichend untersucht sind und zudem stark von der Problemstellung und deren Methodik abhängig sind (Alkharboush und Li 2010).

## 2.2.3 Wissen in Supply Chains

In diesem Abschnitt wird der Begriff Wissen in den Kontext der SC gestellt. Dazu wird das SCM als stellvertretender Adressat des Wissens in SCs bestimmt. Im Anschluss werden aufzufindende Formen des Wissens in SC-Daten anhand häufiger Fragestellungen des SCM diskutiert. Es wird jedoch darauf hingewiesen, dass es im Kontext der Fragestellung keine spezifische Abgrenzung zwischen den Begrifflichkeiten der SC und der Logistik in der Literatur gibt (Cooper et al. 1997).

### 2.2.3.1 Wissen im Supply Chain Management

Die Ressource Wissen ist von ausgezeichneter Bedeutung im Kontext der SC und des SCM (Hult et al. 2006). Wissen muss im Kontext des Adressaten betrachtet werden, denn dieser bestimmt letztendlich die relevanten Aufgabenstellungen der Wissensentdeckung, die es zu beantworten gilt. Stellvertretend für spezifische Aufgaben, die in der Methodenanwendung an Bedeutung gewinnen, soll hier das SCM mit seinen Aufgabenstellungen und Zielen vorgestellt werden. Der Begriff SCM ist eng mit dem Begriff der SC verbunden. Oftmals werden beide Begriffe zusammen behandelt. So verweist beispielsweise ten Hompel in seinem Logistiklexikon unter SC direkt auf SCM (ten Hompel und Heidenblut 2011). Auch Ayers führt in seiner Enzyklopädie des SCM ein „definition problem“ hinsichtlich der SC-Terminologie auf (Ayers 2012).

Die Grundlagen des SCM entwickelten sich in den 1980er Jahren innerhalb der USA. In Deutschland etablierte sich das SCM in den 1990er Jahren. SCM bezeichnet die Planung und Steuerung der Objektflüsse von der Quelle zur Senke (Lambert 2005). Ziel des SCM ist hierbei das Erreichen eines globalen Optimums durch das abgestimmte Verhalten der einzelnen Beteiligten (Günther und Tempelmeier 2011). Werner stellt fest, dass das primäre Ziel im SCM die Integration von Unternehmensaktivitäten ist. Der Ansatz erstreckt sich dabei auf die Optimierung der Effektivität und die Effizienz der Unternehmensaktivitäten (Werner 2010). Wie schon bei der SC selbst gibt es verschiedene Definitionsansätze für das SCM, von denen sich bis jetzt keiner endgültig durchsetzen konnte (Wellbrock

2015). Eine Übersicht über die Begriffsentwicklung ist in Mentzer et al. (2001) und Gibson et al. (2005) dargestellt und soll hier nicht näher fokussiert werden. Nach Corsten und Gössinger finden sich jedoch in den unterschiedlichen Definitionen Gemeinsamkeiten, auf die eine Formulierung aufbauen kann (Corsten und Gössinger 2008). Diese Gemeinsamkeiten sind die Kundenorientierung, die optimale Gestaltung der Geschäftsprozesse, die kooperative Zusammenarbeit der Teilnehmer sowie informationstechnische Interoperabilität der SC-Teilnehmer. Dies wird beispielsweise in der Definition von Kuhn aufgegriffen, der Verbesserung der Kundenorientierung, Synchronisation der Versorgung mit dem Bedarf, Flexibilisierung, bedarfsgerechte Produktion und Abbau der Bestände entlang der Wertschöpfungskette als Ziel des SCM definiert (Hellingrath und Kuhn 2002).

Das in dieser Arbeit zugrunde gelegte Verständnis des SCM basiert auf diesen Gemeinsamkeiten in der Literatur und folgt dem vielzitierten Lexikon APICS.

> **Definition 2.7 Supply Chain Management:** „Design, maintenance, and operation of supply chain processes, including those that make up extended product features, for satisfaction of end-user needs“ (Ayers 2012, S. 10).

#### 2.2.3.2 Formen von Wissen

Häufig sind Beiträge zum SCM inhaltlich auf bekannte logistische Fragestellungen aus dem Unternehmensalltag reduziert. Cooper et al. (1997) führen jedoch an, dass SCM nicht mit Logistik gleichgesetzt werden kann, da das SCM als umfassender als die Logistik anzusehen sei. Eine umfassende Darstellung von Einbindungsmöglichkeiten der KDD-Techniken in die verschiedenen Analyseaufgaben der SC ist in Wannenwetsch und Nicolai (2004) gegeben. Hier liegt der Fokus auf den technischen Einsatzmöglichkeiten in der SC, z. B. im Bereich der Data-Warehouse-Analyse durch Data Mining. Da in der Literatur oftmals nur fachliche Fallbeispiele aufgeführt sind und es keine spezifische Untersuchung von Fragestellungen des SCM im Kontext der Wissensentdeckung gibt, wurden diese im Rahmen von Vorarbeiten am Fachgebiet ITPL untersucht (Gürez 2015). Die exemplarischen Fragestellungen aus Tabelle 2.2 wurden an den Prozesskategorien der obersten Ebene des Supply-Chain-Operations-Reference-Modells (SCOR) ausgerichtet. Die Fragestellungen zeigen, welche Art von fachinhaltlichem Wissen im Kontext der SC entdeckt werden kann.

Bei Fragestellungen des SCM ist häufig die Frage nach Wirkzusammenhängen leitend. Wirkzusammenhänge sind im Allgemeinen von großer Wichtigkeit und wurden bereits als wesentliche Aufgaben sogenannter logistischer Assistenzsysteme adressiert (Kuhn et al. 2008). Die unbekannten und von Assistenzsystemen nicht einfach abzubildenden Wirkzusammenhänge, die sich aus den Daten des Materialflusses (Teil der physikalischen Flüsse) sowie den damit verbundenen Daten der Informationsflüsse in SCs ergeben, sind insbesondere für den Bereich des SCM von

**Tabelle 2.2: Beispielhafte Fragestellungen des SCM nach Gürez (2015)**

| Nummer | SCOR-Kategorie | Fragestellung |
|---|---|---|
| 1 | Planen | Wie hoch werden zukünftige Kundenbedarfe sein? |
| 2 | Planen | Wo entsteht der Bedarf? Bei welchen Kunden bzw. in welchen Regionen liegt dieser Bedarf vor? |
| 3 | Planen | Welche Distributionszentren sollen wo eingerichtet werden und welche Kunden beliefern? |
| 4 | Beschaffen | Bei welchem Lieferanten soll bestellt werden? |
| 5 | Beschaffen | Welche Produkte sollen in welchen Mengen beschafft werden, um eine optimale Lagerverwaltung und Materialflüsse zu gewährleisten? |
| 6 | Herstellen | Wie hoch ist die Qualität der produzierten Teile? |
| 7 | Herstellen | Wie oft müssen Wartungen an Produktionsanlagen durchgeführt um zukünftige Maschinenausfälle zu vermeiden? |
| 8 | Herstellen | Wie lassen sich die Rüstzeiten optimieren? |
| 9 | Liefern | Welches Transportmittel oder welcher Transportweg ist wann am effizientesten? |
| 10 | Liefern | Wie kann die Anzahl der verspäteten Aufträge minimiert und eine fristgemäße Lieferung sichergestellt werden? |
| 11 | Rückliefern | Worauf lässt sich der Defekt bzw. die Beschwerde am rückgelieferten Produkt zurückführen? |

Relevanz (Harland 1996). Wirkzusammenhänge wurden bereits in Vorarbeiten im Kontext der SC untersucht; hierbei sei insbesondere auf Rabe und Scheidler (2014) verwiesen. Dieser Beitrag setzt den Begriff der Wirkzusammenhänge in den Kontext der Wissensentdeckung und zeigt ihren Gewinn für die Modellierung auf. Die Wirkzusammenhänge können im SCM über verschiedene Techniken erfasst und modelliert werden. Diese Techniken sind unter dem Begriff „causal analytic techniques“ in der Literatur zu finden, gehen jedoch in der Regel von bereits bekannten Zusammenhängen aus und fokussieren sich auf die Darstellung mittels Graphen (vgl. Tan et al. 2015).

Der Begriff der Wirkzusammenhänge wird in der Literatur sowie in der Industrie oftmals eingesetzt, um den Aspekt der Wechselwirkung zwischen verschiedenen Elementen zu betonen. Dies erklärt, warum der Begriff der Wirkzusammenhänge teilweise mit Wechselwirkung gleichgesetzt wird (Wagemann 1994). Die synonyme Verwendung der Begriffe gestattet den Rückschluss, dass Wirkzusammenhang etymologisch eine (nicht näher definierte) Kausalität impliziert. Daraus folgt, dass Wirkzusammenhänge im logistischen Kontext oftmals einen kausalen Zusammenhang enthalten. Hinzu kommt, dass der Begriff Wirkzusammenhänge auch explizit benannt wird, um abstrakte Ergebnisse zu veranschaulichen und diese in einen kausalen Kontext zu setzen (z. B. in Weskamp et al. 2014). Im Umfeld der SC findet der Begriff beispielsweise in Wenzel, Weiß et al. (2008) Anwendung. Soll eine Definition für Wirkzusammenhänge angegeben werden, so ist es naheliegend, auf Mathematik oder Statistik zurückzugreifen, denn dort spielt der Begriff Zusammenhang eine ausgezeichnete Rolle. Der Begriff Zusammenhang ist ein zentrales Element der Korrelation, die ein Maß für die Stärke des Zusammenhangs zwischen X und Y ist (Fahrmeir et al. 2010). Jedoch impliziert Korrelation eben nicht zwangsweise einen direkten Kausalzusammenhang zwischen einzelnen Elementen. Der Rückschluss, dass Korrelationen auch Wirkzusammenhänge sind, ist folglich unzulässig. Vielmehr muss die These geprüft werden, ob Korrelation eine mandatorische Eigenschaft von Wirkzusammenhängen ist. Treffender ist die Annäherung aus dem Bereich der Mathematik, die die Wirkzusammenhänge der Klasse der n-stelligen Relationen zuordnet. Hierbei beschreibt eine Relation ($\Re$) die Beziehung zwischen Elementen von verschiedenen Mengen ($A_1 ... A_n$). Eine n-stellige Relation ist in der Mathematik definiert als:

$$\Re \subseteq A_1 \times \ldots \times A_n \tag{2.1}$$

Unter dem Gesichtspunkt des algebraischen Felds kann die statistische Korrelation als Abbildung oder spezifischer Homomorphismus aufgefasst werden. Da auch diese Definition nicht den implizierten Kausalzusammenhang abdeckt, wird an dieser Stelle auf eine tiefere mathematische Diskussion verzichtet. Unter Berücksichtigung des Problems von Zusammenhang und Kausalität aus Sicht der Modellbildung können Wirkzusammenhänge zuerst nur als Hypothese angenommen werden, bis sie durch einen manuellen Prüfschritt bewiesen sind. Hingegen kann ein reiner Zusammenhang vollautomatisch über entsprechende Methoden ohne menschliche Interpretation aufgezeigt werden. Hieraus folgt, dass in erster Annäherung die allgemeingültigen Definitionen nicht passend für das logistische Verständnis von Wirkzusammenhängen sind. Da es in der Logistik keine verbindliche Definition von Wirkzusammenhängen gibt, wird folgende Definition aus Vorarbeiten am Fachgebiet ITPL eingeführt:

> **Definition 2.8 Wirkzusammenhänge:** Wirkzusammenhänge beschreiben „innere Beziehungen zwischen Entitäten, die implizieren, dass Veränderungen“

von ein oder mehreren beteiligten Entitäten „eine Wirkung auf ein oder mehrere anderen Entitäten des Wirkzusammenhangs ausüben“ (Köster 2015, S. 6).

## 2.3 Verfahren im Kontext der Wissensentdeckung

Wissen aus den Datenbeständen zu extrahieren und nachgeschalteten Prozessen zur Verfügung zu stellen, ist lohnenswert. Durch gezielte Analyse der Datenbestände kann nützliches Wissen für sämtliche Unternehmensbereiche gewonnen werden. Derartiges Wissen kann auf lange Sicht einen Vorsprung gegenüber Wettbewerbern bedeuten. Die Informatik fasst die Disziplin des Wissenserwerbs aus großen Datenmengen unter KDD, wörtlich Wissensentdeckung in Datenbanken, zusammen. KDD repräsentiert jedoch nicht eine vollständige Neuentwicklung, sondern liegt vielmehr in der Schnittmenge verschiedener Disziplinen wie Datenbanktechnologien, künstlicher Intelligenz oder Statistik (Küppers 1999). KDD ist ein nicht-trivialer (Fayyad et al. 1996b), iterativer und interaktiver Prozess (Wrobel et al. 1996). Insbesondere ist KDD ein Vorgehensmodell, das aus einzelnen Phasen besteht. Die Phasen werden teilweise auch als Schritte bzw. die Schritte, die die Phasen bilden, als Teilschritte deklariert. In dieser Arbeit wird der Begriff der Phase, die aus einzelnen Schritten besteht, zugrundelegt. In dem KDD-Vorgehensmodell ist die zentrale Phase – das Data Mining – so bedeutend, dass heutzutage die Begriffe KDD und Data Mining oftmals synonym verwendet werden und viele Autoren keine inhaltliche Unterscheidung treffen (Adriaans und Zantinge 1996; Säuberlich 2000). Nach Knobloch (2000) haben Data Mining und KDD identische Ziele, KDD jedoch eine größere Reichweite. Auch weitere Namensvarianten neben KDD und Data Mining wurden von Autoren eingeführt. Hierzu zählen knowledge extraction, database mining oder information harvesting. Küppers (1999) sieht den Grund für die Vielzahl von Namen darin, dass die angewandten Methoden in ihren Forschungsfeldern schon lange bekannt sind und an sich keinen eigenständigen Ansatz darstellen. In der Literatur konnte sich jedoch keine der zuletzt genannten Varianten durchsetzen. Diese Arbeit folgt in der Orientierung und Definition Fayyad et al. (1996b), die das Data Mining als Phase des übergeordneten KDD-Vorgehensmodells verstehen.

**Definition 2.9 KDD:** KDD ist der nicht-triviale Prozess der Identifizierung gültiger, neuartiger, potentiell nützlicher und letztlich verständlicher Muster in Daten (Fayyad et al. 1996b).

**Definition 2.10 Data Mining:** Data Mining ist eine essentielle Phase im KDD-Vorgehensmodell, die aus spezifischen Algorithmen besteht, welche Muster (oder Modelle) aus den Daten extrahieren (Fayyad et al. 1996b).

KDD ist eng verwandt mit dem maschinellen Lernen, das sich mit dem Lernen von Gesetzmäßigkeiten aus Beispielen beschäftigt. Insbesondere kommen eine Vielzahl von Verfahren und Algorithmen sowohl im KDD als auch im maschinellen Lernen zum Einsatz. So wird eine Support Vector Machine sowohl im maschinellen Lernen als auch im KDD (beispielsweise für Prognosen) eingesetzt. Für eine tiefere Beschäftigung mit dem Thema maschinelles Lernen sei auf das Standardwerk von Bishop (2006) verwiesen. Des Weiteren gibt es eine große Überschneidung zu den Techniken der Statistik, insbesondere der explorativen Datenanalyse. Hierbei wird festgestellt, dass die Verfahren eine große Ähnlichkeit aufweisen, die betrachtete Datenmenge bei den KDD-Techniken jedoch erheblich größer ist (Fahrmeir et al. 2010; Rönz und Strohe 1994). Auch Wrobel (1998) konstatiert als wesentlichen Unterschied zwischen KDD und der Statistik sowie dem maschinellen Lernen die Skalierbarkeit der eingesetzten Methoden und die damit verbundene Notwendigkeit, mit großen Datenmengen umzugehen. Ein andere Unterscheidung kann über den Begriff der Hypothese gefunden werden. Hierbei stellen viele KDD-Verfahren hypothesenfreie Verfahren (Bottom-Up-Verfahren) dar, die als Ziel haben, neue Erkenntnisse zu gewinnen. Klassische Verfahren der Statistik, wie die Varianzanalyse, sind überwiegend hypothesengetriebene Verfahren (Top-Down-Verfahren) (Knobloch 2000). Ein weiteres Forschungsgebiet, das im Kontext der SC-Wissensentdeckung von Bedeutung ist, ist das des Process Minings. Ebenso wie die KDD-Techniken wird Process Mining bei großen Datenbeständen eingesetzt, die nicht mehr manuell bearbeitet werden können. Der Hauptfokus beim Process Mining liegt jedoch auf der Entdeckung von Prozesswissen und der anschließenden Modellierung der Prozesse. Die eingesetzten Verfahren unterscheiden sich von den allgemeinen KDD-Techniken, da ihre Datenbasis in Form von Prozessablaufprotokollen sehr spezifisch und daher mit speziellen Anforderungen verknüpft ist (van der Aalst 2011). Es existieren Ansätze, das Process Mining im Kontext der SC anzuwenden, um beispielsweise die SC-Analyse zu unterstützen (Gerke et al. 2009). Im Wesentlichen unterscheiden sich die Techniken jedoch in ihrem Fokus und den Mining-Zielen: Trotz Überschneidungen der beiden Gebiete fokussieren sich KDD-Techniken auf die Mustersuche und nicht auf Prozessrepräsentationen (van der Aalst und Weijters 2004).

In den folgenden Abschnitten wird der Gedanke eines Vorgehensmodells im KDD aufgegriffen und die elementaren Phasen erläutert, die Teil der meisten KDD-Vorgehensmodelle sind. Basierend auf den Analyseerkenntnissen der einzelnen Vorgehensphasen wird im Anschluss aufgezeigt, welche Kombinationsmöglichkeiten für Vorgehensmodelle zur Wissensentdeckung und ereignisdiskreter Simulation zu identifizieren sind.

### 2.3.1 Vorgehensmodelle zur Wissensentdeckung

Es gibt verschiedene Vorgehensmodelle im Bereich des KDD, die trotz unterschiedlicher Ausrichtungen große Überschneidungen in den Kernelementen aufweisen.

Pioniere auf diesem Gebiet sind Frawley et al. (1992), die den Begriff Data Mining als die Wissensentdeckung von impliziten, bisher nicht bekannten und potentiell nützlichem Wissen aus Daten definieren. Fayyad et al. (1996b) wiederum fassten diesen Vorgang als Prozess auf und beschrieben ihn vier Jahre später in ihrem Beitrag „From Data Mining to Knowledge Discovery in Databases“. Ein Vorgehensmodell, das in den letzten Jahren an Bedeutung gewonnen hat, ist das Crisp-DM Model, das von dem gleichnamigen Konsortium initialisiert wurde und permanent weiterentwickelt wird. Crisp-DM steht für Cross Industry Standard for Data Mining (Branchenübergreifender Standard für Data Mining) und unterteilt den Prozess in sechs Kernphasen, die iterativ durchlaufen werden (Gabriel et al. 2009). Hierbei ist der Fokus auf das Projektgeschäft gelegt, was beispielsweise in der ersten Phase, genannt „Business Understanding“, von Crisp-DM deutlich wird. Hier werden Projektziele und Projektpläne mit Fokus auf die Koordination festgelegt. Insbesondere zeigt dieses Modell kaum die konkreten Phasen und Schritte, um die Ergebnisse in die Praxis umzusetzen. Berry und Linoff (2000) haben ein Modell entwickelt, das als Fokus die Weiterverwendung des gewonnenen Wissens und seines betriebswirtschaftlichen Nutzens hat. Hierfür identifizieren sie vier Stufen, die unter dem Begriff „virtuous cycle of data mining“ genutzt werden. Knobloch (2003) entwickelte dieses Modell weiter. Es gibt einige Modelle, die für spezielle Anwendungen entwickelt worden sind, jedoch problemlos in andere Anwendungsfelder übertragen werden können. Exemplarisch ist das Modell „Knowledge Discovery in Industrial Databases“ zu erwähnen, das die Thematik von heterogenen Datenbestände aus der Industrie adressiert. In diesem Modell besteht die wesentliche Innovation aus dem Einführen von Meilensteinen mit dem Ziel, einen Überwachungsmechanismus für den Projektfortschritt zu implementieren. Die fachliche Bedeutung der einzelnen Phasen beruht hierbei auf dem Modell von Fayyad et al. (1996b) und wurde um einzelne projektbezogene Schritte ergänzt (z. B. „IT-Infrastruktur aufnehmen“ oder „IT-Prototyp erstellen“) (Lieber et al. 2013).

Ein spezifisches Vorgehensmodell, das ursprünglich aus dem Marketingumfeld kommt, ist das von Hippner und Wilde (2001). Der Geltungsbereich umfasst die Problemidentifikation bis hin zur Ergebnisumsetzung. Der KDD-Prozess wird hier sowohl als dynamisch als auch als iterativ beschrieben, gliedert sich in sieben Phasen und baut im Wesentlichen auf den Modellen von Fayyad et al. (1996b) und Fayyad und Uthurusamy (1994) auf. Tabelle 2.3 gibt die Phasen und zugehörigen Schritte des Modells von Hippner und Wilde wieder. An diesem Modell kann exemplarisch der Aufbau eines KDD-Vorgehensmodells nachvollzogen werden, da das Modell alle wesentliche Elemente der Wissensentdeckung beinhaltet. Bei der Anordnung der einzelnen Phasen im Modell von Hippner und Wilde handelt es sich um das Standardvorgehen, das in der Modellbeschreibung als Leitfaden vorgeschlagen wird.

**Tabelle 2.3: Modell nach Hippner und Wilde basierend auf Hippner und Wilde (2001)**

| Phase | Schritte | Kurzbeschreibung |
|---|---|---|
| 1. Aufgabendefinition | 1.1 Bestimmung der betriebswirtschaftlichen Problemstellung | Formulierung von Zielkriterien und Beschreibung der Gestaltungsalternativen |
| | 1.2 Ableitung analytischer Ziele für das Data Mining und Projektplanung | Festlegung von Datenanalyseaufgaben und Erfolgskriterien für die Datenanalyse |
| | 1.3 Projektplanung | Entwicklung des Projektplans |
| 2. Auswahl der relevanten Datenbestände | 2.1 Katalogisierung und Bewertung der verfügbaren Datenquellen | Selektion der relevanten Datenbestände und anschließende Prüfung |
| | 2.2 Bestimmung der geeigneten Datenbestände | Basierend auf der Zieldefinition erfolgt eine Verwendung von unternehmensinternen oder -externen Daten |
| 3. Datenaufbereitung | 3.1 Datentransformation in ein geeignetes Datenformat zur Datenanalyse | Überführung der selektierten Datenbestände in ein Standarddatenformat |
| | 3.2 Explorative Datenanalyse | Erfassung von Anhaltspunkten über den Aussagegehalt des Datenmaterials |
| | 3.3 Datenanreicherung | Einbeziehung von Daten aus höheren Aggregationsebenen zur Beschreibung von Informationsobjekten |
| | 3.4 Datenreduktion | Reduktion des vorliegenden Datenbestands |
| | 3.5 Behandlung fehlender Merkmalswerte | Beseitigung von fehlenden Merkmalswerten der Daten durch geeignete Techniken |
| | 3.6 Behandlung von fehlerhaften Merkmalswerten und Ausreißern | Erkennung und Bereinigung fehlerhafter Merkmale und Ausreißer durch geeignete Techniken |

**Tabelle 2.3: Modell nach Hippner und Wilde basierend auf Hippner und Wilde (2001) (Fortsetzung)**

| Phase | Schritte | Kurzbeschreibung |
|---|---|---|
| | 3.7 Kodierung der Merkmale | Spezifische Transformation der Daten durch Techniken wie Skalentransformation oder Normierung |
| 4. Auswahl von Data-Mining-Methoden | 4.1 Bestimmung der Auswahlkriterien für Data-Mining-Methoden | Kriterien der jeweiligen Problemstellung entsprechend aufstellen und gewichten |
| | 4.2 Bewertung der Data-Mining-Methoden | Bewertung der Methode nach vorgegebenen Kriterien wie Robustheit oder Interpretierbarkeit |
| | 4.3 Bestimmung geeigneter Data-Mining-Methoden | Auswertung der unter 4.2. aufgestellten Kriterien |
| | 4.4 Auswahl eines Data-Mining-Werkzeugs | In Rückkoppelung mit 4.3 muss ein geeignetes Data-Mining-Werkzeug bestimmt werden |
| 5. Anwendung der Data-Mining-Methoden | 5.1 Entwicklung von Data-Mining-Modellen | Festlegung von Methoden- und Modellparameter auf Basis von a priori Wissen |
| | 5.2 Test von Data-Mining-Modellen | Anwendung von methodenspezifischen und -unabhängigen Tests |
| | 5.3 Kombination von Data-Mining-Methoden (in Hybridsystemen) | Kombination zweier oder mehrerer Data-Mining-Methoden zum wechselseitigen Ausgleich der Defizite der jeweiligen Methoden |
| 6. Interpretation und Evaluation der Data-Mining-Ergebnisse | 6.1 Extraktion handlungsrelevanter Data-Mining-Ergebnisse | Unter Berücksichtigung der Handlungsrelevanz sind interessante Ergebnisse zu extrahieren |
| | 6.2 Betriebswirtschaftliche Bewertung der Data-Mining-Ergebnisse | Planung und Bewertung betriebswirtschaftlicher Maßnahmen |

**Tabelle 2.3: Modell nach Hippner und Wilde basierend auf Hippner und Wilde (2001) (Fortsetzung)**

| Phase | Schritte | Kurzbeschreibung |
|---|---|---|
| | 6.3 Bewertung des Data-Mining-Prozesses | Qualitätskontrolle der Ergebnisse hinsichtlich betriebswirtschaftlicher Ziele |
| 7. Anwendung der Data-Mining-Ergebnisse | 7.1 Anpassung der operativen Geschäftsprozesse im Marketing | Vorbereitung und eventuelle Änderung der operativen Prozesse |
| | 7.2 Einbindung der Data-Mining-Modelle in die operativen Geschäftsprozesse im Marketing | Anhand der gewonnenen Ergebnisse und Erkenntnisse sind Maßnahmen zu entwickeln |
| | 7.3 Empfehlungen für Führungsentscheidungen im Marketing | Ableiten von Handlungsempfehlungen aus den Erkenntnissen der Wissensentdeckung |
| | 7.4 Aufgabendefinition für weitere Data-Mining-Prozesse | Ableiten von Erkenntnissen, die als Grundlage für weitere Data-Mining-Prozesse genutzt werden können |

Bei der zeitlichen Gewichtung der einzelnen Phasen und Schritte zur Wissensentdeckung finden sich vergleichbare Größenordnungen in der Beschreibung verschiedener KDD-Vorgehensmodelle (vgl. beispielsweise die Beschreibungen von Cabena et al. 1998 oder Küppers 1999). Abbildung 2.6 zeigt exemplarisch die Zeitangaben des Modells von Hippner und Wilde. Wurden in der Originalliteratur mehrere Phasen zeitlich zusammengefasst, wurde dies in der Darstellung übernommen.

In der Literatur gibt es nur wenige vergleichende Untersuchungen von Vorgehensmodellen. Hierbei verfolgt der Modellvergleich primär das Ziel, die Modelle zu vereinigen, um ein generisches, branchenübergreifendes Modell zu erstellen. Da es zudem eine Vielzahl von Modellen gibt, die sich teilweise nur marginal unterscheiden, beschränkt sich die Literatur überwiegend auf eine Untersuchung der bekannten Modelle (vgl. Kurgan und Musilek 2006). Es gibt nur vereinzelt Beiträge zur Klassifizierung oder Strukturierung von Vorgehensmodellen. Exemplarisch kann der Beitrag von Mariscal et al. (2010) angeführt werden, der die Modelle auf Fayyad et al. oder Crisp-DM zurückführt und diese gegeneinander abgleicht. Die wenigen existierenden Klassifizierungen von Vorgehensmodellen zur Wissensentdeckung sind zudem branchenunabhängig und unterstützen in der Folge nicht die Entwicklung eines spezifischen Vorgehensmodells zur Wissensentdeckung im SC-Umfeld.

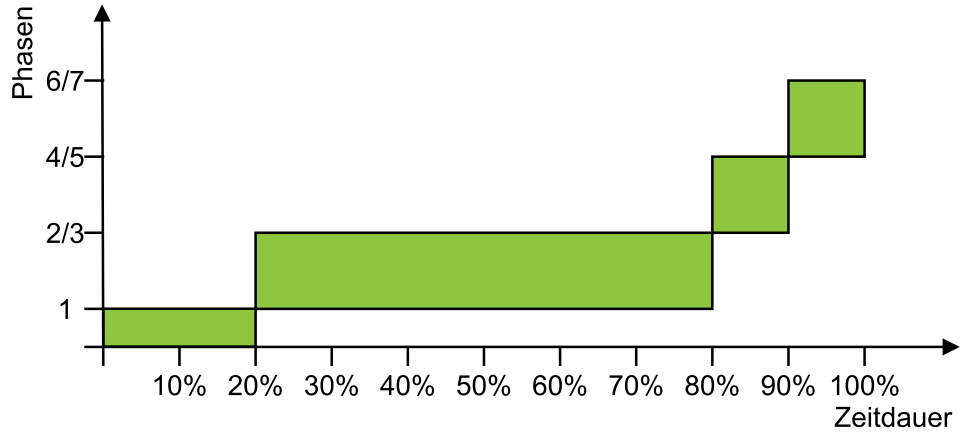



**Abbildung 2.6: Zeitangaben für Phasendauer im Modell von Hippner und Wilde (2001)**

Obwohl sich viele KDD-Vorgehensmodelle in der Anzahl der Phasen wie auch im Fokus deutlich unterscheiden, können Überschneidungen festgestellt werden. So beinhaltet jedes Vorgehensmodell eine Phase der Vorbereitung, die sich sowohl auf die Zieldefinition wie auch die Datenvorbereitung bezieht. Im Mittelteil findet sich immer die Methodenanwendung – das eigentliche Data Mining. Die Abschlussphase der Modelle wird maßgeblich von der Auswertung und Interpretation der gewonnenen Ergebnisse gebildet. Zusätzlich zeichnen sich alle Vorgehensmodelle durch eine sukzessive, iterative Anordnung der einzelnen Schritte aus. Darüber hinaus konnte bei allen Modellen festgestellt werden, dass eine manuelle Steuerung über Aktivitäten als Grundvoraussetzung angenommen wird. Diese gemeinsamen Elemente der Vorgehensmodelle werden in dem Zitat von Wrobel (1998, S. 3) deutlich, der sagt: „Schließlich bleibt als zentrales Merkmal von KDD die Betonung des interaktiven und iterativen Prozesses, bei dem Mensch und Data-Mining-Verfahren gemeinsam verständliches und interessantes Wissen entdecken.“ Die weiteren Gemeinsamkeiten der Modelle können in Tabelle 2.4, die bekannte Vorgehensmodelle des KDD mit ihren Phasen auflistet, nachvollzogen werden. Die Spalte mit der Bezeichnung „generisches Modell“, steht hierbei für ein anwendungsneutrales Modell. Diese neutralen Modelle sind kaum in der vorherrschenden Literatur zu finden und zumeist so einfach aufgebaut, dass sie nur grobe Grundideen vermitteln können. Zudem ist ein praktischer Einsatz der neutralen Modelle in keinem bekannten Anwendungsfeld publiziert, sodass diese für die nachfolgenden Betrachtungen entfallen.

Tabelle 2.4 kann um weitere Vorgehensmodelle ergänzt werden. Tabelle 2.6 gestattet eine Übersicht der gängigen Vorgehensmodelle sowie ihrer Phasen und beruht in Teilen auf der Vorarbeit aus dem Fachgebiet ITPL von Beckmann (2015). Hierbei wurde als wesentliches Selektionskriterium für die Modelle eine gesicherte Publikationslage, ein rudimentärer Bezug zum Bereich Produktion und Logistik sowie eine hinreichende fachliche Divergenz zu bereits etablierten Modellen angesetzt. Hierbei stellt sich heraus, dass einige oft zitierte Modelle überwiegend in Sekun-

**Tabelle 2.4: Wichtige Vorgehensmodelle des KDD nach Kurgan und Musilek (2006)**

| **Modell** | Fayyad et al. (vgl. Fayyad et al. 1996b) | Cabena et al. (vgl. Cabena et al. 1998) | Anand & Buchner (vgl. Anand et al. 1998) | Crisp-DM (vgl. Gabriel et al. 2009) | Cios et al. (vgl. Cios 2001) | generisches Modell |
|---|---|---|---|---|---|---|
| **Jahr** | 1996 | 1998 | 1998 | 2000 | 2000 | |
| **Einsatz** | Wissenschaft | Praxis | Wissenschaft | Praxis | Wissenschaft | |
| **Anzahl an Phasen** | 9 | 5 | 8 | 6 | 6 | |
| **Vorgehen** | 1. Developing and Understanding of the Application Domain | 1. Business Objectives Determination | 1. Human Resource Identification | 1. Business Understanding | 1. Understanding the Problem Domain | 1. Application Domain |
| | 2. Creating a Target Data Set | 2. Data Preparation | 2. Problem Specification | 2. Data Understanding | 2. Understanding the Data | 2. Data Understanding |
| | 3. Data Cleaning and Preprocessing | 3. Data Mining (DM) | 3. Data Prospecting | 3. Data Preparation | 3. Preparation of the Data | 3. Data Preparation and Identification of DM Technology |
| | 4. Data Reduction and Projection | 4. Domain Knowledge Elicitation | 4. Domain Knowledge Elicitation | 4. Modeling | 4. DM | 4. DM |

**Tabelle 2.4: Wichtige Vorgehensmodelle des KDD nach Kurgan und Musilek (2006) (Fortsetzung)**

| **Modell** | Fayyad et al. (vgl. Fayyad et al. 1996b) | Cabena et al. (vgl. Cabena et al. 1998) | Anand & Buchner (vgl. Anand et al. 1998) | Crisp-DM (vgl. Gabriel et al. 2009) | Cios et al. (vgl. Cios 2001) | generisches Modell |
|---|---|---|---|---|---|---|
| | 5. Choosing the DM Task | 5. Assimilation of Knowledge | 5. Methodology Identification | 5. Evaluation | 5. Evaluation of the Discovered Knowledge | 5. Evaluation |
| | 6. Choosing the DM Algorithm | | 6. Data Preprocessing | 6. Deployment | 6. Using the Discovered Knowledge | 6. Knowledge Consolidation and Deployment |
| | 7. DM | | 7. Pattern Discovery | | | |
| | 8. Interpretating Mined Patterns | | 8. Knowledge Postprocessing | | | |
| | 9. Consolidating Discovered Knowledge | | | | | |

därliteratur behandelt werden und thematisch nur noch schwer den Originalquellen zuzuordnen sind. Die Modelle nach John (1997) sowie die in Kurgan und Musilek (2006) beschriebenen Modelle von Edelstein und Haglin wurden aus Gründen der Vollständigkeit aufgeführt, entfallen aber in den weiteren Betrachtungen.

Die Umfragen aus den Jahren 2007 und 2014 verdeutlichen den Einsatz der verschiedenen Vorgehensmodelle in der Wirtschaft (Tabelle 2.5).

**Tabelle 2.5: Einsatz von Vorgehensmodellen in der Praxis nach Piatetsky-Shapiro (2014)**

| Modell | 2007 | 2014 |
|---|---|---|
| Crisp-DM | 42,0 % | 43,0 % |
| Eigene Modelle | 19,0 % | 27,5 % |
| SEMMA | 13,0 % | 8,5 % |
| Allgemeine Modelle | 8,0 % | 4,0 % |
| Modell nach Fayyad et al. | 7,5 % | 7,3 % |
| Organisationsspezifische Modelle | 3,5 % | 5,3 % |
| Domänenspezifische Modelle | 2,0 % | 4,7 % |
| Keine Modelle im Einsatz | 0,0 % | 4,7 % |

### 2.3.2 Phasen in Vorgehensmodellen

Eine häufige Betrachtungsweise des KDD ist, dieses als Vorgehensmodell zu verstehen, in dem eine Gliederung in mehrere Phasen erfolgt. Die Phasen werden als Prozesse verstanden, die sich aus logisch aufeinanderfolgenden Schritten zusammensetzen. Da das bereits diskutierte Modell von Hippner und Wilde (vgl. Tabelle 2.3) sowie eine Vielzahl anderer KDD-Vorgehensmodelle und ein Großteil der aktuell vorherrschenden Literatur das Basismodell von Fayyad et al. (1996b) als Ausgangspunkt festlegen, erfolgt die explizite Erläuterung der einzelnen Phasen anhand dieses Modells. Kennzeichnend für dieses Modell ist neben der interaktiven Struktur der manuelle Anteil im Vorgehensmodell. Dies ist eine gute Ausgangslage für logistische Sachverhalte, die von Kontextwissen geprägt sind und daher die Automatisierung innerhalb der Vorgehensmodelle erschweren. Zudem gestatten die neun Phasen ein detailliertes Verständnis des Vorgehensmodells. Die einzelnen Phasen lauten wie folgt:

1. Identifikation des domänenspezifischen (Vor-)Wissens und Zieldefinition der Wissensfindung
2. Datenauswahl

**Tabelle 2.6: Weitere Vorgehensmodelle des KDD**

| Modell | Jahr | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| 5 A's (Talia und Trunfio 2013) | 2003 | Assess | Access | Analyze | Act | Automate | - | - | - |
| Adriaans und Zantinge (Adriaans und Zantinge 1996) | 1996 | Data Selection | Cleaning Enrichment | Coding | DM | Reporting | - | - | - |
| Berry und Linoff (Berry und Linoff 2000) | 1997 | Identifying the Problem | Analysing the Problem | Taking Action | Measuring the Outcome | - | - | - | - |
| Brachmann und Anand (Fayyad und Uthurusamy 1994) | 1996 | Task Discovery | Data Discovery | Data Cleaning | Model Development | Data Analysis | Output Generation | - | - |

**Tabelle 2.6: Weitere Vorgehensmodelle des KDD (Fortsetzung)**

| Modell | Jahr | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| Cooley et al. (Cooley et al. 1999) | 1999 | Preprocessing | Mining Algorithms | Pattern Analysis | - | - | - | - | - |
| Edelstein (Kurgan und Musilek 2006) | 2001 | Identifying the Problem | Preparing the Data | Building the Model | Using the Model | Monitoring the Model | - | - | - |
| Haglin (Kurgan und Musilek 2006) | 2005 | Goal Identification | Target Data Creation | Data Preprocessing | Data Transformation | DM | Evaluation and Interpretation | Take Action steps | - |
| Hippner und Wilde (Hippner und Wilde 2001) | 2001 | Aufgabendefinition | Auswahl der Daten | Datenaufbereitung | Auswahl der DM Verfahren | Anwendung der DM Verfahren | Interpretation, Evaluation | Anwendung der Ergebnisse | - |
| John (John 1997) | 1997 | Define a Problem | Extract Data | Data Engineering | Algorithm Engineering | Run Mining Algorithm | Analyse Results | - | - |

**Tabelle 2.6: Weitere Vorgehensmodelle des KDD (Fortsetzung)**

| Modell | Jahr | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| KDD Roadmap (Roy 2001) | 2001 | Problem Specification | Resourcing | Data Cleansing | Preprocessing | Data Mining | Evaluation of Results | Interpretation of Results | Exploitation of Results |
| Petersohn (Petersohn 2005) | 2005 | Aufgabendefinition | Datenselektion | Datenaufbereitung | Datenanalyse | Evaluierung des Modells | Anwendung des Analysemodells | Ergebnisinterpretation | - |
| Reinartz und Wirth (Reinartz und Wirth 1995) | 1995 | Requirement Analysis | Knowledge Aquisition | Preprocessing | Pattern Extraction | Postprocessing | Deployment | - | - |
| Runkler (Runkler 2010) | 2010 | Vorbereitung | Vorverarbeitung | Mustererkennung | Nachbearbeitung | - | - | - | - |
| SEMMA (Talia und Trunfio 2013) | 1997 | Sample | Explore | Modify | Model | Assess | - | - | - |

**Tabelle 2.6: Weitere Vorgehensmodelle des KDD (Fortsetzung)**

| Modell | Jahr | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| Wrobel (Wrobel 1998) | 1998 | Anwendung verstehen | Beschaffung & Integration der Daten | Analyseverfahren auswählen | Analysedatensatz erzeugen | Festlegung der Verfahrensparameter | Ergebnisbewertung & -säuberung | Nutzung der Ergebnisse | - |

3. Prepocessing
4. Transformation
5. Auswahl der Data-Mining-Verfahren in Bezug auf die Zieldefinition der Wissensfindung
6. Algorithmen- und Parameteranpassung für die Data-Mining-Verfahren
7. Data Mining – Mustersuche
8. Interpretation der Muster
9. Weiterverwendung der Muster

Abbildung 2.7 zeigt eine abstrahierte Darstellung des Vorgehensmodells von Fayyad et al. (1996b) und der zugehörigen Phasen. In dieser Darstellung repräsentieren Kanten die Phasen und die Ergebnisse der Kanten illustrieren wichtige Zwischenergebnisse im KDD. Hinter den einzelnen Phasen verbergen sich mitunter komplexe, vielschichtige Verfahrensschritte, die im folgenden kurz erläutert werden.

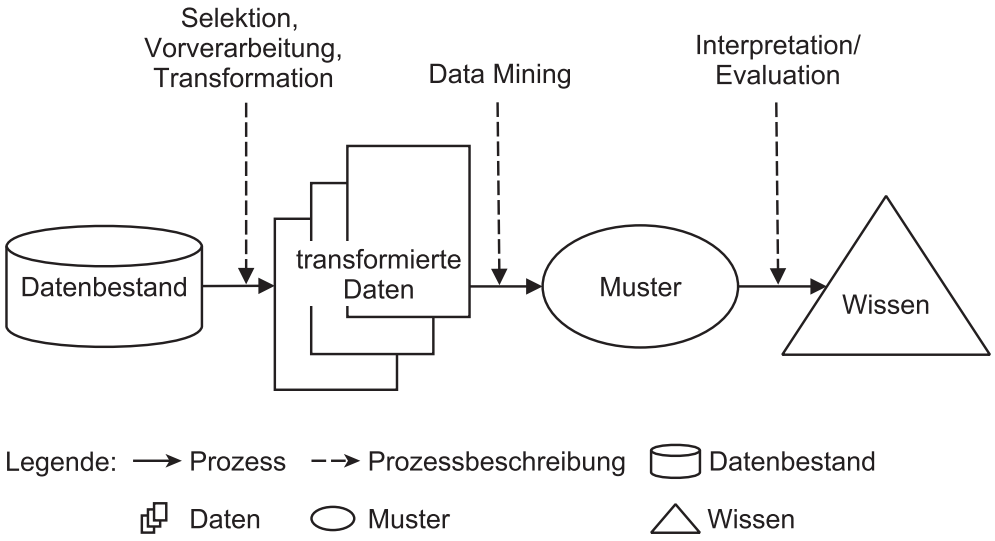



**Abbildung 2.7: KDD-Prozess**

**Identifikation des domänenspezifischen (Vor-)Wissens** Die erste Phase ist die Identifikation des domänenspezifischen (Vor-)Wissens und die Definition der Aufgabenstellung der Wissensfindung. Die klassischen Aufgaben des KDD sind Klassifikation, Segmentierung, Prognose und Abhängigkeitsanalyse (vgl. Alpar und Niedereichholz 2000 und Düsing 2010). Je nach Autor finden dabei beliebige Ergänzungen oder Abwandlungen der wesentlichen Aufgaben statt. So nennen einige Autoren beispielsweise noch die Abweichungsanalyse, bei der Objekte identifiziert werden, die sich in ihren Eigenschaften von den anderen Objekten in den Datenbeständen unterscheiden. Die Zuordnung zu den Kernaufgaben des KDD ist

jedoch umstritten und nicht zentral (Gabriel et al. 2009), weswegen sie in den nachfolgenden Überlegungen nicht berücksichtigt wird. Für eine Diskussion zum Thema Data-Mining-Aufgaben und Einordnungskriterien sei auf Küppers (1999) verwiesen, der sich eingehend mit verschiedenen Ordnungsansätzen in der Literatur beschäftigt. Bei der Klassifikation werden Klassen definiert, denen einzelne Objekte zugeordnet werden. Objekte, die gleiche Ausprägungen von Merkmalen besitzen, können zu einer gemeinsamen Klasse zusammengefasst werden. Eine weitere Aufgabe, die Nennung bei fast allen Autoren findet, ist die Segmentierung. Ziel der Segmentierung ist die Aufteilung der Daten in Gruppen, sodass sich Daten einer Gruppe möglichst ähnlich, Daten zweier verschiedener Gruppen jedoch möglichst unähnlich sind. Abzugrenzen hiervon ist das Aufgabenfeld der Prognosemethoden. Als Prognose wird die Vorhersage unbekannter Merkmalswerte auf Basis anderer Merkmale oder Vergangenheitswerten bezeichnet. Eine Methode, die auch über das KDD hinaus Anwendung findet, ist die Abhängigkeitsanalyse. Die Aufgabe der Abhängigkeitsanalyse ist die Entdeckung von Abhängigkeiten zwischen Merkmalen oder einzelnen Merkmalsausprägungen. In Tabelle 2.7 sind die Kernaufgaben des KDD den Aufgabenstellungen des SCM zugeordnet (vergleiche Tabelle 2.2 sowie Vorarbeiten in Gürez 2015). Hierbei ist zu beachten, dass unterschiedliche Kernaufgaben des KDD zur Beantwortung gleicher SCM-Fragestellung beitragen können. Des Weiteren zeigt die Tabelle A.2 im Anhang eine Zuordnung von spezifischen Aufgaben des SCM zu existierenden Lösungsmöglichkeiten in der KDD-Literatur.

**Tabelle 2.7: Zuordnung der SCM-Fragestellungen zu den KDD-Kernaufgaben**

| **Kernaufgabe** | **Beispielhafte Fragestellung** |
|---|---|
| Klassifikation | • Bei welchem Lieferanten soll bestellt werden?<br>• Welche Produkte sollen in welchen Mengen beschafft werden, um eine optimale Lagerverwaltung und Materialflüsse zu gewährleisten?<br>• Wie hoch ist die Qualität der produzierten Teile?<br>• Welches Transportmittel oder welcher Transportweg ist wann am effizientesten? |
| Segmentierung | • Wo entsteht der Bedarf? Bei welchen Kunden bzw. in welchen Regionen liegt dieser Bedarf vor?<br>• Welche Distributionszentren sollen wo eingerichtet werden und welche Kunden beliefern?<br>• Welche Produkte sollen in welchen Mengen beschafft werden, um eine optimale Lagerverwaltung und Materialflüsse zu gewährleisten?<br>• Wie lassen sich die Rüstzeiten optimieren? |

**Tabelle 2.7: Zuordnung der SCM-Fragestellungen zu den KDD-Kernaufgaben (Fortsetzung)**

| Kernaufgabe | Beispielhafte Fragestellung |
|---|---|
| | • Welches Transportmittel oder welcher Transportweg ist wann am effizientesten?<br>• Wie kann man die Anzahl der verspäteten Aufträge minimieren und eine fristgemäße Lieferung sicherstellen?<br>• Worauf lässt sich der Defekt bzw. die Beschwerde am rückgelieferten Produkt zurückführen? |
| Prognose | • Wie hoch werden zukünftige Kundenbedarfe sein?<br>• Wo entsteht der Bedarf? Bei welchen Kunden bzw. in welchen Regionen liegt dieser Bedarf vor?<br>• Wie oft müssen Wartungen an Produktionsanlagen durchgeführt werden, um zukünftige Maschinenausfälle zu vermeiden? |
| Abhängigkeitsanalyse | • Welche Produkte sollen in welchen Mengen beschafft werden, um eine optimale Lagerverwaltung und Materialflüsse zu gewährleisten?<br>• Wie lassen sich die Rüstzeiten optimieren?<br>• Worauf lässt sich der Defekt bzw. die Beschwerde am rückgelieferten Produkt zurückführen? |

In der ersten Phase wird neben der beschriebenen Aufgabenspezifikation das domänenspezifischen (Vor-)Wissen spezifiziert. Zusätzlich wird festgelegt, auf welcher Datenbasis die Wissenssuche stattfinden soll. Insbesondere ist hier zu prüfen, ob Daten bereits vorliegen, beispielsweise in einer Datenbank oder erst für das KDD beschafft werden müssen.

**Datenauswahl** Die zweite Phase ist eine spezifische Datenauswahl aus der Gesamtmenge der zuvor identifizierten Datenbasis. In komplexen SCs kann die Datenauswahl durch eigene Vorgehensmodelle, wie beispielsweise das prozessorientierte Vorgehensmodell zur Informationsgewinnung im Kontext logistischer Netze, unterstützt werden (Jodin et al. 2009). Das Ergebnis einer spezifischen Datenauswahl wird als Subset bezeichnet und stellt die Datengrundlage für die nachfolgende Schritte im Vorgehensmodell dar. Ein Subset kann erhoben, erzeugt oder ausgewählt werden. Im Kontext der hochdigitalisierten SC liegen oftmals alle wesentlichen Datenbestände in den Systemen vor und werden nur bezüglich konkreter Bedarfe neu erhoben. Das wesentliche Ziel der Subsetbildung ist nach Möglichkeiten, nur relevante Information zu selektieren. Im Zuge der Informationsselektierung haben sich unterschiedliche Techniken bewährt. Entscheidend für die Wahl

der Technik sind der vorliegende Datenumfang und das Format der Daten. So sind beispielsweise für Zeitreihen andere Selektionsmethoden als für Randdaten, d. h. Daten die bei der Nutzung von informationstechnischen Infrastrukturen entstehen, zu bevorzugen. Eine umfassende Übersicht über die Techniken ist in Liu und Motoda (2001) zu finden. Eine besondere Herausforderung ergibt sich im Kontext der SC, da Stichproben, wie sie aus Statistik und Informatik bekannt sind, Zusammenhänge zwischen einzelnen Transaktionen nicht berücksichtigen (vgl. Emergenz der SC in Abschnitt 2.2.2.3). Eine Ausnahme bilden die sogenannten Cluster-Samples, mit denen über geeignete Cluster-Verfahren zusammenhängende Datenuntermengen gebildet werden. Die Stichprobe bezieht sich dann auf die Auswahl von Untermengen und nicht mehr auf die Auswahl von einzelnen Transaktionen (vgl. García et al. 2015). Da Cluster-Sampling jedoch sehr aufwendig ist, ist alternativ die Technik der Fensterung möglich, um Zusammenhänge der ursprünglichen Datenbasis in das Subset zu integrieren. In dieser Technik wird ein Fenster mit festgelegter Schrittgröße über die Daten gelegt. Es lassen sich horizontale (Anzahl der Attribute) und vertikale Partitionierung (Anzahl der Entitäten) im Datenbankkontext unterscheiden. In beiden Fällen wird in variierenden Größen ein Fenster sowie die Schrittgröße definiert. Wird ein Fenster-Verfahren auf einer Datenbanktabelle angewendet, können einzelne, nicht-disjunkte Partitionen gebildet werden (Draisbach 2012). Hierzu muss festgehalten werden, dass jede Partition als Stichprobe fungieren kann, jedoch nicht jede Stichprobe einer Partition entspricht. In einer einfachen Grundform wird ein Fenster der Größe zwei mit Schrittgröße eins über die Datenbasis gelegt und erzeugt eine Partition mit je zwei Entitäten. Abbildung 2.8 stellt eine Zufallsstichprobe und eine Fensterung mit Fenstergröße und Schrittweite drei dar. Es ist zu beachten, dass auch Fensterungen mitunter eine Vorauswahl der Daten benötigen können, da ansonsten beispielsweise irrelevante Datensätze in die Subsets integriert werden.

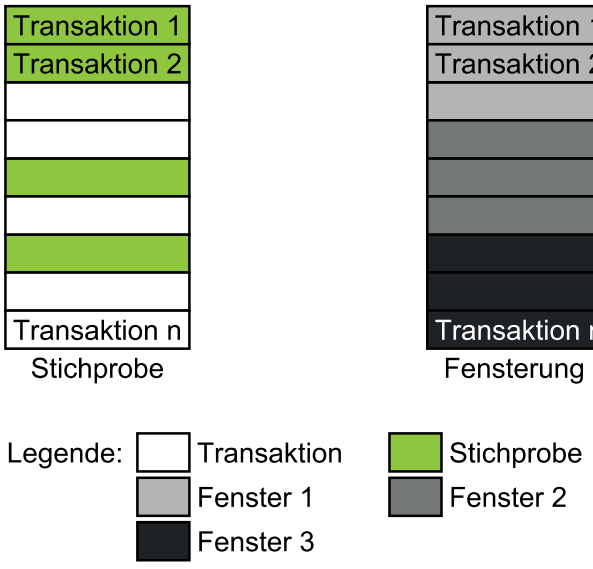



**Abbildung 2.8: Prinzip von Stichprobe und Fensterung auf Datenbank**

Fenster-Techniken gehören zu den Partitionierungsverfahren, da sie im Ergebnis die Datenbank in Partitionen aufteilen. Partitionierungsverfahren können in verschiedene weitere Techniken unterteilt werden (vgl. Nowitzky 2001). Vorarbeiten am Fachgebiet ITPL zum Thema Partitionierung von SC-Datenbanken haben gezeigt, dass horizontale, wertebasierte Partitionierungen auf SC-Datenbanken notwendig sind (Guhl 2014). Diese Partitionierungsverfahren müssen die sachlogischen Zusammenhänge der Daten berücksichtigen, d.h. diese sachlogischen Partitionen sind auf Basis eines semantischen Kontexts aus einem oder mehreren Attributwerten zu bilden. Daraus folgt, dass bei einer sachlogischen Partitionierung das Partitionierungskriterium auf einer sachlichen Relation der einzelnen Datensätze beruht. Demnach können sachlogische Partitionierungen zeitliche Relationen über Zeitstempel oder inhaltliche Relationen z. B. über Lieferantennetzwerke im SC-Kontext berücksichtigen. Der Vollständigkeit halber sei erwähnt, dass die Unterscheidung der Fenster-Techniken und Cluster-Verfahren Schnittmengen aufweisen und bei Parametrierung spezifischer Clusterverfahren über eine Abstandsfunktion die Resultate denen einer geeigneten Fenster-Funktion auf den Daten entsprechen. Es handelt sich jedoch um Spezialfälle und beide Verfahren können sowohl in ihren Grundfunktionen als auch in ihren inhaltlichen Zielsetzungen unterschieden werden. Des Weiteren können Partitionierungsverfahren in unterschiedlichen Phasen des Vorgehensmodells Anwendung finden. Beispielsweise ist es möglich, diese Verfahren in der Vorbereitungsphase von spezifischen Data-Mining-Verfahren einzusetzen.

**Preprocessing** Die dritte Phase im Vorgehensmodell ist das Preprocessing der zuvor ausgewählten Daten. Hierzu zählt zum einen die Datenintegration, die notwendig wird, wenn die Daten aus verschiedenen Quellsystemen stammen. Zum anderen muss jedoch auch die Schemaintegration beachtet werden. Hierunter wird die Integration von sogenannten Metadaten aus verschiedenen Quellsystemen verstanden. Die Daten weisen im Allgemeinen drei Schwachstellen auf: Unvollständigkeit, Verrauschung und Inkonsistenz. Diese müssen im Rahmen der Vorverarbeitung behoben werden, denn ohne ausreichende Datenqualität sind die Ergebnisse des KDD fragwürdig (vgl. Abschnitt 2.2.2.4). Fayyad und Uthurusamy (1994, S. 6) nutzen die Formulierung: „a KDD process cannot succeed without a serious effort to clean or scrub the data“ . Dieser zweite Aspekt des Preprocessings wird unter dem Begriff Data Cleansing behandelt. Data Cleansing bietet für jede Datenschwachstelle ein Repertoire an Lösungsmöglichkeiten. Unvollständigkeit, beispielsweise durch fehlende Attribute, kann durch das Auffüllen von Attributen mittels verschiedener mathematischer Maßzahlen behoben werden. Verrauschungen können mit Verfahren der Ausreißerkorrektur (Binning, Clustering oder Regression) geglättet werden. Die Korrektur von inkonsistenten Daten ist ebenfalls mittels verschiedenartiger Techniken möglich, z. B. mittels Ersetzen durch plausible Werte(-Imputation). Data Cleansing wird jedoch nicht nur auf den Bereich des Data Minings angewendet. Er findet vielmehr immer dort Anwendung, wo große Datenmengen von einer oder mehreren Quellen zu einem oder mehreren Zielen bewegt werden sollen. So

ist Data Cleansing ebenfalls ein Schritt von ETL-Prozessen wie beispielsweise der Daten-Migration.

**Transformation** Die vierte Phase im Vorgehensmodell ist Datentransformation. Zur Transformation zählen die Aspekte Reduktion und Projektion. Die Trennung zwischen der Reduktion und der zweiten Phase des Vorgehensmodells, die die Auswahl der Daten behandelt, ist oftmals nicht sauber in der Literatur aufgeführt und wird häufig unter dem Begriff „Feature Selection“ zusammengeführt (Liu und Motoda 2001). Die unsaubere Trennung resultiert aus der Tatsache, dass beide Teilbereiche die Datenreduktion zum Ziel haben. Unter einem reduzierten Datenbestand ist ein Datenbestand zu verstehen, der bezüglich Datenvolumen und Datenkomplexität oder beider zuvor genannten Aspekte reduziert wurde, aber trotzdem identische Data-Mining-Ergebnisse liefert. Generell stehen zwei Methoden zur Datenreduktion zur Verfügung. Zum einen gibt es die Auswahl von einzelnen Merkmalen mittels Subset-Selection-Algorithmen, zum anderen existieren die Feature-Ranking-Algorithmen. Zusätzlich sind unter Varianz- und Diskriminanzanalyse in der Statistik zwei große Datenanalysegruppen zu finden, die Aufschluss über die Gesetzmäßigkeiten der einzelnen Merkmale liefern und so Hinweise zum Reduktionspotential beisteuern. Die Projektion, als zweiter Aspekt der Datentransformation, lässt sich wiederum in die Bereiche Normalisierung und Aggregation aufteilen. Zur Normalisierung und Aggregation zählen Verfahren wie die Diskretisierung, die Dimensionsreduktion und die numerische Datenreduktion. Insbesondere die Diskretisierung nimmt im Kontext der SC einen hohen Stellenwert ein. Der Stellenwert ist durch die verschiedenen Merkmalsausprägungen begründet, denen im Rahmen der Diskretisierung ein oder mehrere Skalen zugewiesen werden können. Die einzelnen Skalenniveaus unterscheiden sich danach, ob sie in eine Reihenfolge gebracht werden können, in welchen Abständen einzelne Ausprägungen vertreten sind und welchen Wertebereich die Ausprägungen abdecken. In der Literatur sind unterschiedliche Synonyme für die einzelnen Skalen zu finden. In Abbildung 2.9 ist eine Übersicht über die unterschiedlichen Skalenarten und ihre Hierarchie gegeben. Hierbei erfolgte die Orientierung an der Begriffswelt von Fahrmeir et al. (2010).

Etliche Data-Mining-Verfahren, insbesondere die für implizite Wissensrepräsentationen, gehören zur Kategorie der „training-by-examples“. Diese Methoden fordern in ihrer Anwendung kategorisierbare Attribute. In der SC ist jedoch eine Vielzahl von kontinuierlichen Attributen zu finden, wie beispielsweise die Zeitstempel in den Transaktionsdaten (siehe Abschnitt 2.2.2.1). Das Standardverfahren, genannt „global discretization“, wäre nun die Vorgehensweise, alle kontinuierlichen Attribute zu diskretisieren (Frank und Witten 1999). Diese Verfahren sind jedoch für die SC-Daten aufgrund der Attribute, wie zum Beispiel Mengenangaben oder spezifische Zeiteinheiten, und nicht zuletzt der Vielzahl der Attribute, nicht effizient einsetzbar. In der Literatur existieren jedoch viele Speziallösungen. Die „lokale Diskretisierung“ ist eine von diesen die im konkreten Fall erprobt werden können (Liu und Motoda 2001).

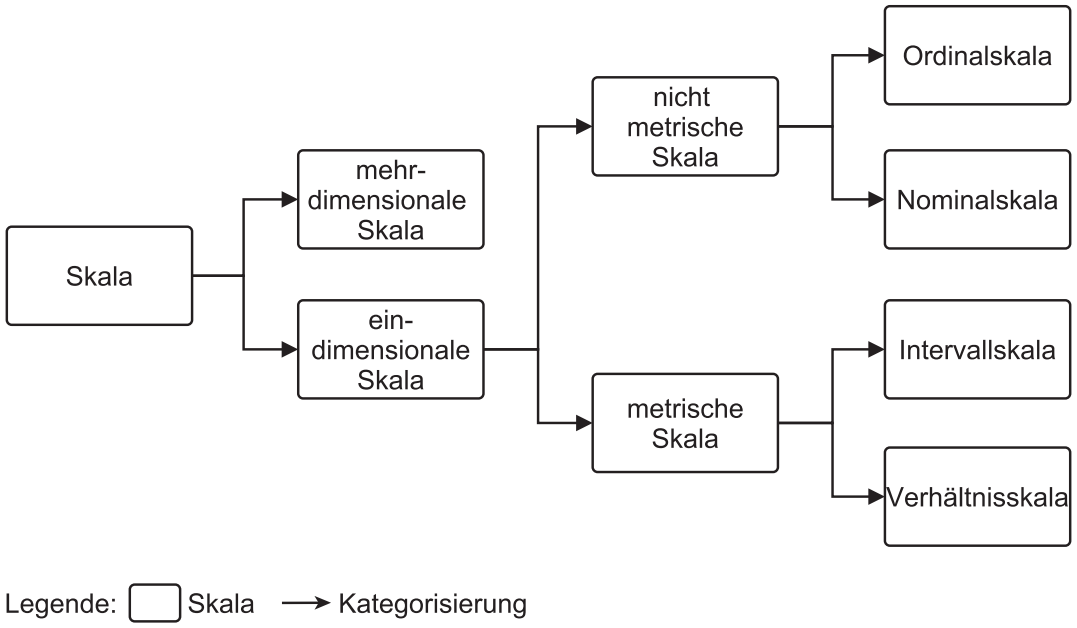



**Abbildung 2.9: Hierarchische Darstellungen der Skalenarten**

In der vorliegenden Arbeit wird die Transformation als separater Teilschritt des KDD diskutiert. Die Separierung vom KDD-Preprocessing ist jedoch fachlicher Natur und wird in den praktischen Umsetzung für die Industrie oftmals nicht unterschieden. Dies wird zum einen in den Vorgehensmodellen zur Wissensentdeckung deutlich, aber auch in der eingesetzten Wissensentdeckungssoftware, die die Operatoren für diese Bereiche nicht separiert. Demnach können die Operatoren aus Preprocessing und Transformation in sechs Basisklassen eingeteilt werden (García et al. 2015):

1. Bereinigung (Frage: Wie kann ich den Datenbestand bereinigen?)
2. Transformation (Frage: Wie kann ich „akkurate“ Daten bereitstellen?)
3. Integration (Frage: Wie können Daten- und Schemata integriert werden?)
4. Normalisierung (Frage: Wie kann ich Daten skalieren und normalisieren?)
5. Datenauffüllung (Frage: Wie kann ich fehlende Daten bereitstellen?)
6. Noise-Bereinigung (Frage: Wie kann ich Verunreinigungen (Noise) in den Daten aufdecken und wie bereinigen?)

Der Bereich der Transformation muss in Bezug auf die in der nächsten Phase eingesetzte Data-Mining-Methode gesehen werden. Aus diesem Grund ist der iterative Charakter zwischen diesen beiden Phasen von besonderer Bedeutung. Wie in Abschnitt 2.2.2 erläutert, benötigen Data-Mining-Verfahren spezifische Eingabeformate. Diese Eingabeformate müssen durch die Transformation, sei es durch Projektion oder Aggregation, erzeugt werden. Das Erzeugen der spezifischen Verfahrenseingaben ist sogar bedeutender für das Data-Mining-Ergebnis als die Wahl

des eigentlichen Data-Mining-Verfahrens (Weiss und Indurkhya 1998). Da in der SC häufig alphabetische Daten vorliegen (vgl. Abschnitt 2.2.2.2) und ein Großteil der Methoden numerische Eingabewerte benötigt bzw. eine metrische Beziehung zwischen den Attributen fordert, sind hier geeignete Metriken zu definieren. Der Begriff Metrik ist definiert als ein Abstandswort, das je zwei Elementen einen nicht negativen reellen Wert zuordnet, die sogenannte Distanz. Je geringer die Distanz zwischen einzelnen Elementen ist, desto ähnlicher sind diese zueinander. Im Allgemeinen werden Distanz- und Ähnlichkeitsmaße auch als Proximitätsmaße zusammengefasst. Gängige Metriken sind in Tabelle 2.8 aufgeführt. Hierbei sei erwähnt, dass die Skalenarten (vgl. Abbildung 2.9) bezüglich einzelner Metriken nicht eindeutig zuzuordnen sind.

**Tabelle 2.8: Übersicht gängiger Metriken nach Bronštejn et al. (2015)**

| Name der Metrik | Beschreibung | Datenkategorie |
|---|---|---|
| Euklidische Distanz | Beschreibt die Länge der Strecke zwischen zwei Elementen | numerisch |
| Hamming-Distanz | Gibt die Anzahl der Binärstellen an, in denen sich zwei Zeichenketten unterscheiden | numerisch, alphabetisch, alphanumerisch |
| Levenshtein-Distanz | Beschreibt die minimale Anzahl von Einfüge-, Lösch- und Ersetz-Operationen zwischen zwei Zeichenketten, um die eine Zeichenkette in die andere umzuwandeln | numerisch, alphabetisch, alphanumerisch |
| Manhattan-Distanz | Gibt die Summe der absoluten Differenzen an | numerisch |
| Maximum-Distanz | Entspricht der größten absoluten Differenz eines Wertepaares | numerisch |

**Auswahl der Data-Mining-Verfahren** In der fünften Phase müssen geeignete Data-Mining-Verfahren in Bezug auf die Aufgabenstellung entwickelt werden. Hierbei können unterschiedliche Methoden den einzelnen Aufgaben zugeordnet werden. Insbesondere kann eine Aufgabenstellung durch mehrere unterschiedliche Methoden gelöst werden. Beispielsweise kann die Klassifikation sowohl mittels fallbasiertem Schließen als auch über Entscheidungsbaumverfahren durchgeführt werden. Auf der anderen Seite ist die Zuordnung der Methoden zu den Aufgaben ebenfalls nicht in jedem Fall eindeutig. Beispielsweise finden Künstliche Neuronale Netze (KNN) im Bereich der Prognose und auch in der Segmentierung Anwendung. Insbesondere können Methoden in verschiedene Unterkategorien gegliedert werden. Dies wird insbesondere bei KNNs deutlich, bei denen eine Vielzahl von Netztypen mittels unterschiedlichen Netztopologien und Verbindungsarten unterschieden

werden können. Tabelle 2.9 gibt einen Einblick in mögliche Zuordnungsrelationen. Ob ein Data-Mining-Verfahren für die vorangestellte Analyseaufgabe geeignet ist, beurteilt Küppers mittels unterschiedlicher Kriterien, die er in drei Bereiche unterteilt: anwendungsorientierte Kriterien, methodenorientierte Kriterien und datensatzorientierte Kriterien. Die Kriterien können zur Bewertung der unterschiedlichen Data-Mining-Verfahren herangezogen werden. Die einzelnen Kriterien können der Tabelle nach Küppers (1999) im Anhang entnommen werden (vgl. Tabelle A.3). Generell lässt sich feststellen, dass drei Kriterien auf den Methodenraum, also die Menge der möglichen Algorithmen, einwirken. Diese sind die Aufgabenstellungen der Wissensentdeckung (vgl. Abschnitt 2.2.3.1), die zugrundeliegende Datenbasis (vgl. Abschnitt 2.2.2.2) sowie die Darstellungsanforderung des Wissens bezüglich der Explizität (vgl. Abschnitt 2.1.2).

**Tabelle 2.9: Beispielhafte Gegenüberstellung von KDD-Aufgaben und zugehörigen Verfahren**

| **Aufgabe** | **Verfahren** | **Beispielalgorithmen / Subtypen** | **Beispielhafte Literatur für Verfahren** |
|---|---|---|---|
| Klassifikation | Entscheidungsbaum | • CHAID<br>• ID3<br>• Random Tree | • Quinlan 1986<br>• Cleve und Lämmel 2014 |
| | Nearest-Neighbour | • k-Nearest-Neighbor | • Cleve und Lämmel 2014<br>• Petersohn 2005 |
| | KNN | • AutoMLP<br>• Perceptron Netz<br>• Self-Organizing Maps (SOM) | • Freund und Schapire 1999<br>• Kohonen 2001 |
| | Regression | • Lineare Regression<br>• Vector Linear Regression<br>• Polynomial Regression | • Yan und Su 2009 |
| Segmentierung | Clusteranalyse | • BIRCH<br>• CHAMELEON<br>• k-Means<br>• x-Means<br>• Support Vector Clustering | • Cleve und Lämmel 2014<br>• Pelleg und Moore 2000 |

**Tabelle 2.9: Beispielhafte Gegenüberstellung von KDD-Aufgaben und zugehörigen Verfahren (Fortsetzung)**

| Aufgabe | Verfahren | Beispielalgorithmen / Subtypen | Beispielhafte Literatur für Verfahren |
|---|---|---|---|
| KNN | • AutoMLP<br>• Perceptron Netz<br>• SOM | • Freund und Schapire 1999<br>• Kohonen 2001 | |
| Prognose | Support Vector Machines | • LibSVM<br>• SVMlight<br>• Fast Large Margin | • Smola und Schölkopf 2004 |
| | KNN | • AutoMLP<br>• Perceptron Netz<br>• SOM | • Freund und Schapire 1999<br>• Kohonen 2001 |
| | Regression | • Lineare Regression<br>• Vector Linear Regression<br>• Polynomial Regression | • Yan und Su 2009 |
| Abhängigkeitsanalyse | Assoziationsanalyse | • Apriori<br>• FPGrowth | • Cleve und Lämmel 2014 |
| | Bayes'sches Netz | • Naive Bayes (Kernel)<br>• Mc-Culloch-Pitts-Netze | • Cleve und Lämmel 2014 |

**Algorithmen- und Parameteranpassung für die Data-Mining-Verfahren** In der sechsten Phase müssen geeignete Algorithmen für die zuvor festgelegten Data-Mining-Verfahren ausgewählt werden. Jedes Verfahren kann durch eine Vielzahl von konkreten Algorithmen gelöst werden. Beispielsweise können Verfahren der Assoziationsanalyse, die der Abhängigkeitsanalyse zuzuordnen sind, mittels Algorithmen wie FPGrowth oder Apriori gelöst werden. In Tabelle 2.9 sind exemplarische Verfahren zur Einordnung dargestellt. Für eine explizite Beschreibung der einzelnen Algorithmen sei auf die beispielhafte Literatur in der Tabelle verwiesen.

**Data Mining – Mustersuche** In der siebten Phase erfolgt das Data Mining, das als wesentliche Aufgabe die Suche nach Mustern beinhaltet. Der Musterbegriff taucht hier erst als Ergebnis des Data Minings auf. Synonym finden bei einigen Autoren Begriffe wie Schema, Vorbild oder Modell (Petersohn 2005), Struktur

(Fahrmeir et al. 2010; Rönz und Strohe 1994) aber auch Hypothese oder Funktion (Liu und Motoda 2001) Anwendung. Diskussionen um die Begrifflichkeit existieren seit geraumer Zeit und fanden keine zufriedenstellende Lösung. Bereits in den 1970er Jahren verwies Niemann auf die Diskussionen um die Terminologie „Muster“ (Niemann 1974) und stellte treffend fest, dass es keinen allgemeingültigen Ansatz gibt. Diese Tatsache spiegelt auch die Tabelle 2.10.

**Tabelle 2.10: Gängige Musterdefinitionen**

| Autor | Definition |
|---|---|
| Bissantz und Hagedorn 2009, S. 139 | „Muster bezeichnen Beziehungen zwischen Datensätzen, zwischen den Daten innerhalb eines Satzes oder bestimmte Regelmäßigkeiten.“ |
| Borchers 2001, S. 7 | „[...] a structured textual and graphical description of a proven solution to a recurring design problem.“ |
| Fahrmeir et al. 2010 | [...] ein Muster kennzeichnet Struktur [...] |
| Fayyad et al. 1996a, S. 30 | „[...] the term pattern goes beyond its traditional sense to include models or structure in data. In this definition, data comprises a set of facts (e.g., cases in a database), and pattern is an expression in some language describing a subset of the data (or a model applicable to that subset).“ |
| Fayyad et al. 1996b, S. 31 | „[...] and pattern is an expression in some language describing a subset of the data or a model applicable to the subset. Hence, in our usage here, extracting a pattern also designates fitting a model to data; finding structure from data; or, in general, making any high-level description of a set of data.“ |
| Fockel 2009 | (Daten-) Muster ist eine Beschreibung von Daten, die einfacher ist als die Aufzählung der Daten selbst. Eine „Menge von Mustern“ (Datenmuster) bezeichnet hierbei also einen Ausdruck in einer formalen Sprache, der eine Objektmenge beschreibt, ohne die einzelnen Datenobjekte einfach nur aufzulisten. Die Datenobjekte beschreiben Objekte eines Realsystems durch ausgewählte Merkmale. Die Datenmuster werden dazu verwendet, reale Zusammenhänge zwischen den beobachteten Merkmalen zu modellieren. |
| Gamma et al. 1995, S. 233 | „[...] patterns are descriptions of communicating objects and classes that are customized to solve a general design problem in a particular context.“ |

**Tabelle 2.10: Gängige Musterdefinitionen (Fortsetzung)**

| Autor | Definition |
|---|---|
| Hanhijärvi 2011, S. 122 | „[...] patterns that are a result of exceptional structure in the data.“ |
| Le et al. 2015, S. 189 | „Real data rarely conform to regular patterns. Finding subtle deviation from the norm, or novelty detection, is an important research topics [...] Novelty detection refers to the task of of finding patterns in data that do not conform to expected behaviors. These anomaly patterns are interesting because they reveal actionable information, the known unknowns and unknown unknowns.“ |
| Liu und Motoda 2001 | [...] ein Muster ist auch bekannt als Hypothese oder Funktion [...] |
| Morik und Klingspor 2006, S. 167 | „Jedes Muster beschreibt ein in unserer Umwelt beständig wiederkehrendes Problem und erläutert den Kern der Lösung für dieses Problem, sodass Sie diese Lösung beliebig oft anwenden können, ohne sie jemals ein zweites Mal gleich auszuführen.“ |
| Niemann 1974, S. 3 | $$\mathbf{f}(\mathbf{x}) = \begin{cases} f_1(x_1, \ldots, x_n) \\ \vdots \\ f_m(x_1, \ldots, x_n) \end{cases} \quad (2.2)$$ |
| Petersohn 2005 | Ein Muster wird auch als Schema, Vorbild oder Modell bezeichnet. |
| Piazza 2010, S. 31 | „Mit Muster wird ein der Form her beliebiger Zusammenhang in den Daten bezeichnet. Das Muster stellt dabei eine zusammenfassende Aussage über Einzelheiten in den Daten dar, d. h. es aggregiert Detailmerkmale des analysierten Datenbestands. Damit wird zunächst eine Komplexitätsreduktion erreicht, die grundsätzlich dem Entscheidungsträger einen vereinfachten Zugang zu den Inhalten des Datenbestands ermöglichen soll.“ |

**Tabelle 2.10: Gängige Musterdefinitionen (Fortsetzung)**

| Autor | Definition |
|---|---|
| Rashidi et al. 2015, S. 215 | „Mining abnormal, i.e., anomalous, patterns is a significant component of many data mining tasks. Numerous methodologies have been developed for detecting anomalous data objects under the assumption that there is no relational information between these objects.“ |
| Rönz und Strohe 1994 | [. . . ] ein Muster ist Struktur [. . . ] |
| Runkler 2010, S. 2 | ”„Unter Wissen verstehen wir interessante Muster, und unter interessanten Mustern schließlich solche, die allgemein gültig sind, nicht trivial, neu, nützlich und verständlich.“ |
| Winn und Calder 2002, S. 59 | „[. . . ] pattern identifies a recurring problem and a solution, describing them in a particular context [. . . ]. Patterns thus aim to capture and explicitly state abstract problem-solving knowledge that is usually implicit and gained only through experience.“ |

In der vorliegenden Arbeit wird der Begriff Muster als Ergebnis des Data Minings verwendet. Ein Muster selbst ist generisch und kann über mehrere Musterinstanzen verfügen. Das Muster beinhaltet freie Variablen und wird durch die Quantifizierung der freien Variablen instanziiert. Das Muster selbst repräsentiert eine Klasse für alle Instanzen, die ihr gemäß einer Klassifikationsaufgabe zugeordnet wurden. Im Kontext der SC kann die Regel „wenn Transportzeit $> 300$ Stunden dann Flugzeug“ als Beispiel für ein Muster dienen. Transaktionen in der Datenbank, die dieser Regel entsprechen, sind Instanzen des Musters. Sie weisen konkrete Werte für die Transportzeit auf, beispielsweise die Merkmalsausprägung „325 Stunden“, und erfüllen demnach die Bedingung der Regel.

Die angeführten Definitionen des Musterbegriffs zeigen deutlich, dass textuelle Beschreibungen die Literatur im Bereich KDD und Logistik dominieren. Einer der wenigen mathematischen Ansätze stammt aus der Fachdisziplin der graphischen Systeme und definiert ein Muster als vektorwertige Funktion (vgl. Gleichung 2.2). In der SC-Literatur gibt es im Gegensatz dazu keine mathematischen Ansätze für die Musterdefinition.

Gemein ist allen bisher angeführten Definitionen, dass der Musterbegriff über die Unterscheidung „einfach“ bzw. „komplex“ hinausgehend nicht weiter differenziert wird. Dies wird der Wissensentdeckung im SC-Umfeld nicht gerecht, denn Muster können in verschiedenen Formen in unterschiedlichen Schritten des Vorgehensmodells existieren - es ist dann von unterschiedlichen Mustertypen die Rede. Klösgen

und Zytkow (1996) führen den Begriff des Musters sowie verschiedene Muster-Kategorien in einer inhaltlichen, nicht-mathematischen Definition ein. Die im Nachfolgenden angeführten Definitionen der Mustertypen orientieren sich stark an dieser Definition.

> **Definition 2.11 Muster:** Ein Muster ist ein generisches Statement, das freie Variablen (die Parameter der Muster) enthält. Durch die Quantifizierung der freien Variablen oder durch das Ersetzen mit konkreten Werten können Muster instanziiert werden (Klösgen und Zytkow 1996).

> **Definition 2.12 Muster-Instanz:** Eine Muster wird instanziiert, indem die Musterparameter mit Konstanten und / oder die Musterparameter-Variablen mit Quantoren ersetzt werden. Das Ergebnis dieser Instantiierung ist die Muster-Instanz (Klösgen und Zytkow 1996).

> **Definition 2.13 Muster-Kategorien:** Muster können in verschiedene Kategorien unterteilt werden. Die drei Hauptkategorien sind logisch-numerische Muster (zu denen auch textuelle und räumliche Muster zählen, die jedoch im Kontext der SCs irrelevant sind), elementare Muster und statistische Muster.
>
> Logisch-numerische Muster repräsentieren Wissen bezüglich logischen oder numerischen Funktionen. Diese Kategorie gestattet weitere Unterteilungen in sogenannte Subtypen. Zu den Subtypen zählen beispielsweise Bäume, Regeln oder Tabellen.
>
> Elementare Muster sind univariate Muster, die durch einfache Erkennungsmethoden (die die Subtypen bilden) für Maximum, Minimum, Ausreißer und dergleichen entdeckt werden können.
>
> Statistische Muster weisen verschiedenen Ereignissen oder Merkmalen und deren Kombination Wahrscheinlichkeiten zu und bilden die Grundlage für probabilistische Modelle. Auch diese weisen Subtypen auf und können beispielsweise in unterschiedlichen Verteilungsfamilien gegliedert werden. (Klösgen und Zytkow 1996)

Hierbei ist auffällig, dass zwischen Mustern als Ergebnis eines komplexen Data-Mining-Verfahrensschritts (logisch-numerische Muster) und einfachen Mustern (elementare Muster), die bereits vor dem Data Mining identifizierbar sind, unterschieden wird. Es fehlt jedoch eine Einbeziehung des Kontextwissens in den Musterbegriff. Dennoch kann zusammenfassend konstatiert werden, dass zu entdeckendes Wissen aus den Fragestellungen des SCM (vgl. Abschnitt 2.2.3.2) mittels der vorgestellten Musterbegriffe abgebildet werden kann. Insbesondere können die vorgestellten Wirkzusammenhänge (vgl. Abschnitt 2.2.3.2) als Sonderform der Regeln aufgefasst werden und folgen damit der Definition 2.13.

**Interpretation der Muster** In der achten Phase erfolgt die Interpretation der Muster. Die Muster werden im Anschluss der V&V unterzogen. Hierbei ist die Grundfrage, ob die gefundenen Muster in Bezug auf das zugrundeliegende Logikmodell verifiziert werden können und ob das angenommene Logikmodell selbst als solches valide ist.

> **Definition 2.14 Verifikation:** „Verifikation ist die Überprüfung, ob ein Modell von einer Beschreibungsart in eine andere Beschreibungsart korrekt transformiert wurde“ (Rabe, Spieckermann et al. 2008, S. 14).

> **Definition 2.15 Validierung:** „Validierung ist die kontinuierliche Überprüfung, ob die Modelle das Verhalten des abgebildeten Systems hinreichend genau wiedergeben“ (Rabe, Spieckermann et al. 2008, S. 15).

Der vollständige Nachweis der Richtigkeit durch die V&V ist nur bei sehr einfachen Modellen möglich. Je nach Definition der einzelnen Schritte ist auch eine Zuordnung der Verifikation zur vorangestellten Phase 7 möglich. Da V&V nicht als Einzelschritt in die KDD-Vorgehensmodelle eingeordnet wird, sind verschiedene Verknüpfungspunkte für diese Vorgehensmodelle möglich (Düsing 2010). Je nach V&V-Ergebnis können iterativ Schritte der Phasen 1 bis 7 wiederholt oder zur letzten Phase 9 übergegangen werden. Die Interpretation der Muster impliziert neben der V&V auch separate Visulalisierungstechniken. Dies wird bei Fayyad et al. deutlich, da sie als Teilaufgabe die Visualisierung benennen: „This step can also involve visualization of the extracted patterns and models or visualization of the data given the extracted models“ (Fayyad et al. 1996b, S. 42). Da Visualisierungstechniken jedoch ebenfalls den V&V-Techniken zugeordnet werden können (Rabe, Spieckermann et al. 2008), wird dieser Aspekt nicht gesondert betrachtet.

Sehr deutlich wird die Isolation der V&V in Maimon und Rokach (2010). Dort wird im System „Data Mining Taxonomy“ die Verifikation als separierte Methodenfamilie aufgeführt und hat keine Verbindung zu den konkreten Data-Mining-Verfahren. Zusammenfassend sei hervorgehoben, dass die V&V in der vorherrschenden Literatur oftmals nur die Phase 7 im Rahmen der Musterentdeckung berücksichtigt. Folglich steht die V&V der Ergebnisse, also der gefundenen Muster, im Fokus. Es erfolgt keine explizite V&V von KDD-Zwischenergebnissen, wie beispielsweise eine Prüfung der vorverarbeiteten Daten. Insbesondere ist in der relevanten Literatur kein Konzept zur modellbegleitenden V&V in den KDD-Vorgehensmodellen aufgeführt.

Eine umfassende Betrachtung der V&V ist in der KDD-Literatur nicht zu finden. Andere Disziplinen, wie beispielsweise die Softwareentwicklung oder die Simulation, verstehen die V&V als modellbegleitenden Prozess und stellen hierfür eine Vielzahl an unterschiedlichen Techniken bereit. Die Techniken reichen vom Codereview bis hin zu mathematischen Modellen. Eine Übersicht geben Balci (1998), Het-

zel (1984) und Rabe, Spieckermann et al. (2008). Im KDD liegt der Schwerpunkt auf der V&V der Modelle zur Wissensentdeckung. Die Bewertung der Modelle erfolgt im Standardverfahren mithilfe von Trainings-, Validierungs- und Testdaten sowie Fehlerkoeffizienten (Fehlermaß). Die Data-Mining-Modellbildung erfolgt auf den Trainingsdaten, die den Großteil des Datenbestandes ausmachen. Das erlernte Modell wird mittels der Validierungsdaten parametriert, d. h. es werden Parametereinstellungen für die Modelle gesucht, die auf den Validierungsdaten die geringsten Fehlerkoeffizienten aufweisen. Die Testdaten wiederum dienen zur Bewertung der Modellqualität. Mittels der Testdaten kann geprüft werden, wie die Generalisierungsfähigkeit des Modells zu bewerten ist. Um verlässliche Aussagen zu erhalten, sind die Trainings-, Validierungs- und Testdatenmengen im Regelfall disjunkt. Diese Technik ist unter dem Begriff Holdout bekannt und kann jedoch nur dort Anwendung finden, wo ausreichend Datenbestände vorhanden sind (Steinlein 2004). In der Literatur schwankt die Angabe für Testdaten zwischen 10 % und 30 % des Gesamtdatenbestandes. Als guten Mittelwert, unter Berücksichtigung der gängigen Literatur, schlagen Hastie et al. (2009) 25 % vor. Für den Trainings- und Testprozess stehen weitere Techniken zur Verfügung, die unterschiedliche Vor- und Nachteile bieten. Eine für den Industrieeinsatz hinreichende Einsicht in das Themengebiet kann bei Weiss und Indurkhya (1998) und Runkler (2010) gefunden werden.

Als abschließende Überlegung zur V&V im KDD-Einsatz kann konstatiert werden, dass in Bezug auf das zu entwickelnde Vorgehensmodell die V&V umfassender integriert werden sollte. Dies umfasst die Berücksichtigung der V&V in allen Phasen des Vorgehensmodells. Dadurch wird ein frühzeitiges Erkennen von fehlerhaften Aufgabenstellungen, ungültigen Vorverarbeitungsschritten oder fehlerhaften Modellbildungen ermöglicht. Die durchgeführten Schritte der V&V sowie deren Ergebnisse müssen zudem systematisiert werden und sollten abschließend über eine strukturierte Dokumentation nachvollziehbar sein (vgl. hierzu auch Rabe, Spieckermann et al. 2008).

**Weiterverwendung der Muster** In der neunten Phase erfolgt der Einsatz der Muster. Die Muster werden nun ihrem Weiterverwendungszweck zugeführt. Dabei ist es sinnvoll, nur die Muster einzubeziehen, die für das SCM eine Relevanz besitzen. Neben individuellen Bewertungskriterien, die beispielsweise von einer Fachseite aufgestellt werden können, gibt es auch technische Maßzahlen. Diese Maßzahlen, genannt „interestingness measures“, geben einen Hinweis darauf, ob bestimmte Muster potentiell relevant seien können und gestatten somit Methoden wie Musterpriorisierungen (Geng und Hamilton 2006). Um die interessanten Muster weiterverwenden zu können, ist eine geeignete Darstellungsform notwendig. Im Kontext des KDD gibt es verschiedene Darstellungsmöglichkeiten für dieses Wissen. Ist das Wissen beispielsweise eine Komponente von Entscheidungssystemen, so kann implizites Wissen Anwendung finden. Häufiger wird für die Entscheidungsfindung des SCM jedoch explizites Wissen benötigt, um dieses beispielsweise unmittelbar zu verwerten (vgl. Abschnitt 2.1.2). Vertreter für explizites Wissen im KDD sind Re-

geln, die über Entscheidungstabellen, Entscheidungsbäume oder instanzenbasierte Darstellungen repräsentiert werden. Häufige Darstellungsmöglichkeiten für implizites Wissen hingegen sind neuronale Netze und Cluster (Cleve und Lämmel 2014). Insbesondere stellt jedes Muster unabhängig von seiner Darstellungsform nach seiner Evaluierung Wissen dar (Klösgen und Zytkow 1996). Die unterschiedlichen Wissensrepräsentationen verdeutlichen ebenfalls den notwendigen iterativen Charakter des KDD-Vorgehensmodells, denn nicht jedes Data-Mining-Ergebnis lässt sich problemlos in andere Wissensdarstellungen überführen und eine alternative Data-Mining-Methode kann notwendig sein, um die gewünschte Darstellungsform zu erhalten.

Da die modellbegleitende V&V nicht Teil der existierenden KDD-Vorgehensmodelle ist, muss eine entsprechende Methodik an geeigneten Stellen im Modell ergänzt werden. Hierfür wird die interdisziplinäre Technik der ereignisdiskreten Simulation eingeführt. Diese zeigt Potentiale zur Unterstützung der V&V auf, die insbesondere bei der Validierung der Muster vielversprechend erscheinen.

### 2.3.3 Simulation

Simulation ist eine Analysemöglichkeit für komplexe Systeme. Hierbei ist ein System eine „Menge miteinander in Beziehung stehender Elemente, die in einem bestimmten Zusammenhang als Ganzes gesehen und als von ihrer Umgebung abgegrenzt betrachtet werden“ (DIN IEC 60050-351 2014, S. 21). Als Modell, spezifischer Systemmodell, wird eine vereinfachte Nachbildung des Systems bezeichnet. Systemmodelle, die mittels Simulation nachgebildet und analysiert werden, werden Simulationsmodelle genannt. Simulationsmodelle finden dann Einsatz, wenn analytische Lösungen nicht angewendet werden können, weil diese zu langsam oder fachlich nicht realisierbar sind. Simulationsmodelle werden häufig im Rahmen einer Simulationsstudie erstellt und ausgeführt. Die Ziele der Simulationsstudie sind vielfältig, sie reichen von der Untersuchung festgelegter Problemstellungen bis hin zur Prozessverbesserung (Wenzel, Weiß et al. 2008). Bei der Durchführung der Simulationsstudien können Assistenzsysteme zum Einsatz kommen, die den Anwender bei wiederkehrenden Aufgaben unterstützen (Mayer et al. 2012). Als Beispiele für solche Assistenzsysteme können AssistSim, das den Anwender in der Konzeption und Ausführung unterstützt, oder EDASim, das sich auf Eingabe-und Ausgabedaten in der Simulation fokussiert, angeführt werden (Bogon et al. 2012). Neben der eigentlichen Modellbildung ist oftmals die statistische Versuchsplanung ein zentrales Element der Simulationsstudie (Law 2014). Eine wesentliche Aufgabe der statistischen Versuchsplanung ist die Bewertung von simulierten Prozessen mittels Daten, die an verschiedenen Messpunkten des Modells oder zu beliebigen Zeitpunkten erhoben werden können. Diese Daten können dann in Dateien oder Datenbanken protokolliert werden (Trace) und gestatten das Zurückverfolgen von Simulationsabläufen über Grafiken oder Tabellen (VDI-Richtlinie 3633 Blatt 3 1997). Da die statistische Versuchsplanung in der vorliegenden Arbeit nur Anwendung findet und keinen Un-

tersuchungsgegenstand darstellt, gibt der Beitrag von Kleijnen et al. (2005) einen ausreichenden Überblick. Nach Law (2014) können Simulationsmodelle in statisch oder dynamisch, deterministisch oder stochastisch und kontinuierlich oder diskret unterschieden werden. Zudem wird unterschieden, ob es sich um terminierende und nicht-terminierende Systeme handelt. Bei nicht-terminierenden Systemen existiert ein prinzipiell unendlicher Zeithorizont. Hier entsteht nach einer Warmlaufphase (Einschwingphase) ein stationärer Zustand, der in einem terminierenden System im Allgemeinen nicht von Bedeutung ist. Die Aufteilung der Simulationsmodelle findet sich in allen gängigen Lehrbüchern aus dem Bereich der Simulation und wurde von verschiedenen Autoren in ähnlicher Weise grafisch aufbereitet. Hierbei ist anzumerken, dass das deterministisch-statische Modell nur von theoretischer Bedeutung ist. Abbildung 2.10 gibt eine Übersicht über die unterschiedlichen Simulationsarten und stellt auf der untersten Ebene konkrete Simulationsmethoden dar. Die genannte Quelle ist als exemplarische Referenz zu verstehen.

Die meisten Autoren beziehen sich in ihren Simulationsdefinitionen aufgrund der Systemkomplexität direkt auf dynamische Systeme. Banks (1998, S. 3) definiert Simulation als: „Simulation is the imitation of the operation of a real-world process or system over time“ . Die in dieser Arbeit verwendete Definition von Simulation richtet sich nach VDI-Richtlinie 3633 Blatt 1 (2014) und beschränkt sich ebenfalls auf dynamische Prozesse, welche im Fokus der vorliegenden Arbeit stehen.

> **Definition 2.16 Simulation:** „Nachbilden eines Systems mit seinen dynamischen Prozessen in einem experimentierbaren Modell, um zu Erkenntnissen zu gelangen, die auf die Wirklichkeit übertragbar sind; insbesondere werden die Prozesse über die Zeit entwickelt“ (VDI-Richtlinie 3633 Blatt 1 2014, S. 3).

Die Einsatzgebiete der Simulation sind vielfältig und umfassen unterschiedliche Forschungsfelder. Im logistischen Kontext findet die Simulation Anwendungsmöglichkeiten bei Aufgaben der Planung, Einführung und Optimierung. Mit Hilfe von What-if-Szenarien lassen sich u. a. klassische Fragestellungen zu Anlagendimensionierungen, Auslastung von Ressourcen und Durchlaufzeiten beantworten. Da die Anwendung der Simulation selbst oftmals keine Lösung für das zugrundeliegende Problem liefert, kann die Simulation mit anderen geeigneten Techniken kombiniert werden, um ihr volles Potential zu entfalten. Die Grundlagen der SC-Modellierung über Netzwerkstrukturen wurden von Mattfeld und Vahrenkamp (2014) umfänglich beschrieben. Die Techniken der SC-Simulation wurden u. a. von Terzi und Cavalieri (2004) und Chandra und Grabis (2007) dokumentiert.

Im Bereich der Produktion und Logistik befinden sich zumeist Modelle im Einsatz, die dynamisches Systemverhalten mit stochastischen Komponenten zu diskreten Zeitpunkten abbilden. Dabei sind fast ausschließlich ereignisorientierte Methoden wie die DES in Anwendung (Hellström und Johnsson 2002; März et al. 2011). Beispielsweise kommt die DES im Bereich der SC-Risikoanalyse (Ghadge et al. 2013) und Materialflussplanung der SC (Rubinstein 1997) zum Einsatz. Weitere For-

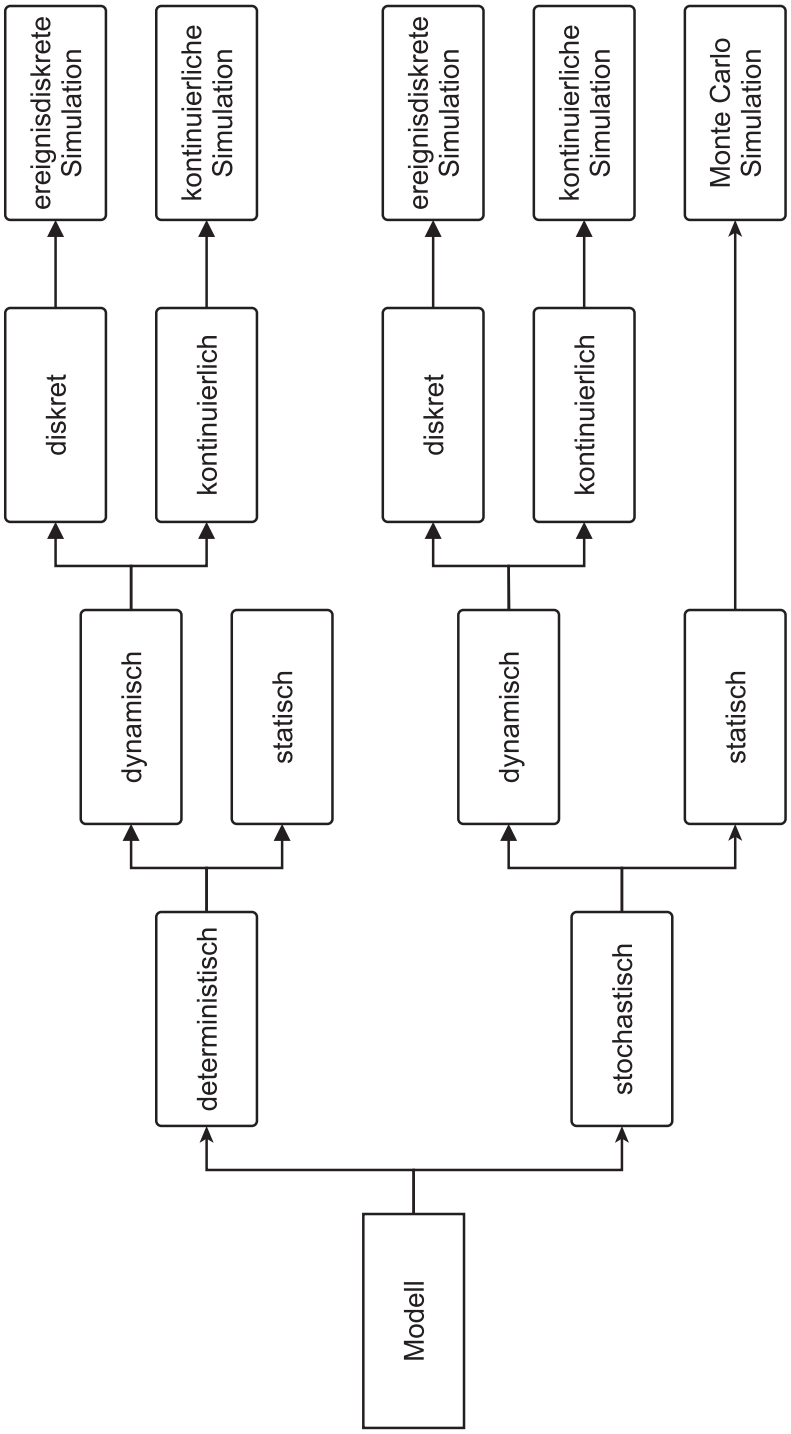



**Abbildung 2.10: Hierarchie der Simulationsarten nach Harrell et al. (2012, S. 71-102)**

schungen haben gezeigt, dass DES zu der Bestimmung von Zeitverhalten in SCs (Meinhardt et al. 2005) sowie zu der Analyse von autonomen Logistikprozessen (Becker et al. 2006) eingesetzt werden kann. Rubinstein und Kroese (2008) führen als Beispiele für den Einsatz der DES im SC-Kontext Bestandsführungssysteme, Fertigungslinien und Lebenszyklusanalysen an. Hellström und Johnsson (2002, S. 2) konstatieren „DES may be confidently used as a technique in supply chain planning“ und führen darüber hinaus aus [S.9]: „DES enabled users to observe and analyze the dynamic behavior of the model, which can be used to support decision making“ . Wenzel, Boyaci et al. (2010) konstatieren, dass die DES im gesamten logistischen Prozess Anwendung findet und auch Prozesse zwischen unterschiedlichen Unternehmen von dieser profitieren können. Für eine weitere Übersicht unterschiedlicher Verbindungsansätze von DES und SC wird auf Rabe und Deininger (2012) und Terzi und Cavalieri (2004) verwiesen.

Die DES wird ebenso wie das KDD mittels eines Vorgehensmodells umgesetzt. Die Vorgehensmodelle sind zahlreich und können deutlich bezüglich ihrer Komplexität unterschieden werden (Banks 1998; Law 2014). Die hiesige Arbeit orientiert sich im weiteren Verlauf an dem Vorgehensmodell des VDI (vgl. VDI-Richtlinie 3633 Blatt 1 2014), das auf Rabe, Spieckermann et al. (2008) zurückgeht, da das Vorgehensmodell des VDI im deutschsprachigen Raum eine hohe Akzeptanz besitzt (vgl. Abbildung 2.11). Selbst für die V&V in der Simulation existiert ein eigenes Vorgehensmodell, das aufgrund seiner visuellen Gestaltung unter dem Name „Dreiecksmodell“ bekannt ist (Rabe, Spieckermann et al. 2008). Die Grundidee dieses Modells geht auf das V&V-Vorgehensmodell von Brade (2003) zurück, der als Erster ein Modell zur schrittweisen V&V von Simulationsmodellen entwickelte.

### 2.3.4 Validierung der Wirkzusammenhänge in der Simulation

Die Wirkzusammenhänge der SC werden in dieser Arbeit unter Zuhilfenahme der DES überprüft. Der grundsätzliche Nutzen der Simulation für V&V wurde bereits von anderen Autoren aufgezeigt. So beschreiben Girard und Pappas (2006) Möglichkeiten zur Verifikation von kompletten Systemen mittels Simulation. Die Grundidee im Kontext der begleitenden V&V für das zu entwickelnde Vorgehensmodell ist die Überprüfung der Zusammenhänge in den Echtdaten, die mittels Mustern beschrieben werden. Muster, wie die diskutierten Wirkzusammenhänge, repräsentieren Informationen über bestimmte Teile der Simulation. Informationen können im Simulationsmodell über Eingriffspunkte im Simulationsmodell, sogenannte Stellgrößen, integriert werden. Diese Stellgrößen ermöglichen das Parametrieren von einzelnen Simulationskomponenten (März et al. 2011).

Eine weitere Möglichkeit, Informationen darzustellen, ist mittels des eigentlichen Simulationsmodells mit seinen Simulationskomponenten und der dazugehörigen Ausführungslogik. Ein Muster stellt jedoch in der Regel nur Informationen über Teile der Simulation dar; so gibt eine gefundene Regel aus dem KDD-Vorgehensmodell mitunter nur Informationen zu ausgewählten Komponenten wieder.

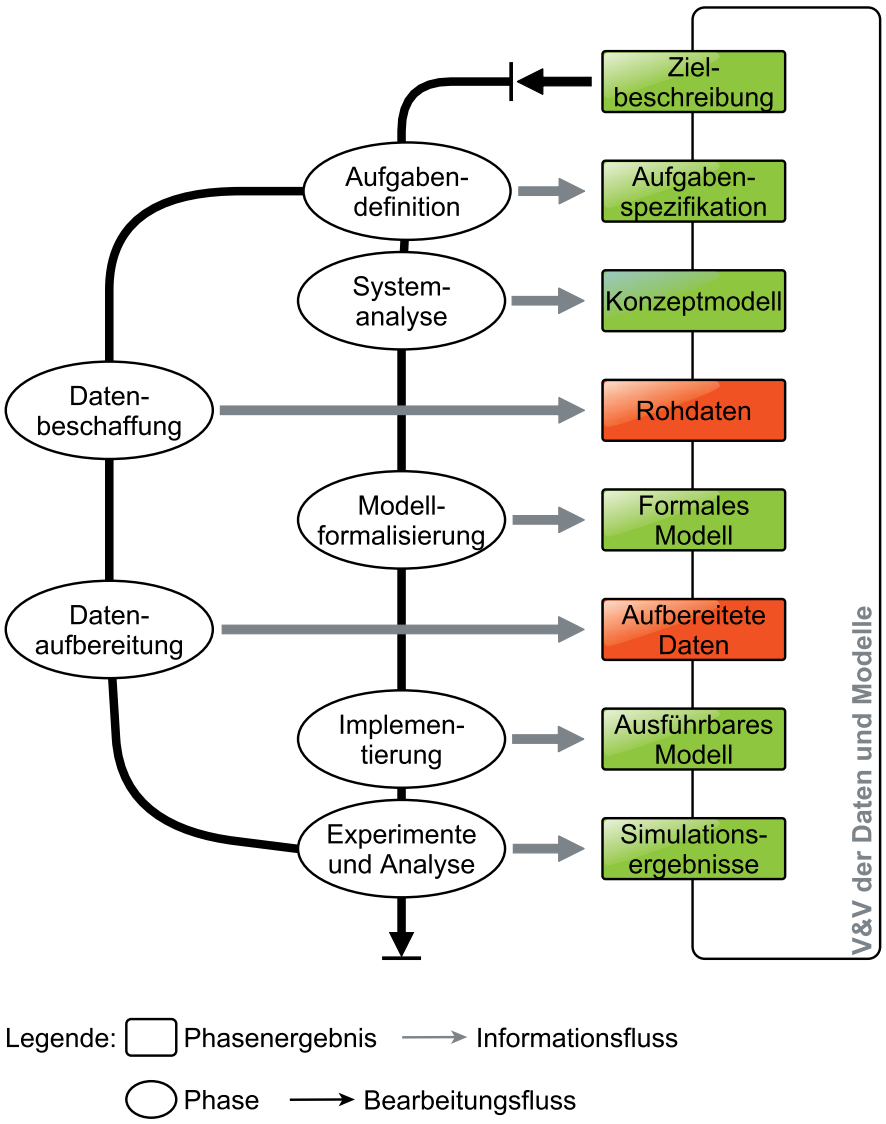



**Abbildung 2.11: ASIM-Vorgehensmodell nach Rabe, Spieckermann et al. (2008, S. 5)**

Dementsprechend reicht ein Muster bzw. eine Menge von Mustern im Regelfall nicht aus, um die vollständige Ausführungslogik der DES festzulegen. Selbst wenn eine Vielzahl von Regeln gefunden wird, so besteht kein Anspruch auf eine vollständige Beschreibung des zu modellierenden Systems (vgl. Emergenz der SC in Abschnitt 2.2.2.3).

Eine Prüfung der identifizierten Wirkzusammenhänge ist jedoch in jedem Fall notwendig, da es sich bei den Mustern um sogenannte Artefakte handeln kann, die keinen realen Zusammenhängen entsprechen (Dasu und Johnson 2003). Um diese Artefakte zu identifizieren, kann das entsprechende SC-Szenario simuliert werden. Danach erfolgt die Prüfung, ob die Muster in dem Szenario erzeugt werden können. Hierbei fungiert die Simulation als globales Modell und das zu prüfende Muster

fungiert als lokales Teilmodell. Diese Konstrukte sind implizit der Standardliteratur zur Simulation zu entnehmen (z. B. in Banks 1998), allerdings fehlt es an konkreten Forschungsarbeiten zu diesem Thema. Eine Technik, die im Kontext der Simulation zum Einsatz kommt, ist das Data Farming (Brandstein und Horne 1998). Die Grundidee besteht in der Generierung von Datenbeständen mittels Simulation. Diese Datenbestände bieten die Möglichkeit, auch bei unzureichender Datenqualität (vgl. Abschnitt 2.2.2.4) Vorgehensmodelle des KDD anzuwenden. Data Farming bedeutet sinnbildlich das „Kultivieren“ von Daten mittels Simulationsszenarien und ein nachfolgendes „Ernten“ der Daten aus den Replikationen oder Simulationsexperimenten. Hierbei bedeutet die Replikation das Durchführen der Simulation mit gleichen Parametern, aber verändertem Startwert (Seed) für die Zufallszahlenerzeugung der Simulation. Das Simulationsexperiment hingegen umfasst neben den Replikationen auch das Durchführen mehrerer Simulationsläufe mit Variation der Stellgrößen. Den Farming-Vorgang beschreibt Sanchez (2014, S. 800-801) treffend mit: „they grow data from their models“ . Eine besondere Herausforderung des Data Farmings im Kontext der Wissensentdeckung liegt in den Aggregationsstufen, in denen das ausführbare Modell seine Ausgabegrößen erzeugt. Für eine intensivere Betrachtung der Datenaggregationsstufen im Data Farming sei auf Rabe und Scheidler (2014) und Rabe und Scheidler (2015) verwiesen.

Der Einsatz des Data Farmings verfolgt im Rahmen dieser Arbeit zwei Ziele. Das erste Ziel ist das Erzeugen einer Datenbasis als Testmenge für die V&V der aus den Echtdaten gewonnenen Muster. Zusätzlich können in der Simulation Effekte innerhalb des zu betrachtenden Systems isoliert untersucht sowie die Abhängigkeiten einzelner Merkmale untereinander analysiert werden. Das zweite Ziel ist das Erzeugen einer Datenbasis für die Anwendung des zu entwickelnden Vorgehensmodells bei unzureichender Datenlage, wie beispielsweise in Planungsphasen der SC. Die generelle Anwendbarkeit dieser Technik für SC-Szenarien wurde in Vorarbeiten am Fachgebiet ITPL erarbeitet und ist in Arndt (2014) sowie Rabe und Scheidler (2015) dokumentiert.

## 2.4 Handlungsbedarf und Abgrenzung

Aus den zuvor dargestellten Beiträgen zum Thema wird deutlich, dass die Entdeckung von Mustern in Form von Wirkzusammenhängen in SC-Datenbanken große Relevanz für eine Vielzahl von logistischen Anwendungen besitzt. Um das Wissen über Wirkzusammenhänge zur Verfügung zu stellen, ist zu untersuchen, welches KDD-Vorgehensmodell für das Umfeld der SC sich zur Wissensentdeckung eignet. Weiterhin ist zu untersuchen, wie Modellanpassungen, beispielsweise im Bereich einer modellbegleitenden V&V, vorgenommen werden können. Insbesondere muss im Vorgehensmodell in den Phasen der Vorverarbeitung und Methodenwahl auf spezifische SC-Problemfelder eingegangen werden. Die spezifischen Problemfelder bezüglich der Datenlage wirken u. a. auf die Segmentierung der Daten.

### 2.4.1 Thematische Abgrenzung und Randbedingungen

Im Rahmen dieser Arbeit werden konkrete Fragestellungen des SCM untersucht und mittels eines spezifischen Vorgehensmodells beantwortet. Dazu wurde der Betrachtungsgegenstand auf spezifische Typen der SC und Datenausprägungen begrenzt (vgl. Abschnitt 2.2.2.2). Das Wissen, das im Rahmen des Vorgehensmodells gewonnen wird, muss in geeignete Darstellungsformen überführt werden. Zu diesem Zweck ist der herkömmliche Musterbegriff der Informatik zu erweitern, um den Besonderheiten der SC gerecht zu werden. Im Anschluss werden zwei logistische Anwendungen vorgestellt und aufgezeigt, wie der Einsatz des zu entwickelnden Vorgehensmodells in der Praxis gestaltet werden kann. Diese Arbeit führt die bestehenden Gebiete SC bzw. SCM, die Techniken des KDD und der Simulation zusammen und schafft eine gemeinsame Begriffswelt.

Nicht Teil dieser Arbeit sind spezifische Datenstrukturen für SC-Daten (siehe Abschnitt 2.2.2.1) sowie deren Datenarchitekturen und Performancebetrachtung. Die Komplexität und der Umfang der vorliegenden Datenbestände erfordert ein angepasstes Vorgehen, jedoch wird bewusst der Begriff Big Data vermieden.

Für Big Data sind Anpassungen der Verfahren über diese Arbeit hinaus notwendig. Ein Ausblick darauf wird im Kapitel 7 gegeben.

Abschließend wird festgehalten, dass bestehende Algorithmen in geeigneter Weise kombiniert und angepasst werden, jedoch keine Neuentwicklung vorgenommen wird. Somit sind die eingesetzten Data-Mining-Verfahren nur Teil des Vorgehensmodells und nicht Schwerpunkt dieser Arbeit.

Sofern nicht anders angegeben, wurden alle Experimente unter RapidMiner (vgl. Mierswa et al. 2006 und RapidMiner Inc. 2016) und Plant Simulation (vgl. Siemens Industry Software GmbH 2016) durchgeführt. Die Experimente fanden auf den Datenbeständen von drei SCs statt, die im Anhang in den Tabellen B.1, B.2 sowie im Evaluierungskapitel unter Tabelle 6.1 schematisch beschrieben wurden. Eine explizite Darstellung der Daten ist aufgrund des Umfangs nur in Auszügen möglich. In Einzelfällen wurden detailliertere Auszüge der Datenbestände dokumentiert, sofern diese für das Verständnis der Modellentwicklung und Praxisevaluierung notwendig sind.

### 2.4.2 Forschungsfragen

Aus dem derzeitigen Stand der Wissenschaft werden Forschungsfragen abgeleitet. Diese Fragen bilden die Essenz einer „Methode zur Erschließung von Wissen aus Datenmustern in SC-Datenbanken“.

In Abschnitt 2.3.2 wurde der Musterbegriff und seine Entwicklung in der Literatur diskutiert. Es wurde aufgezeigt, dass der Musterbegriff überwiegend als Resultat der Wissensentdeckungsphase (Data Mining) von den Autoren verortet wird und

folglich für den SC-Kontext nicht hinreichend ist. Der Wissensbegriff wurde diskutiert (Abschnitt 2.1.2) und in den anschließenden Abschnitten aufgezeigt, dass insbesondere das Kontextwissen in der SC von entscheidender Bedeutung ist (Abschnitt 2.2.3.2) und somit über geeignete Musterbegriffe abgebildet werden muss.

*Forschungsfrage 1:* Welche Musterbegriffe sind im Rahmen der Wissensentdeckung in SC-Datenbanken notwendig?

Die Thematik der unterschiedlichen Vorgehensmodelle zur Wissensentdeckung wurde in Abschnitt 2.3.1 diskutiert. Durch die Diskussion motiviert, stellen sich die Fragen, welches der vorgestellten Vorgehensmodelle im Kontext von SC-Datenbanken geeignet ist und welche Adaptionen des Vorgehensmodells notwendig sind. Um diese Fragen fundiert zu untersuchen, wurde der Adressat des zu entdeckenden Wissens (SCM) sowie damit verbunden, mögliche Fragestellungen des SCM diskutiert, die mit Hilfe des Vorgehensmodells untersucht werden können (Abschnitt 2.2.3.1).

*Forschungsfrage 2:* Wie muss ein geeignetes Vorgehensmodell zur Wissensentdeckung in SC-Datenbanken aufgebaut sein?

Die Wissensentdeckung, insbesondere im Umfeld der Logistik, benötigt geeignete Formen der V&V. Wie unter Abschnitt 2.3.2 gezeigt, ist eine begleitende V&V in den aktuellen KDD-Vorgehensmodellen nicht integriert, obwohl Techniken wie das Dreiecksmodell hierzu Ansätze liefern. Im Kontext der Validierungstechniken wurde die Simulation als interdisziplinäre Technik eingeführt (Abschnitt 2.3.3) und Ansätze aufgezeigt, die Rückschlüsse gestatten, dass die Simulation zur generellen Mustervalidierung geeignet ist (Abschnitt 2.3.4).

*Forschungsfrage 3:* Wie kann eine begleitende V&V in das entwickelte Vorgehensmodell integriert werden?

In der SC existieren verschiedene Aufgabenfelder mit unterschiedlichen Zeithorizonten (Abschnitt 2.2.3.2). Insbesondere sind Teilnehmer häufig mit Situationen (z. B. in der Planung) konfrontiert, in denen Daten nicht oder zumindest nicht in ausreichender Qualität zur Verfügung stehen (Abschnitt 2.2.2.4). In diesem Fall muss das Vorgehensmodell eine Alternative für die fehlenden Daten anbieten, um in möglichst vielen Aufgabenfeldern zum Einsatz zu kommen. In Abschnitt 2.3.4 wurde diskutiert, dass Daten in verschiedenen Anwendungsfeldern mittels Simulation generiert werden können. Es wurde aufgezeigt, dass diese Methode auch in der angestrebten Wissensentdeckung genutzt werden kann, um Daten für die nachfolgenden Schritte des Vorgehensmodells zu generieren.

*Forschungsfrage 4:* Wie können Daten für das entwickelte Vorgehensmodell mit Hilfe der Simulationstechnik generiert und effizient bereitgestellt werden?

Abschließend kann konstatiert werden, dass die vorliegende Arbeit eine Methode zur Erschließung von Wissen entwickelt. Hierzu werden die notwendigen theoretischen Grundlagen (vgl. *Forschungsfrage 1*) sowie die Methode selbst (vgl. *Forschungsfrage 2*) ausgearbeitet. Zusätzlich werden Verfahrensschwächen in der V&V

aufgelöst (vgl. *Forschungsfrage 3*) und die Praxistauglichkeit in Bezug auf die Datenqualität verbessert (vgl. *Forschungsfrage 4*). Abbildung 2.12 zeigt die einzelnen Elemente der Arbeit und ordnet die vier Forschungsfragen den betroffenen Bereichen zu.

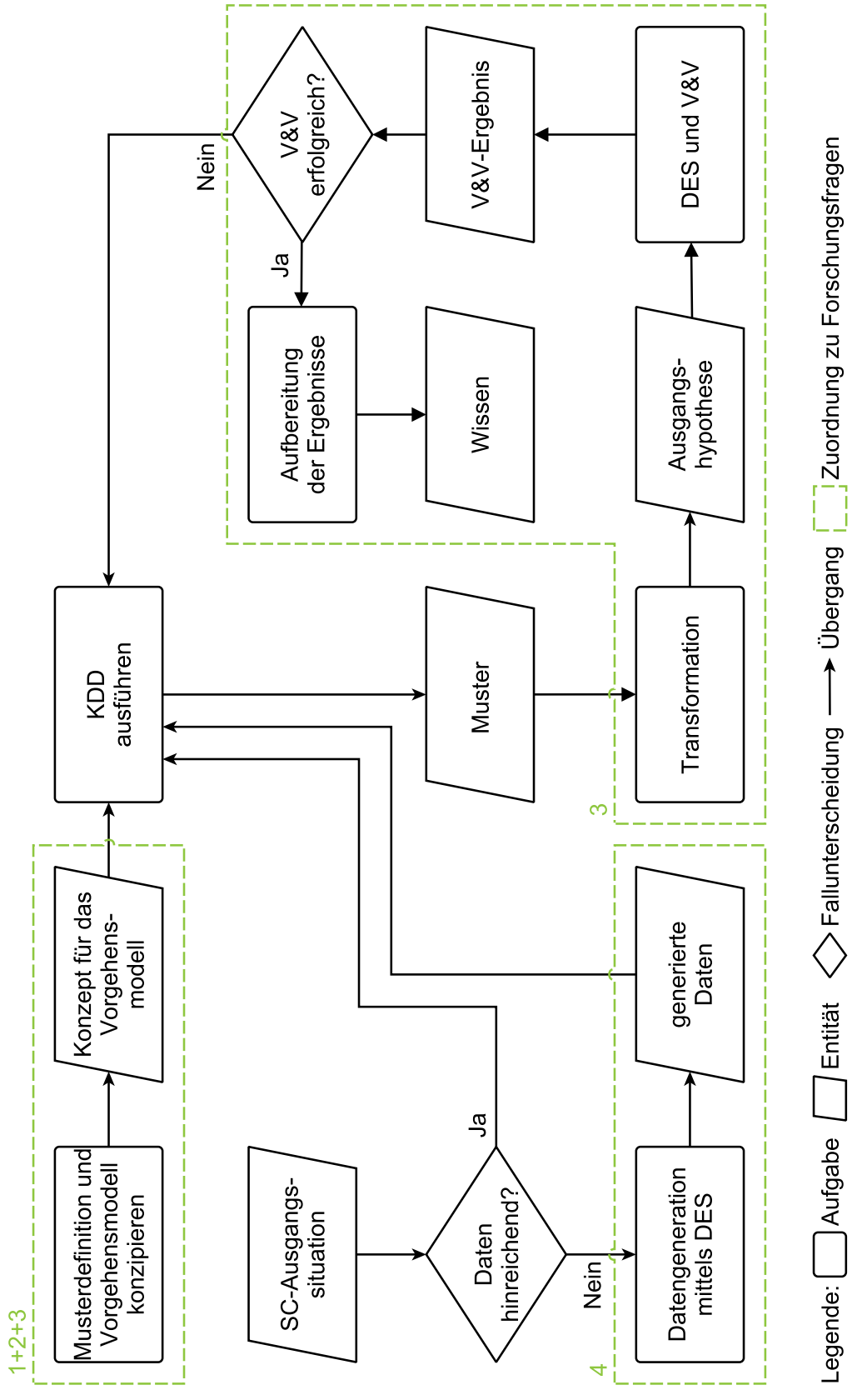



**Abbildung 2.12: Einordnung der vier Forschungsfragen**

# 3 Entwicklung einer Methode zur Wissensentdeckung in Supply-Chain-Datenbanken

Dieses Kapitel beschreibt die Entwicklung eines Vorgehensmodells zur Wissensentdeckung unter Berücksichtigung der Besonderheiten von SC-Datenbanken. Das Kapitel korrespondiert mit dem Abschnitt 2.4.2 und beantwortet die zuvor gestellte *Forschungsfrage 2:* Wie muss ein geeignetes Vorgehensmodell zur Wissensentdeckung in SC-Datenbanken aufgebaut sein? Des Weiteren wird die *Forschungsfrage 3* (Wie kann eine begleitende V&V in das entwickelte Vorgehensmodell integriert werden?) untersucht und ein Lösungsweg aufgezeigt.

Hierzu werden im ersten Schritt die spezifischen Anforderungen an ein Vorgehensmodell im SC-Umfeld aufgestellt. Im Anschluss wird, basierend auf den erhobenen Anforderungen, ein Vorgehensmodell als Grundlage für die eigene Entwicklung ausgewählt. Danach werden spezifische Veränderungen und Erweiterungen aufgezeigt, die für dieses Vorgehensmodell notwendig sind.

## 3.1 Methodenkonzeptionierung

Dieser Abschnitt erläutert die Anwendung des Methodenbegriffs in der vorliegenden Arbeit. Eine Methode besteht aus verschiedenen Methodenelementen, die in ihrem Zusammenwirken ein planmäßiges Vorgehen ermöglichen, um im Vorfeld festgelegte Ziele zu erreichen (vgl. Einleitung zu Kapitel 2). Das Vorgehensmodell als zentrales Element einer Methode bildet auch in der vorliegenden Arbeit den methodischen Kernaspekt. Das Ziel der in dieser Arbeit entwickelten Methode ist die Wissensentdeckung in SC-Datenbanken. Für diese Wissensentdeckung ist jedoch ein Vorgehensmodell nicht ausreichend (Abschnitt 2.4). Aus diesem Grund wird das Vorgehensmodell um zusätzliche Methodenelemente erweitert. Hierzu zählt die Integration einer V&V für die Vorgehensmodell-Phasen sowie eine Möglichkeit der Datengenerierung bei unzureichender Datenlage in der SC (vgl. Forschungsfragen 3 und 4 in Kapitel 2). Es gibt verschiedene Möglichkeiten, Methodenelemente miteinander zu verbinden. Die Entwicklungstheorie hinter der Methodenkonzeption liegt aber nicht im Aufgabenbereich dieser Arbeit und wird demzufolge nicht weiter diskutiert. In dieser Arbeit wird ein integrativer Ansatz in der Konzeptionierung verfolgt, d. h. aufbauend auf dem Kernelement wird die Methode schrittweise um neue Methodenelemente erweitert. Sowohl die V&V als auch die Datengenerierung werden auf die spezifischen Anforderungen des entwickelten Vorgehensmodells angepasst und fungieren demzufolge als Erweiterung des Vorgehensmodells. Für die Konzeption der Methode bedeutet dies, dass das Vorgehensmodell das erste Me-

thodenelement ist. Dieses wird sukzessive um die zuvor aufgeführten Elemente erweitert, sodass am Abschluss der Methodenentwicklung mehrere Methodenelemente integriert wurden.

Wenn am Anfang der Konzeptionierung die Methode nur aus dem zugrundeliegenden Vorgehensmodell besteht, sind die Methode und das Methodenelement des Vorgehensmodells äquivalent. Die Begriffsverwendung von Vorgehensmodell und Methode in dieser Arbeit spiegeln den erläuterten Sachverhalt wieder. So wird, wie der Titel der Arbeit anzeigt, zwar auf übergeordneter Ebene eine Methode entwickelt, jedoch basiert die Entwicklung auf einem Vorgehensmodell. Um diesen Aspekt hervorzuheben, wird auch von einer Erweiterung des Vorgehensmodells gesprochen und nicht von einer Erweiterung der Methode. Die Methode wird entwickelt und ist nach Integration aller Methodenelemente vollständig. Aus diesem Grund wird in einzelnen Abschnitten das Vorgehensmodell in den Vordergrund gestellt und vorausgesetzt, dass hiermit immer das Vorgehensmodell im Bezug auf die entwickelte Methode adressiert wird. Die Methode wird immer dann als Begriff verwendet, wenn eine Entwicklung in dem übergeordneten Strukturrahmen dieser Arbeit angezeigt ist. Dies ist beispielsweise der Fall, wenn neue Methodenelemente integriert werden und das Vorgehensmodell erweitert wird.

Die verwendeten Begriffe werden teilweise kontrovers in unterschiedlichen Disziplinen verwendet. Es wird ausdrücklich darauf hingewiesen, dass die verwendeten Begriffe sich an der Literatur der Einleitung zu Kapitel 2 orientieren.

## 3.2 Ableitung der wissensbezogenen Charakteristika von Supply Chains

Die Anforderungen, die sich aus Sicht der SC an ein Vorgehensmodell ergeben, werden aus verschiedenen Einflussfaktoren abgeleitet. Ein wesentlicher Einflussfaktor sind die zugrundeliegenden Datenbanken der SC, die in Abschnitt 2.2.2 diskutiert wurden. Ein weiterer Einflussfaktor ist das zu entdeckende Wissen selbst (vgl. Abschnitt 2.2.3.2 und dessen Vorbedingung in Form von Kontextwissen). Zusätzlich werden allgemeine Anforderungen berücksichtigt, die sich aus der bisherigen wissenschaftlichen Diskussion der KDD-Vorgehensmodelle ergeben (vgl. Abschnitt 2.3.1).

### 3.2.1 Charakteristika von Supply-Chain-Datenbanken

Die Datenbank zur Wissensentdeckung wird hauptsächlich von den sogenannten Transaktionsdaten gebildet. Die dazugehörigen Stammdaten können Kontextinformationen aufgrund ihrer inhaltlichen Bedeutung liefern, gehen jedoch nicht vollumfänglich in die Datenbasis der Wissensentdeckung ein (vgl. Abschnitt 2.2.2.2). Untersuchungen unterschiedlicher Datenbankbestände haben fünf Charakteristika

in den SC-Daten aufgezeigt, die in dieser Form nicht auf andere Datenbankbestände zu übertragen sind.

Das erste Charakteristikum ist die zwingende Einbindung von *Kontextwissen*, die im SC-Umfeld einen großen Stellwert einnimmt (vgl. Abschnitt 2.1.2). Dies liegt zum einen an dem hohen Automatisierungsgrad der SC-Prozesse, die durch die vollautomatischen Erfassungen über Sensoren eine Vielzahl von technischen Werten protokollieren. Diese technischen Werte, die über separate Attribute gehalten werden, sind ohne Kontextwissen nicht zu interpretieren. Des Weiteren sind an einer globalen SC eine Vielzahl von Partnern beteiligt. Diese Partner verwenden oftmals eigene Standards, sodass das selbe BO in unterschiedlicher Weise über Attribute kodiert wird. Hierzu ist Kontextwissen in Form der unterschiedlichen Kodierungsstandards notwendig. Als letzter Punkt sind manuelle Eingriffe zu nennen, die sich aus den Systemstrukturen ergeben. In der Praxis werden beispielsweise Lieferänderungen in dringenden Fällen über den Telefonweg kommuniziert und der entsprechende Sachbearbeiter veranlasst ein Update der dazugehörigen Datensätze. Das Update wird zwar auf Tabellenebene protokolliert, jedoch findet sich in den Referenztabellen keine entsprechende Auftragsmitteilung. Obwohl das Kontextwissen als eigenes Charakteristikum auf Interpretationsebene aufgeführt wird, sind auch die nachfolgenden Charakteristika der SC-Daten in Verbindung mit Kontextwissen zu sehen.

Das zweite Charakteristikum der SC-Daten ergibt sich aus den Wechselwirkungen der einzelnen Transaktionen über größere Zeiträume. Dieser Aspekt wurde bereits in Abschnitt 2.3.2 unter dem Aspekt der *Datenauswahl* mit dem Stichwort Fensterung thematisiert. Da die Fensterungen variable Parameter aufweisen, muss für jede SC bestimmt werden, in welchen Zeiträumen nach Wechselwirkungen gesucht werden soll. Es gibt Anhaltspunkte, die sich beispielsweise aus dem Produktlebenszyklus oder den Bestellhäufigkeiten ergeben, doch letztendlich muss immer eine separate Vorabanalyse in die Vorgehensmodelle integriert werden.

Das dritte Charakteristikum ist in der Struktur der SC begründet, denn bei komplexeren Netzwerken gibt es nicht mehr den einen Transportweg, der als quasi Transaktionsstandard fungiert. Vielmehr muss mittels Kontextwissen über die SC-Strukturen bestimmt werden, wie die Datenbasis aufgeteilt werden muss, um eine Vergleichbarkeit der Datensätze zu gewährleisten. Der Aspekt der *Gruppierung* steht in enger Wechselwirkung mit der Fragestellung der Wissensentdeckung. Soll beispielsweise eine Verspätung entdeckt werden, so wären mögliche Unterteilungskriterien bei mehreren Alternativrouten der Transportweg oder das benutzte Transportmittel. In Abbildung 3.1 ist eine beispielhafte Gruppierung von Transaktionen nach Transportwegen dargestellt.

Das vierte Charakteristikum ist in der Attributsspeicherung der relationalen Datenbanken begründet. Sofern eine Datenbasis *atomare Attribute* aufweist (wie es im Idealfall der relationalen SC-Datenbanken gegeben sein sollte), werden diese Attribute im Normalfall nicht weiteren Untersuchungen unterzogen, da atomare At-

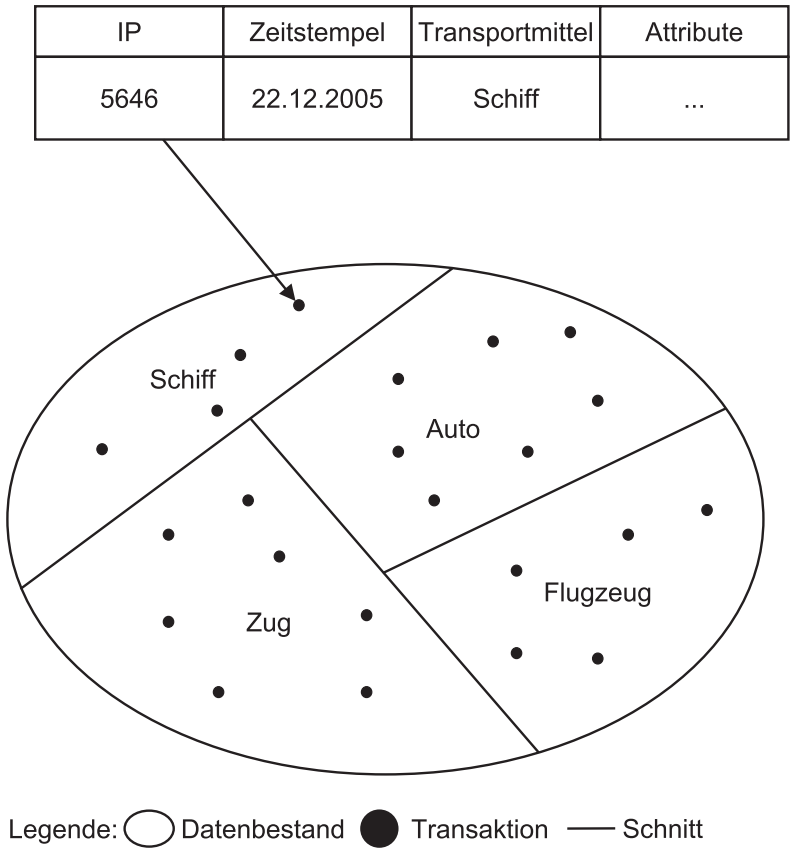



**Abbildung 3.1: Gruppierung eines Datenbestands nach Transportwegen**

tribute per Definition als unteilbar gelten. SC-Datenbanken weisen jedoch oftmals Attribute mit einer inneren Struktur auf, so dass eine weitere Aufspaltung dieser Attribute für die Wissensentdeckung sinnvoll sein kann. Diese Attribute kommen im SC-Kontext beispielsweise als sprechende Artikelnummern vor und kodieren dann Artikelmerkmale. Diese Kodierung stellte insbesondere vor der Verbreitung der relationalen Datenbanken einen Mehrwert für das SCM dar. Im Anhang finden sich Auszüge aus untersuchten Datenbeständen, in denen der Sachverhalt anhand der Fabrikationsnummer nachvollzogen werden kann (Tabellen B.3 und B.4). Die mögliche Aufsplittung der Fabrikationsnummer in den Daten ist in Abbildung 3.2 dargestellt. Dies muss in der Vorverarbeitung berücksichtigt werden, da eine Suche über Fabrikationsnummern andere Zusammenhänge aufzeigen kann als eine Suche über Teile der Fabrikationsnummer.

Das fünfte Charakteristikum lässt sich sich im Kategorisierungssystem der SC-Daten unter dem Kriterium *Format* verordnen (vgl. Abschnitt 2.2.2.2). In den SC-Datenbeständen weisen eine Vielzahl von Attributen alphanumerische Formate auf. Das liegt in den dazugehörigen BOs begründet, da die Merkmale der BOs oftmals nicht über rein numerische Werte ausgedrückt werden (beispielsweise Standorte oder Lieferanten). Die Wertebereiche der Attribute lassen sich mittels verschiedener Techniken auf numerische Werte projezieren, allerdings müssen hier geeignete

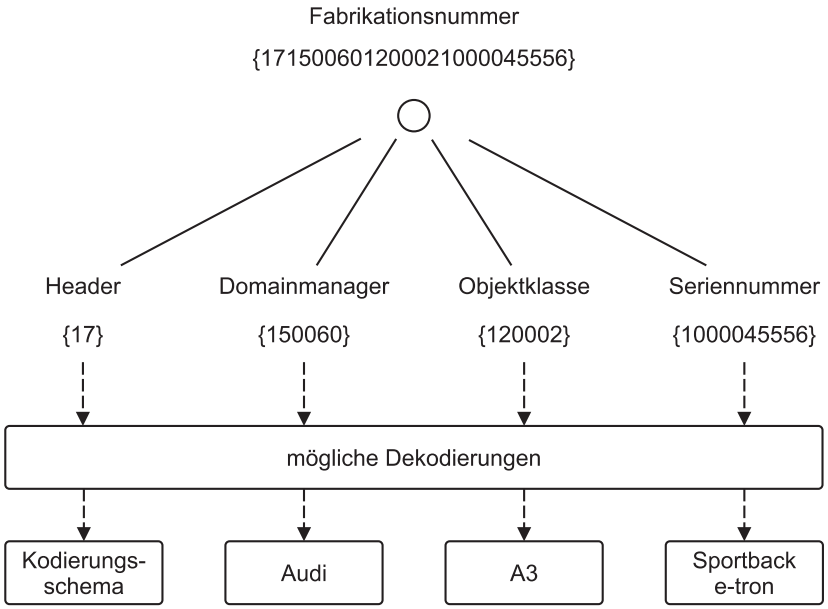



**Abbildung 3.2: Beispielkodierung von Produkten im SC-Umfeld**

Mapping-Verfahren entwickelt werden. Beispielsweise ist das Mappen von 50 000 Produkten auf eine numerische Skala wenig sinnvoll. Die Schwierigkeit der Mapping-Verfahren wird dadurch erhöht, dass auf die meisten numerischen Attribute der SC keine fachlich sinnvolle Metrik angewandt werden kann. Dies wird am Beispiel verschiedener Standorte deutlich. Zwar können künstliche Metriken wie Entfernungen zum Zentrallager erzeugt werden, doch sind diese für die Wissensentdeckung nicht notwendigerweise förderlich. Aus diesem Grund sind die SC-Daten bezüglich der Skalenart unter nicht-metrisch einzuordnen (vgl. Abbildung 2.9), was zu separaten Bearbeitungsschritten und der Definition von geeigneten Metriken für SC-Attribute führt. Innerhalb der Transformationsphase in Abschnitt 2.3.1 wurden bereits Metriken diskutiert. Da die Metriken größtenteils funktionaler Natur sind und die fachspezifischen Zusammenhänge zwischen den Daten nicht erfassen können, sind hier geeignete fachliche Metriken zu entwickeln.

### 3.2.2 Charakteristika des Supply-Chain-Wissens

Die nachfolgenden Charakteristika beziehen sich auf das Wissen, das es zu entdecken gilt sowie die logistischen Anforderungen an ein Vorgehensmodell. Das Wissen, das es in der SC zu entdecken gilt, umfasst sämtliche Kernaufgaben des KDD (vgl. Tabelle 2.7). Daher kann von diesem Standpunkt ausgehend keine Einschränkung getroffen werden. Aus Sicht des SCM als stellvertretender Adressat des Wissens ist jedoch eine *explizite Darstellung* des Wissens förderlich (vgl. Phase 9 in Abschnitt 2.3.2), um das Wissen unmittelbar in das betriebliche Umfeld zu integrieren. Die unmittelbare Integration ermöglicht das Ableiten von Handlungsempfehlungen durch das SCM, ohne dass andere Bewertungssysteme eingebunden

werden müssen. Dies hat unmittelbare Auswirkungen auf die verwendeten Data-Mining-Verfahren, denn nur bestimmte Verfahren gestatten direkt eine explizite Darstellung des gefundenen Wissens. Bei den anderen Verfahren ist eine Überführung des entdeckten Wissens in eine geeignete Darstellungsform notwendig. Im Falle einer Überführung muss ein entsprechender Zwischenschritt in den Vorgehensmodellen integriert werden, um mittels maschinell implementierter Algorithmen die einzelnen Elemente der Wissensdarstellung über ein geeignetes Metamodell in eine andere Darstellungsform zu transformieren. Aus den vorgestellten Gründen wird die explizite Darstellung des Wissens bzw. die notwendige Transformation von implizitem Wissen als sechstes Charakteristikum definiert.

Als weiteres Charakteristikum kann der *Geltungsbereich* des Vorgehensmodells aufgeführt werden. Der Geltungsbereich ergibt sich durch den Hauptnutzer des Vorgehensmodells, der in der vorliegenden Arbeit durch das SCM repräsentiert wird (vgl. Abschnitt 2.2.3.1). Der Geltungsbereich umfasst die Extraktion der Daten aus einem gegebenen System und endet mit der Bereitstellung der evaluierten Muster. Somit wird der Geltungsbereich, der durch das SCM bedingt wird, als siebtes Charakteristikum festgelegt. Dies ist insbesondere wichtig, da einige Vorgehensmodelle aus den Tabellen 2.4 und 2.6 einen diesbezüglich ungünstigen Bereich abdecken. Hierzu zählen beispielsweise die Modelle von Anand oder Cabena, die Positionen wie Mitarbeiteridentifikation oder BO-Bestimmung als Einzelschritte in die Modelle integriert haben.

Als achtes Charakteristikum kann die *Granularität* der Modellphasen angeführt werden. So ist es für das SCM sinnvoll, einen angemessenen Detailierungsgrad für die Phasen des Modells zu wählen. Dieser muss zum einen eine ausreichende Phasenanzahl aufweisen, um eine zeitliche Projektkontrolle sowie eine Überprüfung von Zwischenergebnissen zu gestatten. Zum anderen darf das Modell jedoch nicht zu detailliert konzipiert werden, sodass als Folge die praktische Durchführung der Wissensentdeckung zu komplex für das SCM wäre. Die Modellgranularität ist ein wichtiger Aspekt in der Wissensentdeckung, da es sich bei der Ausführung des KDD-Vorgehensmodells um einen iterativen und interaktiven Prozess handelt (vgl. Abschnitt 2.3). Dies impliziert eine ausreichende Menge an klar definierten manuellen Eingriffspunkten für die Interaktion bzw. Sprungstellen für den iterativen Prozess. Modelle mit beispielsweise nur drei Phasen oder ungünstigen Bündelungen von Interaktionen in den einzelnen Schritten verfügen über keinen ausreichenden Detaillierungsgrad und sind in der Folge für das SCM ungeeignet.

Als neuntes Charakteristikum kann die benötigte *Dokumentation* aufgeführt werden. Sowohl die Schrittbezeichnung als auch die Dokumentation der einzelnen Schritte sind in diesem Kontext von Bedeutung. Da die Wissensentdeckung im Bereich der SC prinzipiell in bestehende Business-Landschaften einzubetten ist, ist eine sinnvolle Bezeichnung der Phasen und Schritte sowie deren Dokumentation zwingend notwendig. Neben unzureichend dokumentierten Modellen, die bereits im Abschnitt 2.3.1 verworfen wurden, wird hier der Fokus auf Integration in die

Praxis gelegt. Generische Bezeichnungen für die Phasen wie „Taking Action“ sind in diesem Zusammenhang weniger geeignet für den Einsatz im Kontext der SC.

## 3.3 Ableiten der Anforderungen an ein Vorgehensmodell zur Wissensentdeckung

Auf Grundlage der aufgestellten Charakteristika werden nun die Anforderungen an ein Vorgehensmodell zur Wissensentdeckung abgeleitet. Die entsprechenden Anforderungen wurden mit den Phasen des Vorgehensmodells von Fayyad et al. verglichen, um die Auswirkungen auf einzelne Phasen des Vorgehensmodells einzugrenzen (vgl. Abschnitt 2.3.2). Die Ergebnisse sind in Tabelle 3.1 dargestellt und dienen als Grundlage für die Auswahl eines Vorgehensmodells im folgenden Abschnitt.

## 3.4 Auswahl eines Vorgehensmodells zur Wissensentdeckung

Die aufgelisteten neun Anforderungen werden nun mit den verbleibenden 15 Modellen aus Tabelle 2.4 und Tabelle 2.6 abgeglichen. Die Ergebnisse können in Tabelle 3.2 nachvollzogen werden.

Das Einbinden von *Kontextwissen* erfordert sowohl seine Identifikation als auch eine passende Kodierung für dieses Wissen. Betrachtet man die Modellphasen, so weisen alle Modelle vor dem eigentlichen Data Mining einen Schritt der Zielbeschreibung oder Anforderungsanalyse sowie geeignete Vorverarbeitungsphasen auf. Lediglich die Modelle von Adriaans und Zantinge, Cooley et al. und Runkler beginnen mit der sofortigen Datenauswahl bzw. Datenvorverarbeitung und erfüllen somit die Anforderung nicht.

Die Anforderung der geeigneten Datenauswahl mit Berücksichtigung auf die SC-Besonderheiten wird von drei Modellen nicht erfüllt. Berry und Linoff benennen keine separate Phase für die Datenauswahl, Cooley et al. beginnen bereits mit der Vorverarbeitung und setzen eine gültige Stichprobe voraus und Reinartz und Wirth adressieren diese Phase nur indirekt unter der Bezeichnung „access data“.

Die Anforderungen 3 und 4, Gruppierung und Atomarität, können in den Modellen über geeignete Vorverarbeitungsschritte der Daten erfüllt werden. Eine Form der Vorverarbeitung weisen alle Modelle auf, jedoch variiert der Umfang der Vorverarbeitung. Des Weiteren liegt bei einigen Modellen der Schwerpunkt der einzelnen Vorverarbeitungsschritte nicht auf möglichen Attributsprüfungen oder Datengruppierungen. Dies wird insbesondere bei den Modellen deutlich, die statt der allgemeinen Vorverarbeitung die Datenbereinigung als Fokus haben. Hierzu zählen die

**Tabelle 3.1: Anforderung an ein Vorgehensmodell zur Wissensentdeckung im SC-Kontext**

| Nr. | Kurz-bezeichnung | Anforderung | Auswirkung auf die Phasen des Modells von Fayyad et al. |
|---|---|---|---|
| 1 | Kontextwissen | Kontextwissen muss zwecks Datenauswahl und Datenkodierung integriert werden | Phase 1-4 |
| 2 | Datenauswahl | Transaktionsdaten müssen aufgrund ihrer zeitlichen Bezüge ausgewählt werden | Phase 2 |
| 3 | Gruppierung | Gruppierung der ausgewählten Daten zur Weiterverarbeitung durch die Data-MiningVerfahren | Phase 3-4 |
| 4 | Atomarität | Prüfung, ob Attribute innere Struktur aufweisen und somit weiter aufgespalten werden können | Phase 4 |
| 5 | Format | Mapping-Verfahren für SC-Daten initialisieren und Metriken definieren | Phase 5 |
| 6 | Explizität | Explizite Darstellungsform der validierten Muster über geeignete Data-Mining-Verfahren oder extra zu initialisierenden Post-Processing-Schritt innerhalb von Schritt 9 | Phase 7 und 9 |
| 7 | Geltungsbereich | Geeigneter Geltungsbereich für das Vorgehensmodell | Phase 1 bis 9 |
| 8 | Granularität | Ausreichende Menge an Schritten, um Interaktion und Iteration zu gewährleisten | Phase 1 bis 9 |
| 9 | Dokumentation | Sinnvolle Bezeichnung der Phasen und Schritte sowie existierende Dokumentation für Schritte | Phase 1 bis 9 |

Modelle von Adriaans und Zantinge mit der Phase „Cleaning Enrichment“ und Brachmann und Anand mit der Phase „Data Cleaning“. Das Modell von Reinartz

**Tabelle 3.2: Überprüfung der gängigen Vorgehensmodelle bezüglich der SC-Anforderungen**

| | Kontextwissen | Datenauswahl | Gruppierung | Atomarität | Format | Explizität | Geltungsbereich | Granularität | Dokumentation |
|---|---|---|---|---|---|---|---|---|---|
| 5 A's | ● | ● | ● | ● | ○ | ○ | ○ | ● | ○ |
| Adriaans und Zantinge | ○ | ● | ◐ | ◐ | ● | ◐ | ● | ● | ◐ |
| Anand und Bucher | ● | ● | ● | ● | ◐ | ◐ | ○ | ● | ● |
| Berry und Linoff | ● | ○ | ● | ● | ○ | ◐ | ● | ○ | ◐ |
| Brachman und Anand | ● | ● | ◐ | ◐ | ○ | ◐ | ● | ● | ● |
| Cabena et al. | ● | ● | ● | ● | ○ | ● | ● | ● | ● |
| Cios et al. | ● | ● | ● | ● | ○ | ● | ● | ● | ● |
| Cooley et al. | ○ | ○ | ● | ● | ○ | ◐ | ● | ○ | ◐ |
| Crisp-DM | ● | ● | ● | ● | ○ | ● | ● | ● | ● |
| Fayyad et al. | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| Hippner und Wilde | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| KDD Roadmap | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| Petersohn | ● | ● | ● | ● | ● | ◐ | ● | ● | ● |
| Reinartz und Wirth | ● | ○ | ○ | ● | ● | ● | ○ | ◐ | ○ |
| Runkler | ○ | ● | ● | ● | ○ | ◐ | ○ | ● | ◐ |
| SEMMA | ● | ● | ● | ● | ○ | ◐ | ● | ◐ | ● |
| Wrobel et al. | ● | ● | ● | ● | ● | ● | ● | ● | ● |

**Legende:**
● Erfüllt ◐ Teilweise erfüllt ○ Nicht erfüllt

und Wirth führt zwar unter Manipulation das Preprocessing an, diskutiert den Begriff aber über eine Nennung hinaus nicht.

Die Anforderung, spezifische Mapping-Verfahren für SC-Daten zu initialisieren, bedeutet einen separaten Vorverarbeitungsschritt für die Data-Mining-Verfahren. Dieser kann in der Auswahl der Data-Mining-Verfahren als Phase integriert sein, so wie es bei beispielsweise bei den Modellen von Hippner und Wilde oder Brachman und Anand der Fall ist. Alternativ kann ein expliziter Schritt für diese Aufga-

be existieren, sowie es im Modell von Petersohn vorzufinden ist. Andere Modelle wie beispielsweise SEMMA, das mit der Phase „Modify“ sowohl die Vorverarbeitung der Daten, als auch die dazugehörende Auswahl, Reinigung, Transformation und Zielformatierung der Daten beschreibt, weisen keine separaten Phasen oder ausgezeichnete Schritte auf.

Die Anforderung der Explizität kann auf zwei Weisen erfüllt werden. Zum einen können die eingesetzten Data-Mining-Verfahren beschränkt werden, um nur solche Verfahren zu integrieren, die explizite Darstellungen erzeugen. Dies ist jedoch eine starke Einschränkung und nimmt dem Vorgehensmodell eine großen Teil seiner Flexibilität bezüglich der zu bearbeitenden Zielstellungen. Zum anderen kann eine explizite Nachbereitung der evaluierten Muster integriert werden. Zu diesem Zweck werden die Vorgehensmodelle auf ihren Abschluss und die Möglichkeit, weitere Analysen zu integrieren, geprüft. Die aufgeführten Modelle besitzen stark unterschiedliche Ausprägungen in Bezug auf die Nachbereitungsphase des Data Minings. Die Modelle nach Fayyad et al., Cabena et al., Cios et al., Hippner und Wilde, Wrobel et al. und Crisp-DM haben gemeinsam, dass nach dem eigentlichen Data Mining noch zwei abschließende Phasen folgen. Diese Phasen umfassen die Interpretation und fachliche Bewertung der Muster und zeigen im Anschluss deren Bereitstellung und Anwendung auf. Die Modelle nach Anand und Bucher, Adriaans und Zantinge, Anand und Brachmann, Berry und Linoff, Cooley et al., Petersohn, Runkler und SEMMA schließen mit einer Interpretationsphase ab, in die theoretisch eine Darstellungstransformation zu integrieren wäre, jedoch nicht explizit vorgesehen ist. Das Modell 5 A’s bietet mit der Abschlussphase „Automate“ keinen verwertbaren Ansatzpunkt.

Der Geltungsbereich der Modelle umfasst im Wesentlichen die Gesamtspanne der Wissensentdeckung. Jedoch fassen einige Modelle den Fokus zu weit, indem sie beispielsweise die Identifikation von Experten integrieren (Anand und Bucher) oder decken nicht die gesamte Spanne ab, indem sie die Nachbereitung vernachlässigen (5 A’s). Das Modell von Reinartz und Wirth wird durch seinen Aufbau aus einer Vielzahl von Einzelkomponenten gekennzeichnet, die auch Nebenaspekte beleuchten. Beispielsweise wird in der Phase „Requirement Analysis“ eine Applikationsbeschreibung mit Unterpunkten wie eine Domainbeschreibung gefordert. Die Anforderung der Modellgranularität zur Ermöglichung von Iterationen und Interaktionen wird von der überwiegenden Anzahl von Modellen erfüllt. Nur die Modelle von Berry und Linoff sowie das von Cooley et al. weisen eine zu geringe Phasenzahl für angemessene Interaktionen auf. Reinartz und Wirth setzen in ihrem Modell den Fokus auf die einzelnen Komponenten und nicht auf deren Interaktion. Diese Komponenten sind auf gleicher Ebene ohne zeitlichen Bezug angeordnet. Diese Darstellungsweise lässt zu viel Raum für Interpretation und vernachlässigt den iterativen Modellcharakter. Die Modelle von Fayyad et al., Runkler, Brachmann und Anand und Crisp-DM führen jedoch explizit die Iteration und die damit verbundenen möglichen Rücksprünge in den Modellen auf. Im Modell nach Hippner und Wilde wird eine Standardreihenfolge für den Ablauf vorgeschlagen, doch ausdrück-

lich auf mögliche Reihenfolgevariation und Iterationen hingewiesen. Rücksprünge sind auch im Modell SEMMA möglich, jedoch erfolgt hier die Einschränkung auf spezifische Schritte und eine Interaktion zwischen einzelnen Schritten ist nicht Teil des Modells.

Die letzte Anforderung beinhaltet die Existenz sinnvoller Bezeichnungen für Phasen und deren Schritte sowie deren vorhandene Dokumentation. Dieses Kriterium wird von den meisten Modellen erfüllt, wobei weit verbreitete Modelle im Allgemeinen besser dokumentiert sind als spezifische, eher unbekannte Modelle. Die sinngemäße Benennung der einzelnen Phasen und die damit verbundenen Erklärungen sind bei einigen Modellen eher schwach ausgeprägt. Aus diesem Grund werden die Modelle von Adriaans und Zantinge, Berry und Linoff, Cooley et al., Reinartz und Wirth und Runkler abgewertet. Das Modell 5 A's genügt mit seinen generischen Bezeichnern wie „Act" den Ansprüchen an die „sprechenden Bezeichnungen" für das fachlich ausgerichtete SCM nicht, da sie zu viel Raum für Interpretation bieten.

Neben den aufgestellten Anforderungen fließen weitere Überlegungen in die Modellfindung ein. So muss berücksichtigt werden, dass die Modelle SEMMA der Firma SAS und 5 A's der Firma SPSS eine unternehmensspezifisch geprägte Sichtweise auf die Wissensentdeckung einnehmen und diese Modelle somit nur bedingt als Grundlage für ein generisches Modell im SC-Umfeld geeignet sind. Letztendlich zeigt die Tabelle, dass nur die Modelle von Fayyad et al., Hippner und Wilde, KDD Roadmap sowie Wrobel et al. die aufgestellten Anforderungen vollständig abdecken. Bezieht man die Tatsache ein, dass die Unterscheidung von Preprocessing und Transformation fachspezifisch in der Informatik begründet ist (vgl. Abschnitt Transformation in Abschnitt 2.3.2) und für das SCM somit in der Praxis irrelevant, werden die Modelle von Fayyad et al. sowie die KDD Roadmap abgewertet. Der hohe Detaillierungsgrad der Modelle ist für den praktischen Einsatz im SC-Umfeld ungeeignet, da die Fachanwender aus der Logistik oftmals die einzelnen Vorverarbeitungsschritte nicht unterscheiden können. Vergleicht man die beiden verbleibenden Modelle von Hippner und Wilde sowie Wrobel et al., muss die Entscheidung zugunsten des Models von Hippner und Wilde gefällt werden. Hippner und Wilde haben ihr Vorgehensmodell mit sieben Phasen (vgl. Abschnitt 2.3.1) in weitere Schritte untergliedert. Dadurch ist eine bessere Strukturierung gegeben und mögliche Modellerweiterungen können im Vorgehensmodell verankert werden. Dieses Vorgehen ist bei Wrobel et al. nicht möglich, da die Phasen des Modells nicht weiter strukturiert sind und im Wesentlichen eine Menge von möglichen Handlungen in nicht näher spezifizierter Reihenfolge darstellen. Dieser Unterschied lässt sich dadurch begründen, dass Hippner und Wilde eines der wenigen Modelle für den konkreten Praxiseinsatz entwickelt haben und somit ihr Modell eine geeignete Grundlage für ein praxisorientiertes Vorgehensmodell zur Wissensentdeckung im SC-Kontext bildet. Tabelle 2.3 gibt das Modell von Hippner und Wilde mit den Phasen und Schritten sowie deren Beschreibung wieder.

## 3.5 Anpassung und Erweiterung des Vorgehensmodells von Hippner und Wilde

Das vorgestellte Modell von Hippner und Wilde aus der Marketingbranche (vgl. Tabelle 2.3) dient als Referenzmodell (RM) für das Vorgehensmodell im SC-Bereich und wird im Nachfolgenden schrittweise untersucht, auf den SC-Bereich transformiert und auf mögliche Erweiterungen geprüft.

Phase 1 wurde im RM in drei Schritte unterteilt, die von der Bestimmung der Problemstellung bis hin zur Projektplanung reichen. Diese Betrachtungsweise setzt den Projektfokus in den Vordergrund und bezieht projekttechnische Randdaten wie beispielsweise Risiken mit ein. Insbesondere ist auffällig, dass die separat aufgelisteten Aspekte, wie die Formulierung der Zielkriterien oder die Festlegung der Datenanalyseaufgabe, eine rein organisatorische Trennung im RM darstellen und nicht fachlich motiviert sind. Das zu entwickelnde Vorgehensmodell für die SC setzt den Fokus aber auf das fachliche Vorgehen und fasst in der Schlussfolgerung den Geltungsbereich enger. In der Folge werden die Schritte aus Phase 1 aggregiert und der Fokus der transformierten Phase auf die Aufgabendefinition und Zielbeschreibung gesetzt.

In Phase 2 beschreibt das RM die Auswahl von relevanten Datenbeständen. Die Schritte 2.1 (Bewertung von Datenquellen) und 2.2 (Bestimmung von Datenbeständen) werden im zu entwickelnden Modell unter dem Schritt 2.1 zusammengefasst. Der aggregierte Schritt 2.1 wird zusätzlich um den Aspekt des Kontextwissens erweitert und wird mit dem neuen Namen „Datenbeschaffung“ bezeichnet. Das Kontextwissen ist bereits bei der eigentlichen Datenauswahl von entscheidender Bedeutung, da sonst keine Bewertungsgrundlage für die einzelnen Datenbestände gegeben ist. Der Name Datenbeschaffung wurde gewählt, um sich deutlich von den Nachfolgeschritten abzugrenzen, die eine spezifische Selektion beinhalten. Die Aggregation erfolgt, da in der Praxis Datenquelle und verfügbare Datenbestände nicht voneinander getrennt werden können sowie hinsichtlich der fachlichen Organisation die Datenbestände oftmals über mehrere Datenquellen verteilt sind. So sollten zuerst die fachlichen Datenbestände definiert und dann potentielle Quellen identifiziert werden, was in die Modellbeschreibung des transformierten Modells zu integrieren ist. Als neuer Schritt 2.2 wird die Datenauswahl angegeben, da das RM diesen Aspekt nicht im Sinne einer reduzierenden Tätigkeit berücksichtigt. Im Bereich der SC-Daten liegen jedoch oftmals so große Datenbestände vor, dass eine Entnahme der gesamten verfügbaren Datenmenge entweder technisch nicht sinnvoll ist, da eine Analyse zu umfangreich wäre oder fachlich nicht sinnvoll ist, da beispielsweise die Zeiträume nicht mit der Aufgabenstellung in Phase 1 korrespondieren.

In Phase 3 erfolgt die Datenaufbereitung, die im RM sinnvollerweise sowohl Vorverarbeitung als auch Transformation beinhaltet und die sechs Basisklassen von Fragen abdeckt (vgl. hierzu Diskussion unter Transformation in Abschnitt 2.3.1).

Der Schritt 3.1 im RM wird in das zu entwickelnde Modell übernommen, da er branchenunabhängig für alle Datenbestände gültig ist. Der Schritt 3.2 im RM, der unter „Explorativer Datenanalyse“ aufgeführt wird, wird nicht übernommen. Der Aussagegehalt der Daten wurde bereits in einer ersten Näherung unter 2.1 geprüft und kann letztendlich in ausreichendem Umfang erst durch Data-Mining-Verfahren untersucht werden. Daher ist eine manuelle Zwischenprüfung bei komplexen Transaktionsdaten im Hinblick auf die unter Phase 1 festgelegten Ziele wenig sinnvoll. Da zudem in Phase 2 die Auswahl der Daten erfolgt, entfällt der Schritt 3.4 „Datenreduktion“ des RM. Die Kodierung der Merkmale, die im RM unter 3.7 aufgeführt ist, wird der Phase 4 zugeordnet, da die Kodierung der Merkmale insbesondere in Rückkoppelung mit den Data-Mining-Verfahren steht und somit im RM für die SC-Daten zu früh angeordnet ist. Zusätzlich wird in der Phase 3 zu Beginn ein Schritt für die fachliche Datengruppierung eingefügt, um die korrespondierende Anforderung der „Gruppierung“ von SC-Daten abzudecken. Diese Phase muss nach der Datenauswahl jedoch vor den Transformationen der Daten erfolgen. Die Begründung liegt darin, dass eine Gruppierung vor der Datenwahl gegebenenfalls zu falschen Gruppen führen würde, wenn beispielsweise irrelevante Zeiträume in die Betrachtung einfließen würden. Die Gruppierung muss zusätzlich zwingend vor den Transformationen erfolgen, da mitunter einzelne Gruppen unterschiedliche Transaktionsschritte aufweisen können. Der Schritt 3.3 des RM wird im zu entwickelnden Modell beibehalten und die Beschreibung um den SC-relevanten Faktor des Kontextwissens ergänzt. Die verbleibenden Schritte des RM mit den Nummern 3.5 und 3.6 werden zu einem neuen Schritt „Transformation“ zusammengefasst und ergänzt. Die Transformation beinhaltet im zu entwickelnden Modell die Prüfung auf Atomarität der Attribute, die Merkmalsreduktion, die Behandlung von fehlenden und fehlerhaften Merkmalen sowie die Ausreißerkorrektur.

Unter Phase 4 werden im RM alle Schritte zur Auswahl und Bewertung von Data-Mining-Verfahren angegeben. Die Verwendung des Begriffs Methode im RM, wie z. B. in Data-Mining-Methode oder Methodenauswahl, wird nicht weiter fortgeführt. Stattdessen wird der Begriff Methode durch den Begriff Verfahren im zu entwickelten Modell ersetzt, um in Analogie zur Tabelle 2.9 eine konstante Begriffswelt beizubehalten. Der Phasenname ist zudem im RM ungünstig gewählt, da beispielsweise auch die Toolauswahl integriert ist, die nicht zur Verfahrensauswahl gehört. Aus diesem Grund erfolgt eine Umbenennung, die den Begriff der allgemeinen Vorbereitung stärker zentriert. Zudem ist die Unterteilung der ersten drei Schritte für die Industrie zu feingranular. Die ersten Schritte umfassen die Auswahl der Bewertungskriterien, den Abgleich verschiedener Verfahren gegen die Bewertungskriterien sowie letztendlich die Bestimmung spezifischer Data-Mining-Verfahren. Die Bewertungskriterien für Data-Mining-Verfahren hängen im SC-Umfeld oftmals von den eingesetzten Tool-Möglichkeiten, der Datenlage, dem verfügbaren Know-how über einzelne Verfahren sowie gegebenenfalls Best-Practice-Techniken ab, um nur einige Faktoren zu nennen. Insofern sind eingesetzte Bewertungskataloge wie der vorgeschlagene Katalog im RM oder der in dieser Arbeit referenzierte Bewertungskatalog (vgl. Tabelle A.3 im Anhang) möglicherweise

zu nutzen, werden aber nicht separat adressiert. Daher werden die aufgeführten Schritte des RM unter einem neuen Schritt Verfahrensauswahl zusammengefasst. Im Anschluss an die Verfahrensauswahl erfolgt als neuer Schritt die Merkmalskodierung, die erst nach Auswahl möglicher Data-Mining-Verfahren sinnvoll umgesetzt werden kann. Die Merkmalskodierung wird in zwei Schritte eingegliedert, da nach fachlicher Kodierung und technischer Kodierung unterschieden wird. Fachliche Kodierung beinhaltet sämtliche Operationen, die Kontextwissen voraussetzen. Hierzu zählt insbesondere die Festlegung von fachlichen Metriken (vgl. Diskussion unter Charakteristikum „Format“ in Abschnitt 3.2) als Grundvoraussetzung für die Definition von Skalen auf einzelnen Attributen oder Attributsaggregationen. Neben der fachlichen Kodierung gibt es die technische Kodierung, die die Grundvoraussetzung für die Mining-Algorithmen erzeugt. Bei der technischen Kodierung spielt die Skalentransformation eine ausgezeichnete Rolle. Die Skalentransformation, im Gegensatz zur Skalendefinition, muss in der Regel unter Zuhilfenahme von Kontextwissen durchgeführt werden. Hierbei müssen die kontinuierlichen Attribute in den Transaktionsdaten in kategorisierbare Attribute umgewandelt werden (vgl. Methoden der Transformation in Abschnitt 2.3.2). Die Auswahl der Data-Mining-Werkzeuge im RM wird in das neu zu entwickelnde Modell übernommen. Die Phasenzugehörigkeit ist sinnvoll gewählt, da die Auswahl des Data-Mining-Werkzeugs in enger Wechselwirkung zu den anderen Phasenschritten, wie beispielsweise der technischen Kodierung, steht. Die Auswahl des Data-Mining-Werkzeuges ist essentiell für den Data-Mining-Prozess und beeinflusst die Verfahrenswahl maßgeblich. Die Auswirkung auf die Verfahrensauswahl begründet sich in der Tatsache, dass Data-Mining-Werkzeuge im Regelfall nur eine begrenzte Anzahl von Verfahren und Algorithmen zur Verfügung stellen. Zusätzlich gibt es werkzeugabhängige Implementierungen der eingesetzten Data-Mining-Algorithmen, z. B. bestimmte Programmiersprachen, die auf die Verfahren einwirken können. Die Unterschiede in der Implementierung bedingen in manchen Fällen sogar eine unterschiedliche technische Kodierung von Attributen, z. B. je nach Werkzeug in Ganzzahlen mit unterschiedlicher Genauigkeit (vgl. Tabelle A.1). Aus diesem Grund muss die Werkzeugauswahl vor den Kodierungsschritten der Attribute durchgeführt werden.

Die Phase 5 beinhaltet die Anwendung der zuvor ausgewählten Data-Mining-Verfahren. Das Ergebnis des Auswahlprozesses in der vorangestellten Phase sind ein oder mehrere Verfahren, die im Regelfall ein spezifisches Modell erzeugen. Es gibt Techniken, die mehrere Modelle zu Prüfzwecken erstellen. Diese Techniken stellen aber eine Verfahrensform dar und erzeugen nach der Prüfphase im Resultat immer noch ein einzelnes Modell. Daher wird im folgenden abweichend zum RM der Singular des Modellbegriffs verwendet. Schritt 5.1, der die Entwicklung des Modells beinhaltet, wird übernommen. Er bildet eine Konstante in allen Vorgehensmodellen und wird aus diesem Grund in das zu entwickelnde Modell integriert. Schritt 5.1 wird um den Aspekt der Testdaten angereichert, da diese von den Trainingsdaten der Modellerstellung separiert werden müssen. Zusätzlich wird im zu entwickelnden Vorgehensmodell der Schwerpunkt auf das Modelltraining in Schritt 5.2 und dessen Interaktion mit Schritt 5.1 gesetzt. Der Test des Modells mittels

der separierten Testdaten wird ausgelagert, da er ein klassisches Element der V&V repräsentiert. Als letzter Schritt wird unter 5.3 im RM die Kombination von zwei oder mehr Data-Mining-Verfahren innerhalb von Hybridsystemen angeführt. Da dies auch als Teilaspekt der Verfahrenswahl des Data-Minings unter Phase 4 und seiner Ausführung unter 5.1 und 5.2 betrachtet werden kann, entfällt der Schritt 5.3.

In der Phase 6 des RM werden die Schritte zur Ergebnisevaluation und -interpretation vorgestellt. Unter 6.1 beschreiben Hippner und Wilde die Extraktion von handlungsrelevanten Data-Mining-Ergebnissen. Der Begriff handlungsrelevant muss an dieser Stelle mit geeigneten technischen Verfahren unterstützt werden, da eine manuelle Bewertung der Data-Mining-Ergebnisse mitunter nicht möglich ist. Dies trifft beispielsweise auf den Fall zu, in dem ein Data-Mining-Verfahren in Abhängigkeit der Parametrierung eine unüberschaubare Menge an möglichen Wirkzusammenhängen zurückliefert. In der Folge wird der RM-Schritt 6.1 in das zu entwickelnde Modell übernommen und um einen technischen Aspekt angereichert. Zusätzlich wird im Anschluss ein weiterer Schritt benötigt, der die Ergebnisse der Data-Mining-Verfahren in eine explizite Darstellungsform überführt (vgl. Charakteristika „explizite Darstellung“ in Abschnitt 3.2). Dieser Schritt ist optional, da mitunter ein geeignetes Data-Mining-Verfahren bereits eine explizite Darstellungsform als Ausgabeergebnis liefert (vgl. Weiterverwendung der Muster in Abschnitt 2.3.2). Der ursprüngliche Schritt 6.2, der im RM die betriebswirtschaftliche Planung und Bewertung der Maßnahmen beschreibt, entfällt. Die Planung und Bewertung von konkreten Handlungsanweisungen fällt organisatorisch in andere Kompetenzbereiche und kann somit nicht im Geltungsbereich des Vorgehensmodells liegen. Unter Schritt 6.3 wird im RM die Bewertung des Data-Mining-Prozesses hinsichtlich Verbesserungs- und Rationalisierungsmöglichkeiten adressiert. Dieser Schritt umfasst insbesondere methodische Prüfungen und eine Qualitätskontrolle der vorherigen Phasen. Ein solcher Punkt ist auch für das zu entwickelnde Referenzmodell notwendig, allerdings ist die Zusammenführung von methodischen Prüfungen in V&V und Qualitätskontrolle im RM nicht zielführend. Aus diesem Grund werden die Aspekte der V&V ausgelagert und an geeigneter Stelle aufgegriffen. Der Einsatz von Techniken zur Qualitätskontrolle ist in die nachfolgende Phase zu integrieren. Um den veränderten fachlichen Schwerpunkt der Phase Rechnung zu tragen, erfolgt eine abschließende Umbenennung der Phase in „Weiterverarbeitung der Data-Mining-Ergebnisse“.

Die Phase 7 des RM beinhaltet die Anwendung der Data-Mining-Ergebnisse im betriebswirtschaftlichen Alltag und adressiert somit einen zu weit gefassten Fokus für den Geltungsbereich des SC-Vorgehensmodells (vgl. Charakteristika „Geltungsbereich“ in Abschnitt 3.2). Die Schritte 7.1, „Veränderung von Geschäftsprozessen“, 7.2 „Entwickeln von operativen Maßnahmen“ und 7.3 „Empfehlung für Führungsentscheidungen“ sind im neu zu entwickelnden Vorgehensmodell nicht integriert. Die beschriebenen Maßnahmen können im Nachgang an die Wissensentdeckung vom SCM eingeleitet werden, sind aber losgelöst vom hier entwickelten Vorgehens-

modell zur Wissensentwicklung zu betrachten. Dafür wird in Phase 7 die bereits zuvor adressierte Qualitätskontrolle integriert und als erster Schritt in Phase 7 aufgeführt. Der letzte Schritt 7.4 des RM, der das Ableiten von Erkenntnissen für neue Data-Mining-Prozesse beschreibt, wird übernommen. Hierbei wird der Fokus im SC-Kontext auf die Weiterverwendung der Data-Mining-Ergebnisse als Hypothese für neue Data-Mining-Prozesse gesetzt. Dieser Punkt ergibt sich aus der Komplexität der SC-Datenbestände, in der eine Vielzahl von Data-Mining-Aufgaben im ersten Schritt hypothesenfrei gefunden werden und erst die gefundenen Ergebnisse Anhaltspunkte für weitere Data-Mining-Prozesse liefern können.

Aus den aufgeführten Überlegungen ergibt sich das Vorgehensmodell für die Methode zur Musterextraktion in SCs (MESC), das in Tabelle 3.3 dargestellt ist. Es ist zu berücksichtigen, dass das RM selbst eine Variation der Schrittreihenfolge gestattet. Das macht das Modell zum einen hochflexibel, birgt aber den Nachteil von Anwendungsfehlern. Dies begründet sich darin, dass die Autoren keine Einschränkungen bezüglich der Flexibilität machen und so auch diverse ungültige Durchlaufreihenfolgen entstehen können, die dem Anwender keine zufriedenstellenden Ergebnisse liefern. Aus diesem Grund wird die Reihenfolge in MESC als gesetzt vorgegeben und in den folgenden Abschnitten bei Bedarf der Aspekt der Phaseniteration näher erläutert.

**Tabelle 3.3: Vorgehensmodell zur Musterextraktion in SCs**

| **Phase** | **Schritte** | **Kurzbeschreibung** |
|---|---|---|
| 1. Aufgabendefinition | 1.1 Bestimmung der Aufgabenstellung | Formulierung der Aufgabenstellung des SCM unter Berücksichtigung von gegebenen Randbedingungen und Festlegung der Zielkriterien |
| 2. Auswahl der relevanten Datenbestände | 2.1 Datenbeschaffung | Bestimmung und Zugang zu den Datenquellen und den zugehörigen Datenbeständen gemäß Zieldefinition |
| | 2.2 Datenauswahl | Auswahl der Datenbestände mittels Kontextwissen zwecks Datenreduktion |
| 3. Datenaufbereitung | 3.1 Formatstandardisierung | Überführung der selektierten Datenbestände in ein Standarddatenformat |
| | 3.2 Gruppierung | Fachliche Gruppierung der Datenbestände unter Berücksichtigung der Aufgabenstellung |

**Tabelle 3.3: Vorgehensmodell zur Musterextraktion in SCs (Fortsetzung)**

| Phase | Schritte | Kurzbeschreibung |
|---|---|---|
| | 3.3 Datenanreicherung | Datenanreicherung unter Einbeziehung von Kontextwissen |
| | 3.4 Transformation | Prüfung auf Atomarität der Attribute, Merkmalsreduktion, Behandlung von fehlenden und fehlerhaften Merkmalen sowie Ausreißerkorrektur |
| 4. Vorbereitung des Data-Mining-Verfahrens | 4.1 Verfahrenswahl | Auswahl des einzusetzenden Verfahrens in Abhängigkeit zur Aufgabenstellung |
| | 4.2 Werkzeugauswahl | Auswahl eines geeigneten Data-Mining-Werkzeuges |
| | 4.3 Fachliche Kodierung | Fachliche Auswahl und Kodierung geeigneter Attribute |
| | 4.4 Technische Kodierung | Technische Auswahl und Kodierung geeigneter Attribute |
| 5. Anwendung der Data-Mining-Verfahren | 5.1 Entwicklung eines Data-Mining-Modells | Modellentwicklung und Trennung der Datenbestände in Trainings-, Validierungs- und Testdaten |
| | 5.2 Training des Data-Mining-Modells | Training des Data-Mining-Modells mittels der Validierungsdaten aus 5.1 |
| 6. Weiterverarbeitung der Data-Mining-Ergebnisse | 6.1 Extraktion handlungsrelevanter Data-Mining-Ergebnisse | Unter Berücksichtigung der Handlungsrelevanz sowie der technischen Maßzahlen sind die für das SCM interessanten Ergebnisse zu extrahieren |
| | 6.2 Darstellungstransformation der Data-Mining-Ergebnisse | In Abhängigkeit der eingesetzten Data-Mining-Verfahren sowie der Aufgabenstellung müssen die Ergebnisse in eine explizite Darstellungsform überführt werden |

**Tabelle 3.3: Vorgehensmodell zur Musterextraktion in SCs (Fortsetzung)**

| Phase | Schritte | Kurzbeschreibung |
|---|---|---|
| 7. Bewertung des Data-Mining-Prozesses | 7.1 Qualitätskontrolle des Data-Mining-Prozesses | Qualitätskontrolle mittels geeigneter Maßnahmen und Abschluss V&V-Maßnahmen |
| | 7.2 Rückführung von Data-Mining-Ergebnissen | Rückführen von Ergebnissen, die als Grundlage für weitere Data-Mining-Prozesse genutzt werden können |

Zum Abschluss müssen die prozentualen Zeitangaben des RM (vgl. Abbildung 2.6) für die MESC angepasst werden, da die ursprünglichen Angaben der MESC nicht gerecht werden. Die notwendigen zeitlichen Adaptionen der MESC sind hierbei maßgeblich von drei Faktoren beeinflusst. Erstens wurde die Phasenzugehörigkeit einzelner Schritte angepasst sowie SC-spezifische Schritte eingefügt. Dies wirkt sich auf die Zeitangaben der einzelnen Phasen aus. Des Weiteren hat sich die IT-Infrastruktur und Hardware im Vergleich zu den ursprünglichen Angaben aus den letzten 20 Jahren verbessert, sodass das Ausführen von IT-Prozessen, wie beispielsweise das Data Mining, kürzere Laufzeiten als vor einigen Jahren aufweisen. Zusätzlich bedingen die SC-Organisationsstrukturen eine Verkürzung von einigen Bearbeitungsschritten. Durch den notwendigen Einsatz von Kontextwissen verlängern sich jedoch im Gegenzug einzelne Abläufe in MESC. Abbildung 3.3 zeigt die einzelnen MESC-Phasen und deren geschätzten Zeitspannen in Prozentangaben in Bezug auf die anzusetzende Gesamtzeit des Vorgehensmodells.

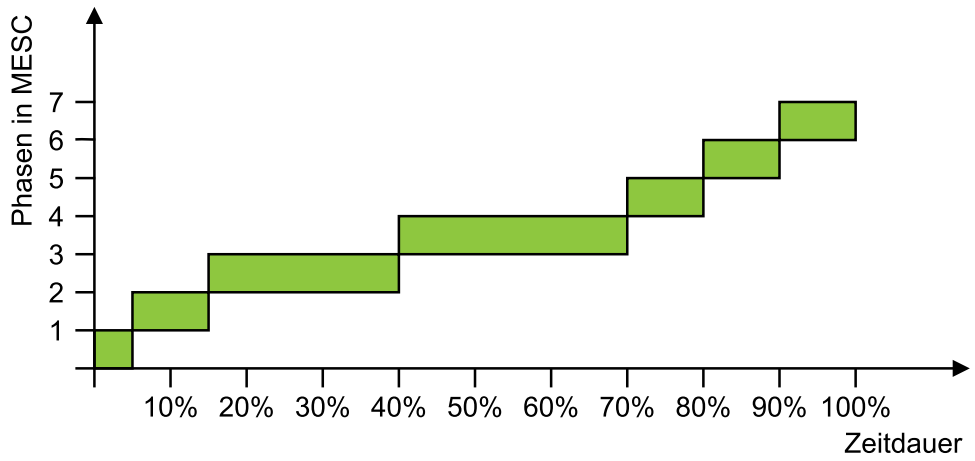



**Abbildung 3.3: Zeitangaben für Phasendauer in MESC**

Folgende Annahmen sind im Kontext der MESC denkbar: Die Aufgabendefinition könnte im SC-Kontext nur noch 5 % der Gesamtzeit einnehmen, da zum einen

in einer SC spezifische Fragestellungen existieren und zum anderen der organisatorische Strukturrahmen globaler SCs Vorgaben und Randbedingungen für die Wissensentdeckung aufzeigt. Hingegen ist die Auswahl der relevanten Datenbestände bedingt durch die SC-Struktur mitunter ein komplexer Vorgang, der mit 10 % angesetzt werden könnte. Im Idealfall liegen die Daten aufbereitet an einem zentralen Speicherort, doch im Normalfall gibt es eine heterogene IT-Landschaft mit unterschiedlichen Verantwortungsbereichen und Wissensträgern. So ist es nicht selten, dass die Identifikation von geeigneten Datensätzen eine zeitintensive Aufgabe darstellen kann. Die Datenaufbereitung in Phase 3 ist, wie auch in den ursprünglichen Literaturangaben, ein potentiell zeitintensiver Schritt und könnte mit 25 % angesetzt werden. Daran anschließend erfolgt die längste Phase, die eigentliche Vorbereitung des Data-Mining-Verfahrens. Hierbei ist die fachliche Kodierung der Attribute ein zeitlicher Hauptfaktor, sodass in der Folge die Phase mit 30 % der Gesamtzeitdauer geschätzt werden könnte. Die Phase 5, die die Anwendung der Data-Mining-Verfahren beschreibt, könnte aufgrund der Weiterentwicklung in Hard- und Software nur noch mit 10 % deklariert werden. Die Weiterverarbeitung der Data-Mining-Ergebnisse könnte wegen der möglichen Notwendigkeit von Kontextwisseneinsatz ebenfalls mit potentiellen 10 % veranschlagt werden. Auch die Abschlussphase könnte mit einem Richtwert von 10 % geschätzt werden, da letztendlich eine Gesamtbewertung des Vorgehensmodells durchgeführt werden müsste.

Es muss ausdrücklich darauf hingewiesen werden, dass die Zeitangaben nur für den SC-Kontext gelten und nur als Richtwerte fungieren. Bei Anwendung der MESC in einem konkreten SC-Kontext müssen die Werte mitunter aufgrund projektspezifischer Faktoren korrigiert werden. Bis zum heutigen Zeitpunkt fehlen Untersuchungen, die ein Übertragen der Werte auf andere Bereiche der Wissensentdeckung im produktionslogistischen Kontext gestatten. Es muss aber konstatiert werden, dass globale logistische Richtwerte ausgeschlossen werden können. Dies wird am Bereich der Produktions- und Montageplanung deutlich. In diesem Kontext kann beispielsweise die Zeitdauer für die Aufgabendefinition 20 % der Wissensentdeckung einnehmen. Dies begründet sich darin, dass die Aufgabenstellung oftmals unklar ist und der Datenbestand nicht in Datenbanken gespeichert ist. Insbesondere in den zuvor genannten Bereichen sind Exceltabellen mit Texten die Quelle für mögliches Wissen, sodass zeitintensive Verfahren wie Text Mining die Aufgabendefinition begleiten können.

## 3.6 Integration eines Vorgehensmodells zur V&V aus der Simulation

Die Erkenntnisse aus Abschnitt 2.3.1, die besagen, dass alle identifizierten Vorgehensmodelle sich durch eine iterative, sukzessive Schrittanordnung sowie eine manuelle Steuerung der Aktivitäten auszeichnen, werden übernommen und gelten demnach auch für das zu entwickelnde Vorgehensmodell. Die Begriffe iterativ in

Verbindung mit manueller Steuerung implizieren einen Entscheidungsprozess, der besagt, wann welcher Schritt auszuführen ist. Dieser Entscheidungsprozess ist im allgemeinen an den Vorgang der V&V gebunden. Wie bereits in Abschnitt 2.3.2 diskutiert, erfolgt in den bisherigen Vorgehensmodellen die V&V, wenn überhaupt, nur als separater Schritt. Um den Ansprüchen der Wissensentdeckung in der SC im Projekteinsatz gerecht zu werden, ist jedoch eine modellbegleitende V&V notwendig (vgl. Forschungsfrage 3 in Abschnitt 2.4.2). Um eine modellbegleitende Prüfung der einzelnen Schritte zu gewährleisten, wird das Vorgehensmodell zur V&V in der Simulation, das Dreiecksmodell (vgl. Abschnitt 2.3.3), auf das Vorgehensmodell in der MESC transformiert und die Initiierung der V&V unter Einsatz des Dreiecksmodells als initiativer Schritt eingeführt.

Abbildung 3.4 zeigt das dahingehend transformierte Dreiecksmodell. Hierbei sind die Schritte des Originalmodells durch die entsprechenden Phasen der MESC ersetzt worden. In der Folge findet sich in MESC eine größere Kumulation von Prüfelementen, da im transformierten Dreiecksmodell die Betrachtung auf Phasenebene und nicht auf Schrittebene erfolgt. Dies ist jedoch sinnvoll, da oftmals erst am Phasenende der MESC ein überprüfbares Ergebnis vorliegt und die V&V in jedem Schritt eine unangemessene Granularität für den praktischen Einsatz zeigt.

Das transformierte Dreiecksmodell führt alle erforderlichen Prüfschritte auf Phasenebene auf. Dazu dienen die Ergebnisse der korrespondierenden MESC-Phasen als Grundlage. Der Schwerpunkt des transformierten Dreieckmodells liegt auf der Beschreibung der Prüfvorgänge für die einzelnen Phasenergebnisse. Die für die Prüfvorgänge eingesetzten Verfahren der V&V sind nicht spezifisch für die Wissensentdeckung auf SC-Transaktionsdaten. Aus diesem Grund können geeignete Standardverfahren beispielsweise aus dem Softwaretest Anwendung finden (vgl. Phase 8 in Abschnitt 2.3.2). Tabelle A.4 im Anhang gestattet eine Zuordnung von exemplarischen V&V-Techniken zu den einzelnen Prüfschritten der MESC.

Das Vorgehen ist für alle Phasen des Dreiecksmodells identisch und orientiert sich an den Phasenbeschreibungen des Originalmodells (vgl. Abschnitt 2.3.3). Die Prüfungen finden immer zum Ende einer MESC-Phase statt und fungieren als Phasenabschluss. Dazu werden im ersten Schritt die Ergebnisse der korrespondieren MESC-Phasen intrinsisch geprüft und im Anschluss gegen die vorherigen Phasenergebnisse der MESC geprüft (vgl. Abbildung 3.4). Sobald ein Fehler in den intrinsischen Prüfungen oder in der Prüfung gegen andere Phasenergebnisse vorliegt, wird dieser Fehler behoben und die Phase wird erneut sowohl intrinsisch als auch gegen die vorhergehenden Phasen geprüft. Tabelle 3.4 zeigt die einzelnen Prüfschritte für die V&V der MESC. Die Notation der einzelnen Prüfschritte orientiert sich an der Notation des Vorgehensmodells zur V&V in der Simulation (vgl. Abschnitt 2.3.3).

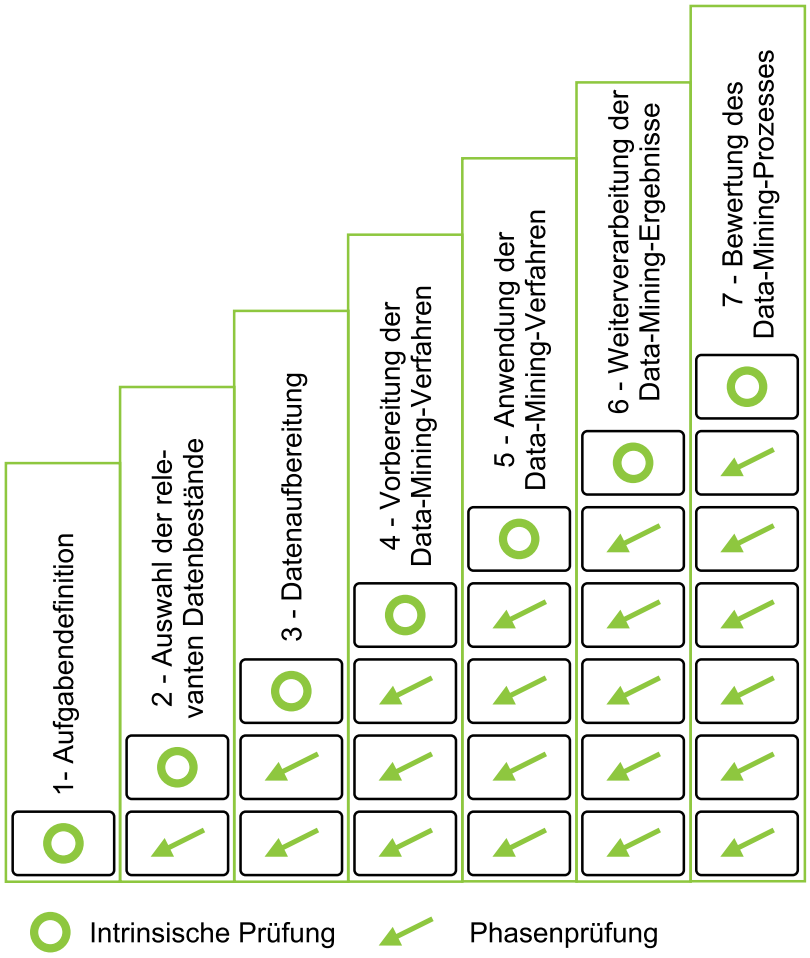



**Abbildung 3.4: Transformiertes Dreiecksmodell**

**Tabelle 3.4: Dreiecksmodell in MESC**

| Phase | intrinsische Prüfung | Prüfung gegen die Vorphase |
|---|---|---|
| 1. Aufgabendefinition | 1,1: Prüfung auf Vollständigkeit und Plausibilität der gegebenen Randbedingungen und festgelegten Zielbedingungen | - |
| 2. Auswahl der relevanten Datenbestände | 2,2: Prüfung auf Relevanz der verwendeten Datenquellen und Datenbestände | 2,1: Prüfung, ob die ausgewählten Datenquellen und Datenbestände für das Erreichen der Zielbedingungen geeignet sind |

**Tabelle 3.4: Dreiecksmodell in MESC (Fortsetzung)**

| Phase | intrinsische Prüfung | Prüfung gegen die Vorphase |
|---|---|---|
| 3. Datenaufbereitung | 3,3: Prüfung, ob die vorliegenden Daten entsprechend transformiert worden sind | 3,2: Prüfung der Datenaufbereitung gegen die relevanten Datenbestände |
| | | 3,1: Prüfung, ob die aufbereiteten Daten für das Erreichen der Zielbedingungen geeignet sind |
| 4. Vorbereitung des Data-Mining-Verfahrens | 4,4: Prüfung auf geeignete Auswahl des Data-Mining-Verfahrens | 4,3: Prüfung, ob die Datenaufbereitung als Vorbereitung für das Data-Mining-Verfahren ausreichend ist |
| | | 4,2: Prüfung, ob die Auswahl der relevanten Daten den Anforderungen des ausgewählten Data-Mining-Verfahrens entspricht |
| | | 4,1: Prüfung, ob das Data-Mining-Verfahren für die Erfüllung der Aufgabenstellung geeignet ist |
| 5. Anwendung des Data-Mining-Verfahrens | 5,5: Prüfung auf richtige Anwendung des Data-Mining-Verfahrens | 5,4: Prüfung ob das Data-Mining-Verfahren für seine Anwendung zuvor richtig vorbereitet worden ist |
| | | 5,3: Prüfung, ob die Datenaufbereitung für die Anwendung des Data-Mining-Verfahrens geeignet ist |
| | | 5,2: Prüfung, ob die Datenselektion eine fachgerechte Anwendung des Data-Mining-Verfahrens ermöglicht |

**Tabelle 3.4: Dreiecksmodell in MESC (Fortsetzung)**

| Phase | intrinsische Prüfung | Prüfung gegen die Vorphase |
|---|---|---|
| | | 5,1: Prüfung, ob durch die Anwendung des Data-Mining-Verfahrens die Zielbedingungen erfüllt werden |
| 6. Weiterverarbeitung der Data-Mining-Ergebnisse | 6,6: Prüfung auf ordnungsgemäße Weiterverarbeitung der Data-Mining-Ergebnisse | 6,5: Prüfung, ob die Anwendung des Data-Mining-Verfahrens interpretierbare Daten als Resultat liefert<br>6,4: Prüfung, ob das ausgewählte Data-Mining-Verfahren interpretierbare Daten liefern kann<br>6,3: Prüfung, ob die Daten für die Interpretation fachlich richtig aufbereitet worden sind<br>6,2: Prüfung, ob die Datenselektion für die Interpretation ausreichend ist oder ob andere Daten für die Interpretation selektiert werden müssen<br>6,1: Prüfung, ob durch die Interpretation gewonnenen Erkenntnisse den vordefinierten Zielen der Aufgabenstellung genügen |
| 7. Bewertung des Data-Mining-Prozesses | 7,7: Prüfung, ob die Qualitätskontrollen des Data-Mining-Prozesses richtig initiiert worden sind | 7,6: Prüfung, ob die vorhandenen Data-Mining-Ergebnisse ausreichend für die Qualitätskontrolle dokumentiert sind<br>7,5: Prüfung, ob die Anwendung des Data-Mining-Verfahrens genügend dokumentiert ist |

**Tabelle 3.4: Dreiecksmodell in MESC (Fortsetzung)**

| Phase | intrinsische Prüfung | Prüfung gegen die Vorphase |
|---|---|---|
| | | 7,4: Prüfung, ob die Dokumentation für den Auswahlprozess des Data-Mining-Verfahrens ausreichend ist |
| | | 7,3: Prüfung, ob der Datenaufbereitungsprozess und die Prozessergebnisse ausreichend dokumentiert sind |
| | | 7,2: Prüfung, ob die Dokumentation des Datenauswahlprozesses und der Prozessergebnisse ausreichend ist |
| | | 7,1: Prüfung, ob die Aufgabenstellung unter Berücksichtigung von Randbedingungen und Zielkriterien ausreichend dokumentiert sind |

Die Phase 1 der Aufgabendefinition beinhaltet die intrinsische Prüfung der Aufgabe des SCM unter Angabe von Randbedingungen. Als Grundlage hierfür dient die Projektdokumentation der entsprechenden MESC-Phase. In der Phase 2 wird zunächst intrinsisch die Korrektheit der verwendeten Datenquellen und Datenbestände geprüft. Anschließend wird diese Phase gegen die Aufgabendefinition geprüft, wobei untersucht wird, ob die ausgewählten Daten für das Erreichen der Aufgabenziele geeignet sind. In der Phase 3 wird in der intrinsischen Prüfung kontrolliert, ob die vorliegenden Daten entsprechend transformiert worden sind. Wenn keine intrinsischen Fehler vorliegen, wird die Datenaufbereitung zunächst gegen die relevanten Datenbestände geprüft, sodass anschließend kontrolliert werden kann, ob die aufbereiteten Daten für die Ziele der Phase 1 genügen. In der Phase 4 erfolgt innerhalb der intrinsischen Prüfung die Kontrolle der Data-Mining-Verfahrensauswahl. Anschließend wird kontrolliert, ob die in Phase 2 und 3 ausgewählten und aufbereiteten Daten als Vorbereitung für das Data-Mining-Verfahren ausreichend sind. Schließlich wird noch gegen die Phase 1 der Aufgabendefinition geprüft, indem kontrolliert wird, ob das ausgewählte Data-Mining-Verfahren für die Erfüllung der Aufgabenstellung geeignet ist. In der Phase 5 erfolgt die Anwendung des vorbereiteten Data-Mining-Verfahrens. Hierbei wird intrinsisch geprüft,

ob das Data-Mining-Verfahren richtig angewendet wurde. Diese Phase inkludiert die bereits diskutierte V&V-Phase der KDD-Vorgehensmodelle, die aus dem RM nicht in das Vorgehensmodell zur Wissensentdeckung übernommen wurde (vgl. Abschnitt 3.5). Die Prüfung kann unter Zuhilfenahme der Testdaten durchgeführt werden, die dafür in der MESC-Phase 5 separiert wurden. Nach erfolgreicher Prüfung erfolgt im Anschluss die Kontrolle, ob das Data-Mining-Verfahren für die Anwendung geeignet vorbereitet worden ist. Daran anschließend wird kontrolliert, ob durch die Datenaufbereitung und Datenauswahl eine korrekte Anwendung der Data-Mining-Verfahren ermöglicht wird und ob durch die Anwendung des Data-Mining-Verfahrens die in Phase 1 definierten Ziele erreicht werden können. Die Phase 6 umfasst die Weiterverarbeitung der Data-Mining-Ergebnisse. Nach der intrinsischen Prüfung auf angemessene Weiterverarbeitung der Data-Mining-Ergebnisse wird in der Prüfung gegen die Vorphase zunächst untersucht, ob das Data-Mining-Verfahren interpretierbare Daten als Resultat liefert. Weiterhin wird geprüft, ob die Daten für die Weiterverarbeitung der Ergebnisse fachlich richtig aufbereitet und ausreichend selektiert worden sind. Anschließend wird kontrolliert, ob die durch die Weiterverarbeitung der Data-Mining-Ergebnisse gewonnenen Erkenntnisse für die Aufgabenstellung geeignet sind. In der letzten Phase (Bewertung des Data-Mining-Prozesses) wird intrinsisch überprüft, ob eine richtige Initiierung der erforderlichen Qualitätskontrollen des Data-Mining-Prozesses vorliegt. Im Anschluss erfolgt die Prüfung gegen die Vorphasen, die kontrollieren muss, ob die vorhandenen Ergebnisse des Data-Mining-Prozesses ausreichend für die Qualitätskontrolle dokumentiert worden sind. Daran anschließend wird überprüft, ob die gewünschte Dokumentation ebenfalls in der Anwendung und Vorbereitung des Data-Mining-Verfahrens vorliegt. Außerdem muss geprüft werden, ob die Datenaufbereitung, die Datenauswahl und die Aufgabenstellung unter Berücksichtigung von allen Randbedingungen und Zielkriterien ausreichend dokumentiert worden sind. Der Schritt 7.2, die Rückführung der Data-Mining-Ergebnisse (vgl. Tabelle 3.3), ist im Dreiecksmodell unberücksichtigt, da eine praktische Prüfung dieser Ergebnisse oftmals nicht im zeitlichen Rahmen des Projektes liegt. Dies ist beispielsweise in der Regel der Fall, wenn die Ergebnisse erst für ein Nachfolgeprojekt Anwendung finden.

Eine Konzeptalternative zur Struktur in Tabelle 3.4 bestand in der Integration des Dreiecksmodells in einzelne MESC-Phasen. Dies hat jedoch den Nachteil, dass dann die V&V ihre eigene Initiierung prüfen muss und der Selbstbezug ein vermeidbares theoretisches Problem konstruiert. Wenn das Dreiecksmodell beispielsweise in Phase 1 initiiert wird, so lautet die intrinsische Phasenprüfung des Dreiecksmodells „Prüfe mit dem Dreiecksmodell, ob das Dreiecksmodell gemäß der Projektanforderungen korrekt initiiert wurde“.

Das Dreiecksmodell wird in Schlussfolgerung als MESC-Methodenelement neben der Wissensentdeckung in Tabelle 3.3 verstanden und komplettiert diese im Bereich der V&V. Das integrierte Dreiecksmodell stellt die korrekte Ausführung des

Vorgehensmodells in MESC sicher, sodass im Folgenden eine Referenzierung der MESC das Dreiecksmodell inkludiert.

Abbildung 3.5 zeigt die MESC und seine Bestandteile der Wissensentdeckung (vgl. Tabelle 3.3) in schwarz sowie den Dreiecksmodellbaustein in seiner V&V-Funktionalität (vgl. Tabelle 3.4) in grün.

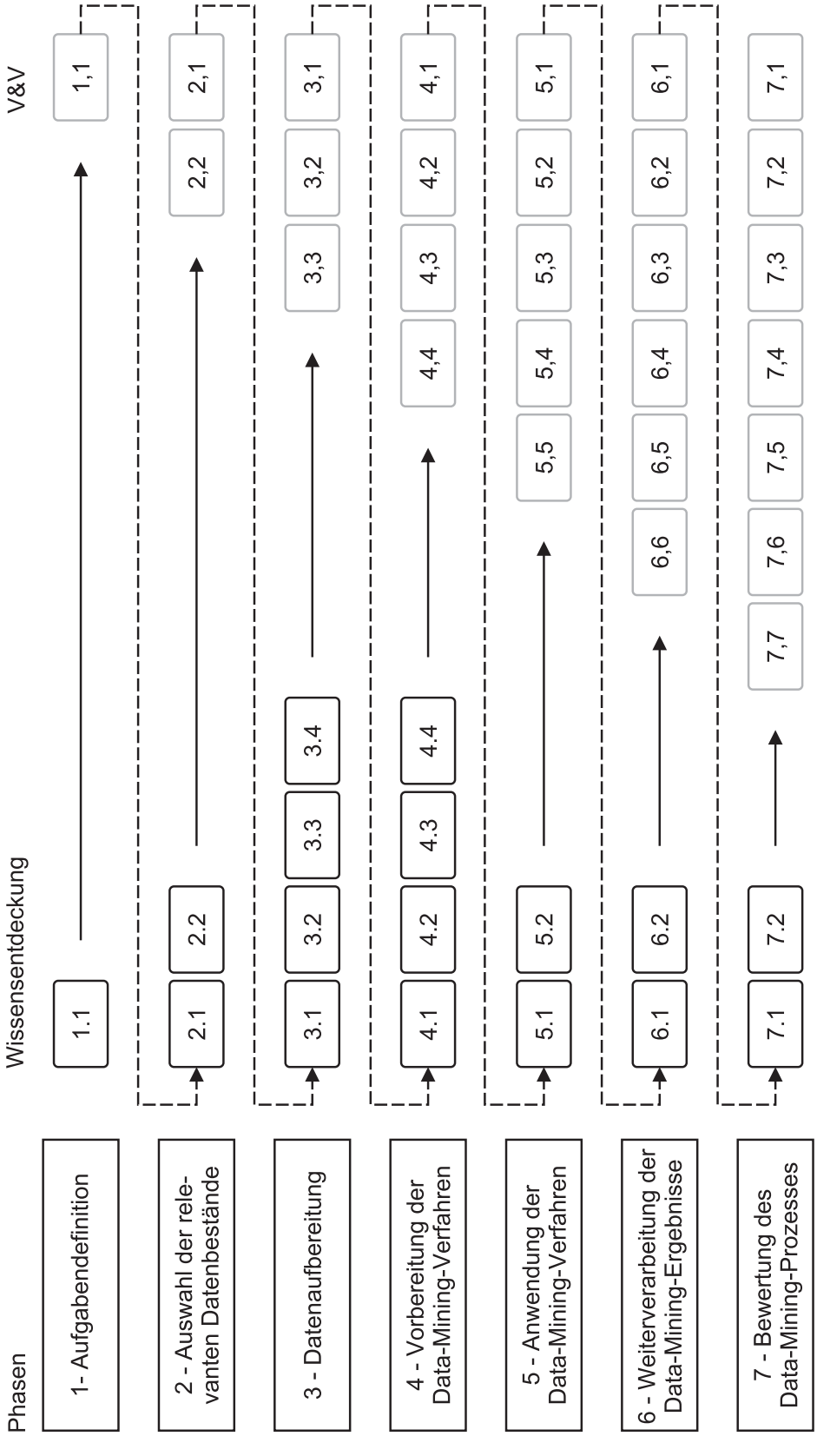



**Abbildung 3.5: Gesamtübersicht der Methodenelemente der MESC**

# 4 Detaillierte Untersuchung der einzelnen Phasen des Vorgehensmodells

Basierend auf der im vorherigen Kapitel erarbeiteten strukturellen Gestaltung für das Vorgehensmodell (vgl. Abschnitt 3.5) werden in diesem Kapitel die Schritte der Phasen einzeln untersucht und konzipiert. Zusätzlich werden erforderliche Verfahren benannt, die zur Durchführung der Schritte eingesetzt werden können. Die Ausarbeitung des Vorgehensmodells findet sowohl deduktiv, indem Erkenntnisse aus der Literatur (vgl. Kapitel 2) sowie Anforderungen und notwendige Erweiterungen aus der Modelluntersuchung in Kapitel 3 berücksichtigt werden als auch induktiv durch die Integration logistikspezifischer Aspekte statt. Insbesondere wird der Musterbegriff aufgegriffen und eine verallgemeinerte Definition erarbeitet. Der entsprechende Abschnitt korrespondiert mit dem Abschnitt 2.4.2 und bezieht sich auf die *Forschungsfrage 1:* Welche Musterbegriffe sind im Rahmen der Wissensentdeckung in SC-Datenbanken notwendig? Zur Verwendung des Wortes Prozess im Rahmen der MESC (z. B. in Data-Mining-Prozess) wird auf die Diskussion in Kapitel 2 verwiesen.

## 4.1 Vorphase und Aufgabendefinition

In diesem Abschnitt erfolgt die Beschreibung der Vorphase für die MESC, deren Aufgabe die Modellinitiierung ist. Im Anschluss wird die erste Phase der MESC diskutiert und aufgezeigt, welche Aspekte im Kontext von SC-Datenbanken zu berücksichtigen sind.

### 4.1.1 Initiierungsphase

Als Vorphase erfolgt die Initiierung der MESC. Da das integrierte Dreiecksmodell mit den einzelnen Phasen der Wissensentdeckung interagiert (vgl. Abschnitt 3.6), ist eine ausgezeichnete Initiierung vor der Ausführung des Vorgehensmodells notwendig. Abbildung 4.1 zeigt die MESC mit entsprechend vorgelagerter Initiierungsphase.

Die hier beschriebene Initiierungsphase darf nicht mit den Aufgabenstellungen der allgemeinen Projektinitiierung gleichgesetzt werden. Die Aufgabenstellungen der allgemeinen Projektinitiierung wie beispielsweise Projektorganisation und Kontextanalyse werden im nächsten MESC-Schritt eingebunden.

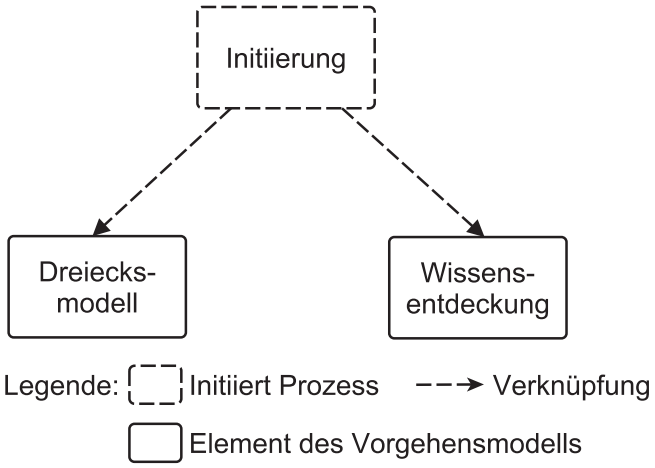



**Abbildung 4.1: Initiierungsphase in MESC**

### 4.1.2 Bestimmung der Aufgabenstellung

Diese Phase stellt den Beginn der MESC dar und beinhaltet die Festlegung der Aufgabenstellung zur Wissensentdeckung unter Berücksichtigung von gegebenen Randbedingungen und Definition der Zielkriterien. Die Aufgabenstellungen der Wissensentdeckung im Kontext der SC sind vielfältig. Tabelle 2.7 gestattet eine Übersicht der gängigen Fragen im SC-Kontext, die im Rahmen einer Wissensentdeckungsaufgabe gestellt werden können. Zu jeder Aufgabe müssen mögliche Randbedingungen als gegebene Größen berücksichtigt werden. Die Randbedingungen lassen sich in Data-Mining-Projekten in folgende Hauptkategorien unterteilen:

1. Zeit - Startzeitpunkt, Meilensteine, Deadlines (z. B.: Wann sind Daten verfügbar?)
2. Organisation - Unternehmensorganisation, Erfahrung im Data-Mining, interne Prozesse (z. B.: Welche Abteilungen sind beteiligt?)
3. Personal - Personal pro Schritt in einzelnen Phasen (z. B.: Welches Personal steht für die Erhebung von Kontextwissen zur Verfügung?)
4. Recht - Normen, Rechtsvorschriften, unternehmensinterne Vorschriften in Bezug auf Daten (z. B.: Welche Daten müssen anonymisiert werden?)
5. Technik - IT-Landschaft, Hardware, Zusatz-Software, Data-Mining-Werkzeuge (z. B.: Welche Data-Mining-Werkzeuge können eingesetzt werden?)
6. Fachlich - Datenqualität, Nachvollziehbarkeit einzelner Phasen (z. B.: Ist eine ausreichende Datenqualität für den Einsatz von Data-Mining-Verfahren gesichert?)
7. Ökonomie - Budget für Data-Mining-Projekt, das sich beschränkend auf die vorherigen Phasen auswirkt (z. B.: Gibt es eine Personalkostenlimitierung für externe Mitarbeiter?)

Um das Data-Mining-Projekt abzuschließen, muss das Ziel der Wissensentdeckung in Bezug auf die Aufgabe erfüllt sein. Um die Erfüllung zu messen, müssen in dieser Phase Kriterien festgelegt werden, die an die Zielerfüllung zu stellen sind. Diese Kriterien werden zumeist von unternehmensinternen Richtlinien beeinflusst. Bei der Zielformulierung der Wissensentdeckung sind des Weiteren folgende Punkte zu berücksichtigen:

1. Realistisch - Ziele sollen unter Berücksichtigung der Randbedingungen formuliert werden
2. Spezifisch - Ziele der Wissensentdeckungen müssen konkret und problemspezifisch sein
3. Eindeutig - Ziele sollen keinen Platz für Interpretation lassen
4. Terminiert - Ziele sollten immer an einen konkreten Zeitpunkt gekoppelt sein
5. Messbarkeit - Ziele sollten in geeigneter Weise zu erfassen sein, z. B. Fehlerkorridor bei Modellen

## 4.2 Auswahl der relevanten Datenbestände

Diese Phase gliedert sich in die Schritte Datenbeschaffung und Datenauswahl, die eng miteinander verknüpft sind. Je nach Komplexität der vorherrschenden IT-Landschaft sind in diesem Schritt eine Vielzahl von beteiligten Personen zu integrieren. Die Vielzahl der Personen führt dazu, dass diese Phase als kommunikationsintensiv betrachtet werden muss.

### 4.2.1 Datenbeschaffung

Die Ziele der Datenbeschaffung liegen in der Bestimmung einer geeigneten Datenquelle sowie der Ermöglichung des Zugangs zur identifizierten Datenquelle. Im Kontext der globalen SC ist der Akteur mit einer komplexen Systemlandschaft konfrontiert, die aus einer Vielzahl von Einzelsystemen mit teilweise redundanten oder aggregierten Informationen besteht. Da das zu entdeckende Wissen im Allgemeinen keine einzelnen Unternehmensteile und deren Datenbanken betrifft (vgl. Abschnitt 2.2.2.3), steht der SC-Akteur vor der Herausforderung, dass die benötigten Informationen über verschiedene Systeme verteilt sind. Die Identifikation der benötigten Ressourcen ist von spezieller Bedeutung, da nun die Treppe aus Daten, Information und Wissen (vgl. Definition 2.2) in Richtung des strategischen Wissensmanagements durchlaufen werden muss (vgl. Abbildung 4.2). Dies bedeutet, dass basierend auf Aufgabenstellung und Ziel, die das Wissen bilden, die notwendigen Informationen, die dieses beinhalten können, identifiziert werden müssen. Diese Informationen wiederum sind über unterschiedliche Datenbestände

und Systeme verteilt und können auf Datenebene eine Vielzahl von Attributen umfassen. Sowohl für die Transformation von Wissen zu Information als auch für die Transformation von Information zu Daten ist Kontextwissen (vgl. Abschnitt 2.1.2) notwendig, denn nur mittels Kontextwissens, beispielsweise in Form von unternehmensinternen Kodierungen für Artikelgruppen, ist eine fachliche Zuordnung der zumeist rein technischen Bezeichnungen von Attributen und Tabellen der SC-Datenbanken möglich. Dieses Kontextwissen kommt organisatorisch in seltenen Fällen von nur einer Person, denn die Ebene Wissen - Information umfasst einen größeren Unternehmensblickwinkel, wo hingegen die Transformation von Informationen auf Ebene von konkreten Daten (Systeme, Datenbestände, Tabellen und deren Attribute) häufig von Systemspezialisten vollzogen werden kann.

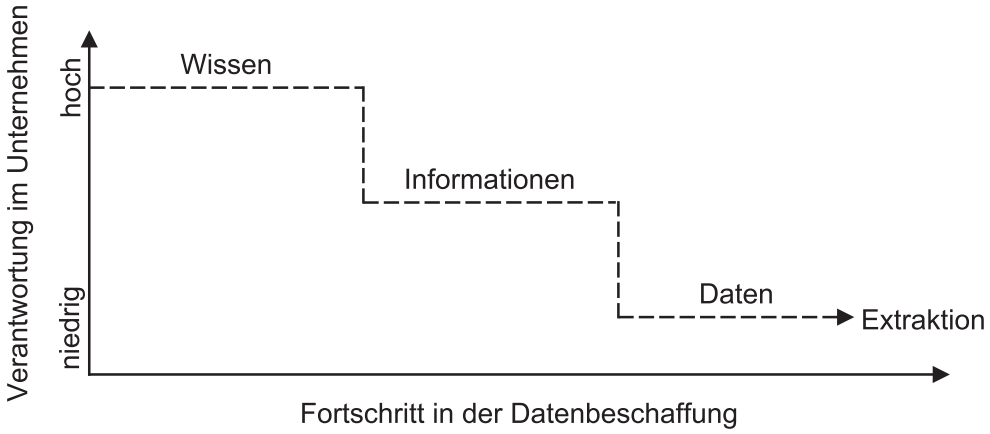



**Abbildung 4.2: Stufen der Datenextraktion in der SC**

### 4.2.2 Datenauswahl

Wenn die notwendigen Datenbestände identifiziert worden sind, erfolgt die Extraktion der Daten. Da in diesem Schritt nur die relevanten Informationen für das Data Mining extrahiert werden sollen, ist oftmals eine Reduktion der identifizierten Datenbestände notwendig. Hierfür existieren verschiedene Techniken wie Stichproben, Fensterung oder Cluster-Sample (vgl. die Phase der Datenauswahl in Abschnitt 2.3.2). Wie bereits bei der Analyse der Charakteristika von SC-Daten festgestellt, wird ein Großteil der vorliegenden Datenbestände von den Transaktionsdaten gebildet. Für die Transaktionsdaten sind die Zeiträume ausschlaggebend (vgl. Abschnitt 3.2). Die Folge ist, dass aufgrund des zeitlichen Bezuges Stichprobenverfahren als ungeeignet eingestuft werden, da diese die temporären Zusammenhänge zwischen einzelnen Transaktionen nicht berücksichtigen. Ein weiterer Kandidat sind die Cluster-Samples, die sich jedoch in praktischen Versuchen dieser Arbeit als ungeeignet für die SC-Datenreduktion herausstellten. Die Cluster-Verfahren wiesen hierbei hohe Laufzeiten auf den SC-Datenbeständen auf, sodass ein praktischer Unternehmenseinsatz für die SC-Datenreduktion nicht zu empfehlen

ist. Neben dem angesprochenen Laufzeitproblem ist zudem nach der automatischen Clusterbildung eine Stichprobe aus den einzelnen Clustern zu ziehen. In der Folge bleiben die Zusammenhänge zwischen den Transaktionen sowohl in der Clusterbildung als auch bei den einzusetzenden Stichprobenverfahren unberücksichtigt. Als drittes wurde die sachlogische Partitionierung adressiert, die wertebasiert und horizontal auf den SC-Datenbeständen arbeitet. Da jedoch zuvor die relevanten Daten entnommen werden müssen, um fachlich unsinnige Daten auszuschließen (vgl. die Phase der Datenauswahl in Abschnitt 2.3.2), wird in diesem Schritt nur die eigentliche Entnahme der Daten adressiert. Die Reduktion der Datenbestände durch die gezielte Datenentnahme kann sowohl das Volumen als auch die Komplexität betreffen (vgl. Abschnitt 2.2.2). Die Beschränkung auf die reine Datenreduktion ist auch im Hinblick auf das geforderte Standarddatenformat sinnvoll (vgl. Punkt 3.1 in Tabelle 3.3), denn die Entnahme von fachlich sinnvollen Datenbeständen ermöglicht ein parametervariables Partitionieren in späteren Phasen.

## 4.3 Datenaufbereitung

Diese Phase behandelt die Datenaufbereitung und beinhaltet die Techniken, die in den Phasen des Preprocessings und der Transformation in Abschnitt 2.3.2 eingeführt wurden. Diese Phase erfordert sowohl Kontextwissen über die Daten als auch technisches Wissen über die Techniken der Datentransformation.

### 4.3.1 Formatstandardisierung

Die Überführung in ein Standarddatenformat ist insbesondere dann notwendig, wenn die Daten aus verschiedenen Quellen stammen und in der Folge zusammengeführt werden müssen. Dieser Vorgang ist auch als Technik des Preprocessings in Abschnitt 2.3.2 unter den Fachbegriffen Daten- und Schemaintegration referenziert. Die Extraktion der Daten, die in der Datenauswahl (vgl. Abschnitt 4.2.2) diskutiert wurde, kann auf verschiedenen Wegen erfolgen. Die häufigsten Extraktionen in der industriellen Praxis im Bereich SC sind das blockweise Auslesen der Daten sowie das unmittelbare Überführen von einzelnen Entitäten in ein Zielformat. Aus diesen beiden Extraktionsmechanismen resultieren unterschiedliche Ausgangsformate, die im Folgenden einer Struktur- und Formatvereinheitlichung unterzogen werden. Hierzu sind gegebenenfalls die extrahierten Datenbestände der SC in eine spaltenorientierte Darstellung umzuwandeln. Dabei müssen die extrahierten Datenbestände aus verschiedenen Quellen vereinheitlicht werden und existierende Relationen, z. B. zwischen einzelnen Tabellen, dürfen nicht entfallen. Gegebenenfalls müssen neue Attribute in das Standarddatenformat integriert werden, um bestehende Relationen zu erhalten. Die Aggregationsstufe (vgl. Abschnitt 2.2.2.1) ist bei der Integration von Daten aus verschiedenen Quellsystemen von besonderer Bedeutung, denn Daten aus verschiedenen Aggregationsstufen las-

sen sich nur schwer miteinander verknüpfen. Die Begründung liegt in den fehlenden Relationen und Schlüsseln zwischen unterschiedlichen Aggregationsstufen. Existieren beispielsweise in einem Data Warehouse die Einzelposten eines Auftrages nicht mehr, sondern nur noch die Auftragsnummer, ist eine Verknüpfung über die eindeutigen Artikelnummern aus den Stammdaten kaum möglich. Häufig tritt im SC-Umfeld auch der Fall auf, das Attribute in unterschiedlichen Systemen geführt werden. Dann müsste in das Standardformat ein Attribut für die Repräsentation des Quellsystems eingefügt werden, so dass die Eindeutigkeit der existierenden Attribute gewährleistet wird. Die Darstellung der Daten kann auf mehrere Arten geschehen z. B. in Tabellen-, Graph- und Textdarstellung. Wichtig dabei ist, dass die Darstellungsart keine zusätzlichen Relationen anhand der Reihenfolge der Datensätze erzeugt oder unterschiedlichen Datensätzen anhand der Position eine veränderte Gewichtung gibt. Um in nachgelagerten Schritten die Daten in entsprechende Data-Mining-Werkzeuge zu importieren, müssen diese nun in ein passendes Format übertragen werden. Die meisten Data-Mining-Werkzeuge können mit verschiedenen Datenformaten arbeiten. Zu den gängigsten Formaten für diese Werkzeuge gehören Comma-separated values (CSV) und Extensible Markup Language (XML). Des Weiteren wird eine Anbindung zu relationalen Datenbanken unterstützt, die den Datenaustausch zwischen den Servern und dem Werkzeug ermöglicht. Die Formate CSV und XML können direkt in die Data-Mining-Werkzeuge importiert werden. CSV ist ein spaltenorientiertes Dateiformat, wohingegen XML ein hierarchisch strukturiertes Dateiformat ist. Eine relationale Datenbank unterstützt verschiedene Dateiformate für die Exportvorgänge.

Soweit nicht projektspezifische Gründe für XML sprechen, empfiehlt sich für den SC-Datenbestand die spaltenorientierte Darstellung in CSV oder eine direkte Anbindung an relevante SQL-Tabellen. Da die Data-Mining-Werkzeuge zumeist XML in Tabellenform überführen, bietet sich die direkte Darstellung in einem spaltenorientierten Standardformat für SC-Datenbestände an. Der Vorteil liegt insbesondere in der konstanten Darstellung über mehrere Vorgehensmodellschritte hinweg, da die Daten in der SC zumeist in Datenbanken gespeichert werden (vgl. Abschnitt 2.2.2) und auch die Darstellung in den meisten Data-Mining-Werkzeugen tabellarisch erfolgt. Zusätzlich bedeutet die fehlende Umwandlung von SC-Datenbanken zum Standarddatenformat sowie das Entfallen der internen Umwandlung in den Data-Mining-Werkzeugen eine Zeitersparnis.

### 4.3.2 Gruppierung

In diesem Schritt erfolgt die fachliche Gruppierung der Datenbestände unter Berücksichtigung der Aufgabenstellung. Die getesteten Gruppierungsverfahren entsprechen den diskutierten Hauptverfahren der Datenauswahl in Abschnitt 2.3.2. Die Auswahl von Untermengen mittels Kontextwissen wurde bereits unter den Anforderungen diskutiert (vgl. z. B. Abbildung 3.1). Die manuelle Auswahl, die mittels SQL-Abfragen direkt auf der Datenbank vollzogen wurde, benötigt Kon-

textwissen, um die Auswahlkriterien festzulegen. Die Fensterung benötigt das Kontextwissen in der Parametrierung, da hier Schrittgröße und Schrittweite den sachlogischen Zusammenhang der Daten berücksichtigen müssen. Da in der SC auch Zusammenhänge zwischen zeitlich versetzten Transaktionen aufzufinden sind (vgl. Abschnitt 2.2.2.3), bleibt die Forderung nach möglichst großen Untermengen des ursprünglichen Datenbestandes bestehen. Beide Verfahren können in Abhängigkeit zur Aufgabendefinition (vgl. Abschnitt 4.1) auch in Kombination eingesetzt werden. Überraschend waren die Ergebnisse der Clusterung, die auf dem Datenbestand B.2 durchgeführt wurden. Hierbei stellte sich die Clusterung, trotz Einsatz unterschiedlicher Cluster-Verfahren, als ungeeignet für die Gruppierung von SC-Datenbanken heraus.

Um zu verdeutlichen, warum die Clusterung zur Gruppierung der SC-Daten ungeeignet ist, wird an dieser Stelle eines der relevanten Experimente auszugsweise dargestellt. Das Experiment wurde auf den Transaktionen eines Zulieferernetzwerkes durchgeführt, um verschiedene Verfahren für die Datengruppierung zu untersuchen. Abbildung 4.3 zeigt ein Experimentergebnis, in dem zu erkennen ist, dass das eigentlich ausschlaggebende Attribut der Transportmittel keine verwertbare Zuordnung zu Clustern aufweist und andere Kriterien, wie beispielsweise Herstellernetzwerke, als stärkere Einflussfaktoren identifiziert wurden. Die Abszisse enthält hierbei die zwölf unterschiedlichen Transportmittel, die durch Nummern repräsentiert werden. Auf der Ordinate sind die Herstellernetzwerke abgetragen und die Applikate gibt die Liefermengen der einzelnen Bestellungen an. Die Farben der einzelnen Elemente repräsentieren die Cluster. Die Clusteraufzählung ist von links nach rechts absteigend bezüglich der Größe, d. h. der Menge der zugeordneten Objekte pro Cluster, sortiert. Beachtet man den Farbverlauf in der Abbildung, so fällt auf, dass die Transportmittel über die unterschiedlichen Cluster verteilt sind, so dass eine Zuordnung nicht mehr möglich ist.

Die unterschiedlichen Clusterverfahren wurden auf zwei Datenbeständen verifiziert und haben in der Zusammenfassung gezeigt, dass je nach verwendetem Cluster-Verfahren ein spezifischer Datensatz in verschiedene Cluster fällt (vgl. Tabellen B.5 und B.6 im Anhang).

Die Begründung für das Experimentergebnis liegt in den fehlenden Metriken, die bei der Zuordnung der Transaktionen keine sinnvolle Ähnlichkeitsbeziehung zwischen einzelnen Clusterelementen gestattet. Im Rückschluss entfällt die Clusteranalyse für den Vorverarbeitungsschritt der Gruppierung, da eine spezifische Merkmalskodierung für eine Großzahl an Attributen den Nutzwert der Cluster nicht rechtfertigt.

### 4.3.3 Datenanreicherung

Die Aufgabe dieses Schrittes besteht in Datenanreicherung unter Zuhilfenahme von Kontextwissen. Die Formulierung „unter Einbeziehen von Wissen aus höheren

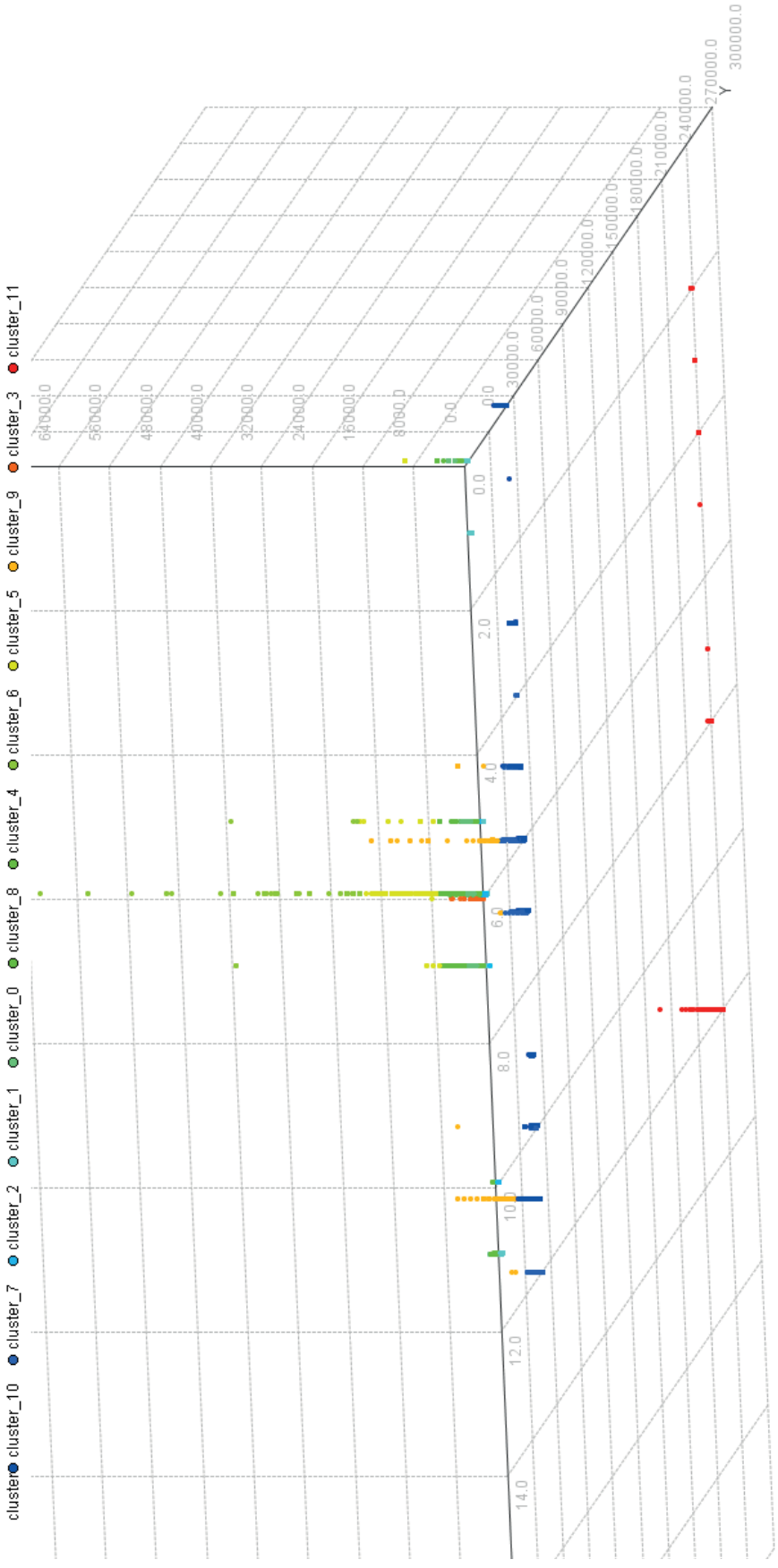



**Abbildung 4.3: Clusterung mit k-means und k = 12 auf 100 000 Transaktionen eines Zuliefernetzwerkes**

Aggregationsebenen“ des RM in Tabelle 2.3 spiegelt das Einbeziehen von Kontextwissen wider, denn in der SC-Landschaft gibt es eine Vielzahl von Systemen auf unterschiedlichen Aggregationsstufen und mit unterschiedlichem Informationsgehalt (vgl. Abschnitt 2.2.2.1). Der Begriff Kontextwissen geht allerdings über die Aggregationsstufen hinaus, denn er umfasst auch Wissen, das mitunter nicht in der bestehenden SC-Landschaft technisch erfasst ist. Solches Wissen muss dann von Experten beigesteuert werden, die diesen Schritt begleiten. Das Kontextwissen wird genutzt, um die Datenanreicherung mittels Operationen auf Attributsebene durchzuführen. Die Datenanreicherung kann mittels dreier unterschiedlicher Operationen auf dem SC-Datenbestand durchgeführt werden:

1. Aggregation - zwei oder mehr bestehende Attribute werden aggregiert, um ein neues Attribut zu erzeugen. Die Attribute der Operation werden nach dem Ausführen der Aggregation entfernt.
2. Addition - dem Datenbestand wird ein neues Attribut hinzugefügt. Die bestehenden Attribute bleiben unverändert.
3. Transformation - die Merkmalsausprägung eines bestehenden Attributes wird unter Zuhilfenahme von Kontextwissen oder Informationen aus höheren Aggregationsstufen transformiert. Der Datenbestand bleibt ansonsten unverändert.

Tabelle 4.1 zeigt die drei Operationen auf einem Auszug von Transaktionsdaten aus der Automobilbranche. Das Attribut „Änderung“ wurde durch die Aggregation von Attributen erzeugt, die Auskunft über den Bestellstatus einer Lieferung gestatten. Das Attribut „Lieferdauer“ wurde aus der Differenz des Versand- und Lieferdatums der jeweiligen Transaktion erzeugt und als zusätzliches Attribut in die spezifische SC-Datenbank eingefügt. Das Attribut „Zielort“ wurde durch die Anwendung von Kontextwissen generiert. Mittels unternehmensspezifischen Wissens des SCM zu Vertriebsstrukturen konnten einzelne Adressen auf übergeordnete Vertriebsregionen transformiert werden.

**Tabelle 4.1: Auszug der Datenanreicherung auf den Automobil-SC Daten**

| Transaktionsnummer | Änderung | Lieferdauer | Zielort |
|---|---|---|---|
| 37 721 426 | 1 | 3 | Italien 3 |
| 37 678 834 | 1 | 5 | Italien 4 |
| 45 920 982 | 0 | 8 | Italien 2 |
| 46 389 098 | 0 | 3 | Italien 2 |
| 63 070 080 | 0 | 20 | Thailand 1 |
| 65 854 764 | 1 | 5 | Italien 1 |
| 66 757 958 | 0 | 10 | Italien 1 |
| 66 758 094 | 1 | 20 | Italien 2 |

### 4.3.4 Transformation

Die Transformation stellt eine Sammlung von Techniken zur Veränderung von Attributen und Attributsausprägungen in den Datenbeständen dar (vgl. Abschnitt 2.2.2.1). Da es bezüglich der Ausführungsreihenfolge keine Untersuchungen gibt, gestaltet sich die Reihenfolge flexibel und kann bei Bedarf abgewandelt werden. Des Weiteren ist jede der aufgeführten Techniken optional und wird maßgeblich von den vorliegenden Datenbeständen sowie den Aufgabenstellungen und Verfahren des Data Minings beeinflusst.

Bei den Transformationstechniken im SC-Kontext ist die Behandlung von fehlenden und fehlerhaften Merkmalen im Rahmen des Data Cleansing (vgl. Data Cleansing im Preprocessing in Abschnitt 2.3.2) ein Sonderfall, der nur eine untergeordnete Bedeutung für MESC hat. Da in der SC die anfallenden Daten im Regelfall über Datenbanken, Data Marts oder Data Warehouses verwaltet werden (vgl. Abschnitt 2.2.2.1), die alle über interne Prüfmechanismen zum Erhalt der rudimentären Datenqualität verfügen, sind offensichtliche Qualitätsmängel im Praxisalltag selten. Data Cleansing übernimmt als Folge in MESC nur die Aufgabe einer rudimentären Datenqualitätssicherung und stellt einen optionalen Schritt dar. Rudimentär umfasst in diesem Kontext nur den Aspekt der reinen Datenqualität und nicht den Aspekt der Qualitätssicherung von Strukturelementen der SC-Daten (vgl. Abschnitt 2.2.2.4). Die Techniken der Ausreißerkorrektur, die dem Data Cleansing zugeordnet werden, können mittels unterschiedlicher Ansätze vollautomatisiert auf SC-Datenbanken genutzt werden.

Um die Ausreißerkorrektur auf den SC-Datenbeständen in dieser Arbeit zu erforschen, wurden verschiedene Experimente durchgeführt, die im Folgenden auszugsweise dargestellt werden. In Experimenten auf den unterschiedlichen SC-Datenbeständen (vgl. Tabellen B.1 und B.2) stellte sich heraus, dass unterschiedliche Verfahren der Ausreißerkorrektur eine auffällige Menge an Werten als Ausreißer in SC-Datenbanken klassifizierten. Die Folge wäre bei einer Standardanwendung der Data-Cleansing-Techniken eine Elimination oder Glättung der entsprechenden Werte (vgl. Data Cleansing im Preprocessing in Abschnitt 2.3.2). Die besagten Ausreißer stellten sich jedoch in den nachfolgenden manuellen Untersuchungen und Analysen als interessante Information für die Wissensentdeckung heraus. Beispielsweise wurde durch die Ausreißerkorrektur eine große Abweichung von zeitlichen Durchschnittswerten – ein klassischer Ausreißer – geglättet. Beim Vergleich der korrigierten Daten mit den Rohdaten stellte sich jedoch heraus, dass die korrigierten Ausreißer teilweise Bestandteil von Regeln auf den SC-Datenbeständen waren. Als Folge wurden die weiteren Transformationstechniken nicht nur auf ihre technische Anwendungsmöglichkeit geprüft, sondern auch die resultierenden Ergebnisse in einem iterativen, fachlich geführten Prozess hinterfragt. In den prozessbegleitenden Analysen hat sich gezeigt, dass konkrete Schritte der Datenaufbereitung, die Kontextwissen benötigen, nicht automatisierbar sind. Tabelle 4.2 zeigt konkrete Operationen der Vorverarbeitung, die auf einem Datenbestand von 1 570 121 Da-

tensätzen mit 67 Attributen durchgeführt wurden (vgl. Tabelle B.1). Nach Schritt 7 konnte der Datenbestand auf 1 569 499 Datensätze und 58 Attribute reduziert werden.

**Tabelle 4.2: Datenaufbereitungsoperationen auf konkretem Datensatz mit Automatisierungspotential**

| Schritt | Operation | Automatisierbarkeit | SC-Eignung |
|---|---|---|---|
| 1 | Attribute ohne Attributswerte entfernen | ja | ja |
| 2 | Entfernen aller Attribute mit nur einer Ausprägung | ja | ja |
| 3 | Vereinigung oder Elimination abhängiger Attribute | teilweise | ja |
| 4 | Data Cleansing | ja | ja |
| 5 | Umwandeln der Datumsformate zu Integer-Werten | ja | ja |
| 6 | Entfernen von Ausreißern | nein | teilweise |
| 7 | Mapping von Attributen, um die Laufzeit zu verbessern | ja | ja |
| 8 | Fensterung | teilweise | teilweise |
| | Clusterung | teilweise | nein |
| | Manuelle Auswahl mittels Kontextwissens | nein | ja |

In den Experimenten konnte die Vereinigung sowie Elimination von Attributen teilweise automatisiert werden. Hier sind in Schlussfolgerung nur die Operationen automatisierbar, die sich auf reine Attributsabhängigkeiten beschränken. Die Abhängigkeit umfasst hierbei zum einen hochkorrelierte Attribute und zum anderen redundante Informationsgehalte der Attribute. Tabelle 4.3 verdeutlicht den Sachverhalt auf SC-Daten aus dem Experimentzyklus. Die Attribute „Bewegungsart“ und „Gewichtseinheit“ sind korreliert. Der Informationsgehalt des Attributs „Rück-Menge“ ist redundant, da sich dieses als Differenz der Attribute „Ist-“ und „Soll-Menge“ errechnet.

Als Ergebnis der Experimente kann konstatiert werden, dass in SC-Datenbanken der Einsatz von Techniken zur Dimensionsreduktion (vgl. Transformationstechniken unter Abschnitt 2.3.2) zielführend ist, um die irrelevanten oder redundanten Attribute zu identifizieren. Darüberhinausgehende fachliche Beziehungen können nur über Kontextwissen identifiziert werden und diskutierte Verfahren aus dem Be-

**Tabelle 4.3: Redundanzen und Korrelationen in SC-Datenbank aus der Lebensmittelbranche**

| TA-Nummer | Bewegungsart | Gewichtseinheit | Soll-Menge | Ist-Menge | Rück-Menge |
|---|---|---|---|---|---|
| 230 454 | 601 | KG | 3 | 3 | 0 |
| 230 874 | 601 | KG | 10 | 11 | -1 |
| 230 875 | 601 | KG | 24 | 24 | 0 |
| 230 876 | 601 | KG | 2 | 2 | 0 |
| 230 877 | 601 | KG | 25 | 27 | -2 |
| 230 878 | 601 | KG | 51 | 51 | 0 |

reich der Transformationen wie beispielsweise der Korrelationstest, können keine Anwendung finden.

Tabelle 4.2 beinhaltet mit den Operationen 1-3 die Standardoperationen für die Merkmalsreduktion in SC-Datenbanken. Hierbei stellen Schritte 1 und 2 sehr einfache Schritte dar, die mittels SQL-Befehlen direkt auf den SC-Datenbanken ausgeführt werden können. Diese Schritte entfallen bei idealer Auswahl der Datenbestände (vgl. Abschnitt 4.2.2), da nur Attribute mit Informationsgehalt selektiert werden. Schritt 3, der die Vereinigung oder Elimination von Attributen beinhaltet, ist von komplexerer Natur. Es steht eine Vielzahl von Techniken bereit, um die reine Attributsabhängigkeit zu erfassen (vgl. Transformationstechniken in Abschnitt 2.3.2). Diese Techniken sind im Kontext der SC von spezieller Bedeutung, da in seltenen Fällen Datenbestände aus nur einem System genutzt werden und sich so die Abhängigkeiten und Redundanzen potenzieren. Dies lässt sich insbesondere mit der manuellen Analyse, Identifikation und Entnahme der Datenbestände in den ersten Phasen begründen. Die heterogenen IT-Landschaften globaler SCs, die eine Daten- und Schemaintegration innerhalb des Standdatenformats vorsehen (vgl. Abschnitt 4.3.1), erhöhen den Stellenwert der Merkmalsreduktionsoperationen. Für diese Operationen sind im Kontext der SC geeignete technische Unterstützungen in das Vorgehensmodell zu integrieren, da beispielsweise Korrelationen zwischen Merkmalen unterschiedlicher Kodierung nur schwer manuell zu erfassen sind (z. B. hat jeder SC-Teilnehmer ein eigenes Zeiterfassungssystem).

Die letzte Transformationsoperation, die im Kontext der SC von besonderer Bedeutung ist, ist die Prüfung auf Atomarität und die geeignete Aufteilung von nicht-atomaren Attributen. Die Notwendigkeit der Prüfung und entsprechende Beispiele wurden bereits in Abschnitt 3.2 vorgestellt. Auch diese Operation setzt Kontextwissen voraus, denn ohne Branchenkenntnisse oder Wissen von Unternehmensstandards sind in den Attributsausprägungen keine inneren Strukturen zu erkennen. Da das Auflösen der inneren Struktur im SC-Kontext keine gesonder-

ten technischen Verfahren voraussetzt, wird im Rahmen der MESC auf weitere Untersuchungen der Attributsaufteilung verzichtet.

## 4.4 Vorbereitung des Data-Mining-Verfahrens

In dieser Phase werden die Vorbereitungen für die Data-Mining-Anwendung auf den vorverarbeiteten SC-Daten getroffen. Die Phase beinhaltet die Auswahl von geeigneten Verfahren sowie sämtliche fachlichen und technischen Vorbereitungsschritte für die letztendliche Verfahrensanwendung innerhalb eines Data-Mining-Werkzeugs.

### 4.4.1 Verfahrensauswahl

Das Ziel dieses Schrittes ist die Auswahl eines geeigneten Data-Mining-Verfahrens für die konkrete Aufgabenstellung der ersten MESC-Phase. Da es für die Data-Mining-Verfahren prinzipiell keine Einschränkungen, jedoch Empfehlungen im SC-Umfeld gibt, wird auf die allgemeine Abhandlung zum Data Mining im Kontext der Logistik unter Abschnitt 2.3.2 verwiesen. Als Phasenresultat muss ein geeignetes Verfahren (oder wenn notwendig eine Liste von geeigneten Verfahren) dokumentiert werden.

Jedes Verfahren repräsentiert die gefundenen Erkenntnisse mittels Mustern. Die Muster stehen in direkter Wechselwirkung mit dem gewählten Verfahren, denn die Ausgabe, also das Muster, ist verfahrensabhängig. So ist neben der technischen Anwendbarkeit eines bestimmten Verfahrens auch explizit zu prüfen, ob das Muster und seine Aussagekraft zur Beantwortung der Aufgabenstellung einsetzbar sind. Wie bereits in der Forschungsfrage 1 (vgl. Abschnitt 2.4.2) ersichtlich, ist die Frage nach einem klar abgegrenzten Musterbegriff für das Vorgehensmodell tragend. Denn für den praktischen Einsatz im Unternehmen sowie eine potentielle Erweiterbarkeit des hier entwickelten Modells ist ein eindeutiges Verständnis der Begrifflichkeiten notwendig. Als Basis für die Begriffsentwicklung der Muster dient die Niemann-Funktion (vgl. Gleichung 2.2) aus der Tabelle 2.10, da diese den allgemeinen, mathematischen Ansatz repräsentiert. Die Eigenschaft vektorwertig bleibt erhalten, da die Ausgabe im Allgemeinen ein komplexes Muster und keinen Einzelwert darstellt. Dieser Musterbegriff wird auf die Definition 2.13 angewandt, sodass sich folgender Ausdruck ergibt:

$$\begin{aligned} &\mathbf{f}(x) \in \mathcal{G} \cup \mathcal{H} \cup \mathcal{P} \text{ mit} \\ &\quad \mathcal{G} = \{\mathbf{g}(\mathbf{x}) \mid \text{logisch-numerisch}\}, \\ &\quad \mathcal{H} = \{\mathbf{h}(\mathbf{x}) \mid \text{statistisch}\} \text{ und} \\ &\quad \mathcal{P} = \{\mathbf{p}(\mathbf{x}) \mid \text{elementar}\}. \end{aligned} \tag{4.1}$$

Dabei repräsentieren die Mengen $\mathcal{G}$, $\mathcal{H}$ und $\mathcal{P}$ die sogenannten Subtypklassen, die wiederum einzelne Subtypen enthalten. Das Data Mining, in seiner Bedeutung als wesentliche Phase im KDD, liefert als Ergebnis Muster der Subtypklasse logisch-numerisch (vgl. Definition 2.13). Je nach eingesetztem Data-Mining-Verfahren kann das Muster verschiedenen Subtypen zugeordnet werden. So liefert beispielsweise eine Klassifikation mittels Entscheidungsbäumen (vgl. Tabelle 2.9) als Muster den Subtyp Baum zurück.

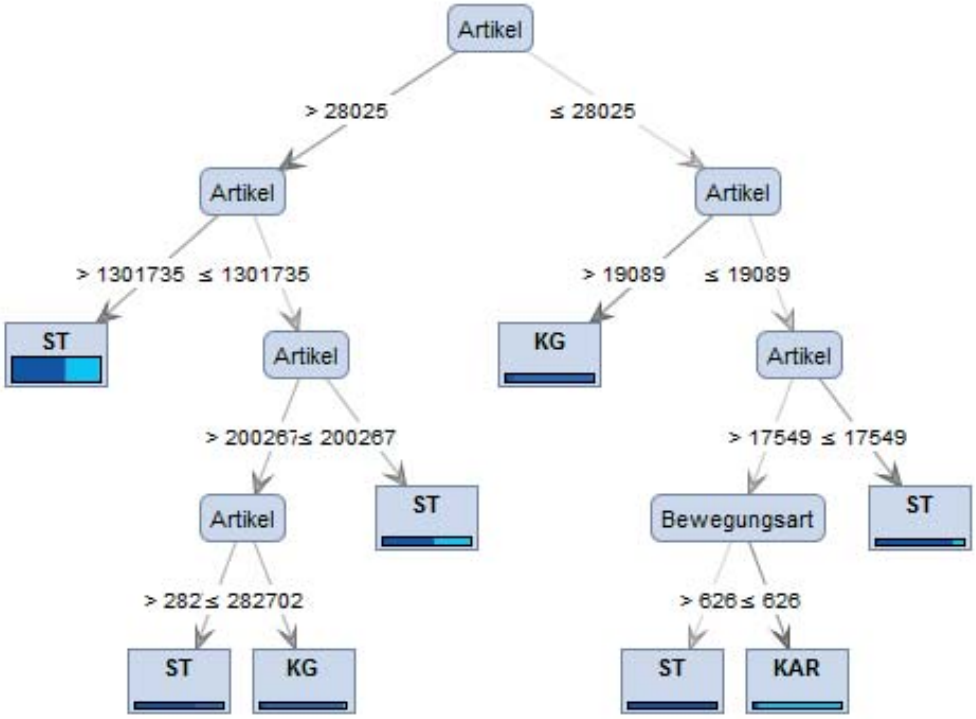
Im Rahmen von durchgeführten Experimenten für diese Arbeit wurden verschiedene Subtypen und ihre Darstellungsformen visualisiert. Abbildung 4.4 zeigt einen Entscheidungsbaum, der auf Transaktionsdaten aus der Lebensmittelbranche gebildet wurde (vgl. Tabelle B.2). Die Knoten des Baums repräsentieren verschiedene Nummernbereiche von SC-Artikeln sowie unterschiedliche Bewegungsarten der Artikel innerhalb der SC. In den Blättern des Baums sind verschiedene Maßeinheiten der Artikel aufgeführt. Es wurden die Originalmaßeinheiten aus den Attributen übernommen, die in der Legende des Entscheidungsbaums aufgeführt sind. Diese Bezeichnung wird auch in weiteren Darstellungen übernommen. Das Entscheidungsbaummodell dient zur Konsolidierung von Artikeln und zeigt die Möglichkeiten der Bildung von Transporteinheiten in Bezug auf verschiedene SC-Merkmale.

Die Darstellung des Entscheidungsbaums wurde im Anschluss in ein Regelwerk überführt (Regelwerk 4.1). Das Regelwerk stellt die Zuordnung von Artikelnummernbereichen zu den Maßeinheiten in acht Regeln dar.

Das Regelwerk lässt sich in verschiedene weitere Darstellungsformen überführen. Beispielsweise kann das Regelwerk in eine Entscheidungstabelle überführt werden. Diese enthält die identischen Informationen, allerdings in anderer Form aufbereitet (vgl. Tabelle 4.4). Die referenzierte Entscheidungstabelle enthält keine konsolidierten Regeln, um die Nachvollziehbarkeit der Darstellungsüberführung zu erleichtern. Des Weiteren handelt es sich um keine vollständige Entscheidungstabelle, da nicht zu allen Bedingungskombinationen eine Regel entdeckt werden kann (vgl. Begriff der Lokalität in Abschnitt 2.3.4). Es gibt weitere Darstellungsformen für Regeln, bei denen insbesondere die unternehmensinternen tabellarischen Darstellungen von Bedeutung für die praktische Anwendung sind.

**Regelwerk 4.1: Regelwerk auf Lebensmittel-SC-Daten**

```
1 if Artikel > 1301735 then ST
2 if Artikel <= 1301735 and Artikel > 282702 then ST
3 if Artikel > 200267 and Artikel <= 282702 then KG
4 if Artikel > 28025 and Artikel <= 200267 then ST
5 if Artikel <= 28025 and Artikel > 19089 then KG
6 if Artikel <= 19089 and Artikel > 17549 and Bewegungsart >
    626 then ST
7 if Artikel <= 19089 and Artikel > 17549 and Bewegungsart <=
    626 then KAR
8 if Artikel <= 17549 then ST
```




**Legende: KG = Kilogramm, ST = Stück, KAR = Volumeneinheit**

**Abbildung 4.4: Entscheidungsbaum auf Lebensmittel-SC-Daten**

Betrachtet man die verschiedenen Subtypen, die die Muster repräsentieren, so kann man vier Grundformen für die explizite Wissensdarstellung im SC-Kontext ableiten:

1. Grafische Repräsentation (z. B. Graphen)
2. Textuelle Repräsentation (z. B. Regeln)
3. Mathematische Ausdrücke (z. B. Regressionsgleichungen)
4. Tabellen (z. B. Clusterrepräsentation über Matrizen)

Die Darstellungsformen können, wie bereits in der letzten KDD-Phase erläutert (vgl. Abschnitt 2.3.2), prinzipiell ineinander überführt werden. Die Umformung ist mitunter direkt möglich oder bedient sich geeigneter Zwischendarstellungen. Eine Diskussion der Darstellungsumformungen soll an dieser Stelle nicht vertieft werden, da für das Verständnis der MESC die Kenntnis der wesentlichen Darstellungsformen ausreichend ist.

Nachdem die Darstellungsformen der Subtypen diskutiert wurden, stellt sich die Frage, ob sich genau ein Subtyp in jeder SC-Datenbank entdecken lässt. Betrach-

**Tabelle 4.4: Auszug aus der Entscheidungstabelle zu dem Regelwerk 4.1**

| **Entscheidungsregel** | **R1** | **R2** | **R3** | **R4** | **R5** | **R6** | **R7** | **R8** |
|---|---|---|---|---|---|---|---|---|
| Artikel > 1 301 735 | ja | nein | nein | nein | nein | nein | nein | nein |
| Artikel > 282 702 | nein | ja | nein | nein | nein | nein | nein | nein |
| Artikel > 200 267 | nein | nein | ja | nein | nein | nein | nein | nein |
| Artikel > 28 025 | nein | nein | nein | ja | nein | nein | nein | nein |
| Artikel > 19 089 | nein | nein | nein | nein | ja | nein | nein | nein |
| Artikel > 17 549 | nein | nein | nein | nein | nein | ja | ja | nein |
| Bewegungsart > 626 | – | – | – | – | – | ja | nein | – |
| | | | | Entscheidungen | | | | |
| Artikeleinheit = KG | | | X | | X | | | |
| Artikeleinheit = ST | X | X | | X | | X | | X |
| Artikeleinheit = KAR | | | | | | | X | |

**Legende:**
– irrelevant X erfüllt

tet man einen spezifischen Datenbestand, so lassen sich bezüglich eines Subtyps oftmals mehrere Ausprägungen feststellen. Diese konkreten Ausprägungen entsprechen der Muster-Instantiierung, die in Definition 2.12 beschrieben ist. Die Subtyp-Instantiierung wird explizit gegen die gegebenen Definitionen der Muster-Instantiierung abgehoben, da ein Muster immer eindeutig zu einem Subtyp gehört. Dies korrespondiert mit der Grundaussage aus Abschnitt 2.3.2, in dem ein Muster (und somit auch seine Subtypen) jeweils eine Klasse für alle Entitäten repräsentiert, die ihr gemäß einer Klassifikationsaufgabe zugeordnet wurden. Entitäten sind hierbei die einzelnen, gegebenenfalls vorverarbeiteten Datensätze (vgl. Abschnitt 2.2.2.1), die als Grundlage für die Mustersuche mittels Data Mining dienen.

Somit gibt es in den Datenbeständen mitunter nicht nur eine Ausprägung des Subtyps, sondern eine disjunkte Menge von Subtypausprägungen, die die einzelnen Instanzen darstellen. Diese Subtypausprägungen können wie folgt erfasst werden: Wenn auf $x$ eine Klassifikationsaufgabe $K$ angewendet wird, gilt:

$$K(x) = \bigcup_i \mathbf{f}_i(x) \subseteq \mathcal{G} \cup \mathcal{H} \cup \mathcal{P} \text{ mit } i \in \{0 \ldots n\}. \tag{4.2}$$

Der Ausdruck $\mathbf{f}_i(x)$ repräsentiert somit eine konkrete Ausprägung eines Subtyps, eine Instanz. Findet sich nur eine Musterinstanz, dann ist $\mathbf{f}_i(x) = \mathbf{f}(x)$. Sollte

dies jedoch nicht der Fall sein, liegen mehrere Instanzen eines Subtyps vor. Die Zusammenhänge der einzelnen Begrifflichkeiten sind im Beispiel 4.1 dargestellt.

**Beispiel 4.1:** In einer SC werden interessante Zusammenhänge über einen Wissensentdeckungsprozess aufgezeigt. In der Data-Mining-Phase wird ein Regellerner angewandt. Dieser Regellerner liefert als Ergebnis ein Regelwerk. Ein Regelwerk besteht aus einzelnen Regeln, die in unterschiedlicher Ausprägung vorliegen. Eine Regel ist eine spezifische Ausprägung einer Subtypklasse. Da das Entdecken von Regeln einen komplexen Lernprozess voraussetzt, ist die zugehörige Subtypklasse logisch-numerisch. Alle Beispiele aus dem Datensatz (in diesem Fall Zeilen einer Tabelle), die eine spezifische Regel aus dem Regelwerk erfüllen, sind Klassenbeispiele für eben diese Subtypausprägung. Die Anzahl der gefundenen Regeln entspricht der Anzahl der Subtypausprägung des Subtyps Regel in der dazugehörigen Subtypklasse.

In den bisherigen Überlegungen konnten zwar der Musterbegriff sowie seine Subtypen und Darstellungsformen definiert werden. Allerdings fehlt in diesen Überlegungen eine Berücksichtigung des zugrundeliegenden Datenbestands. Des Weiteren ist auch das Kontextwissen, dass für das SCM notwendig ist, in den bisherigen Definition nicht integriert (vgl. Abschnitt 2.2.3.2). Aus diesem Grund erfolgt als letztes die Betrachtung des Datenbestandes, der die Datensätze beinhaltet. Dieses wird über das Argument $x$ der Gleichungen 4.1 und 4.2 gebildet. Die Teilmenge d der Grunddatenmenge $D$ repräsentiert das Ergebnis der Datenauswahl (vgl. Phase 2 in Abschnitt 2.3.2). Das Kontextwissen, das zwingend in den Musterbegriff kodiert werden muss (vgl. Phase 7 in Abschnitt 2.3.2), wird in den Transformationen aufgegriffen. Die Transformationen sind als Komposition dargestellt und stellen neben den Standardoperationen des Preprocessings und der Datentransformation (vgl. Phase 3 und 4 in Abschnitt 2.3.2) das Einbeziehen des Kontextwissens dar. Das Kontextwissen weist den Attributen des Datenbestandes eine Bedeutung zu und kann so neben rein automatischen Operationen, wie beispielsweise dem Entfernen von Attributen ohne Ausprägung, sowohl das Preprocessing als auch Datentransformation beeinflussen (vgl. Beispiel 4.2). Insbesondere ist die fachliche Kodierung, die in Abschnitt 4.4.3 beschrieben wurde, eine Transformation, welche Kontextwissen zugrundelegt. Die Komposition wurde gewählt, um die Hintereinanderausführung der einzelnen Transformationen zu verdeutlichen. Da die Ausführreihenfolge in der Literatur unbestimmt ist und nähere Untersuchungen nicht im Fokus dieser Arbeit liegen, ist die Bedeutungszuordnung der $t_1, \ldots, t_n$ irrelevant und es wird auf die allgemeinen Abschnitte zur Vorverarbeitung und Datentransformation in Abschnitt 2.3.2 verwiesen.

**Definition 4.1 Definition von $x$:** Sei $D$ eine Datenbank und $d \in D$ eine Teilmenge und $T = \{t_1, \ldots, t_n\}$ die Menge aller Transformationen, dann ist $x = t_n \circ \ldots \circ t_1(d)$.

**Beispiel 4.2:** In einer Zulieferer-SC liegt Kontextwissen in Form von Kenntnissen über die IT-Prozesse vor. Im Datenbestand wird festgestellt, dass alle Transaktionen mit einem Attribut „step-counter“ versehen sind. Dieses Attribut weist die Ausprägungen 1-6 auf. Das Kontextwissen liefert wertvolle Zusatzinformationen. Diese besagen u. a., dass der „step-counter“ bei Update-Operationen auf dem Datenbestand erhöht wird. Eine Möglichkeit der Kodierung im preprocessing wäre, nur die Transaktionen mit dem höchsten „step-counter“ in die Selektion aufzunehmen und das durchgeführte Transaktionsupdate in einem neuen Attribut zu speichern. Damit müsste potentiell nur jede sechste Transaktion untersucht werden. Es muss jedoch festgehalten werden, dass es in dieser Kodierung Informationsverlust gibt. Es kann zum Beispiel nicht mehr nachvollzogen werden, welche Attribute (z. B. Liefermengenwechsel, Lieferantenwechsel, Zeitplanänderungen) durch das Transaktionsupdate betroffen sind.

Es kann konstatiert werden, dass mit den hier aufgeführten Definitionen der Musterbegriff und seine unterschiedlichen Aspekte im SC-Kontext definiert wurden. Die Definitionen gestatten eine eindeutige Identifikation der Muster und verhindern in der Folge bei der Durchführung der MESC potentielle Ungenauigkeiten oder Missverständnisse, die sich aus unzureichenden textuellen Musterbeschreiben begründen.

### 4.4.2 Werkzeugauswahl

In Rückkoppelung mit der Auswahl eines geeigneten Data-Mining-Verfahrens muss in diesem Schritt ein geeignetes Data-Mining-Werkzeug bestimmt werden. Es gibt eine Vielzahl von unterschiedlichen Werkzeugen, die für die Wissensentdeckung in der SC zum Einsatz kommen können. Um geeignete Beurteilungskriterien festlegen zu können, wurden für diese Arbeit die gängigsten Data-Mining-Werkzeuge analysiert. Hierbei wurden die Softwaredokumentationen der Werkzeuge verglichen und aus den einzelnen Spezifikationsaspekten geeignete Kriterien abgeleitet. In diesem Kontext wurden allgemeine Kriterien identifiziert, wie beispielsweise Anschaffungs- und Wartungskosten, die auf jede Form der Softwareanschaffung zutreffen. Eine Diskussion dieser allgemeinen Kriterien ist jedoch für die Entwicklung der MESC nicht relevant; dennoch können auch eine Reihe von spezifischen Kriterien für Data-Mining-Werkzeuge festgelegt werden. Die praxisbezogenen Auswahlkriterien für Data-Mining-Werkzeuge sind:

1. Systemabhängigkeit
2. Grafikprozessor-Unterstützung
3. Rechencluster-Unterstützung
4. Anpassungsfähigkeit
5. Problemspezifische Lösungen

6. Unternehmensinternes Wissen
7. Datenschutz und Sicherheit
8. Reifegrad des Unternehmens
9. Schnittstellen
10. Interne Lieferantenkriterien

Die Auswahl der Werkzeuge wird von den zur Verfügung stehenden Betriebssystemen des Unternehmens beeinflusst. Da eine große Anzahl der Werkzeuge systemabhängig ist, entfallen einige Werkzeuge im Auswahlprozess. Zusätzlich sollte in diesem Punkt berücksichtigt werden, dass manche Data-Mining-Werkzeuge auf bestimmten Systemarchitekturen eine bessere Leistung erzielen. Da einige Werkzeuge speziell darauf ausgelegt sind die Grafikprozessoren (GPU) des Systems für das Data Mining zu nutzen, stellt dies ein weiteres Auswahlkriterium da. Durch das Einbinden von GPU können im Gegensatz zu Standardprozessoren viele Berechnungen parallelisiert werden. Um diesen Effekt jedoch effektiv nutzen zu können, wird eine spezielle Rechnerarchitektur benötigt, die das Unternehmen bereitstellen muss. Das dritte Auswahlkriterium adressiert ebenfalls die Zeitdauer der Data-Mining-Prozesse in den Werkzeugen. Hierfür bieten einige Werkzeuge die Unterstützung von Rechenclustern an, um den Rechenaufwand für das Data Mining auf mehrere Computer zu verteilen. Die Anpassungsfähigkeit des Werkzeugs an aktuelle und zukünftige Unternehmensbedürfnisse ist gerade beim Data Mining in der SC ein herausragender Punkt. Hierbei umfasst die Anpassungsfähigkeit sowohl die Erweiterbarkeit als auch die Personalisierung der Werkzeuge und wird nicht weiter unterteilt. Die Anpassungsfähigkeit wird mittels Plugins oder Erweiterungen innerhalb der Werkzeuge realisiert. Trotz der möglichen Anpassung muss berücksichtigt werden, dass es eine Vielzahl von hochspezialisierten Werkzeugen für spezifische Problemstellungen gibt. Der Vorteil dieser Werkzeuge liegt zumeist in einer geringeren Laufzeit oder der niedrigen Systemressourcenauslastung. Anschaffungs- und Wartungskosten sind ein Überbegriff für unterschiedliche betriebswirtschaftliche Kosten, die im Folgenden jedoch nicht näher erläutert werden, da diese nicht spezifisch für das Data-Mining-Werkzeug oder die SC sind. Ein wesentlicher Aspekt im Auswahlprozess ist das unternehmensinterne Wissen sowie die Erfahrung, die bereits mit einem Werkzeug bestehen. Da die Data-Mining-Werkzeuge sehr unterschiedlich von den möglichen Bedienungselementen sind, ist eine zeit- und kostenintensive Einarbeitung ein wichtiges Entscheidungskriterium.

Ein Auswahlkriterium, das im Rahmen der globalen MESC zunehmend an Bedeutung gewinnt, ist der Datenschutz. Hierbei ist das Mindestkriterium, dass das gewählte Werkzeug keine Lücken in der Systemsicherheit erzeugt oder unbeabsichtigt Daten an Dritte weiterleitet. Um darüber hinaus die Sicherheit vertraulicher Daten zu garantieren, bieten verschiedene Werkzeuge in unterschiedlichem Umfang Unterstützung an. Der Reifegrad eines Unternehmens ist ebenfalls bei der Werkzeugwahl zu berücksichtigen. Da verschiedene Reifegradanalysen auch die

Prozessebene miteinbeziehen, ist hier zu prüfen, inwieweit ein Werkzeug eventuell mit bereits zertifizierten Prozessen kommuniziert oder ob es andere Einflussfaktoren in diesem Bereich gibt. Die verfügbaren Schnittstellen zum Export der Data-Mining-Ergebnisse beeinflussen ebenfalls die Werkzeugwahl. Im günstigen Fall existieren bereits geeignete Programme zur Auswertung der Schnittstellen oder die Umwandlung in ein Standardformat (vgl. Abschnitt 4.3.1) wird unterstützt. Sollte das Werkzeug nur über proprietäre Ausgabemöglichkeiten verfügen, kann es je nach Verbreitung des Werkzeugs zu Interoperabilitätsbarrieren kommen. Zuletzt wird die Auswahl von spezifischen Lieferantenkriterien bestimmt, die auch für den Einsatz von Data-Mining-Werkzeugen im Unternehmenskontext gelten. Sollte der Hersteller des Data-Mining-Werkzeugs unternehmensinternen Kriterien nicht entsprechen, kann dies zum Ausschluss des Werkzeugs führen. Je nach Branche, vorliegenden Daten oder Randbedingungen kann es weitere spezifische Kriterien zur Auswahl des Werkzeugs geben. Beispielsweise unterscheiden sich die Algorithmen der Data-Mining-Verfahren bezüglich der möglichen Implementierungen in den verschiedenen Werkzeugen, sodass Laufzeitvariationen entstehen können. Solche Kriterien sind schwer zu erfassen und noch schwerer ohne geeignete Vorstudien zu bewerten. Die aufgeführten Aspekte haben folglich keinen Anspruch auf Vollständigkeit, bieten jedoch im praktischen Einsatz eine Richtlinie für den Werkzeug-Entscheidungsprozess.

### 4.4.3 Fachliche Kodierung

Die fachliche Kodierung der Attribute beschreibt die Attributsumwandlung unter der Zuhilfenahme von Kontextwissen für das Data Mining. Dieses Kontextwissen ist notwendig, da SC-Attributsausprägungen oftmals sinnvoll zusammengefasst werden müssen, um geeignete Data-Mining-Verfahren anzuwenden (vgl. Abschnitt 3.2). Das Zusammenfassen benötigt geeignete Kriterien, die bestimmen, welche Attributsausprägungen oder Attributskombinationen ähnlich sind. Die Definition einer Ähnlichkeit ist ohne Kontextwissen jedoch kaum möglich, denn gängige Metriken können im SC-Umfeld keine Anwendung finden (vgl. Abschnitt 2.3.2). Um Kontextwissen sinnvoll einzusetzen, müssen zu Beginn die Attribute bestimmt werden, die kodiert werden müssen. Darauffolgend muss geprüft werden, ob diese Attribute eine fachliche Kodierung benötigen oder ob der Schritt übersprungen werden kann und eine technische Kodierung ausreichend ist.

Sollten mehrere Attribute für die Kodierung ausgewählt werden, so ist für jedes Attribut separat eine geeignete fachliche Kodierung zu definieren. Die fachliche Kodierung bedeutet im Kontext der SC-Daten die Bestimmung einer geeigneten Metrik. Die Metrik verfolgt das Ziel, eine Ähnlichkeit zwischen Attributsausprägungen quantifizierbar zu gestalten und somit die Anzahl der Attributsausprägungen zu reduzieren. In der Tabelle 4.1 ist das Attribut Zielort aufgeführt. Dieses Attribut könnte beispielsweise fachlich kodiert werden, indem die unterschiedlichen Lokalitäten zu einem Zielort pro Land kumuliert werden. In dem angegebe-

nen Tabellenauszug würde dies die Anzahl der Merkmalsausprägungen auf zwei reduzieren (Italien und Thailand). Somit stellt die fachliche Kodierung die bereits erwähnte Skalentransformation auf Attributsebene dar (vgl. Abschnitt 3.5). In der Tabelle 4.5 wurden Lagermerkmale aufgeführt und mögliche Metriken ergänzt. Dieser Vorgang muss separat für jedes zu kodierende Attribut durchgeführt werden und es ist ersichtlich, dass ein Großteil der vorliegenden Attribute aus SC-Beständen verschiedenartige Metriken verwendet.

**Tabelle 4.5: Beispielhafte Lagermerkmale nach ten Hompel und Schmidt (2008) und mögliche Metriken**

| Merkmal | Beschreibung | mögliche Metrik |
|---|---|---|
| Durchschnittlicher Lagerbestand | Gibt an, wie hoch der Bestand der gelagerten Güter durchschnittlich an jedem Tag der Periode ist | Die durchschnittlichen Lagerbestände von eingelagerten Gütern sind ähnlich, wenn die Tageswerte der jeweiligen Lagerbestände in einer vergleichbaren, numerischen Größenordnung liegen. |
| Durchschnittliche Lagerdauer | Gibt den Zeitraum zwischen Lagereingang und -ausgang eines Lagerguts während einer Periode an | Die durchschnittlichen Lagerdauern von eingelagerten Gütern sind ähnlich, wenn die jeweiligen Lagerdauern in einer vergleichbaren, numerischen Größenordnung liegen. |
| Lagereinrichtung | Umfassen z. B. ortsfeste sowie verfahrbare Regale und Schränke | Die Lagereinrichtungen von verschiedenen Lagergütern sind ähnlich, wenn die Art, das Gewicht, das Volumen und auch die Menge der einzulagernden Gütern ähnlich zueinander sind. |
| Lagerhaltungskostensatz | Gibt an, wie hoch die Lagerhaltungskosten in Abhängigkeit vom gelagerten Warenwert sind | Zwei gelagerte Güter haben einen ähnlichen Lagerhaltungskostensatz, wenn auch deren Lagerbestände ähnlich zueinander sind. |

**Tabelle 4.5: Beispielhafte Lagermerkmale nach ten Hompel und Schmidt (2008) und mögliche Metriken (Fortsetzung)**

| Merkmal | Beschreibung | mögliche Metrik |
|---|---|---|
| Lagerreichweite | Gibt den Zeitraum an, für den der durchschnittliche Lagerbestand bei einem geplanten Materialverbrauch ausreicht | Die Lagerreichweiten von gelagerten Gütern sind ähnlich, wenn die Einlagerungszeit der Güter bei einem geplanten Materialverbrauch eine ähnliche Größenordnung besitzt |
| Lagerstandort | Standpunkt von Lagereinrichtungen | Zwei Lagerstandorte sind ähnlich, wenn sie dieselben Kunden beliefern und gleiche Transportkosten aufweisen. |
| Lagerumschlagshäufigkeit | Gibt an, wie oft sich das Lagergut innerhalb einer Periode umschlägt bzw. verbraucht, verkauft oder durch ein neues Lagergut ersetzt wird | Die Lagerumschlagshäufigkeiten von gelagerten Gütern sind ähnlich, wenn sie im gleichem Maße verbraucht, verkauft oder ersetzt werden. |

Es kann konstatiert werden, dass die Definition der individuellen Metriken in SC-Datenbanken von dem Kontextwissen abhängig ist. Aus diesem Grund muss die Fachseite in diesen Entwicklungsschritt einbezogen werden. Da es sich im Regelfall um individuelle Metriken für einzelne Attribute handelt, besteht kein sichtbares Automatisierungspotential und der Schritt muss als zeitintensive Bearbeitung in MESC berücksichtigt werden.

### 4.4.4 Technische Kodierung

Der Fokus der technischen Kodierung liegt auf der Attributsumwandlung und nicht auf der Format- oder Strukturumwandlung wie im Abschnitt Standardformat (vgl. Abschnitt 4.3.1). Die Attributsumwandlung in SC-Datenbeständen benötigt im Gegensatz zur fachlichen Kodierung in Abschnitt 4.4.3 kein Kontextwissen. Das Ziel der technischen Kodierung ist die Umwandlung der einzelnen Attribute in ein geeignetes Format. Das geeignete Format bezieht sich hierbei auf den Datentyp (vgl. Tabelle A.1), der sowohl von dem Data-Mining-Verfahren wie auch von dem einzusetzenden Data-Mining-Werkzeug bestimmt wird. Die Möglichkeiten der technischen Kodierung beschränken sich hierbei auf die Datentypumwandlung, die

im Kontext der SC eine große Herausforderung darstellt. Da eine Vielzahl der Attribute eine zu variantenreiche Attributsausprägung aufweisen (vgl. Abschnitt 3.2) und aufgrund der Beschaffenheit von Transaktionsdaten im Regelfall unterschiedliche Identifikator (IDs) für Artikel, Bestellungen, SC-Teilnehmer sowie Zeitstempel mitgeführt werden, muss eine technische Kodierung genau geprüft werden. Tabelle 4.6 zeigt eine Übersicht von Attributen, Datentypen und möglichen Ausprägungen aus einem realen SC-Bestand.

**Tabelle 4.6: Auswertung der Lebensmittel-SC Daten bezüglich Merkmalsausprägung und Datentyp**

| Attributsname | Ausprägungsanzahl | Datenbereich | Datentyp nach Tabelle A.1 |
|---|---|---|---|
| TA-Nummer | 109 579 | 227 390 - 341 297 | MEDIUMINT |
| TA-Position | 87 | 1 - 87 | TINYINT |
| Artikel | 4 259 | 835 - 6 574 760 | SMALLINT |
| Artikeleinheit | 11 | BTL - UMP | CHAR(3) |
| Quit-Datum | 161 | 01.10.2014 – 31.03.2015 | DATE |
| Quit.-Zeit | 39 855 | 00:00:01 - 22:32:03 | TIME |
| Mitarbeiter | 37 | ANER - ZEIT05 | CHAR(5) |
| Von-Platz | 39 650 | 0 - ZAW O.WARE | CHAR(11) |
| An-Platz | 50 904 | 0 - ZAW O.WARE | CHAR(11) |
| Bewegungsart | 26 | 101 - 999 | TINYINT |
| Gewicht | 4 013 | 0 - 996 | DOUBLE |
| Gewichtseinheit | 1 | KG | CHAR(2) |
| Soll-Menge | 1 255 | 0 - 999 | DOUBLE |
| Ist-Menge | 1 235 | 1 - 999 | DOUBLE |
| Rück-Menge | 227 | -1 493 - 0 | DOUBLE |

Es wird deutlich, dass ein rein technisches Überführen der SC-Daten nur für einige wenige Werte sinnvoll erscheint und ein Großteil der Attribute bereits in dem vorgelagerten Schritt eine zusätzliche fachliche Kodierung benötigt. Nach der Attributsauswahl können die so selektierten Attribute einer technischen Kodierung unterzogen werden. Für die Kodierung eignen sich im Wesentlichen folgende Funktionen:

1. Discretize by Size:
   Weist ganzzahlige Werte einem Bereich zu. Die Größe der Bereiche wird durch Parameter gesetzt.

2. Discretize by Binning:
   Weist ganzzahlige Werte einem Bereich zu. Die Anzahl der Bereiche wird durch Parameter gesetzt.
3. Discretize by Frequency:
   Weist alle verfügbaren ganzzahligen Werte einem Bereich zu. Die Anzahl der Bereiche wird durch Parameter gesetzt und die Größe der Bereiche wird anhand der Häufigkeit der Merkmalsausprägungen erzeugt.
4. Discretize by Entropy:
   Weist ganzzahlige Werte einem Bereich zu. Der Bereich wird automatisch so gewählt, dass die Entropie möglichst gering ist, das heißt die Daten je nach Datentyp möglichst gleichartig und geordnet sind. Die Entropie ist ein Begriff der Informationstheorie, die mathematische Grundlagen zur Messung des Informationsgehalts von Daten liefert.
5. Nominal to Numerical, Text, Date:
   Ersetzt alle Werte durch ganzzahlige Werte, Wortketten oder Datumswerte. Diese Operation kann vom Benutzer parametrisiert werden.
6. Numerical to Binominal, Polynominal:
   Ersetzt alle ganzzahligen Werte durch Binomialwerte. Hierbei lassen sich Schranken angeben, die die Wertebereiche von true oder false bestimmen.
7. Date to Numerical:
   Weist jedem Datum einen ganzzahligen Wert abhängig von einem benutzerdefinierten Datum zu.

Insbesondere bei den Discretize-Funktionen gibt es weitere Varianten, für die sich jedoch in der praktischen Analyse keine Einsatzmöglichkeiten zeigten und die somit nicht weiter untersucht werden.

## 4.5 Anwendung des Data-Mining-Verfahrens

Basierend auf den ausgewählten Data-Mining-Verfahren (Abschnitt 4.4.1) erfolgen nun die Verfahrensausführung und das Modelltraining. Hierbei besteht ein wesentlicher Schritt in der Festlegung der Trainings-, Validierungs- und Testdaten.

### 4.5.1 Entwicklung des Data-Mining-Modells

In Abhängigkeit von dem gewählten Verfahren und dem Werkzeug erfolgt die Durchführung des eigentlichen Data-Mining-Verfahrens. Jedes Verfahren kann durch eine Vielzahl von spezifischen Algorithmen umgesetzt werden. In der Tabelle 2.9 wurde bereits eine Auswahl vorgestellt. Das Ergebnis der konkreten Algorithmenanwendung eines Verfahrens auf einem spezifischen Datenbestand ist das

Data-Mining-Model (DMM). Um das DMM bewerten zu können, ist eine Aufteilung der Daten in Trainings-, Validierungs- und Testdaten notwendig (vgl. Abschnitt 2.3.2). Die Festlegung von Verfahrens- und Modellparametern erfolgt auf den Trainingsdaten und Validierungsdaten, während die Testdaten zur Modellvalidierung verwendet werden (vgl. Begriffe der V&V in der Ausführung zur Interpretation der Muster in Abschnitt 2.3.2). Abbildung 4.5 zeigt die Validierung der Data-Mining-Ergebnisse mittels Testdaten (vgl. Phase 5 in Tabelle 3.4).

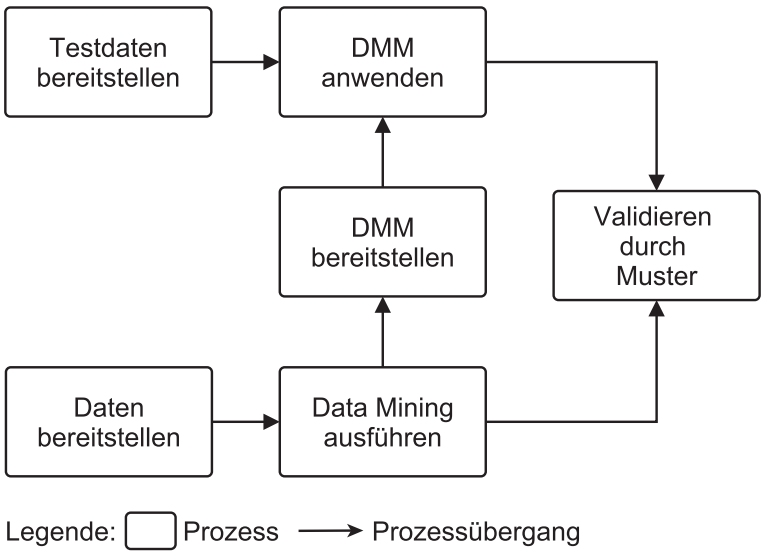



**Abbildung 4.5: Validierung der Muster mittels Testdaten**

Die Aufteilung der Daten muss zwingend vor dem nachgelagerten Test des DMM erfolgen, da die Testdaten streng von den Trainingsdaten getrennt werden müssen und Überschneidungen zu falschen Fehlerraten führen können. Bei der Aufteilung der Trainings- und Testdaten muss erneut der Zusammenhang zwischen einzelnen Transaktionen berücksichtigt werden (vgl. Abschnitt 4.2.2). Aus diesem Grund bieten sich identische Verfahren wie bei der Datenauswahl an und sachlogische Partitionierungen können basierend auf Zeitstempeln ein geeignetes Separierungsverfahren darstellen. In welcher prozentualen Form die Aufteilung erfolgen muss, hängt von den genutzten Datenbeständen ab.

Der Betrachtungsgegenstand dieser Arbeit ist die SC, die über eine Vielzahl technischer Unterstützungsmaßnahmen, z. B. die Speicherung in Datenbanken, verfügt. Daher existieren oftmals grundlegende Qualitätssicherungsmechanismen und -prozesse in der SC. Ob jedoch die Mindestqualitätsansprüche für den Einsatz von Data-Mining-Verfahren ausreichend sind, kann nicht pauschal beantwortet werden. Dies liegt insbesondere daran, dass die vorhandenen Qualitätssicherungsmechanismen im Allgemeinen andere Zwecke verfolgen, als die Eingangsqualität für das Data-Mining sicherzustellen. Sollte die Datenqualität jedoch ausreichend sein, können Holdouttechniken mit 25 % Testdaten (vgl. Abschnitt 2.3.2) Einsatz finden. Trotz verschiedener Techniken, die auch bei geringer Anzahl an Datensätzen

Anwendung finden können, ist ein unzureichender Datenbestand ein Problem in der Wissensentdeckung. Auch für den Fall, dass der Datenbestand ausreichend für das Training der DMMs ist, kann ein prinzipiell unzureichender Datenbestand zu einer fragwürdigen Aussagekraft der DMM-Tests führen.

### 4.5.2 Training des Data-Mining-Modells

Neben der Anwendung eines geeigneten Verfahrens (vgl. Abschnitt 4.4.1) steht das Training des Modells (vgl. Phase 6 in Abschnitt 2.3.2) im Vordergrund. Aufgrund der Vielzahl von Attributen in SC-Datenbanken sowie Modellparametern ist hier mit einer mehrfachen Iteration zu rechnen, bevor der Prozess beendet ist. Das DMM wird auf die Trainingsdaten angewandt und die Ergebnisse geprüft. Danach folgt eine Parameteranpassung des DMM, bis das Fehlermaß einen akzeptablen Wert aufweist. Die Techniken zur Wahl eines Fehlermaßes sind nicht spezifisch für die SC-Datenbanken, sondern hängen von einer Vielzahl von Faktoren wie eingesetzte Data-Mining-Verfahren, Datenumfang oder gewünschte Generalisierungsfähigkeit ab. Daher wird auf die Standardliteratur unter Phase 8 in Abschnitt 2.3.2 verwiesen. Welche Genauigkeit für das Modell im SC-Einsatz gefordert ist, hängt eng mit der Frage zusammen, was letztendlich die vereinbarten Zielkriterien (vgl. Abschnitt 4.1.2) für das Data Mining sind und welche Modellgenauigkeit mittels Festlegung von Fehlerkorridoren für diese gefordert wurde. Aufgrund des Stellenwerts der Trainingsphase wurde diese als separater Schritt in der MESC integriert und erfolgt nicht implizit im Data Mining. Die Begründung liegt in der Wichtigkeit des Schritts für die Modellfindung, denn ein untrainiertes Modell liefert im Regelfall Ergebnisse, die weder zu interpretieren sind noch für das SCM einen Mehrwert darstellen. Abbildung 4.6 zeigt ein untrainiertes Modell des Entscheidungsbaums, der auf den Transaktionsdaten aus der Lebensmittelbranche erlernt wurde (vgl. hierzu auch trainiertes Modell unter Abbildung 4.4). Im Gegensatz zum trainierten Modell ist keine Aussagekraft des Modells gegeben. Dies wird ersichtlich, wenn man den Baum in ein Regelwerk überführt. Im Gegensatz zum trainierten Modell, in dem die Maßeinheit der untersuchten Artikel zur Konsolidierung dienten, ist hier keine Bildung von Transporteinheiten möglich. Die Begründung liegt in der Tatsache, dass in den Knoten ein zu hoher Schwellwert für das Erreichen der Homogenität bezüglich der Maßeinheiten vorliegt. Das andere Extrem stellt einen zu niedrigen Schwellwert für die Homogenität dar. In diesem Fall besteht der Baum nur aus einem Knoten und ist nicht verwertbar. Dieses Beispiel zeigt, dass eine Vielzahl von untrainierten und trainierten DMMs in den SC-Datenbeständen entdeckt werden können.

Das Modelltraining stellt einen iterativen Verbesserungsprozess in MESC dar. Die für diesen Schritt benötigten Validierungsdaten wurden bereits im vorherigen Schritt, der in Abschnitt 4.5.1 beschrieben ist, separiert. Es sei explizit darauf hingewiesen, dass es Verfahren gibt, die auf die Verwendung von Validierungsdaten verzichten (vgl. Phase 8 in Abschnitt 2.3.2). Diese Verfahren setzen jedoch andere

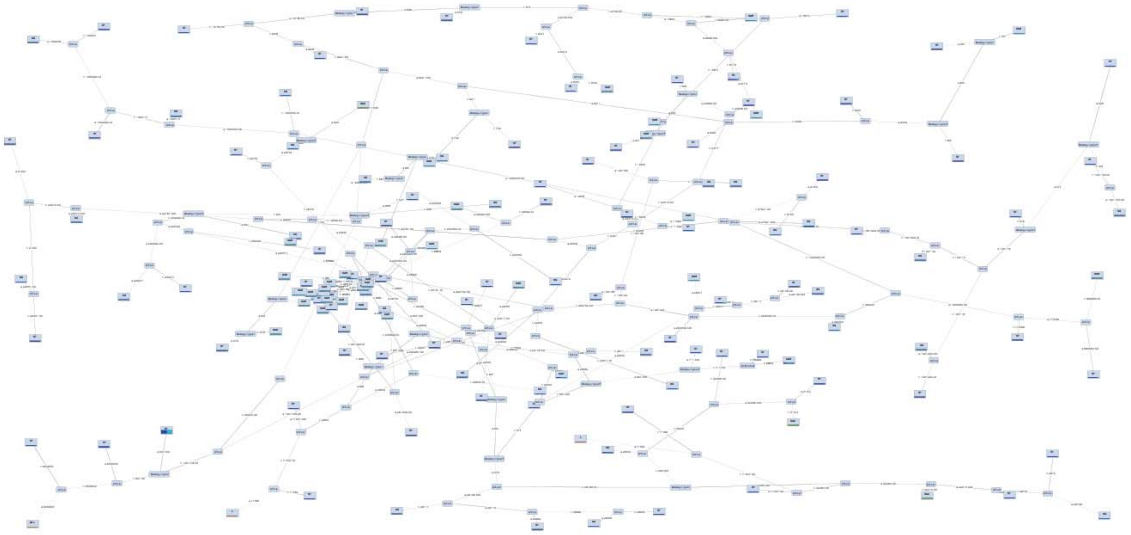


**Abbildung 4.6: Komplexität eines untrainierten DMM in der SC**

Techniken ein, sodass die Berechtigung dieses Schrittes an das Training der Modelle für den Einsatz in der SC und nicht an die Existenz von Validierungsdaten gekoppelt ist.

## 4.6 Weiterverarbeitung der Data-Mining-Ergebnisse

In dieser Phase erfolgt die Weiterverarbeitung der Data-Mining-Ergebnisse. Hierzu werden zunächst die relevanten Ergebnisse extrahiert und nach Bedarf transformiert, um letztendlich die so aufbereiteten Ergebnisse als Wissen für das Unternehmen zu nutzen.

### 4.6.1 Extraktion handlungsrelevanter Data-Mining-Ergebnisse

Aus der Gesamtmenge der verifizierten Muster müssen diejenigen Muster ausgewählt werden, die unter Berücksichtigung der Handlungsrelevanz und technischer Maßzahlen für den Anwender interessante Ergebnisse darstellen (vgl. interestingness measures in Abschnitt 2.3.2). Der Schritt ist in diesem Modell bewusst nach Anwendung der Data-Mining-Verfahren angeordnet, obwohl er fachlich keiner spezifischen Phase im Vorgehensmodell zuzuordnen ist. Die Begründung liegt zum einen im Prozess der Wissensentdeckung selbst, denn sollten Muster als interessant eingestuft werden, die noch nicht verifiziert wurden, so könnten diese mitunter ungültig sein. Dieser Umstand hätte unnötige Iterationen im vorherigen Schritt zur Folge. Zum anderen ist für die Auswahl der Muster das Kontextwissen des SCM notwendig. Gerade im Bereich der SC, wo die Datensätze von einer Vielzahl an Merkmalen geprägt sind und demnach auch die Muster über große Parametervariationen verfügen, ist das Einbeziehen des SCM unumgänglich.

### 4.6.2 Darstellungstransformation der Data-Mining-Ergebnisse

Die zuvor ausgewählten Muster werden in diesem Schritt in das Zielformat überführt. Das Zielformat hängt von der Weiterverwendung der Muster ab. Soll das Wissen, das durch das Muster repräsentiert wird, in Handlungsempfehlungen überführt werden, so ist oftmals eine explizite Darstellungsform von Vorteil. Die Begründung hierfür ist, dass bei der expliziten Wissensdarstellung in vielen Fällen auf aufwendige Darstellungstransformationen oder technische Unterstützung verzichtet werden kann. Dies wirkt sich zumeist positiv auf die benötigte Umwandlungszeit in ein geeignetes Zielformat aus. Soll das Wissen einem System zur Verfügung gestellt werden, sind theoretisch implizite und explizite Darstellungsformen möglich. Da sich MESC auf die Interpretierbarkeit der Ergebnisse durch das SCM stützt, ist die Weiterverwendung impliziter Darstellungen nicht im Fokus des Vorgehensmodells (vgl. hierzu die getroffenen Annahmen in Abschnitt 3.2). Infolgedessen sind zwei Fälle zu unterscheiden:

1. Implizit zu explizit, implizite Wissensrepäsentationen müssen in explizite Darstellungsformen überführt werden
2. Explizit zu explizit, explizite Darstellungsform wird in eine andere explizite Darstellungsform überführt.

## 4.7 Bewertung des Data-Mining-Prozesses

Zum Abschluss des MESC-Vorgehensmodells erfolgt die Bewertung des Gesamtprozesses. Dieser Schritt unterscheidet sich erheblich von der Bewertung der Data-Mining-Ergebnisse mittels des Dreiecksmodell-Bausteins (vgl. Abschnitt 3.6). Die Ausführung des Dreiecksmodellbausteins führt zu einer kontinuierlichen Anwendung der V&V. Dadurch ist die Validität des Data-Mining-Prozesses gesichert und der Fokus kann auf die Qualitätskontrolle der Prozesskennzahlen und die Identifikation von möglichen Optimierungspotentialen gerichtet werden.

### 4.7.1 Qualitätskontrolle des Data-Mining-Prozesses

Die Qualitätskontrolle des Data-Mining-Prozesses ist eine interdisziplinäre Aufgabe, da hier technisches, fachliches und unternehmensbezogenes Wissen notwendig ist. Im Vordergrund steht eine Kontrolle aller Phasen und Schritte im Hinblick auf mögliche Schwachstellen sowie Verbesserungspotentiale. Der Data-Mining-Prozess kann sowohl qualitativ wie auch quantitativ bewertet werden. Bei der qualitativen Bewertung liegt der Fokus auf der dokumentierten Prozessbeschreibung sowie möglichen Optimierungspotentialen. Die quantitative Bewertung des Data-Mining-Prozesses konzentriert sich auf die Prozessleistung und eine Evaluation, ob die Ziele aus Phase 1 erreicht wurden. Das Ergebnis der Prozessbewertung ist jedoch

unabhängig von den vorherigen Prüfergebnissen. Zur Überprüfung der einzelnen Schritte kann jedoch das Dreiecksmodell nützlich sein (vgl. Abschnitt 3.6). Hierbei können die Ergebnisse der Prüfvorgänge als Nachweis für die Prozesskorrektheit dienen und die Protokolle der einzelnen Prüfschritte Aufschluss über Schwachstellen und Verbesserungspotential liefern. Wenn notwendig, kann diese Ausgangslage um unternehmensspezifische Qualitätskontrollmethoden ergänzt werden. Eine große Schwierigkeit ist, dass traditionelle Planungsmethodiken aus dem SCM (TQM, Scorecard) im Kontext der MESC versagen, da KDD in der SC eine nur bedingt planbare Domäne darstellt. Die bedingte Planbarkeit ergibt sich aus der Tatsache, dass zwar grundlegende Fragestellungen mit der Wissensentdeckung verbunden werden, aber KDD im Gegensatz zur Statistik eine hypothesenfreie Methode darstellt (vgl. Abschnitt 2.3). Wenn das SCM also bereits die Ergebnisse der MESC kennen würde, würde keine Wissensentdeckung betrieben werden. Wenn aber Ergebnisse unbekannt sind, kann keine Planung erfolgen und traditionelle Planungsmethodiken sind ungeeignet.

In der Wissensentdeckung auf SC-Datenbeständen kommt das iterativ-inkrementelle Vorgehen am besten zum Tragen: In jedem Schritt (und mit jedem Durchlauf der MESC) wird neues Wissen gewonnen, das vorher noch nicht vorhanden war. Erst aus diesem Wissen lässt sich ableiten, was mögliche Verbesserungspotenziale sind. Dazu gehört nicht nur die Gewinnung von weiterem neuen Wissen, sondern auch der Umgang mit dem vorhandenen Wissen. Hierzu zählen Veränderungen an bestehenden Algorithmen, beispielsweise mit dem Ziel der Performancesteigerung, aber auch Reduktion von Datenquellen und extrahierten Daten oder eine generelle Optimierung der MESC-Durchführung in Bezug auf eingesetzte Personen, Arbeitsweise, Kosten oder Zeit. Um eine künstliche Komplexität zu vermeiden, sollte die Qualitätskontrolle mittels eines iterativen Lernansatzes mit ausreichenden Rückmeldemöglichkeiten initiiert werden. Ansonsten besteht die Gefahr, ein hochkomplexes „System im System“ zu schaffen, das nur mit großer zeitlicher Verzögerung Ergebnisse liefert, die Projektkosten erhöht und im schlimmsten Fall mehr Schaden durch falsche Optimierungsansätze und überfrachtete Prozesse bringt als letztendlich dem Unternehmen zu helfen.

### 4.7.2 Rückführung von Data-Mining-Ergebnissen

Der letzte Schritt im Vorgehensmodell hat die Aufgabe, mögliche Ergebnisse, die während der aktuellen Ausführung gewonnen wurden, in geeigneter Form zu dokumentieren und für weitere Wissensentdeckungsaufgaben bereitzustellen. Bei den anschließenden Wissensentdeckungen können drei Grundformen unterschieden werden:

1. Wiederholung der Wissensentdeckung, um zusätzliche oder genauere Ergebnisse zu erhalten

2. Wiederholung der Wissensentdeckung, um wiederkehrende Unternehmensaufgaben mit neuen Anforderungen und Zielen zu lösen (z. B. monatliche Prognose)
3. Wiederholung der Wissensentdeckung, um neue Unternehmensaufgaben mit abweichenden Anforderungen und Zielen zu lösen

Die Dokumentation nimmt im SC-Umfeld einen besonderen Stellenwert ein, denn nur so kann mögliches Wissen, das aus dem Vorgehensmodell gewonnen wurde, in geeigneter Form für alle SC-Teilnehmer verfügbar gemacht werden.

# 5 Integration der Simulation in das Vorgehensmodell

Vorarbeiten am Fachgebiet ITPL haben gezeigt, dass die Kombination von Simulation und KDD ein großes Synergiepotential beinhaltet (vgl. Abschnitt 2.3.4). In diesem Kapitel wird daher diskutiert, welchen methodischen Anteil die Simulation zur Methode der Wissensentdeckung in SC-Datenbanken beisteuern kann. Die erarbeiteten Ansätze werden im Anschluss in das Vorgehensmodell der MESC integriert.

## 5.1 Transaktionsdatengenerierung durch Simulation

In diesem Abschnitt wird aufgezeigt, wie die Simulation zur Transaktionsdatengenerierung im Kontext der Wissensentdeckung angewandt werden kann und welche grundsätzlichen konzeptionellen Aspekte in diesem Zusammenhang zu klären sind. Zu diesem Zweck wird zuerst erläutert, wie die DES und das Data-Farming-Konzept (vgl. Abschnitt 2.3.3) für die Generierung von Transaktionsdaten eingesetzt werden können. Im Anschluss wird diskutiert, welche besonderen Aspekte bei der Transaktionsdatengenerierung für die Wissensentdeckung zu beachten sind. Die Diskussion korrespondiert mit dem Abschnitt 2.4.2 und bezieht sich auf die *Forschungsfrage 4:* Wie können Daten für das entwickelte Vorgehensmodell generiert und effizient bereitgestellt werden?

### 5.1.1 Ansatz zur Transaktionsdatengenerierung

Die Datenbereitstellung ist eine Grundvoraussetzung für die Anwendung der MESC. Im Kontext der SC gibt es jedoch verschiedene Gründe, aus denen keine Daten bereitgestellt werden können. Die Gründe lassen sich auf zwei wesentliche Ausgangssituationen zurückführen. Zum einen können Daten nicht verfügbar sein, beispielsweise weil Bereiche der SC neu geplant werden oder für relevante Systeme der IT-Landschaft keine Daten protokolliert werden (vgl. Abschnitt 2.3.3). Zum anderen können die vorliegenden Daten nicht ausreichend qualitätsgesichert sein. Das ist zum Beispiel der Fall, wenn die vorliegenden Daten veraltet sind und somit nicht mehr den Ist-Zustand der SC wiedergeben oder wenn beteiligte IT-Systeme keine hinreichende Datenqualität für die Weiterverarbeitung der Daten aufweisen (vgl. Abschnitt 2.2.2.4). Ohne Datengrundlage lassen sich jedoch die Beschaffung und Selektion der Daten, die zu Beginn der MESC ausgeführt werden müssen (vgl. Abschnitt 4.2), nicht durchführen. Im Rückschluss ist der Einsatz der MESC für potentielle Aufgaben der Wissensentdeckung des SCM nur beschränkt anwendbar.

Um den Einsatz der MESC auch bei fehlenden oder nicht ausreichend qualitätsgesicherten Daten zu ermöglichen, müssen die hierfür notwendigen Transaktionsdaten der SC anderweitig erzeugt werden. Generell können Daten auf unterschiedliche Art und Weise generiert werden. Die Grundvoraussetzung für jede Form der Datengenerierung in der SC ist, dass das lokale Systemverhalten der einzelnen SC-Komponenten bekannt ist. Daten können beispielsweise durch viele Data-Mining-Werkzeuge erzeugt werden (vgl. Abschnitt 4.4.2). Die Einsatzmöglichkeit der Datengenerierung durch einfache Berechnungsvorschriften ist im Rahmen der Wissensentdeckung auf SC-Datenbeständen allerdings nicht zielführend. Da sich die globalen Wechselwirkungen der SC nicht unmittelbar aufgrund des lokalen Verhaltens erschließen lassen, bilden die mittels einfacher Berechnungsvorschriften generierten Daten des operativen Geschäfts nicht das reale SC-Verhalten ab (vgl. Emergenz in Abschnitt 2.2.2.3). Aus diesem Grund müssen Techniken zur Datengenerierung betrachtet werden, die komplexe Zusammenhänge der SC auf Datenebene wiedergeben können.

Eine Technik, bei der das Verhalten des Gesamtsystems beobachtet werden kann, ist die DES. Diese hat sich bereits für die Simulation von komplexen SCs bewährt und ist Bestandteil der Konzepte zur Datengenerierung mittels Data Farming (vgl. Abschnitt 2.3.4). Zudem bieten sich mit den Trace-Daten bereits geeignete Anknüpfungspunkte für die Generierung von Transaktionsdaten (vgl. Abschnitt 2.3.3). Dies begründet sich darin, dass ein Trace, der den Zustand bestimmter Simulationskomponenten protokolliert, bei geeigneter Wahl der beteiligten Komponenten die Struktur von Transaktionsdaten nachbildet. Abbildung 5.1 zeigt einen typischen Trace in einer DES. Die Trace-Daten werden in diesem Fall werkzeugspezifisch, insbesondere bei Eintritt und Austritt aus den Knoten (vgl. Techniken der SC-Simulation in Abschnitt 2.3.3), erzeugt. Diese Trace-Daten können direkt analysiert oder wie in der vorliegenden Arbeit als Datengrundlage für die Wissensentdeckung genutzt werden. Trace-Daten sind im Allgemeinen fachlich von den Ausgabegrößen der Simulation zu unterscheiden, weil die Ausgabegrößen und die Trace-Daten unterschiedliche Zwecke in einem Simulationsmodell verfolgen und folglich keine Gemeinsamkeiten aufweisen müssen (vgl. auch Literatur zu Trace-Daten in Abschnitt 2.3.3). Da im Fall der Transaktionsdatengenerierung jedoch die Trace-Daten den gewünschten Ausgabegrößen der Simulation entsprechen, wird nachfolgend die Formulierung Trace-Ausgabegrößen verwendet um diese von der allgemeinen Definition der Ausgabegrößen im Simulationsumfeld abzuheben. Um die Simulationskomponenten zu bestimmen, die für die Generierung von Transaktionsdaten in der SC modelliert werden müssen, muss näher untersucht werden, welche Bereiche der SC in der DES abgebildet werden müssen.

Um Transaktionsdaten zu generieren, muss der operative Geschäftsbereich der SC abgebildet werden. Dieser befasst sich mit der mengenmäßigen und inhaltlichen sowie der zeitlichen Abstimmung der Beschaffungs-, Produktions- und Distributionseinheiten bezüglich der einzelnen Akteure in einer SC (vgl. Abschnitt 2.2.1). Um das operative Geschäft in einem Simulationsmodell abzubilden, ist es notwen-

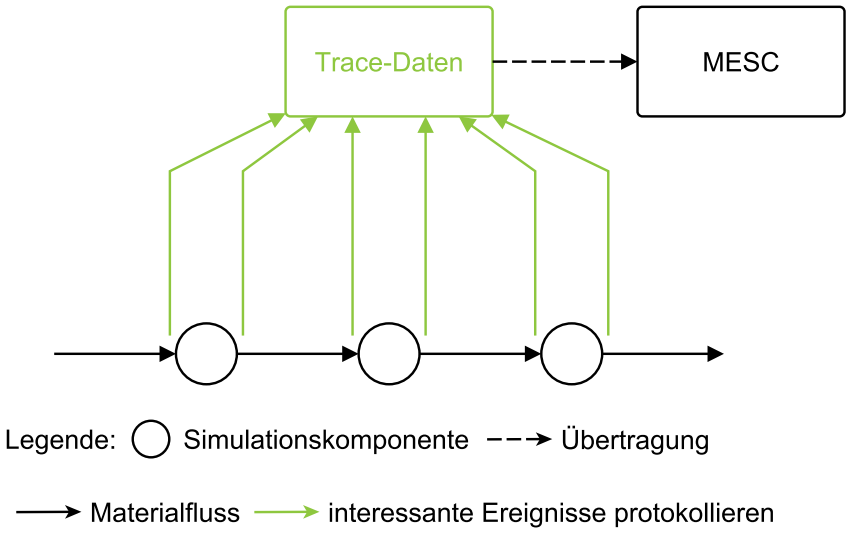



**Abbildung 5.1: Trace-Daten einer Lieferverfolgung durch die simulierte SC**

dig, die beteiligten Objekte und Akteure der SC zu modellieren. Abbildung 5.2 stellt den grundlegenden Aufbau eines SC-Simulationsmodells zur Transaktionsdatengenerierung dar.

Dabei werden die einzelnen Knoten ebenenweise den entsprechenden SC-Akteuren zugeordnet. Die durchgezogenen Pfeile zeigen in der exemplarischen Darstellung den Materialfluss an. Welche Flüsse in der SC simuliert werden, hängt von einer Vielzahl von Faktoren ab. Insbesondere gibt es verschiedenartige Werkzeuge im Bereich der Simulation, mit deren Hilfe neben dem Materialfluss auch der Informationsfluss dargestellt werden kann. Ob die Generierung von Transaktionsdaten mittels DES nur an den Materialfluss oder auch an den Informationsfluss gekoppelt ist, ist stark werkzeugabhängig und für das hier diskutierte Vorgehen nur von untergeordneter Bedeutung. Zusätzlich gibt es verschiedene Steuerungsprinzipien im Materialfluss, die bei der Erstellung des Simulationsmodells berücksichtigt werden müssen. Da die hier beschriebenen Ansätze für alle Steuerungsprinzipien gelten, wird auf eine weitere Diskussion verzichtet und die nachfolgenden Erläuterungen zur Simulation beziehen sich auf eine exemplarische Materialflusssimulation mit Push-Prinzip.

Beginnend bei den Rohstofflieferanten, welche in der Simulation als Quellen abgebildet werden, werden mittels entsprechender Transportwege die nötigen Elemente für die Lieferanten bereitgestellt. Die Teilelieferanten verarbeiten die Elemente und lösen anschließend die nächsten Transporte aus. In den nächsten Stufe werden die Elemente durch die Komponenten- und Systemlieferanten weiter verarbeitet und für die Herstellung des Endprodukts dem Endproduzenten zugeliefert. Nachdem die Elemente die Bearbeitungsschritte bei dem Endproduzenten durchlaufen haben, werden sie dem Großhandel übergeben, der sie im Anschluss dem Einzelhandel ausliefert und dort schließlich für den Endkunden bereitstellt. Die Endkunden stellen die Senken der Simulation dar. In der konkreten SC und somit auch im

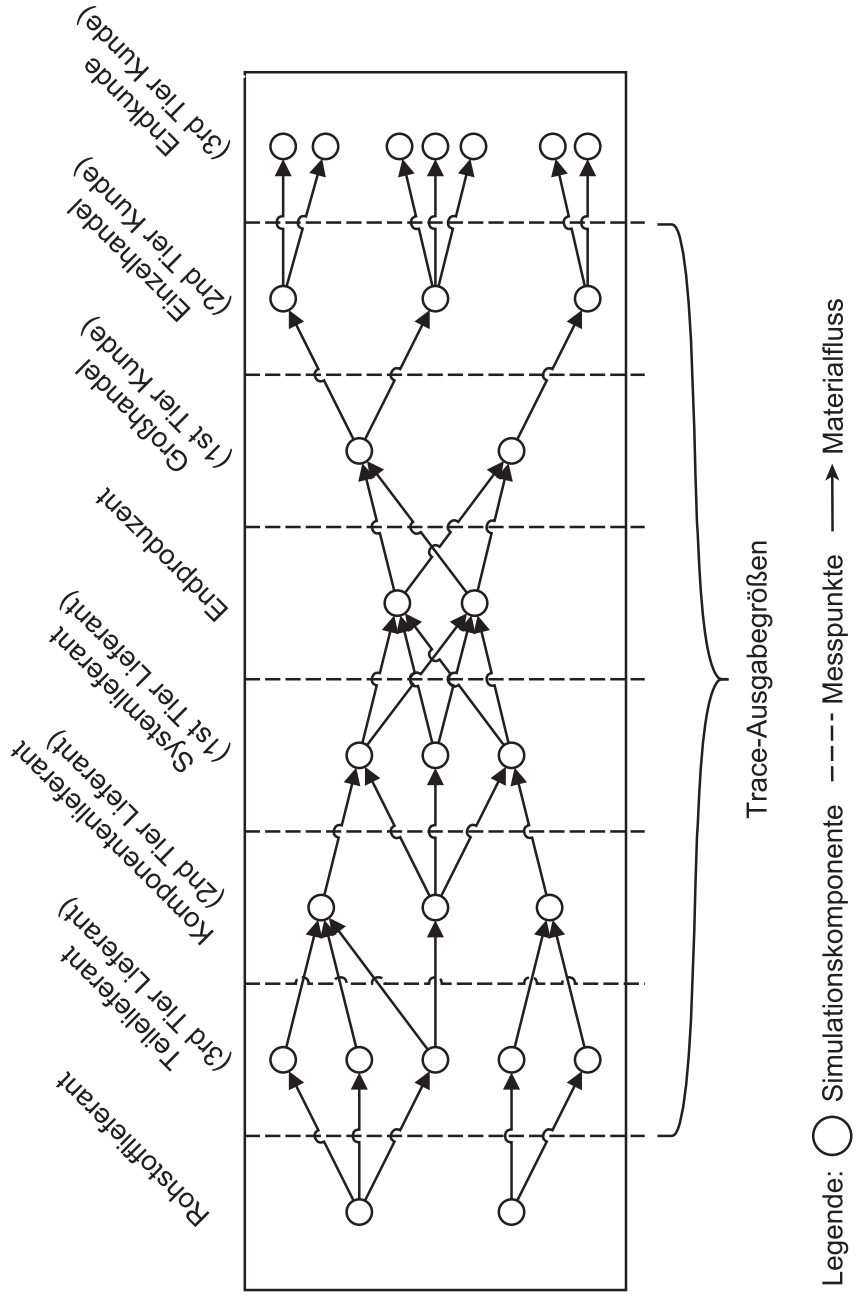



**Abbildung 5.2: Exemplarischer Aufbau eines SC-Simulationsmodells**

korrespondierenden Simulationsmodell können mitunter auch nur Ausschnitte der hier diskutierten exemplarischen SC von Relevanz sein, sodass einige der aufgelisteten Akteure an der SC nicht beteiligt sind und im spezifischen SC-Modell entfallen können. Die Trace-Ausgabegrößen sind SC-abhängige Daten und können demnach projektbezogen variieren. Trotz der unternehmensspezifischen Auswahl dieser Größen sollten die generellen Kennzeichen der Transaktionsdaten (vgl. hierzu Transaktionsaktivitätsdaten in Abschnitt 2.2.2.3) in geeigneter Form in den Trace-Ausgabegrößen abgebildet werden. In Folge liegt ein Schwerpunkt im Konzeptmodell der Simulationsstudie (vgl. Abbildung 2.11) auf der Definition der Trace-Ausgabegrößen der DES. Tabelle 5.1 zeigt mögliche Ausgabegrößen von Transaktionsdaten mit exemplarischen Geltungsbereichen und zugehörigen Datentypen. Jede Trace-Ausgabegröße weist neben einer Vielzahl von kennzeichnenden Kriterien (z. B. Zugehörigkeit zu BO) immer einen Geltungsbereich und einen Datentyp auf.

**Tabelle 5.1: Typische Ausgabegrößen der Transaktionsdatengenerierung**

| **Ausgabegröße** | **exemplarischer Geltungsbereich** | **Datentyp nach Tabelle A.1** |
|---|---|---|
| Artikelnummer | 000000 - ZZZZZZ | Zeichenkette - eindeutig |
| Auftragsnummer | 1 - 10 000 | Ganzzahl - eindeutig |
| Menge | 10 - 10 000 | Ganzzahl |
| Datum | 01.01.2000 - 31.12.9999 | Format: dd.mm.yyyy |
| Lieferantenkürzel | AAAA- ZZZZ | Zeichenkette |
| Lagerkürzel | AAA - ZZZ | Zeichenkette |

Unabhängig vom Einsatzgebiet der Transaktionsdaten sind im Allgemeinen eine gewisse Anzahl von Transaktionen für die Erfüllung von unterschiedlichen Aufgaben notwendig. Um einen hinreichend großen Datenbestand an Transaktionsdaten zu generieren, muss also sichergestellt werden, dass in dem durchzuführenden Simulationsexperiment ausreichend Transaktionsdaten erzeugt werden. Alternativ können mehrere Experimente durchgeführt werden, um eine für die Anforderung der Data-Mining-Verfahren ausreichende Menge an Transaktionsdaten zu generieren. Da dies von vielen projektspezifischen Kriterien abhängt, werden verschiedene Fälle in den praktischen Anwendungsfeldern dieser Arbeit untersucht, aber aufgrund des Fokus der vorliegenden Arbeit auf eine allgemeine Empfehlung verzichtet.

In der nachfolgenden Diskussion wird für die Ausführung eines Simulationsmodells nur noch die Formulierung Experiment verwendet, da diese die mitunter notwendige Parameterveränderung im Simulationsmodell einbezieht (vgl. hierzu Replikation und Experiment in Abschnitt 2.3.3). Wird der Begriff Simulationslauf

verwendet, so ist das einmalige Ausführen eines Simulationsmodells Gegenstand der Betrachtung.

Jeder generierte Datenbestand besteht aus $n$ Datensätzen, die jeweils $w$ Attribute enthalten (vgl. Abschnitt 2.2.2.1). Die Attribute entsprechen den Trace-Ausgabegrößen und ihre konkreten Ausprägungen den Werteausprägungen der Trace-Ausgabegrößen in den generierten Datensätzen. Das Ergebnis ist eine Tabelle, die dem Standarddatenformat in MESC entspricht (vgl. Abschnitt 4.3.1). Abbildung 5.3 zeigt die Ergebnisstruktur der Transaktionsdatengenerierung mit $m$ Experimenten, die jeweils $n$ Datensätze mit $w$ Attributen erzeugen. Es ist anzumerken, dass $m$ je nach Anwendungsfall auch nur einen Simulationslauf enthalten kann.

| Index | 1 | 2 | ... | w-1 | w |
|---|---|---|---|---|---|
| 1 | | | | | |
| ... | | | | | |
| n-1 | | | | | |
| n | | | | | |
| 2 | | | | | |
| ... | | | | | |
| (m-1)*n | | | | | |
| ... | | | | | |
| m*n | | | | | |

**Abbildung 5.3: Ergebnisstruktur generierter Transaktionsdaten**

Zusammenfassend kann festgestellt werden, dass die Transaktionsdatengenerierung eine spezifische Ausprägung des Data Farmings mittels DES darstellt. Hierbei kann sowohl das Vorgehen zum Aufbau einer Simulationsstudie im Bereich Data Farming als auch der generelle Aufbau von SC-Simulationsmodellen über die Erläuterungen in dem Abschnitt 2.3.3 nachvollzogen werden, und ist nicht Teil der weiteren Untersuchung. Die Besonderheit liegt in der Verknüpfung der beiden Aspekte, denn die Technik des Data Farmings wird nun erstmals für die Generierung von Transaktionsdaten ausgelegt. Hierbei ist die Generierung von Transaktionsdaten in erster Annäherung ein eigenständiges Verfahren. Das Ergebnis des Verfahrens, die generierten Daten, können zu unterschiedlichen Zwecken eingesetzt werden. Beispielsweise können die generierten Daten als Testdaten (vgl. Abschnitt 2.3.2) fungieren. Das Vorgehen beruht hierbei im Wesentlichen auf dem Grundgedanken, die generierten Datenbestände auf die Existenz der Muster zu überprüfen und somit eine Aussage über die allgemeine Gültigkeit der Muster zu treffen. In der Abbildung 5.4 ist der Ablauf der Validierung in MESC (vgl. hierzu auch Abbildung 4.5) mittels generierten Daten dargestellt.

In der Abbildung 5.4 ist die Testdatenbereitstellung aus Abbildung 4.5 durch das Data Farming (farblich markierter Bereich) ersetzt worden. Für das Verständnis des Vorgehens ist es unerheblich, ob die Simulation einmal oder mehrfach durchge-

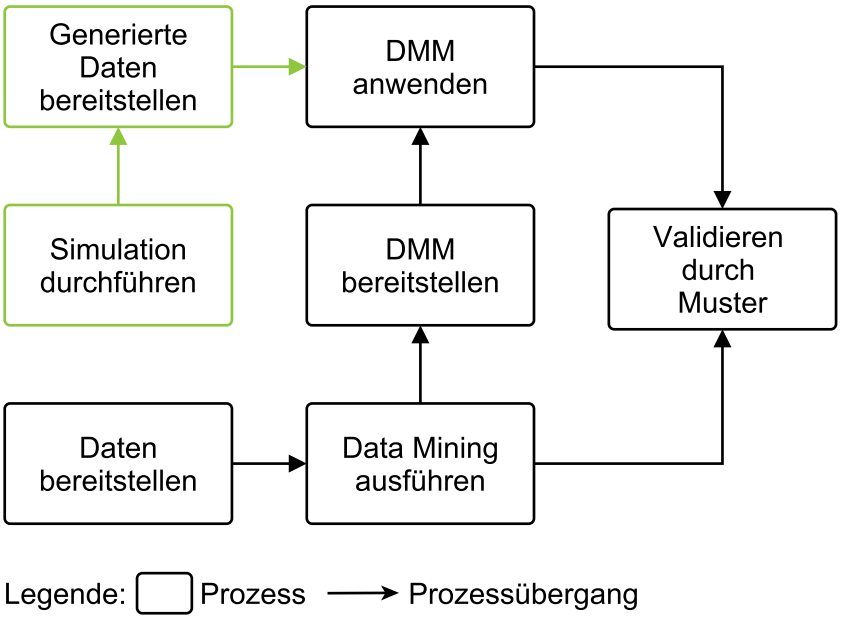



**Abbildung 5.4: Validierung mittels generierter Testdaten**

führt wird, um einen hinreichenden Datenbestand zu erzeugen. Aus diesem Grund wurde auf die Darstellung der Iteration in der Abbildung verzichtet. Es wird jedoch explizit festgehalten, das eine Vielzahl von generierten Datenbeständen benötigt wird, um mit statistischer Sicherheit eine Aussage über die Validität der Muster zu treffen. Jeder generierte Datenbestand fungiert in diesem Vorgehen als unabhängige Stichprobe. So sind bei der Validierung mittels generierter Testdaten in jedem Fall die Grundlagen der statistischen Versuchsplanung einzuhalten (vgl. Abschnitt 2.3.3) und sollten vor der hier diskutierten Validierungsphase schriftlich dokumentiert werden. Insbesondere ist im Vorfeld zu definieren, wie viele Simulationsexperimente mit welchen Stellgrößenvariationen durchgeführt werden sollen und welche statistischen Maßzahlen zugrunde gelegt werden. Das Ende des Validierungsprozesses ist nach der Prüfung aller Muster aus den generierten Daten erreicht. Das Beispiel 5.1 demonstriert den Prozess an einem Fallbeispiel. Die Verwendung der Variablen folgt den Notationen dieses Abschnitts.

**Beispiel 5.1:** In einer SC wurde bei der Durchführung der MESC ein Muster in den SC-Datenbanken erlernt. Das Muster gehört zum Subtyp Wirkzusammenhänge und besteht aus zwei Ausprägungen, die mit einer gewissen Wahrscheinlichkeit in einer Anzahl von Datensätzen enthalten sind (vgl. Abschnitt 4.4.1). Um das SCM zu unterstützen, sollen nun diese Subtypausprägungen durch den Prozess aus Abbildung 5.4 validiert werden. Hierzu wird in der statistischen Versuchsplanung die Anzahl der Experimente auf $m = 50$ festgelegt. Jedes Experiment erzeugt über einen Simulationszeitraum von 14 Tagen $n = 500$ Transaktionsdatensätze. Jeder Transaktionsdatensatz besteht aus $w = 15$ Transaktionsattributen. Nach der Durchführung der 50 Experimente werden die resultie-

renden Datensätze mittels MESC auf das Vorhandensein der Muster untersucht. Die separate statistische Auswertung der 50 Datenbestände zeigt, dass mittels des DMMs in 42 Datenbeständen das Vorhandensein der beiden Subtypausprägungen bestätigt werden kann. In der MESC-Phase 1 wurde festgelegt, dass ein Wert über 70 %, unter Berücksichtigung von Faktoren wie Testdatenmenge, Parameterkombination und Auftrittswahrscheinlichkeit, ausreichend für den Einsatz des aus den Mustern gewonnenen Wissens ist. Da der Wert mit 84 % die Kriterien erfüllt, ist das Ergebnis der Mustervalidierung positiv und es darf erwartet werden, dass die Muster tatsächlich Wissen über die zugrundeliegende SC repräsentieren.

Ob die Testdatengenerierung vorgelagert, zeitgleich oder nachgelagert zur eigentlichen Wissensentdeckung ist, hängt von projektspezifischen Kriterien ab. In jedem Fall wird auf den generierten Daten keine vollständige Wissensentdeckung durchgeführt, denn sowohl die Auswahl der Data-Mining-Methoden (vgl. Abschnitt 4.4.1) als auch Schritte in der Anwendung, wie beispielsweise die Parametrierung der Modelle (vgl. Abschnitt 4.5), entfallen. Die Testdatengenerierung stellt eine Einsatzmöglichkeit des Data Farmings in MESC dar und ermöglicht eine Alternative zu den aus dem KDD bekannten Validierungstechniken mittels Testdaten. Obwohl die Validierung immer noch ressourcen- und zeitintensiv ist, wird die Problematik der Testdatenbegrenzung durch diesen Ansatz gelöst (vgl. Abschnitt 2.3.2).

### 5.1.2 Transaktionsdaten für die Wissensentdeckung

Wie bereits in den Grundlagen der Transaktionsdatengenerierung erläutert, geht die Modellnutzung der Simulation zur Datengenerierung mit der Entscheidung einher, zu welchem Zweck die Daten generiert werden sollen (vgl. Abschnitt 5.1.1). In dem vorliegenden Abschnitt wird untersucht, welche Besonderheiten sich bei der Transaktionsdatengenerierung für den Zweck der Wissensentdeckung ergeben, um den Einsatz der MESC auch bei unzureichender Datenlage zu ermöglichen.

Da die generierten Daten nun für die Wissensentdeckung genutzt werden, ist es notwendig, eine ausreichende Datenqualität zu erzeugen. Aufgrund der Tatsache, dass die unternehmensbezogenen Ursachen für mangelnde Datenqualität (vgl. Abschnitt 2.2.2.4) in einer Simulationsstudie nur eine untergeordnete Rolle spielen, ist dieser Aspekt der Datenqualität in der Regel überwiegend von der Eignung des Simulationsmodells zur Datengenerierung beeinflusst.

Es muss beachtet werden, dass bei Simulationsmodellen, die eine Einschwingphase besitzen, die Daten mitunter erst nach der Einschwingphase für die Wissensentdeckung genutzt werden können. Die Einschwingphase ist in der SC-Simulation von großer Bedeutung, da die SC als typisch nicht-terminierendes System angesehen wird. Im Regelfall sind die Daten der Einschwingphase für die Wissensentdeckung in der SC unbrauchbar, wenn beispielsweise Warteschlangen vor Hubs

oder Lagern in der SC integriert sind (vgl. Abschnitt 2.3.3). Die Berücksichtigung der Einschwingphase hängt in diesem Zusammenhang maßgeblich von dem Ziel der Wissensentdeckung und den dafür benötigten Transaktionsdaten ab (vgl. Abschnitt 4.1.2). Ist beispielsweise der Durchsatz einer spezifischen Warengruppe Teil der Transaktionsdaten, so führt die Nichtbeachtung der Einschwingphase zu fehlerhaften SC-Datenbeständen. Ein weiterer Punkt, der im Kontext der Generierung von Transaktionsdaten in der SC von Bedeutung ist, ist die Frage nach der Betrachtungsdauer des Simulationsmodells. Da es sich bei der Simulation von SCs, wie bereits ausgeführt, um nicht-terminierende Systeme handelt, kann die Simulation über einen beliebig langen Zeitraum Daten generieren. Es ist folglich von wesentlicher Bedeutung für die Wissensentdeckung, die Anzahl der generierten Transaktionen festzulegen und daraus einen Endzeitpunkt für den Simulationslauf zu bestimmen.

Neben den vorgestellten Konzepten zur Transaktionsdatengenerierung ergeben sich aufgrund der Koppelung von Simulationsmodellen und Methoden der Wissensentdeckung spezifische Überlegungen zur Spezifikation der Trace-Ausgabegrößen. Diese Überlegungen werden in den folgenden Abschnitten näher in Bezug auf das entwickelte Vorgehensmodell in MESC ausgeführt.

#### 5.1.2.1 Spezifizierung der Datenaggregation

Um das Vorgehensmodell zur Wissensentdeckung effizient zu gestalten, ist es in erster Näherung zielführend, möglichst viele Phasenschritte der MESC durch geeignete Konfiguration der Trace-Ausgabegrößen einzusparen. Durch dieses Vorgehen müssen die generierten Daten nicht mehr vollumfänglich für den Einsatz der Data-Mining-Verfahren vorbereitet werden und verschiedene Schritte, die im Zusammenhang mit der Datenbearbeitung stehen, können bei der Durchführung der MESC ausgelassen werden. Durch die Auswahl geeigneter Datenaggregationsstufen in der Simulation (vgl. Abschnitte 2.3.4 und 4.3.1) können entsprechend Schritte insbesondere aus den ersten Phasen der MESC entfallen. Welches Einsparpotential durch das Aggregationskonzept von Data Farming und Wissensentdeckung realisierbar ist, hängt von der Konfiguration der Trace-Ausgabegrößen ab.

Die Konzeption der Trace-Ausgabegrößen wird von verschiedenen Faktoren beeinflusst. Zuerst muss unterschieden werden, ob Daten z. B. für die Planung neu generiert werden müssen oder ein bereits bestehender Datenbestand angereichert werden muss. Durch die vorhandenen Echtdaten existieren bei der Datenanreicherung Restriktionen für die Konzeption der Trace-Ausgabegrößen. Da die generierten Daten mit den Echtdaten zusammengeführt werden müssen, kommen Prinzipien der Integration von Daten aus verschiedenen Quellsystemen, wie z. B. die Aggregationsstufe, zum Tragen. Dieses Thema wurde bereits in MESC bei der Integration von Echtdaten aus verschiedenen Systemen und Überführung in ein Standarddatenformat behandelt (vgl. Abschnitt 4.3.1). Des Weiteren fällt sowohl bei der Anreicherung eines bestehenden Datenbestands als auch bei der Generierung von

neuen Daten speziell für die Wissensentdeckung ein vergleichbarer konzeptioneller Aufwand an. Zusätzlich ist in beiden Fällen die Ausführung von Simulationsexperimenten zur Datengenerierung notwendig, so dass die beiden Punkte im Folgenden nicht weiter unterschieden werden.

Die zwei wesentlichen Einflussfaktoren auf die Konzeption der Trace-Ausgabegrößen sind wie folgt: Der erste Einflussfaktor ist durch die Eignung des zu nutzenden Simulationsmodells begründet. In diesem Kontext muss unterschieden werden, ob ein bereits vorhandenes SC-Simulationsmodell für Data-Farming-Zwecke erweitert oder ein neues Simulationsmodell entwickelt wird. Im ersten Fall sind Einschränkungen durch das existierende SC-Modell und dessen Verwendungszweck im Unternehmen gegeben. Beispielsweise kann ein Modell zur SC-Risikoanalyse nicht einfach parallel die Aufgabe der Datengenerierung übernehmen, da die entsprechenden Simulationsergebnisse nicht den benötigten Transaktionsdaten entsprechen, sondern vielmehr risikobezogene Kennzahlen beinhalten. Bei einem vorhandenen Simulationsmodell ist abzuwägen, ob das Modell für die Transaktionsdatengenerierung angepasst werden kann oder ob der Aufwand hierzu aus Sicht des Projektes nicht zu vertreten ist. Falls kein Simulationsmodell vorhanden ist, muss ein neues erstellt werden, bei dem die Konzeption der Trace-Ausgabegrößen ganz auf den Zweck der Datengenerierung ausgerichtet werden kann.

Der zweite Faktor begründet sich in den unterschiedlichen Datenverarbeitungs- und Aggregationsstufen. Diese reichen in der Wissensentdeckung von den Rohdaten über verschiedene Vorverarbeitungsstufen bis hin zu den für das Data-Mining-Verfahren vorbereiteten Datenbeständen. In der Folge können bei der Konzeption der Trace-Ausgabegrößen verschiedene Datenverarbeitungs- und Aggregationsstufen berücksichtigt werden. Hierbei ist es eine projektbezogene Grundsatzentscheidung, in welcher Phase die generierten Daten in MESC Anwendung finden sollen und somit welche konzeptionellen Vorgaben sinnvoller Weise zu beachten sind. Sinnvoll ist im Allgemeinen entweder eine Datengenerierung auf Ebene der Rohdaten oder eine Datengenerierung auf Ebene der vollständig vorverarbeiteten Daten. Zu diesem Zweck werden zwei Schnittstellen definiert. Die erste Schnittstelle ist zwischen den MESC-Phasen 2 und 3 eingeordnet. Hier liegen die Daten als Rohdaten vor und wurden noch nicht für das Data Mining bearbeitet. Die zweite Schnittstelle ist zwischen Phase 4 und 5 der MESC eingeordnet. Hier liegen die Daten in bearbeiteter Form vor und die Data-Mining-Verfahren können in Phase 5 angewandt werden. Es muss also in Folge zwischen diesen beiden Schnittstellen für das Bestimmen der Trace-Ausgabegrößen in der Simulationsstudie unterschieden werden.

Um die Wirkung der Einflussfaktoren auf die Trace-Ausgabegrößen zu verdeutlichen, sind in der Abbildung 5.5 drei ausgewählte Fälle dargestellt. Die Einflussfaktoren bilden die Achsen des Koordinatensystems. An dieser Stelle wurde bewusst die Möglichkeit von Abstufungen zwischen den Ausprägungen der Einflussfaktoren eingeräumt. Diese Variante wurde gewählt, da in der Praxis häufig keine binäre Ausprägung der Faktoren anzutreffen ist. Der Einflussfaktor der Eignung des Si-

mulationsmodells vereinfacht die zuvor angestellte Überlegung, in dem lediglich die Eignung eines Simulationsmodells betrachtet wird und die Ausprägung „ungeeignet“ sowohl den Fall eines vollständig ungeeigneten Simulationsmodells wie auch den Fall eines nicht vorhandenen Modells abdeckt. Die positive Ausprägung „geeignet“ geht von einem Simulationsmodell aus, das keine Anpassung der Trace-Ausgabegrößen für die Generierung von Transaktionsdaten benötigt. Der Einflussfaktor Datenaggregation kumuliert die Datenaggregation mit den Schritten der Vorverarbeitung, da diese Sichtweise für das Verständnis zielführend ist. Es wird jedoch explizit darauf hingewiesen, dass Datenaggregationsstufen und Vorverarbeitungsschritte nicht identisch sind. Es gibt Vorverarbeitungsschritte wie z. B. das Data Cleansing, die nicht mit einem höheren Aggregationsniveau einhergehen (vgl. hierzu auch die Diskussionen zu den Aggregationsstufen z. B. in Abschnitt 2.3.4 oder 4.3.1). Die einzelnen Fälle stellen sich wie folgt dar:

**Fall 1:**

In diesem Fall soll die Interaktion des Simulationsmodells mit dem Vorgehensmodell in der ersten Schnittstelle, also zwischen Phase 2 und 3 stattfinden. Zusätzlich existiert bereits ein Simulationsmodell, das zur Generierung von Transaktionsdaten geeignet ist. Bei der Konzeption der Trace-Ausgabegrößen muss nun geprüft werden, ob die generierten Transaktionsdaten des vorhandenen Simulationsmodells in ihrer Struktur den Rohdaten entsprechen oder ob hier eine konzeptionelle Veränderung zielführend ist. Sollte das Simulationsmodell Daten auf einer hohen Aggregationsebene generieren, so lässt sich aus diesen Daten mittels Vorverarbeitung nicht einfach eine niedrigere Aggregationsstufe erzeugen. An diesem Beispiel wird deutlich, dass die beiden Faktoren sich in der Konzeptionsphase der Trace-Ausgabegrößen gegenseitig beeinflussen.

**Fall 2:**

In diesem Fall soll die Interaktion des Simulationsmodells mit dem Vorgehensmodell ebenfalls in der ersten Schnittstelle, also zwischen Phase 2 und 3 stattfinden. Hier existiert jedoch kein Simulationsmodell, sodass bei der Erstellung des Modells eine vollständige Konzeption der Trace-Ausgabegrößen stattfinden kann. Die Konzeption der Trace-Ausgabegrößen folgt in ihrer Struktur den Rohdaten der SC-Systeme.

**Fall 3:**

In diesem Fall soll die Interaktion des Simulationsmodells mit dem Vorgehensmodell in der zweiten Schnittstelle, also zwischen Phase 4 und 5 stattfinden. Zusätzlich existiert bereits ein Simulationsmodell, das zur Generierung von Transaktionsdaten allerdings nur bedingt geeignet ist. Bei der Konzeption der Trace-Ausgabegrößen muss nun geprüft werden, welche Veränderungen im Simulationsmodell notwendig sind, um mit dem Modell Transaktionsdaten zu generieren, die im Anschluss ohne Vorverarbeitung mittels geeigneter Data-Mining-Verfahren analysiert werden können.

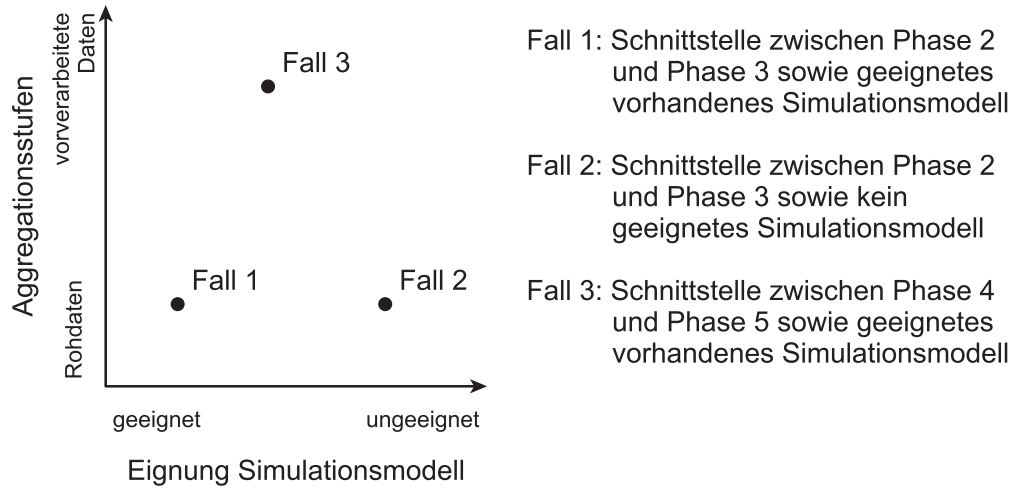



**Abbildung 5.5: Einordnung von exemplarischen Fällen zur Trace-Ausgabegrößenkonzeption anhand gegebener Einflussgrößen**

Die Konstellationen zeigen, dass bei geeigneter Konzeption der Trace-Ausgabegrößen von Simulationsmodellen Schritte der Vorverarbeitungsphasen der MESC entfallen können. Tabelle 5.2 verdeutlicht das Einsparpotential in MESC durch zielführende Konzeption von Trace-Ausgabegrößen im Rahmen der Simulationsstudie.

#### 5.1.2.2 Datenaggregation und Modellflexibilität

Im Kontext des Einsparpotentials von MESC-Phasen durch geeignete Konzeption der Trace-Ausgabegrößen muss das Thema der Modellflexibilität diskutiert werden. Im Abschnitt 5.1.2.1 erscheint es vorteilhaft, Transaktionsdaten zu generieren, die direkt von den Data-Mining-Verfahren verarbeitet werden können. Durch dieses Konzept lassen sich zeitintensive Vorverarbeitungsschritte in der Wissensentdeckung einsparen. Doch dieses Konzept geht auch mit Risiken einher. Insbesondere führt das Festlegen auf hochaggregierte, spezifische Simulationsausgabegrößen zu einer technischen und fachlichen Einschränkung in der Auswahl der einsetzbaren Data-Mining-Verfahren (vgl. hierzu Datenaggregation im Abschnitt 2.3.4). Viele Data-Mining-Verfahren benötigen spezielle Eingabedaten, beispielsweise sind Datentypen oder Anzahl der Attribute nicht beliebig wählbar. In den Vorgehensmodellen des KDD, so auch in MESC, ist daher ein separater Schritt für die technische Eingabekodierung vorgesehen (vgl. Abschnitt 4.4.4). Aus fachlicher Sicht führt das Generieren von hochaggregierten Daten zu einer Einschränkung der möglichen Fragestellungen in der Wissensentdeckung. Dies ist in dem Umstand begründet, dass Informationen (z. B. Lieferzeiten pro Transportmittel pro Händler) nur noch aggregiert generiert werden und folglich Fragestellungen nur noch auf den aggregierten Ebenen möglich sind (z. B. Lieferzeiten pro Händler).

**Tabelle 5.2: Einsparpotential auf Schrittebene in MESC**

| Phase | Schritte | Auslagerung in Simulation |
|---|---|---|
| 3. Datenaufbereitung | 3.1 Überführung in ein Standarddatenformat | Abdeckung über geeignete Konzeption der Trace-Ausgabegrößen |
| | 3.2 Gruppierung | Abdeckung über geeignete Parametrierung und separate Simulationsläufe |
| | 3.3 Datenanreicherung | Abdeckung über geeignete Konzeption der Trace-Ausgabegrößen |
| | 3.4 Transformation | Abdeckung über geeignete Konzeption der Trace-Ausgabegrößen |
| 4. Vorbereitung des Data-Mining-Verfahrens | 4.3 Fachliche Kodierung | Abdeckung über geeignete Konzeption der Trace-Ausgabegrößen |
| | 4.4 Technische Kodierung | Abdeckung über geeignete Konzeption der Trace-Ausgabegrößen |

In dem entwickelten Vorgehensmodell der MESC ist für die Aufgabe der geeigneten fachlichen Kodierung von Eingabegrößen ein Schritt explizit eingeführt worden (vgl. Abschnitt 4.4.3). Zusammenfassend kann konstatiert werden, dass hochaggregierte generierte Simulationsdaten im Rahmen der Wissensentdeckung nur schwer zu verändern sind und in der Folge die MESC an Flexibilität im Bereich von Fragestellung und Einsatz von Data-Mining-Verfahren verliert. Im Rückschluss ist die Kodierung von Trace-Ausgabegrößen für einen spezifischen Zweck der Wissensentdeckung nicht in allen SC-Projekten zu empfehlen.

Wenn die generierten Daten für verschiedene Data-Mining-Aufgaben eingesetzt werden (vgl. Tabelle 2.9), sollte im Allgemeinen ebenfalls eine potentiell niedrigere Aggregationsstufe gewählt werden. Dies hat den Vorteil, dass Daten bei Bedarf während der Ausführung der MESC aggregiert werden können. Ein weiterer Aspekt ist in der Wissensentdeckung selbst begründet. Wenn die Fragestellung neuartig für das Unternehmen ist und daher mit einer ausgedehnten Erprobung von unterschiedlichen Data-Mining-Verfahren zu rechnen ist, ist eine spezifische Kodierung ebenfalls nicht zielführend (vgl. Abschnitt 4.4.1).

Zusammenfassend lässt sich feststellen, das eine problemspezifische Konzeption von Trace-Ausgabegrößen in der Simulation eine Zeitersparnis für die Durchführung der MESC bieten kann. Allerdings geht ein solches Konstrukt zu Lasten des

flexiblen Einsatzes der MESC. Daher ist dieses Konzept eher in wiederkehrenden, zeitkritischen SC-Prozessen wie beispielsweise der regelmäßigen Prognose von Risiken, Lagerbeständen oder Nachfragen zu empfehlen.

## 5.2 Validierung der Data-Mining-Ergebnisse mittels Simulation

Mit dem MESC-Methodenelement des Dreiecksmodellbausteins wurde ein Ansatz zur modellbegleitenden V&V in das entwickelte Vorgehensmodell integriert (vgl. Abschnitt 3.6). Im Ergebnis bietet MESC geeignete Maßnahmen zur Ergebniskontrolle an und etabliert so erstmals die V&V als phasenübergreifende Technik im KDD-Umfeld. Jedoch kann die MESC in der inhaltlichen Umsetzung der Maßnahmen zu der modellbegleitenden V&V nur auf bekannte Techniken zurückgreifen. Insbesondere die Validierung der gewonnenen Ergebnisse gestaltet sich weiterhin als manueller Prozess, mit dem die Eignung der gefundenen Muster aus dem Data Mining (vgl. Abschnitt 4.5), die Auswahl der interessanten Muster (vgl. Abschnitt 4.6.1) und die Darstellungstransformation festgestellt werden können. Die Validitätsprüfung des Dreiecksmodell-Bausteins entscheidet, ob die ausgewählten Muster und ihre Darstellung tatsächlich potentiell nützliches und neues Wissen über das zugrundeliegende System wiedergeben. Diese Prüfung muss im SC-Umfeld mittels geeigneter fachseitiger Unterstützung durch das SCM durchgeführt werden. Da mitunter mehrere Muster, beispielsweise ein ganzes Regelwerk, in dieser Phase zu validieren sind, ist der Schritt sowohl ressourcen- als auch zeitintensiv. Mit dem nachfolgend vorgestellten Ansatz werden Möglichkeiten zur simulationsunterstützten Validierung aufgezeigt und die sich daraus ergebenen Potentiale für die Validierungsphase in der Wissensentdeckung diskutiert.

### 5.2.1 Ansatz zur simulationsunterstützten Validierung

Der zentrale Aspekt bei der Validierung der Muster ist die Frage, ob die gefundenen Muster auch wirkliches Wissen über die SC repräsentieren. Ausgangspunkt der Diskussion ist die Auslagerung der Mustervalidierung aus der MESC, die in Abschnitt 3.5 ausgeführt wurde. Die verbreitete KDD-Mustervalidierung mittels Testdaten (vgl. Abschnitt 2.3.2) wurde in die Phase 5 des Dreiecksmodell-Bausteins integriert (vgl. Abschnitt 3.6). Zusätzlich zu dieser allgemeinen Validierungstechnik wurden in Phase 6 des Dreiecksmodell-Bausteins SC-spezifische Validierungsschritte für Muster initiiert, die die Auswahl von spezifischen Mustern und möglichen Darstellungstransformationen dieser Muster beinhalten. Die Validierungsprüfung der Muster aus den Phasen 5 und 6 der MESC beinhaltet folglich die drei Aspekte:

1. Eignung der Muster

2. Auswahl der interessanten Muster

3. Darstellungstransformation der Muster

Die Validierung konzentriert sich in diesem Abschnitt auf die reine Mustervalidierung und diskutiert Aspekt 2 und 3 nicht. Aspekt 2 eignet sich nicht vollständig für die simulationsunterstützte Validierung, da es zwei Arten von Maßzahlen gibt, die Einfluss auf die Interessantheit nehmen. Zum einen gibt es objektive Maßzahlen, die auf der Anzahl der Datensätze sowie bestimmten Wahrscheinlichkeiten beruhen (vgl. interestingness measures in Abschnitt 2.3.2). Solche objektiven Maßzahlen lassen sich mit dem Validierungsprozess, der in Abbildung 5.4 dargestellt ist, untersuchen. Die generierten Daten sowie die darauf erlernten Muster können auf Wahrscheinlichkeiten oder Zahlenverhältnisse geprüft werden. Anders verhält es sich mit den unternehmensbezogenen Maßzahlen, die beispielsweise aufgrund einer bestimmten Attributskombination ein Muster als interessant einstufen. Diese Art der Interessantheit benötigt Kontextwissen, das beispielsweise durch die SCM-Akteure beigesteuert werden kann. Hier ist der Einsatz der Simulation nicht zielführend, da auch eine Vielzahl von generierten Datensätzen das benötigte Kontextwissen nicht enthält. Der Validierungsprozess (vgl. Abbildung 5.4) ist auch für die Prüfung der korrekten Darstellungstransformation ungeeignet. Dies ist durch die Tatsache begründet, dass die Darstellungstransformation unmittelbar mit der Weiterverwendung der Muster korrespondiert (vgl. Abschnitt 4.6.2). Die Weiterverwendung geht einher mit einem geeigneten Zielformat, das theoretisch technisch überprüfbar ist. Es muss jedoch berücksichtigt werden, dass die Darstellungstransformation kein Prozess ist, der von den generierten Transaktionsdaten profitieren kann, weil die Transformation einen in Bezug auf die Wissensentdeckung nachgelagerten Prozess darstellt. Da die Transformation in ein Zielformat einen einmaligen Schritt der MESC darstellt, ist die manuelle Prüfung durch das SCM im Unternehmensalltag zu bevorzugen. Die Simulation kann hier nicht zur Automatisierung beitragen und darüber hinaus ist die Automatisierung in diesem Schritt aufgrund der geringen Schrittkomplexität für den praktischen Einsatz nicht notwendig. Zusammenfassend kann konstatiert werden, dass Aspekt 2 und 3 einen Großteil von manuellen Aktionen beinhalten, sodass eine weitergehende Untersuchung im Rahmen dieser Ausarbeitung nicht zielführend ist.

Wie bereits im Abschnitt zum wissenschaftlichen Stand der Simulation dargelegt, eignet sich die DES zur Überprüfung von SC-Systemverhalten (vgl. Abschnitt 2.3.3). Insbesondere Muster, wie die Wirkzusammenhänge, die spezifische SC-Systemverhalten darstellen, können durch den Einsatz von Simulationstechnik näher untersucht werden (vgl. Abschnitt 2.3.4). Im Folgenden soll in den theoretischen Grundzügen erläutert werden, wie die Simulation die Mustervalidierung unterstützen kann. Der Fokus liegt hierbei auf dem Aufzeigen der Rahmenbedingungen, die in nachgelagerten Forschungsarbeiten näher untersucht werden müssen. Die grundlegende Idee der simulationsunterstützten Validierung ist, dass ein Muster, welches mit Hilfe der Simulation validiert werden soll, bereits einen Mehrwert darstellt. Das Muster beinhaltet Aussagen über spezifische SC-Komponenten

(Parameterbelegung der Instanz, vgl. hierzu Abschnitt 4.4.1). Demnach können diese Aussagen dazu genutzt werden, die Datengenerierung gezielt zu steuern. In Abbildung 5.6 ist die Verwertung der Muster zur Datengenerierung sowie die betroffenen Tätigkeiten in der simulationsunterstützten Validierung in der Fabe rot markiert (vgl. auch Abbildung 5.4).

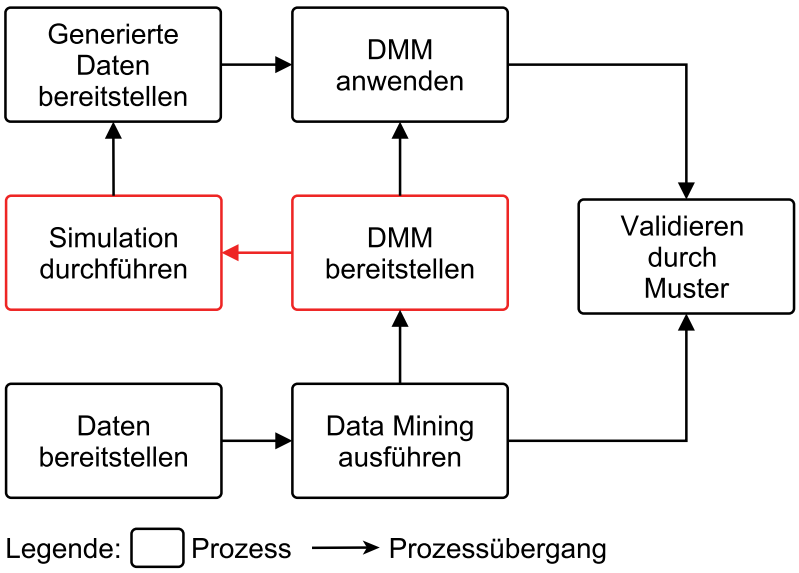



**Abbildung 5.6: Steuerung der Datengenerierung**

Auf der gezielt veränderten Datenlandschaft wird dann mittels des DMM des zu validierenden Musters geprüft, in wieweit das Muster wieder in den Daten zu entdecken ist. Die kontrollierte Veränderung des Datenbestandes aufgrund des existierenden Musters unterscheidet die simulationsunterstützte Validierung von der reinen Testdatengenererierung (vgl. hierzu auch Abbildung 5.4), bei der die Datenlandschaft zufällig erzeugt wird. Die bewusste Steuerung der Datengenerierung und das anschließende Überprüfen des zu validierenden Musters auf dem veränderten Datenbestand bietet neue Möglichkeiten in der Validierung. Hierbei muss berücksichtigt werden, dass das zu validierende Muster auf den Transaktionsdaten erlernt wurde, welche durch die Simulation erzeugt wurden. Diese genierten Transaktionsdaten entsprechen somit der Simulationsausgabe (vgl. Abschnitt 5.1). Um jedoch die Simulation gezielt anzusteuern, müssen Eingangsgrößen wie die Parameterbelegung bekannt sein und diese lassen sich aufgrund der SC-Komplexität nicht aus den Transaktionsdaten berechnen (vgl. z. B. Abschnitt 2.2.2.3). Es ist jedoch möglich, mittels Kontextwissens potentielle Einflussparameter in der SC-Simulation zu identifizieren und diese in den Simulationsexperimenten gezielt zu verändern. Da der Rückschluss von Mustern zu Simulationseingangsgrößen ein manueller Prozess ist, wird der Vorgang im Folgenden nicht näher beschrieben, sondern stattdessen die Einsatzmöglichkeiten erläutert.

### 5.2.2 Einsatzmöglichkeiten und Potentiale der simulationsunterstützten Validierung

Das vorrangige Potential der simulationsunterstützten Validierung besteht in der Tatsache, dass durch die gezielte Veränderung der Datenbasis Hypothesen zu dem zu validierenden Muster geprüft werden können. Wenn beispielsweise ein Wirkzusammenhang zwischen den Aktionen von zwei SC-Akteuren in dem Muster zu erkennen ist, könnte eine mögliche Hypothese lauten, dass die anderen Akteure keinen Einfluss auf diesen Wirkzusammenhang haben. Sind in den generierten Daten nun verschiedene Konstellationen der anderen Akteure enthalten, so dürften diese nicht das Muster und dessen Auftrittswahrscheinlichkeit verändern. Dies lässt sich durch die Anwendung des zu dem zu validierenden Muster gehörenden DMMs auf den generierten Daten prüfen. Die gezielte Veränderung des Datenbestandes zu Validierungszwecken kann in Bezug auf das Muster drei Auswirkungen haben:

1. Auftrittswahrscheinlichkeit des Musters häufiger
2. Auftrittswahrscheinlichkeit des Musters seltener bis nicht vorhanden
3. Auftrittswahrscheinlichkeit des Musters gleichbleibend

Je nach Ausprägung der Hypothesen im konkreten Fall validieren die genannten Auswirkungen mit gewissen Wahrscheinlichkeiten die Hypothesen oder widerlegen diese.

Das Vorgehen der Steuerung von Datenlandschaften über Wissen aus Mustern ist auch dann von Relevanz, wenn ein selten vorkommendes Muster validiert werden soll. Dieses Muster kann beispielsweise bei der ungerichteten Transaktionsdatengenerierung, wie sie in Abbildung 5.4 dargestellt ist, mitunter gar nicht auffindbar und folglich nicht zu validieren sein. Die Parameter des Musters ermöglichen nun, unter Einsatz von Expertenwissen geeignete Konstellationen in der Simulation zu erzeugen und somit das Muster mittels der Wissensentdeckung unter verschiedenen Konstellationen zu untersuchen.

Ein weiteres Potential der simulationsunterstützten Validierung besteht in dem Wissen um die SC-Ausgangssituation, in der das zu validierende Muster entdeckt wurde. In den Echtdaten und ihren Systemen ist deren Ausgangssituation im Regelfall nicht oder zumindest nicht eindeutig zu bestimmen. Auch die Transaktionsdaten der Systeme, die als Basis für die Wissensentdeckung dienen, bieten keine Optionen für eine Berechnung der Datenausgangslage und der zugehörigen Systemzustände (vgl. z. B. Abschnitt 5.2.1). Ohne Einsatz von Simulation ist die zeitintensive Analyse des SCM, um Muster und deren Zusammenhang zu den zugrundeliegenden Datenkonstellationen zu bewerten, in der Validierungsphase unumgänglich. Im Gegensatz hierzu bietet der Einsatz der Simulation in der Validierungsphase die Möglichkeit, die SC-Ausgangssituation, die letztendlich zu spezifischen Mustern geführt hat, zu untersuchen. Einerseits ist in diesem Analyseprozess das unternehmensbezogene Kontextwissen für die Bewertung von Mustern

und SC-Ausgangsituation weiterhin notwendig. Andererseits entfällt eine zeitaufwendige Analyse von Echtsystemen und es liegt eine eindeutige Ausgangssituation vor. Das Wissen um die SC-Ausgangssituation in der simulationsunterstützten Validierung bietet in Folge verschiedene neue Optionen für die Validierung.

Als erstes bietet das Wissen um die Ausgangssituation, die zum Auftreten des zu validierenden Musters geführt hat, die Option, die Ausgangssituation gezielt zu verändern. Die vermuteten Auswirkungen auf die Musterauftrittswahrscheinlichkeit fungieren wie zuvor als Hypothesen. Allerdings ist nun der wesentliche Unterschied, dass nicht das Muster als Ausgangspunkt der gezielten Datengenerierung dient, sondern die Ausgangssituation der SC-Simulation. Zusätzlich können Muster, deren Ausgangssituation für die Wissensentdeckung irrelevant sind, unmittelbar ausgeschlossen werden. Solche Ausgangssituationen in der Simulation sind in ihrer konkreten Ausprägung unternehmensspezifisch. Hierzu müssen jedoch durch die Simulation keine aufwendigen Rekonstruktionen einzelner Systeme oder Protokolle analysiert werden, denn die Ausgangssituation in der SC-Simulation kann direkt von den SCM-Akteuren untersucht und bewertet werden.

### 5.2.3 Weiterführende Anwendungsmöglichkeiten der simulationsunterstützten Validierung

Neben den zuvor aufgeführten unmittelbaren Potentialen ergeben sich weitere Anwendungsmöglichkeiten der simulationsunterstützten Validierung. Diese Möglichkeiten sind zeitlich und fachlich der Validierungsphase in MESC nachgelagert. Es handelt sich hierbei um Potentiale in der Entwicklung und Prüfung von Handlungsempfehlungen, wie sie im ursprünglichen Modell von Hippner und Wilde unter Schritt 7.3 aufgeführt wurden (vgl. Tabelle 2.3). Das Ableiten von Handlungsempfehlungen ist nicht mehr Teil des im Rahmen dieser Arbeit entwickelten Methode MESC (vgl. Argumentationslinie in Abschnitt 3.5), soll aber dennoch wegen der Potentiale der Simulation im Kontext der Wissensentdeckung kurz erläutert werden.

Aus den validierten Mustern der MESC werden im Anschluss an die Wissensentdeckung Handlungsempfehlungen für das Unternehmen abgeleitet. Oftmals kann aus den Mustern aber keine eindeutige Handlungsempfehlung abgeleitet werden. Das begründet sich in der Tatsache, dass Muster, wie die Wirkzusammenhänge, nur Aussagen über spezifische SC-Komponenten gestatten (vgl. Abschnitt 2.3.4). Es bleibt offen, wie die anderen SC-Komponenten zu parametrieren sind, was die spezifische Handlungsempfehlung ungewiss gestaltet. Hier kann das Prinzip der gezielten Manipulation der Simulationsausgangssituation aus Abschnitt 5.2.2 angewendet werden. Die Handlungsempfehlungen werden als Hypothesen formuliert, die mittels neu zu generierender Daten des Simulationsmodells überprüft werden können. Um geeignete Daten zu generieren, muss eine entsprechende SC-Ausgangssituation für die Simulationsläufe festgelegt werden. Die SC-Ausgangssituation wird

in der Simulation über die Stellgrößen des Modells erzeugt (vgl. Abschnitt 2.3.3). Die zu einer spezifischen Ausgangssituation generierten Daten können mit der Wissensentdeckung untersucht werden, um die Veränderungen der Auftrittswahrscheinlichkeiten des Musters bzw. die Veränderungen im Muster selbst zu analysieren. In diesem Szenario ist für die Bildung der entsprechenden Hypothesen zwar immer noch Kontextwissen notwendig, aber die Simulation bietet hierbei neue Unterstützungsmöglichkeiten für die Akteure des SCM.

Das Beispiel 5.2 demonstriert die Verwendung im unternehmerischen Kontext und zeigt, wie die Akteure des SCM auch in Bezug auf die Validierung der Handlungsempfehlungen durch die simulationsunterstützte Validierung profitieren können.

**Beispiel 5.2:** In einer SC wurde bei der Durchführung der MESC ein Muster in den SC-Datenbanken erlernt, das durch den Prozess in Abbildung 5.4 validiert wurde. Das Muster gehört zum Subtyp Wirkzusammenhang und zeigt, dass es ab einer bestimmten Lagerauslastung in Konstellation mit einem bestimmten Lieferanten mit einer gewissen Wahrscheinlichkeit zu Verspätungen bei dem Transport einer Artikelgruppe kommt. Das SCM möchte aus diesem Muster eine Handlungsempfehlung ableiten (vgl. Abschnitt 4.6.1) und steht vor der Wahl verschiedener Umstrukturierungsmöglichkeiten. Exemplarische Möglichkeiten der Umstrukturierung in dieser SC sind Lagererweiterung, Veränderung der Transportrouten oder Änderung der Liefermodalitäten. Da das SCM zwar den Wirkzusammenhang kennt, aber keine Hinweise für die Ursache findet, können mittels der Simulation die unterschiedlichen Konstellationen in What-if-Szenarien erprobt werden (vgl. Abschnitt 2.3.3). Der Unterschied zu einem Simulationslauf in der Planungsphase liegt hier in den generierten Daten. Jedes What-if-Szenario erzeugt bei seiner Ausführung in Abhängigkeit der Replikationsanzahl einen oder mehrere Datenbestände (vgl. Abschnitt 5.2.1). Diese Datenbestände können im Anschluss im Rahmen der Durchführung des Prozesses in Abbildung 5.4 untersucht werden. Im Gegensatz zur Mustervalidierung ist nun das Ziel, auf den Datenbeständen des jeweiligen What-if-Szenarios genau dieses Muster nicht mehr zu finden. Das SCM sieht im Rückschluss, welche Konstellationen für eine Handlungsempfehlung geeignet sind.

Diese Information kann das SCM nutzen, um daraus weitere Rückschlüsse für das gefundene Wissen und die resultierenden Handlungsempfehlungen zu treffen. Zusätzlich bietet die simulationsunterstützte Validierung Ansatzmöglichkeiten für über die Validierung von Handlungsempfehlungen hinausgehenden Optionen. In dem zuvor aufgeführten Beispiel können die Datenbestände, die das relevante Muster nicht enthielten, beispielsweise mit weiteren Data-Mining-Verfahren analysiert werden, um zusätzliche Muster zu entdecken. Diese Muster können wiederum Hinweise auf weitere interessante Konstellationen in der SC liefern. An dieser Ausführung wird der iterative Charakter des Verfahrens deutlich. Dieser birgt insbeson-

dere in der unternehmensspezifischen Ausführung der MESC weitere, über die in dieser Arbeit beschriebenen Möglichkeiten hinausgehende, Potentiale.

Es kann konstatiert werden, dass die simulationsunterstützte Validierung eine sinnvolle Ergänzung für die Wissensentdeckung mittels MESC in SC-Datenbanken darstellt. Hierzu wurde aufgezeigt, dass die Simulation in erster Näherung geeignet ist, Testdaten zu erzeugen. Darüber hinaus wurden die Potentiale in der Mustervalidierung diskutiert. Hier wurde aufgezeigt, dass mittels der Musterausprägung und der bekannten Simulationsausgangssituation neue Hypothesen zu den Mustern überprüft werden können. Des Weiteren wurde ein Ansatz aufgezeigt, die simulationsunterstützte Validierung auch für die Validierung von möglichen Handlungsempfehlungen außerhalb der MESC zu nutzen. Somit zeigt sich, dass die simulationsunterstützte Validierung die herkömmlichen Validierungsmechanismen im KDD zielführend ergänzen kann. Zudem bietet sie den Vorteil der unbegrenzten Testdatenbereitstellung in allen Konstellationen. Als eigenständiges Analysewerkzeug bietet sie verschiedene Unterstützungsmöglichkeiten für das SCM und hilft so, Ressourcen und Zeit einzusparen.

## 5.3 Eingliederung der Simulation in die Phasen des Vorgehensmodells

Um die vorgestellten Ansätze erfolgreich im Rahmen der Wissensentdeckung in SC-Datenbanken zu nutzen, ist eine Eingliederung in MESC notwendig. Hierfür ist zu prüfen, auf welche Phasen der MESC sich die vorgestellten Techniken der Transaktionsdatengenerierung (vgl. Abschnitt 5.1) sowie der simulationsunterstützten Validierung (vgl. Abschnitt 5.2) auswirken.

Die Transaktionsdatengenerierung kann bestehende Rohdaten ergänzen oder diese bei Bedarf ersetzen. Die entsprechende MESC-Phase ist die der Datenauswahl (vgl. Abschnitt 4.2), denn die Möglichkeit der Datengenerierung kann alle Schritte der Phase ergänzen oder ersetzen. Im Rückschluss wird die Phase 2 der MESC (vgl. Tabelle 3.3) erweitert, in dem der Aspekt der Modellkomponente in die Beschreibung der einzeln Schritte aufgenommen wird. Das Ergebnis ist in Tabelle 5.3 dargestellt. Die Veränderungen in den Phasenbeschreibungen wurden in dieser und der folgenden Tabelle kursiv gesetzt, um eine Unterscheidung zu den Originaltabellen der MESC zu erleichtern.

Die praktischen Überlegungen, welche Datenaggregationsstufen für eine konkrete Ausführung der MESC notwendig sind (vgl. Abschnitt 5.1.2.1) sollten zu Beginn der Datenauswahl dokumentiert werden. In Abhängigkeit der Aggregationsstufe der SC-Daten ist der Einsatz der generierten Daten in MESC variabel. Da die direkte Verwendung der Daten jedoch nicht notwendigerweise Teil der Datenauswahl ist, kann der Einsatz der Transaktionsdatengenerierung innerhalb der MESC auf Phase 2 beschränkt werden.

**Tabelle 5.3: Erweiterung der MESC-Phase 2**

| Phase | Schritte | Kurzbeschreibung |
|---|---|---|
| 2. Auswahl der relevanten Datenbestände | 2.1 Datenbeschaffung | Bestimmung und Zugang zu den Datenquellen und den zugehörigen Datenbeständen gemäß Zieldefinition *sowie Bestimmung und Zugang zum Simulationsmodell oder Neukonzeption eines Simulationsmodells* |
| | 2.2 Datenauswahl | Auswahl der Datenbestände mittels Kontextwissen *und Bestimmung von Aufbau und Umfang der Ausgabedaten der Simulation mit dem Ziel der Datenreduktion* |

Die simulationsunterstützte Validierung stellt eine innovative V&V-Technik für die Wissensentdeckung in SC-Datenbanken dar. Ihr Einsatz in MESC muss folglich im Methodenelement des Dreiecksmodells verortet werden (vgl. Tabelle 3.4). Die Validierung der Handlungsempfehlung wird nicht in die MESC eingegliedert, da sie eine über die Grenzen der MESC hinausgehende Möglichkeit des Simulationseinsatzes darstellt (vgl. Abschnitt 5.2.3). Da die simulationsunterstützte Validierung für die Überprüfung der Muster eingesetzt wird, erfolgt die Integration nach dem Data Mining. Die betroffenen Phasen der MESC sind in Folge 5, 6 und 7. Die betoffenen Schritte sind in Tabelle 5.4 aufgeführt.

Abschließend muss darauf hingewiesen werden, dass sowohl im Fall der Transaktionsdatengenerierung als auch in der simulationsunterstützten Validierung das Erstellen und Durchführen eines Simulationsmodells einen erheblichen Zeitfaktor darstellt. Daher ist insbesondere der Einsatz der simulationsunterstützten Validierung eher bei wiederholenden SCM-Aufgaben wie der regelmäßigen Prognose zu empfehlen (vgl. Abschnitt 4.1). Darüber hinaus kann der Einsatz der Simulation bei kritischen Entscheidungen zielführend sein, denn hier kann das zu erwartende Risiko den Mehraufwand rechtfertigen.

Die Konzeption der Simulationsmodelle wurde diskutiert und die beeinflussenden Faktoren dargestellt. Es wurde dargelegt, dass insbesondere zielführende Aggregationsstufen und die Eignung eines möglicherweise existierenden Simulationsmodells großen Einfluss auf die konzeptionellen Entscheidungen haben (vgl. Abbildung 5.5). Die Vorteile und Nachteile der Transaktionsdatengenerierung und der simulationsunterstützten Validierung wurden erörtert und konstatiert, dass unternehmensspezifische Faktoren die Einsatzentscheidung der entwickelten Techniken beeinflussen. Bei geeigneten unternehmens- und projektspezifischen Randbedingungen können jedoch beide Methoden zur Verbesserung der MESC in den Be-

reichen Ausführzeit, Ausführunterstützung und Einsatzmöglichkeit beitragen. Zusammenfassend kann festgestellt werden, dass die Transaktionsdatengenerierung und die simulationsunterstützte Validierung eine Ergänzung für existierende Phasenschritte der MESC darstellen und diese mitunter sogar ersetzen können.

Abbildung 5.7 zeigt den Gesamtablauf der MESC (auf Basis von Abbildung 3.5) mit der Integration der in diesem Kapitel vorgestellten Methodenelemente. Die Beschreibung der einzelnen Schritte folgt hierbei der Tabelle 3.3 (dargestellt in schwarz), der Tabelle 3.4 (dargestellt in grün), den Ergänzungen aus Tabelle 5.3 (dargestellt in rot) sowie den simulationsunterstützten Validierungsschritten aus Tabelle 5.4 (dargestellt in blau). Die Entwicklung der Methode ist durch die Integration der in diesem Kapitel vorgestellten Methodenelemente abgeschlossen (vgl. hierzu auch Abschnitt 3.1).

**Tabelle 5.4: Einsatz der simulationsunterstützten Validierung in der V&V**

| **Phase** | **intrinsische Prüfung** | **Prüfung gegen die Vorphase** |
|---|---|---|
| 5. Anwendung des Data-Mining-Verfahrens | 5,5: Prüfung auf richtige Anwendung des Data-Mining-Verfahrens *durch simulationsunterstützte Validierung* | - |
| 6. Weiterverarbeitung der Data-Mining-Ergebnisse | 6,6: Prüfung auf ordnungsgemäße Weiterverarbeitung der Data-Mining-Ergebnisse *durch simulationsunterstützte Validierung* | 6,5: Prüfung *durch simulationsunterstützte Validierung*, ob die Anwendung der Data-Mining-Verfahren interpretierbare Daten als Resultat liefern |
| 7. Bewertung des Data-Mining-Prozesses | - | 7,6: Prüfung *durch simulationsunterstützte Validierung*, ob die vorhandenen Data-Mining-Ergebnisse ausreichend für die Qualitätskontrolle dokumentiert sind<br>7,5: Prüfung *durch simulationsunterstützte Validierung*, ob die Anwendung des Data-Mining-Verfahrens genügend dokumentiert ist |

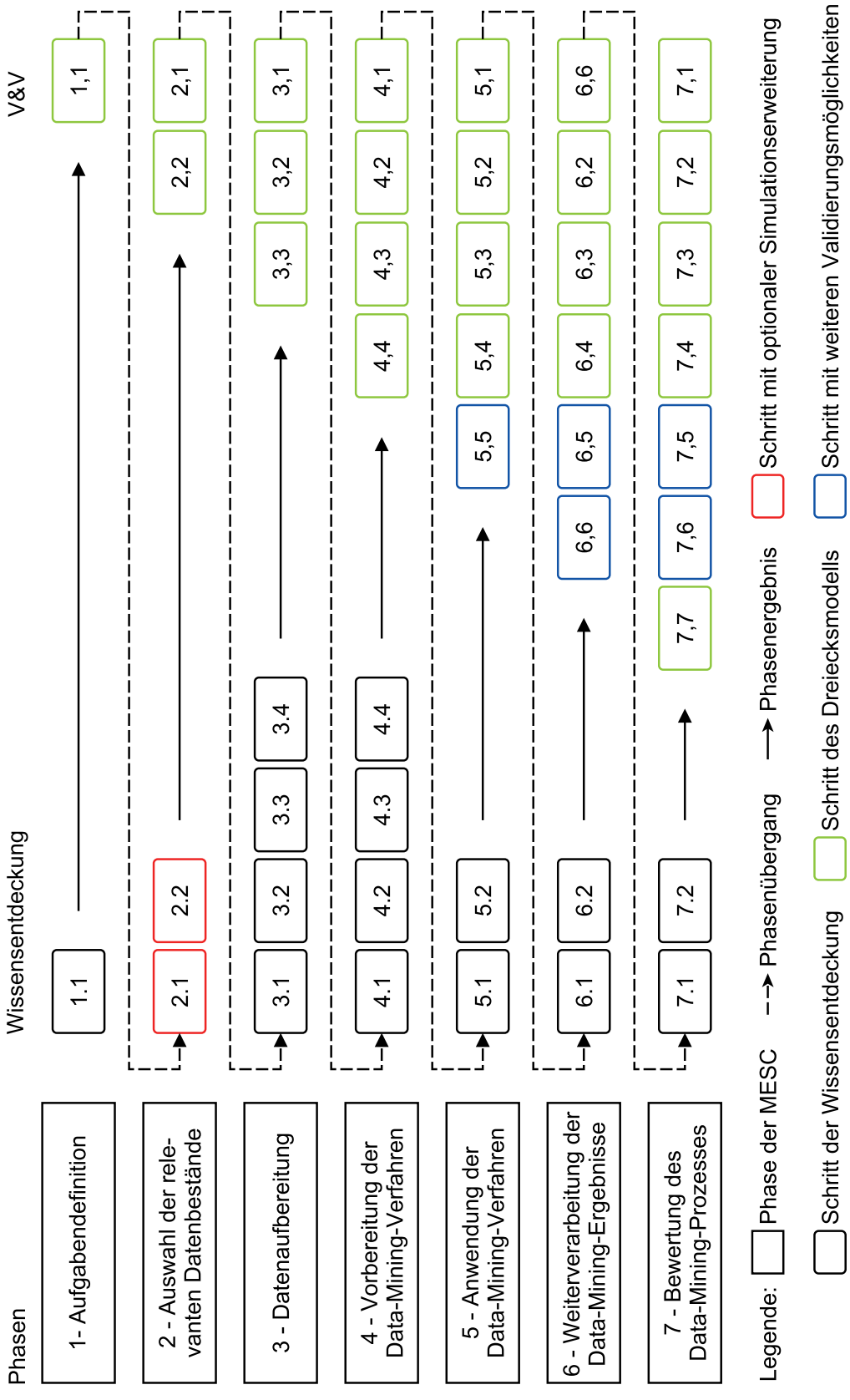



**Abbildung 5.7: Gesamtübersicht der Methodenelemente der MESC mit Integration der Simulation**

# 6 Übertragung in die Praxis

In diesem Kapitel wird die Praxistauglichkeit der zuvor entwickelten Methode MESC an zwei Anwendungsfeldern demonstriert. Dafür wird zunächst das Evaluierungskonzept erläutert sowie dargelegt, in welchem Umfang eine Evaluierung stattfinden kann. Im Anschluss erfolgt die Evaluierung der in dieser Arbeit entwickelten Methoden. Im letzten Abschnitt werden die Evaluierungsergebnisse vorgestellt und die sich daraus ergebenen Rückschlüsse diskutiert.

## 6.1 Evaluierungskonzept

Die Evaluierung der entwickelten Methode wird in zwei Anwendungsfeldern durchgeführt. Da die Methode aus verschiedenen Methodenelementen besteht (vgl. z. B. Abschnitt 5.3), ist eine separate Evaluierung möglich.

Das erste Anwendungsfeld zeigt den Einsatz der MESC in der SC. Hierzu wird das entwickelte Vorgehensmodell der MESC auf die Produktionsdaten eines internationalen Konzerns aus dem Sektor der Elektronikkleingeräte angewandt, der bis zu diesem Zeitpunkt noch keine Prozesse zur Wissensentdeckung in den SC-Datenbanken etabliert hatte. Die Erkenntnisse aus der Durchführung der MESC auf den produktionslogistischen Daten dienen dem Konzern als Unterstützung für die Konzeptionierung eines globalen Datencockpits. Ein Ziel der Evaluierung ist, die Praxistauglichkeit der einzelnen Phasen zu untersuchen. Hierbei ist von zentraler Bedeutung, ob die definierten Phasenschritte der MESC zur Erfüllung einer Fragestellung aus dem SC-Bereich angemessen sind. Dies geht einher mit der Untersuchung, ob die gewählte Anordnung der Phasen zielführend ist (vgl. auch Modellaufbau in Abschnitt 3.5). Ein wesentlicher Aspekt in der Durchführung der MESC stellt die V&V der einzelnen Phasen dar. Neben der generellen Eignung des Dreiecksmodells zur V&V im Bereich KDD steht hier der konkrete Einsatz aus den beispielhaft aufgeführten V&V-Techniken der Literatur im Vordergrund (vgl. Tabelle A.4).

Das zweite Anwendungsfeld stellt ein Data-Farming-Modell zur Generierung von Transaktionsdaten vor. Hier liegt der Fokus der Diskussion auf der Funktionalität der MESC mit dem integrierten Methodenelement der Simulation. Der Schwerpunkt der Betrachtung liegt auf der Prozessabfolge von Simulation, Wissensentdeckung und simulationsunterstützter Validierung. Dieser Fokus unterscheidet sich vom Anwendungsfeld 1, in dem die einzelnen Phasen der MESC und ihre V&V von zentraler Bedeutung sind, insbesondere durch die Evaluierung der Simulationsmethodenelemente (vgl. Abschnitt 5.3). Die Evaluierung erfolgt an einem Simulationsmodell, das spezifisch für die Generation von Transaktionsdaten entwickelt wurde. Die Evaluierung umfasst somit nicht die generelle Eignung von Simulationsmodellen im Bezug auf unterschiedliche Aggregationsstufen der Trace-Ausgabegrößen

(vgl. z. B. Abschnitt 5.1.2.1). Ziel der Evaluierung ist vielmehr, aufzuzeigen, dass die in dieser Arbeit entwickelten Ansätze der simulationsunterstützten Validierung sowie der Transaktionsdatengenerierung zielführende Methodenelemente des MESC darstellen. Die Anwendbarkeit innerhalb der MESC wird hierbei mittels ausgewählter Experimente belegt und demonstriert, wie die grundlegenden Konzeptionsschritte in der Methode umgesetzt werden können.

Der in diese Arbeit entwickelte Musterbegriff (vgl. Abschnitt 4.4) wird konsequent in der Evaluierungsphase verwendet. Ob dieser Begriff außerhalb der entwickelten Methode Verbreitung findet und die heterogenen Begriffe aus Mathematik, KDD und Logistik ersetzen kann (vgl. z. B. Tabelle 2.10), ist nicht Teil des Evalutionskonzepts dieser Arbeit.

Als Basis für die in diesem Kapitel durchgeführten Experimente der Anwendungsfelder dient die vollständige MESC inklusive Dreiecksmodell und Integration der Simulation (vgl. Abbildung 5.7).

## 6.2 Anwendungsfeld 1: Wissensentdeckung in SC-Transaktionsdaten

Um die Durchführbarkeit des in dieser Arbeit entwickelten Vorgehensmodells zu demonstrieren, wird die MESC auf die Rohdaten verschiedener Quellsysteme eines internationalen Konzerns angewandt. In Tabelle 6.1 sind die Konzerndaten in einem Datenblatt beschrieben (vgl. hierzu auch die Datenblätter in den Tabellen B.1 und B.2).

**Tabelle 6.1: Beschreibung des Datenbestands für das Anwendungsfeld 1**

| Kriterien | Beschreibung |
|---|---|
| **Fachliche Kriterien** | |
| Branche | • Hersteller für Elektronikkleingeräte |
| Fachliche Beschreibung | • Produktionsdaten<br>• Zeitraum vom 01.04.2013 bis zum 07.04.2016<br>• Zuordnung von Bauteilen zu Endprodukten<br>• Zuordnung von Endprodukten zu Produktionslinien<br>• Zustand von Bauteilen und Endprodukt zusätzlich enthalten |
| Exemplarische Attributsbeschreibung | • OrderID: Bestellnummer<br>• ProductID: Produktnummer<br>• BeginOfManufacturing/EndOfManufacturing: Bearbeitungszeitraum |

**Tabelle 6.1: Beschreibung des Datenbestands für das Anwendungsfeld 1 (Fortsetzung)**

| Kriterien | Beschreibung |
|---|---|
| | • Result: Statusmeldung<br>• Class: Automatische oder manuelle Bearbeitung<br>• WorkplaceID: Bearbeitungsstation |
| **Technische Kriterien** | |
| Quelle | • Data Warehouse |
| Umfang | • Ca. 225 Millionen Datensätze<br>• 670 Attribute in zwei Datenbanken und 45 Tabellen |
| Technische Attributsbeschreibung | Datentypen nach Tabelle A.1:<br>• Bestellnummer: Zeichenkette<br>• Produktnummer: Zeichenkette<br>• Bearbeitungszeit: Datetime<br>• Statusmeldung: Zeichenkette<br>• Automatische/manuelle Bearbeitung: Boolean<br>• Bearbeitungsstation: Ganzzahl |
| **Untersuchungskriterien** | |
| Analysenotizen | • Fast alle Tabellen existieren in beiden Datenbanken<br>• Einige Tabellen sind in beiden Datenbanken identisch oder haben nur kleine Unterschiede |
| Primärfrage | • Welche Wirkzusammenhänge können unter Einbeziehung von Kontextwissen entdeckt werden und welche Wirkzusammenhänge lassen sich ohne das Einbeziehen von Kontextwissen aufzeigen? |

Die Datenlage erweist sich für die Validierung des entwickelten Modells als geeignet, da in den zugrundeliegenden Datenbanken wesentliche Charakteristika von SC-Datenbanken aufzufinden sind (vgl. Abschnitt 3.2.1). Tabelle 6.2 gibt einen Überblick über die Voruntersuchung bezüglich der Dateneignung des Unternehmens.

Die MESC wird im Anwendungsfeld 1 in seiner Grundform ohne Simulationserweiterung ausgeführt (vgl. Abschnitte 3.5 und Abschnitt 3.6). Da der Konzern sich aktuell in der Entwicklung einer homogenen Datenbasis für Analysen befindet und keine Simulationsansätze existieren, wäre der Aufbau eines Simulationsmodells mit den benötigten Konzepten und Voranalysen ein eigenständiges Großprojekt. Der vorliegende Abschnitt gliedert sich in drei thematische Bereiche, die

**Tabelle 6.2: Voruntersuchung der Charakteristika von SC-Datenbanken in Konzerndatenbanken zur Eignungsprüfung**

| SC-Datenbank Charakteristika | Erfüllt in Konzerndatenbanken | Erklärung |
|---|---|---|
| Kontextwissen | ja | Kontextwissen zur Attributsklärung, Zielstellung und Vorverarbeitung notwendig |
| Datenauswahl | ja | Großer Datenbestand mit steigender Datenqualität und zusätzlichen Attributsausprägungen mit fortschreitender Zeit |
| Gruppierung | unbestimmt | Gruppierung eventuell für Teilprodukte notwendig, da diese in einer Datenbank gehalten werden |
| Atomarität | ja | Existenz innerer Struktur beispielsweise bei Prozess-ID ist bekannt |
| Format | ja | Vielzahl alphanumerischer Formate vorhanden |

die Untersuchungen der einzelnen Phasen aus Kapitel 4 aufgreifen. Alle Phasen der MESC werden auf die Konzerndaten angewandt, um die Praxistauglichkeit sowohl der einzelnen Phasen als auch der Phasenanordnung zu evaluieren. Die relevanten Ergebnisse werden im Rahmen der Abschnitte zur Aufgabenstellung, der Datenauswahl- und aufbereitung, der Data-Mining-Verfahren sowie der V&V der durchgeführten Phasen vorgestellt. Obwohl die V&V die Phasen der MESC begleitet (vgl. beispielsweise Phasenanordnung in Tabelle 3.4), wurde diese in einem separaten Abschnitt dokumentiert. Aus Gründen der Nachvollziehbarkeit der praktischen Evaluierung orientiert sich die Dokumentation in diesem Punkt nicht an dem zeitlichen Ablauf der MESC, sondern fasst alle zu dokumentierenden Schritte der V&V in einem eigenen Abschnitt zusammen.

### 6.2.1 Aufgabendefinition, Datenmodell und Vorverarbeitung

In diesem Abschnitt werden die Ergebnisse der Durchführung der ersten drei MESC-Phasen auf den Datenbeständen (vgl. Tabelle 6.1) der Produktionsunternehmens für Elektronikkleingeräte vorgestellt. Zu diesem Zweck wird zuerst die Bestimmung der Aufgabenstellung diskutiert und im Anschluss werden die zentralen Bearbeitungsschritte in der Datenvorverarbeitung vorgestellt.

#### 6.2.1.1 Bestimmung der Aufgabenstellung

In der ersten Phase wurde die MESC initiiert und die Aufgabenstellung bestimmt (vgl. Abschnitt 4.1). Im Bereich der Aufgabenstellung sind für die Nachvollziehbarkeit der hier dokumentierten Ergebnisse zum einen die eingesetzte Technik und zum anderen die konkrete Ausformulierung der Fragestellung relevant (vgl. Abschnitt 4.1.2). Tabelle 6.3 beschreibt die IT-Konfiguration, die die Ausgangsbasis für die beschriebenen Experimente in diesem Kapitel darstellt.

**Tabelle 6.3: Technische Merkmale**

| Technische Merkmale | Ausprägung |
|---|---|
| IT-Landschaft | • Microsoft SQL Server 2014<br>• ODBC Remoteverbindung |
| Hardware | • Windows 7 Professional 64 Bit-Betriebssystem<br>• Achtkernprozessor mit 3,5 $GHz$<br>• 32 Gigabyte Arbeitsspeicher |
| Zusätzliche Software | • Microsoft SQL Server Management Studio<br>• HeidiSQL |
| Data-Mining-Werkzeuge | • RapidMiner |

Das Ziel der MESC-Durchführung wurde mit der Fachseite des Unternehmens definiert, die in der nachfolgenden Durchführung auch für die Bereitstellung des Kontextwissens verantwortlich war. Die Durchführung der MESC wurde in diesem Anwendungsfeld von zwei Fragestellungen und korrespondieren Aufgabenfeldern bestimmt. Die erste Teilaufgabe sollte die Frage beantworten, ob die Datengranularität der relevanten IT-Systeme das Entdecken von interessanten Wirkzusammenhängen gestattet. Um die Menge der potentiellen Wirkzusammenhänge einzuschränken, wurde die Fragestellung des Unternehmens spezifiziert. Von besonderem Interesse für das Unternehmen waren die Zusammenhänge zwischen nicht bestandenen technischen Prüfungen der produzierten Produkte und Prozessmerkmalen, wie beispielsweise Produktionslinie, Prozessnummer oder Zeitpunkt der Fertigung. Die zweite Teilaufgabe behandelt die Frage, ob der Einsatz von Clusterverfahren (vgl. Tabelle 2.9) eine Unterteilung der produzierten Elektronikkleingeräte liefern kann, die einer Ähnlichkeitsstruktur folgen (vgl. Abschnitt 2.3.2). Diese Ähnlichkeitsstruktur sollte als Basis für weitere Verfahren der Wissensentdeckung in einzelnen Clustern sowie als mögliche Strukturvorgabe für automatische Produktionsüberwachungen dienen. Als Alternative zur Clusteranalyse wurde mit der Fachseite die Möglichkeit der Klassifikation für das Datencockpit diskutiert. Da für diese jedoch ein spezifisches Attribut als Klassifikator bestimmt werden muss, erfolgte eine erste Datensichtung mit der Fachseite. Da nach Datensichtung kein

sinnvoller Klassifikator für den zugrundeliegenden Datenbestand benannt werden konnte, wurde diese Option für die erste Durchführung einer Wissensentdeckung auf Unternehmensdaten ausgeschlossen.

#### 6.2.1.2 Datenauswahl und -aufbereitung

Die Auswahl der relevanten Datenbestände (vgl. Abschnitt 4.2) wurde mit IT- und fachseitiger Unterstützung des Unternehmens durchgeführt. Da das Unternehmen keine produktspezifischen Datenstrukturen nutzt, sind neben einer Vielzahl von ungenutzten Attributen auch leere Tabellen in den IT-Systemen vorhanden. Aufgrund der Komplexität der Datenstrukturen wurde der Entschluss getroffen, zu Projektbeginn einen nahezu vollständigen Bestand an verfügbaren Daten als Basis zu verwenden. In einer Voranalyse wurden dann die von der Fachseite identifizierten fünf Haupttabellen mit notwendigen Informationen aus den weiteren Tabellen angereichert. Im Anschluss erfolgte die Phase der Datenaufbereitung (vgl. Abschnitt 4.3). Die praktische Ausführung der Experimente zur Datenaufbereitung wurden mit Unterstützung von studentischer Zuarbeit durchgeführt (Klein 2017; Li 2017). Sofern Ergebnisse direkt verwendet wurden, erfolgt eine direkte Referenzierung der entsprechenden Zuarbeit.

Die Datenaufbereitung als dritte Phase der SC besteht aus vier Schritten (vgl. Tabelle 3.3). Der erste Schritt in der Datenaufbereitungsphase ist die Formatstandardisierung (vgl. Abschnitt 4.3.1). Diese ist ein notwendiger Schritt, da die Daten aus drei verschiedenen IT-Systemen stammen. Alle Datenbestände wurden als Datenbanktabellen in MSSQL exportiert. Die resultierende MSSQL-Datenbank wurde im nächsten Schritt weitgehend normalisiert. Die vollständige Überführung in die dritte Normalform war aufgrund der Komplexität der Datenbestände in der vorgegebenen Projektzeit nicht durchführbar. Abbildung 6.1 stellt die fünf Haupttabellen der Datenbank dar. Für jede Tabelle sind beispielhafte Attribute angegeben, die eine Übersicht über die Tabelleninhalte ermöglichen. Die Attribute der Haupttabellen wurden für die Erstellung des Datenmodells bereits reduziert. Dies bedeutet, dass Attribute, die keinen Mehrwert für die Wissensentdeckung bringen, in einer Voranalyse aus den Datenbeständen entfernt wurden. Hierzu zählten insbesondere Attribute, die keine Attributsausprägungen aufwiesen (vgl. hierzu auch Abschnitt 6.2.1). Die Beziehungen zwischen den Tabellen werden in der Abbildung über Relationen dargestellt. Zusätzlich wird in jeder Relation das Attribut angegeben, das in Funktion von Primär- und Fremdschlüssel die Tabellen miteinander in Beziehung setzt. Das Attribut „Seriennummer“, das die Tabelle „Werkstück“ mit der Tabelle „Scandaten“ verknüpft, fungiert beispielsweise in der Tabelle „Werkstücke“ als Primärschlüssel und in der Tabelle „Scandaten“ als Fremdschlüssel. Die dargestellten Haupttabellen und ihre Attribute stellen die Datengrundlage für die weiteren MESC-Phasen dar.

Obwohl eine Gruppierungsmöglichkeit der Datenbestände in der Initiierungsphase der MESC nicht identifiziert werden konnte (vgl. Abschnitt 6.2), zeigten die

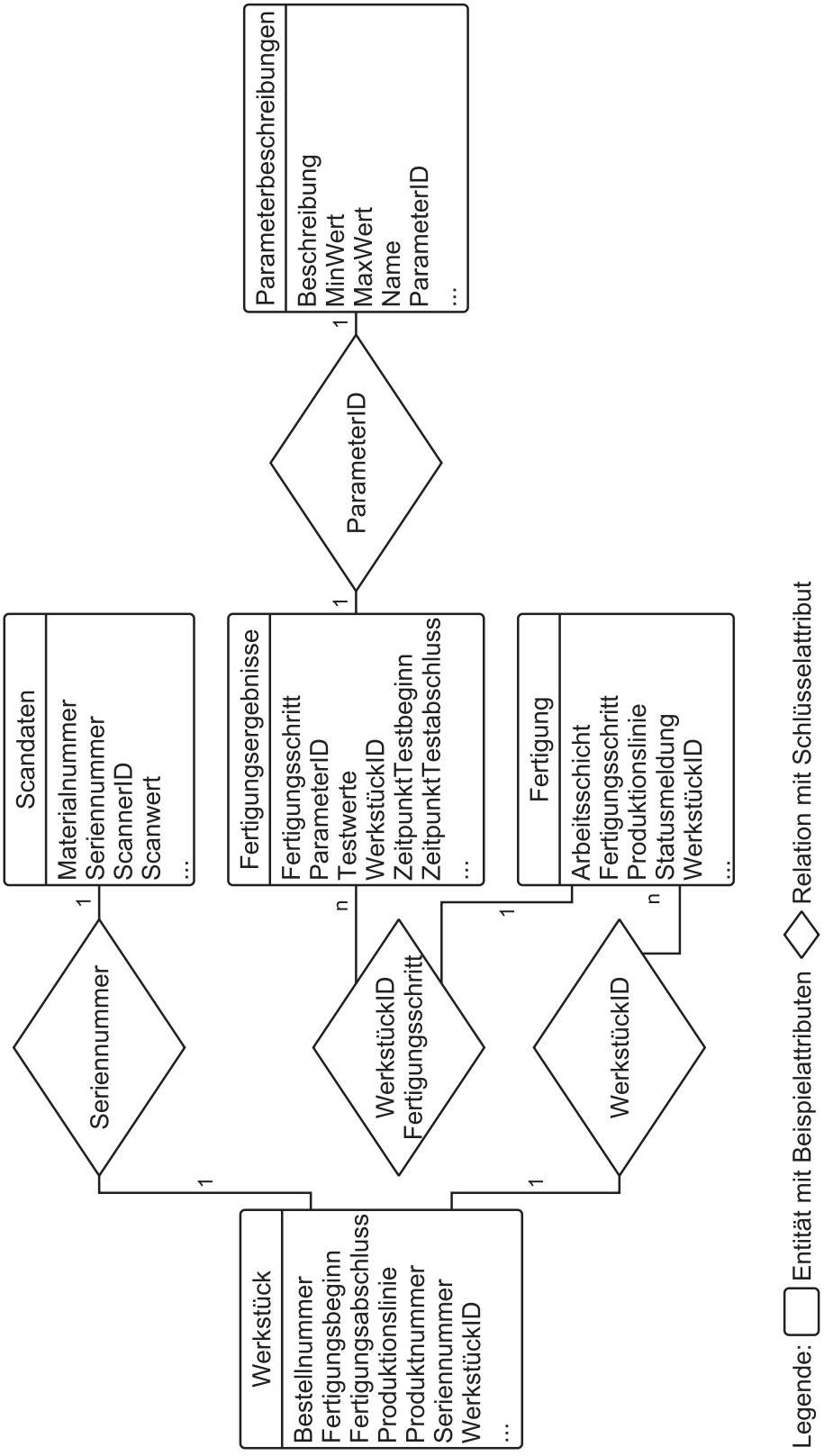



**Abbildung 6.1: ERM der Haupttabellen mit Attributsauszug in der Chen-Notation (Chen 1976)**

Untersuchungen im Bereich der Datenaufbereitung verschiedene Gruppierungsmöglichkeiten auf. Mögliche Gruppierungen ergaben sich beispielsweise durch die Zugehörigkeit von Werkstücken zu einzelnen Produktionslinien und die Zuordnung von Fertigungschritten zu spezifischen Arbeitsplätzen. Eine zielführende Gruppierungsmöglichkeit konnte im Bereich der Messwerte identifiziert werden. In den Unternehmensdatenbeständen repräsentiert ein einzelnes Attribut die Messwerte von verschiedenen Produkttests aus dem Fertigungsbereich. Die Produkttests unterscheiden sich jedoch grundlegend in den eingesetzten Verfahren und den resultierenden Messergebnissen. Da die Attributsausprägungen ohne den Kontextbezug der Messeinheit zu fehlerhaften Regeln in der Wissensentdeckung führen würden (vgl. auch Beispiel Transportweg in Abschnitt 3.2.1), ist eine Gruppierung nach Messeinheiten zielführend. Für das hier diskutierte Anwendungsfeld 1 wurden die Daten wie folgt gruppiert:

- °C = Grad Celsius: Temperatur
  Beispielhafter Anwendungsfall: Motortemperatur
- $mA$ = Milliampere: Elektrische Stromstärke
  Beispielhafter Anwendungsfall: Motorstrom
- $V$ = Volt: Elektrische Spannung
  Beispielhafter Anwendungsfall: Netzspannung
- $Hz$ = Hertz: Frequenz
  Beispielhafter Anwendungsfall: Netzfrequenz
- $mOhm$ = Milliohm: Elektrischer Widerstand
  Beispielhafter Anwendungsfall: Durchgangsprüfung
- $nF$ = Nanofarad: Elektrische Kapazität
  Beispielhafter Anwendungsfall: Kondensatorkapazität
- $1/min$ = Umdrehungen pro Minute
  Beispielhafter Anwendungsfall: Motorgeschwindigkeit
- $g$ = Gramm: Gewicht
  Beispielhafter Anwendungsfall: Gesamtgewicht des Produktes
- Sonstige: Verschiedene Produkttests
  Beispielhafte Anwendungsfälle: Manuelle Prüfung der Seriennummer, Funktionstest Knopf, Funktionstest Bildschirm, Kalibrierungstest

Es muss erwähnt werden, dass die Gruppe „Sonstige“ über 130 verschiedene Produkttests mit teilweise identischen Skalen (z. B. Funktionstests Knöpfe) beinhaltet. Diese Daten könnten mit Unterstützung des hier entwickelten Vorgehensmodells nachklassifiziert werden. Da dieser Bereich jedoch der zugrundeliegenden Datenqualität und nicht der Validierung der MESC zugeordnet werden muss, wird auf eine weitere Diskussion verzichtet.

Im Anschluss erfolgt die Anreicherung des Unternehmensdatenbestands mittels der Attributsoperationen Aggregation, Addition und Transformation (vgl. Abschnitt 4.3.3). In diesem Schritt konnte insbesondere die Bedeutung des Kontextwissens für die MESC validiert werden. Das Kontextwissen wurde zum einen für die abschließende Auswahl der zu transformierenden Attribute benötigt. Zum anderen war auch Kontextwissen für das Ziel der Attributstransformation notwendig, da nur die Fachseite über Informationen der höheren Aggregationsstufen der Attributsausprägungen verfügte.

Tabelle 6.4 demonstriert beispielhaft die Datenanreicherung mittels Kontextwissen. Die Attribute Schicht und Zeitstempel wurden den Unternehmensdatenbeständen entnommen. Die Tabelle zeigt einen Datenauszug aus dem Originalbestand, der verdeutlicht, dass das Attribut „Schicht“ fehlerbehaftet ist. Gemäß den Unternehmensspezifikationen muss eine eindeutige Zuordnung zwischen Zeitstempel und korrespondierender Schicht existieren. Für die Attributsausprägung 9 ist dies jedoch in dem Datenauszug nicht gegeben. Da in den SC-Datenbanken auch identische Zeitstempel unterschiedlichen Schichten zugeordnet wurden, wurde als Basis für die Datenanreicherung nur der Zeitstempel genutzt. Die beiden in diesem Schritt erzeugten künstlichen Attribute sind Wochentag und Teilschicht. Die Teilschicht wurde mittels des fachseitigen Kontextwissens über die Schichtfolge im Unternehmen kodiert.

**Tabelle 6.4: Kodierung der Zeitstempel in Anwendungsfeld 1**

| Schicht | Zeitstempel | Wochentag | Teilschicht |
|---|---|---|---|
| 7 | 21.01.2015 07:20:13 | Mittwoch | morgens |
| 5 | 31.03.2015 14:35:20 | Dienstag | nachmittags |
| 4 | 21.04.2015 10:44:07 | Dienstag | morgens |
| 9 | 26.11.2015 03:54:37 | Donnerstag | nachts |
| 9 | 02.12.2015 23:10:55 | Mittwoch | nachts |
| 2 | 22.02.2016 16:54:46 | Montag | nachmittags |
| 14 | 01.04.2016 13:32:39 | Freitag | nachmittags |

Der resultierende Datenbestand wurde im nächsten Schritt den Transformationsoperationen unterzogen. Da bereits Schritte der Vorverarbeitung, wie das Entfernen von Attributen ohne Attributswerte, im Rahmen der Erstellung des Datenmodells durchgeführt wurden, schränkte sich die Zahl der durchzuführenden Operationen ein. Insbesondere ist hier zu erwähnen, dass auch Attribute mit mehr als einer Merkmalsausprägung in Rücksprache mit der Fachseite entfernt wurden. Dies zeigt, dass das Kontextwissen nicht nur zur Identifikation von fachlichen Attributsbeziehungen dient, sondern auch für die Identifizierung von zu eliminierenden Attributen genutzt werden kann (vgl. Abschnitt 4.3.4). Die Ausführungsreihenfol-

ge der Transformationsoperationen entsprach in diesem Fall der vorgeschlagenen Reihenfolge der MESC (vgl. Tabelle 3.3). Es hat sich jedoch in Experimenten gezeigt, dass auch andere Ausführungsreihenfolgen das gleiche Ergebnis liefern. Dies wurde bereits in Abschnitt 4.3.4 diskutiert, als darauf hingewiesen wurde, dass die Ausführungsreihenfolge der Transformationen flexibel ist und bei Bedarf abgewandelt werden kann. Die Durchführung von Transformationen auf den Unternehmensdaten hat diese Aussage bestätigt.

Die in diesem Abschnitt diskutierten Schritte der Phase 2 und 3 der MESC sind für beide Aufgabenstellungen des hier behandelten Anwendungsfelds (vgl. Abschnitt 6.2.1) identisch. Dies zeigt, dass Verfahren wie beispielsweise das Data Cleansing unabhängig von spezifischen Data-Mining-Verfahren Einsatz finden und somit die gewählte Anordnung der beiden Phasen sowie die in ihnen enthaltenen Schritte sinnvoll gewählt wurden.

## 6.2.2 Vorbereitung und Anwendung der Data-Mining-Verfahren

In diesem Abschnitt wird die Vorbereitung und Anwendung der eingesetzten Data-Mining-Verfahren diskutiert. Da die Teilaufgaben des hier behandelten Anwendungsfelds (vgl. Abschnitt 6.2.1.1) verschiedene Kodierungen sowie den Einsatz von unterschiedlichen Data-Mining-Verfahren erzwingen, werden relevante Auszüge der durchgeführten Schritte in jeweils einem eigenen Abschnitt vorgestellt. Zumal alle Experimente mittels RapidMiner durchgeführt wurden, ist die Werkzeugwahl nicht Teil der Validierung (vgl. Schritt 4.2 in Tabelle 3.3). Es wird auf die allgemein getroffenen Grundvoraussetzungen der vorliegenden Forschungsarbeit verwiesen (vgl. Abschnitt 2.4.1).

### 6.2.2.1 Clusteranalyse

Der Einsatz der Clusteranalyse stellt im Umfeld von SC-Datenbanken eine Herausforderung dar, da ein Ähnlichkeitsmaß oftmals nur mittels Kontextwissen bestimmt werden kann. Diese Problematik wurde bereits im Rahmen der Vorverarbeitung auf verschiedenen Datensätzen verdeutlicht (vgl. Abschnitt 4.3.2). Um eine Clusteranalyse auf den SC-Datenbanken des Unternehmens durchzuführen, erfolgte eine fachliche und technische Kodierung der Daten. Hierbei lag der Schwerpunkt der Kodierung auf der technischen Umwandlung der Attribute (vgl. Abschnitt 4.4.4). Insbesondere die Umwandlung der Attribute in verfahrensspezifische Datentypen (vgl. Tabelle A.1) war für die erfolgreiche Durchführung der Data-Mining-Verfahren notwendig. Da die technische Kodierung ohne Einsatz von Kontextwissen durchgeführt werden konnte, ergab sich in diesem Projektabschnitt ein hoher Autonomiegrad von der Unternehmensfachseite. Das Kontextwissen war in Bezug auf die Clusteranalyse nur dann notwendig, wenn beispielsweise innere Strukturen in den einzelnen Attributen (vgl. Tabelle 3.1) wichtige Informationen für die Cluster-

analyse beinhalteten. In den durchgeführten Experimenten wurde die Attributsanzahl variiert, im Durchschnitt wurde die Clusteranalyse jedoch auf 30 Attributen durchgeführt. Tabelle 6.5 zeigt beispielhaft einen Auszug aus den vorverarbeiteten Daten des Unternehmens.

**Tabelle 6.5: Beispielhafte Kodierung der Daten für RapidMiner k-Medoids und DBScan**

| SerialNumber | ParamDescription | ProductId | WorkPlace | Line |
|---|---|---|---|---|
| 15 325 513 | PDES00000164 | PROD00000010 | WP80 | 2 |
| 15 325 522 | PDES00000176 | PROD00010010 | WP94 | 2 |
| 15 325 532 | PDES00000130 | PROD00000011 | WP80 | 2 |
| 15 325 512 | PDES00000124 | PROD00000012 | WP93 | 3 |
| 11 552 666 | PDES00000179 | PROD00000013 | WP93 | 2 |
| 15 325 554 | PDES00000165 | PROD00000055 | WP17 | 3 |

Mit den vorverarbeiteten Datenbeständen erfolgte die Durchführung der Experimente für die Wissensentdeckung. In den Experimenten kamen unterschiedliche Verfahren zur Clusterung sowie verschiedene Implementierungen der ausgewählten Clusteranalysen zum Einsatz (vgl. Tabelle 2.9). Eine große fachliche Herausforderung im Bereich der MESC-Clusteranalyse lag in der Ergebnisinterpretation der Experimente. Das ist darin begründet, dass bei Verfahren wie der Clusteranalyse die Eigenschaften der Cluster erst durch nachträgliche Analyse mittels Kontextwissens zu interpretieren sind. Zudem hängen die gefundenen Cluster stark von den genutzten Verfahren und Parametrierungen ab. Aus diesem Grund wurde eine Vielzahl von Clustermodellen erstellt und eine Vorauswahl durchgeführt (für Experimentaufbauten siehe auch Vorarbeiten zu dieser Arbeit von Li (2017)). Bei der Modellerstellung wurden sowohl die Clusterverfahren und Parametrierung variiert als auch die Gruppierung der zugrundeliegenden Daten verändert (vgl. Abschnitt 6.2.1.2). Die erstellten Modelle wurden mit der Fachseite diskutiert. Als Ergebnis wurden eine Auswahl von Modellen und daraus resultierenden Mustern für das Datencockpit definiert. Hierbei muss erwähnt werden, dass auch verworfene Modelle einen Wissensgewinn für das Projekt bedeuteten. In Abbildung 6.2 ist das Ergebnis einer Clusteranalyse auf gruppierten Daten nach Messeinheit = Milliohm dargestellt.

Um die einzelnen Cluster hervorzuheben, wurden die Clusterübergänge durch Linien abgegrenzt. Die Position der Trennlinien wurde so gewählt, dass diese den Beginn eines neuen Clusters repräsentieren. Die Auswahl richtet sich nach den Farbübergängen und folgt keiner mathematischen Berechnung. Beispielsweise wurde die erste Trennlinie bei der Position 380 000 auf der Ordinatenachse so festgelegt, dass oberhalb der Trennlinie ausschließlich grüne Punkte (Cluster 0) und unterhalb möglichst viele lilafarbene Punkte (Cluster 4) positioniert sind. Demzufolge

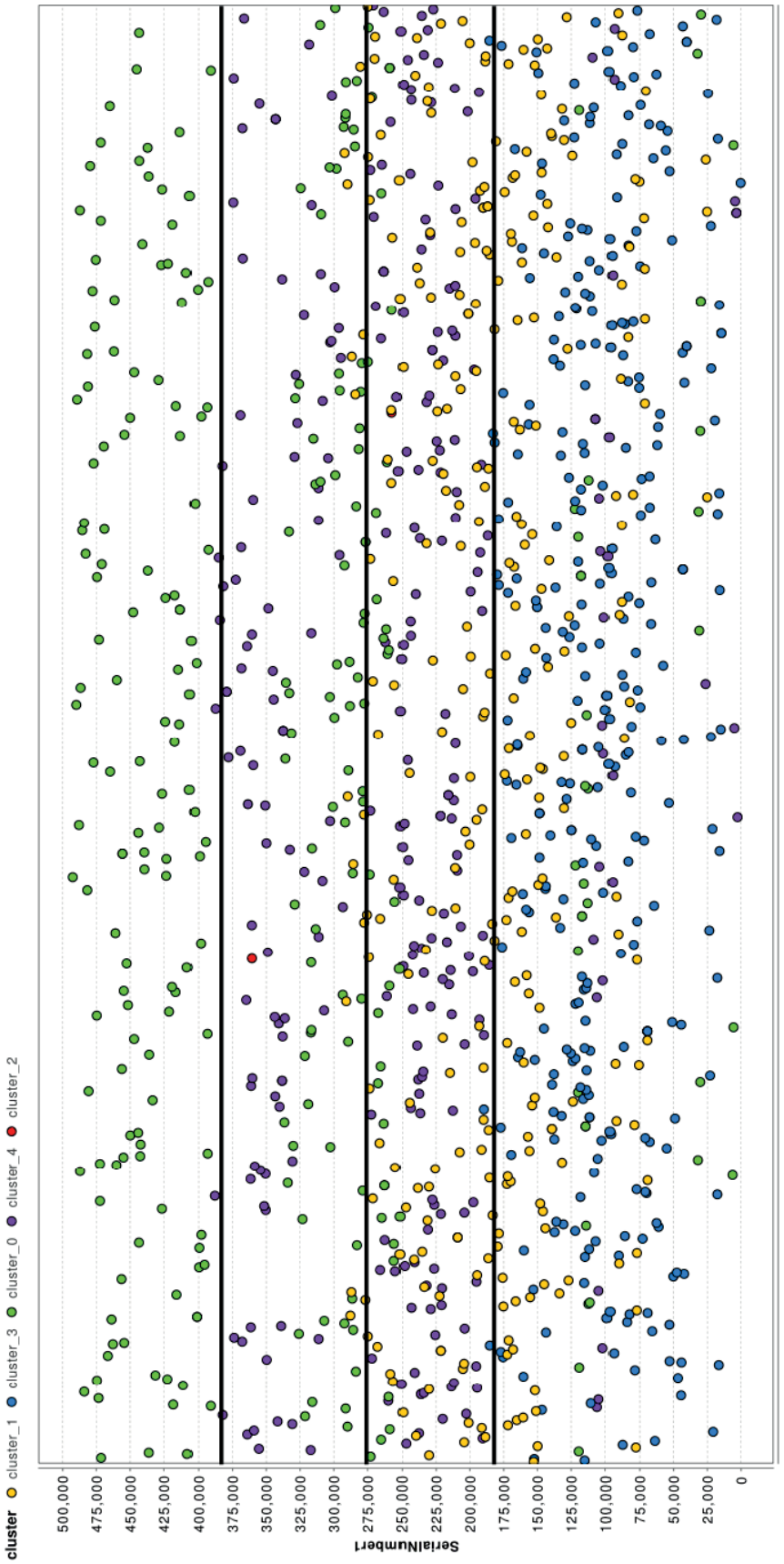



**Abbildung 6.2: Clusteranalyse auf gruppierten Daten nach Messeinheit = Milliohm mit k-means, mixed measures, 150 max runs, 1 000 Optimierungsschritten und k = 5**

zeigt die oberste Linie den Übergang zu Cluster 4, die mittlere Linie den Übergang zu Cluster 1 und die unterste Linie den Übergang zu Cluster 3. Die Clusternummerierung und -färbung ist algorithmisch bedingt und gestattet keine fachliche Interpretation.

Auf den Achsen sind ausschnittsweise die Materialnummern (SerialNumber 1) sowie die Seriennummern der gefertigten Produkte (GUID) abgebildet. Die GUID wurde auf der Abszisse aufgrund von Rückschlussmöglichkeiten auf unternehmenensspezifische Informationen nicht abgebildet. Diese konkrete Zuordnung zu spezifischen GUID ist jedoch auch nicht notwendig, denn die Clusterbildung kann aufgrund der Farben und künstlichen Trennlinien nachvollzogen werden. Das Experiment wurde im Anschluss auf gruppierten Daten nach Messeinheit = Milliampere wiederholt. In Abbildung 6.3 ist zu sehen, dass das erlernte Muster aus nahezu identischen Clustern besteht. Dies kann nachvollzogen werden, wenn die Anordnung und der Verlauf der Trennlinien der ersten Clusterung (Abbildung 6.2) mit der zweiten Clusterung (Abbildung 6.3) verglichen wird. Insbesondere die Position des kleinsten Clusters (unterhalb der obersten Trennlinie, linker Quadrant), in der Farbe rot dargestellt, ist in beiden Mustern kongruent.

In Experimenten auf zwei weiteren Gruppen war die Zuordnung der Transaktionsdaten zu den einzelnen Clustern vergleichbar zum ersten Experiment. Dies gestattet den Rückschluss, dass die ausgewählte Gruppierung für die Clusteranalyse nicht zielführend ist, da keine gruppenspezifische Clusterung gefunden werden konnte. Dies bedeutet, dass sich aufgrund des Produkttests keine gruppenspezifischen Produkte (GUID) und entsprechende Materialien (SerialNumber 1) sowie zugehörigen Zuordnungen der korrespondieren Transaktionen zu den Clustern finden lassen. Die fachliche Erklärung hätte sein können, dass bei der Fertigung eines Elektronikkleingeräts alle Produkttests durchlaufen wurden und sich das Verhalten des Gesamtdatenbestands in der Folge nur in den gruppierten Daten widergespiegelt hätte. Diese Hypothese wurde von der Fachseite bestätigt, da am Ende der Fertigung alle Produkte die vollständige Prüflinie durchlaufen und in der Folge jedes Produkt auch alle Produkttests durchläuft.

Darüber hinaus wurden weitere Experimente zu unterschiedlichen Clusterverfahren mit variierender Clusteranzahl im Anwendungsfeld 1 durchgeführt. Insbesondere die Zuordnung von fehlerhaften Fertigungsprozessen zu möglichen Clustern war von besonderem Interesse für das Unternehmen. Das Streudiagramm in Abbildung 6.4 zeigt das Ergebnis eines Experiments, das auf einer 100 000-Stichprobe des Gesamtdatenbestands durchgeführt wurde. Bei dem Streudiagramm handelt es sich um einen „Jittered Scatterplot“, d. h. in der grafischen Darstellung werden auf die einzelnen Elemente der Cluster Zufallszahlen addiert, so dass die Clusterpunkte um den ursprünglichen Clustermittelpunkt streuen. In den zugrundeliegenden Clustermodellen sind die Daten jedoch unverändert. Es handelt sich lediglich um eine grafische Aufbereitungsart, da ansonsten die Clusterpunkte überlagern würden und die Anzahl der Transaktionen in den einzelnen Clustern, d. h. die Clustergrößen, nicht darstellbar wären. Auf der Ordinate sind die Teilschichten

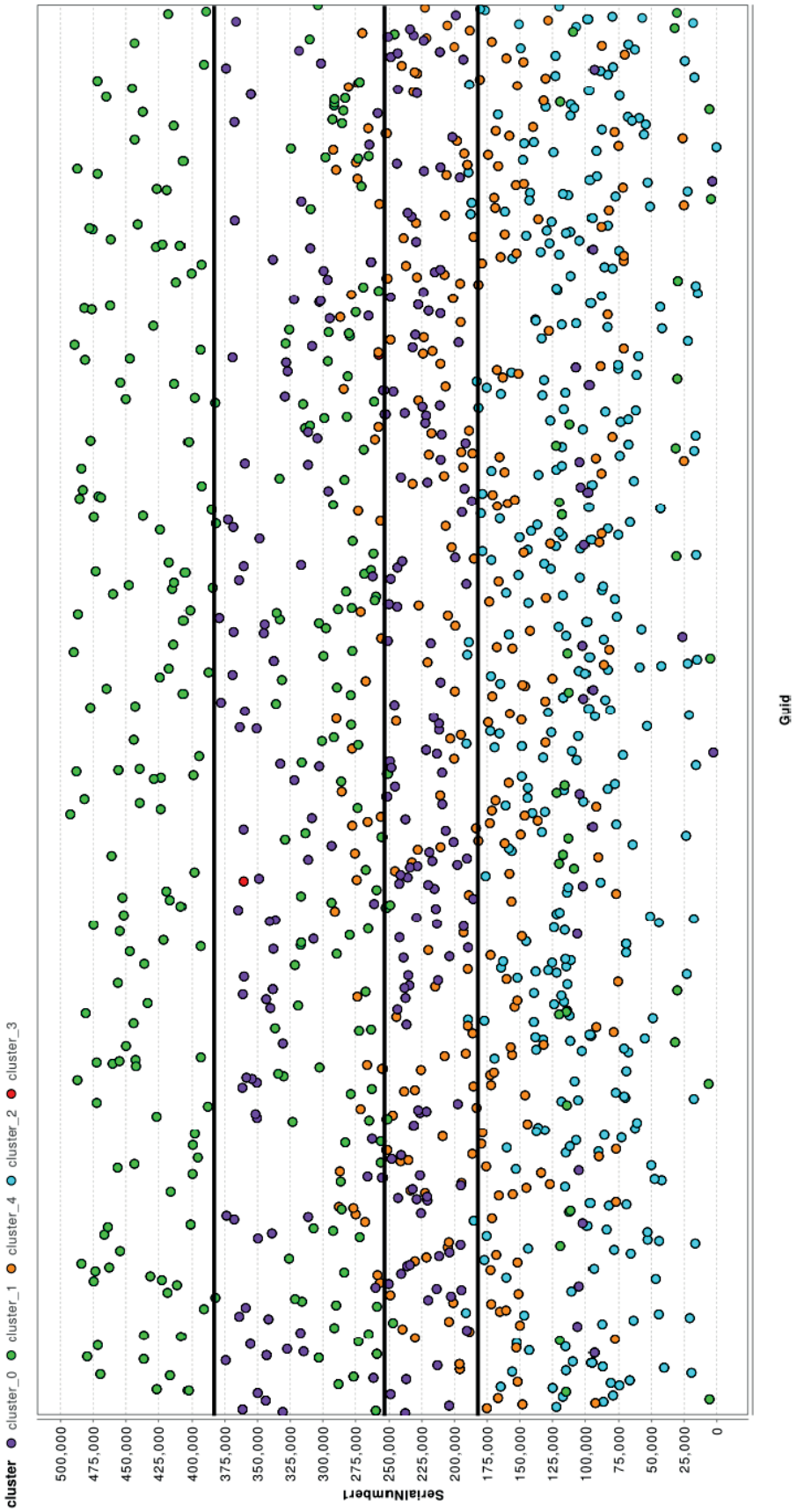



**Abbildung 6.3: Clusteranalyse auf gruppierten Daten nach Messeinheit = Milliampere mit k-means, mixed measures, 150 max runs, 1 000 Optimierungsschritten und k = 5**

verzeichnet, die Abszisse stellt die möglichen Ergebnisse der Fertigung dar. Neben „pass“ und „fail“ ist auf dieser Achse die Merkmalsausprägung „leer“ verzeichnet. Diese Merkmalsausprägung umfasst diejenigen Datensätze, die in den Originaldaten einen anderen Status als „pass“ oder „fail“ aufwiesen. Zu den Ausprägungen zählten fehlende Attributsausprägungen, „null“ oder auch „abort“. In Rücksprache mit der Fachseite wurden diese Werte auf den Status „leer“ gemappt und nicht separat für die erste Durchführung der Wissensentdeckung berücksichtigt.

Da jedoch sowohl der Gesamtdatenbestand als auch die gruppierten Daten ohne Messeinheiten (vgl. Abschnitt 6.2.1.2) jeweils auf einer 100 000-Stichprobe abweichende Clusterergebnisse aufwiesen, wurde die These aufgestellt, dass es spezifische Produkttests zu gibt, die nur für bestimmte Fertigungsteile und Materialien zum Einsatz kommen. Diese These wurde von der Fachseite des Unternehmens jedoch nicht bestätigt, da keine separaten Fertigungsteile innerhalb der Prüflinie getestet werden. Vielmehr wird das Produkt am Ende der Fertigung allen notwendigen Produkttests unterzogen. Ob das abweichende Verhalten im Bezug auf die Gruppierung von Interesse ist, muss im Anschluss an diese Forschungsarbeit von der Fachseite geklärt werden. Bei Interesse bestünde der nächste Schritt in einer fachbezogenen Gruppierung der Datenbestände und einer erneuten Clusterung. Dieser Gruppierungsschritt ist jedoch stark vom Kontextwissen des Unternehmens abhängig, da die Produkttest nach fachlichen Kriterien (z. B. Position in der Prüflinie) gruppiert werden müssen.

An der Abbildung 6.4 ist festzustellen, dass die gewählte Clusteranalyse keine Zuordnung der Datensätze zu Clustern gestattet, die Ähnlichkeiten bezüglich der Fertigungsergebnisse aufweisen. Dies ist nachzuvollziehen, wenn man die Verteilung der Cluster 1 und Cluster 2 über die Schichten und Fertigungsergebnisse verfolgt. Lediglich die Datensätze mit dem Fertigungsergebnis „leer“ werden auch bei steigendem k einem spezifischen Cluster zugeordnet. Des Weiteren ist ersichtlich, dass die fehlerhaften Fertigungsprozesse in keiner der drei Schichten überproportional vertreten sind. Bei mehrfacher Wiederholung des Experiments mit verschiedenen Stichproben sowie unterschiedlicher Stichprobengröße war die Zuordnung der Datensätze mit der Merkmalsausprägung „fail“ zu den einzelnen Teilschichten nahezu gleichverteilt. Für weitere Experimente zu unterschiedlichen Clusterverfahren im Anwendungsfeld 1 wird auf die im Rahmen dieser Forschungsarbeit entstandene Zuarbeit des Fachgebiets ITPL von Li (2017) verwiesen.

Als Resultat von Phase 5 wurde eine Auswahl von verschiedenen Clustermodellen als Ergebnis der Data-Mining-Verfahren festgelegt. Abbildung 6.5 zeigt eine Clusterung, die auf einer 1 000 000-Stichprobe der Subgruppe Produktionslinie 1-3 erlernt wurde. Auf der Abzisse sind die Materialnummern dargestellt (auf numerische Werte gemappt), die Ordinate zeigt die Startzeitpunkte für die einzelnen Fertigungsschritte (auf numerische Werte gemappt, 30-Minuten-Intervalle). Die Cluster zeigen die Zuordnung von bestimmten Materialien zu Fertigungszeiträumen. Am Verlauf der Trennlinien wird deutlich, dass spezifische Materialnummern zu bestimmten Zeitpunkten im Fertigungsprozess genutzt werden. Da sich die Cluster

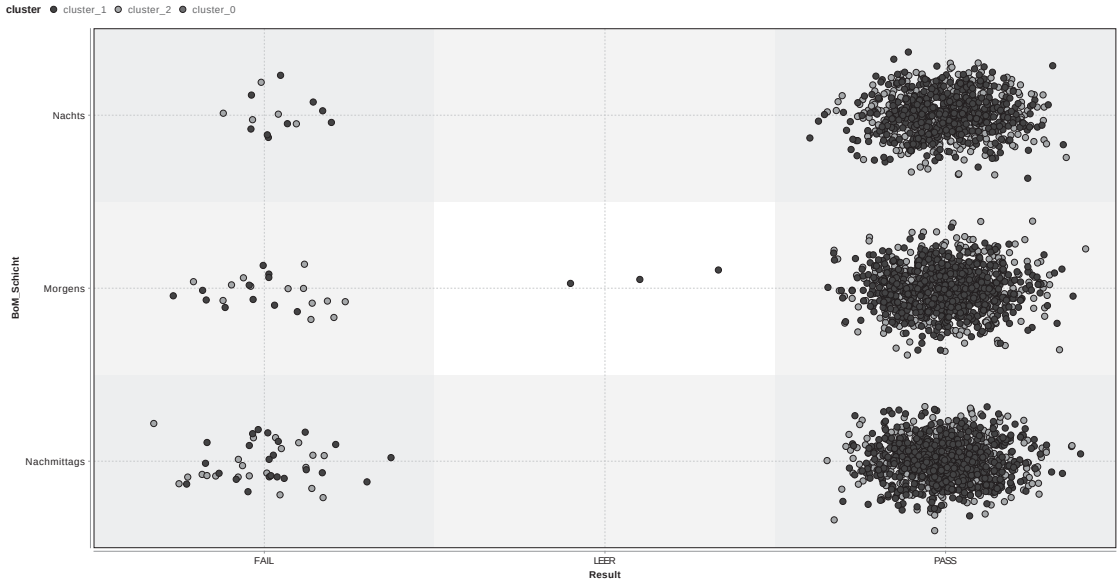



**Abbildung 6.4: Pseudometrisches Streudiagramm einer Clusteranalyse mit k-means, mixed measures, 150 max runs, 1 000 Optimierungsschritten und k = 3**

2 und 3 grafisch nicht voneinander abgrenzen lassen, sind in der Abbildung nur 3 Bereiche zu erkennen. Die Zuordnung von bestimmten Materialien zu Fertigungszeiträumen wurde von der Fachseite validiert, da im Rahmen der standardisierten Fertigung die Materialnutzung an festgelegte Fertigungsschritte und in der Folge auch an spezifische Zeitpunkte gekoppelt ist. Die Anzahl der Cluster wurde bei dem gewählten Verfahren in den Algorithmusparametern festgelegt (k = 4).

Für das zu konzipierende Datencockpit muss im Anschluss an diese Forschungsarbeit festgelegt werden, ob die Parametrierung für das Unternehmen angemessen ist. Abbildung 6.6 zeigt die Materialunterteilung mit der Parametrierung k = 5. Hier ist an der neu entstandenen Trennlinie ersichtlich, dass sich mit angepassten Parametern gegebenenfalls weitere Unterteilungen anbieten (Cluster 1/2 und 3). Da auch eine mögliche Abgrenzung zwischen Cluster 1 und 2 untersucht werden sollte, wurden weitere Iterationen mit variierendem k durchgeführt. Bis k = 7 ergab sich in den Experimenten jedoch keine Abgrenzung der Cluster 1 und 2. Des Weiteren muss diskutiert werden, ob die Fertigungszeitpunkte oder die einzelnen Fertigungsprozesse Berücksichtigung in der Analyse finden sollen. Diese Schritte sind jedoch konzeptioneller Natur und im Unternehmen angesiedelt. Zudem basiert die Konzeption auf den hier gewonnen Ergebnissen und ist somit zeitlich erst im Anschluss an die vorliegende Forschungsarbeit anzusetzen.

In Bezug auf die Aufgabenstellung der Wissensentdeckung in Abschnitt 6.2.1.1 konnte aufgezeigt werden, dass eine Materialgruppierung entlang der Fertigungsprozesse eine vielversprechende Möglichkeit zur Gruppenbildung darstellt und die daraus resultierenden Cluster beispielsweise für Einzelanalysen im Datencockpit Anwendung finden können.

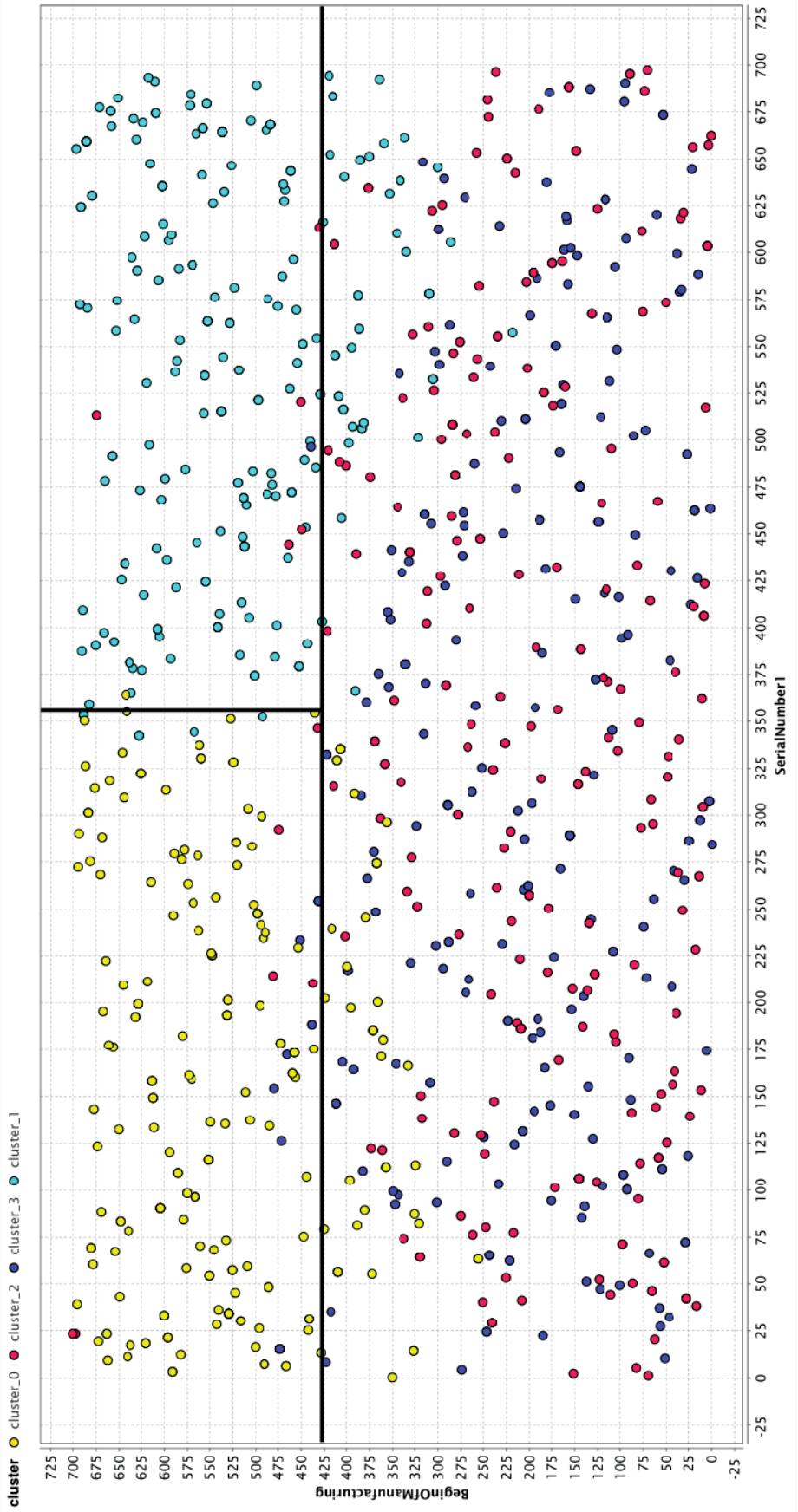



**Abbildung 6.5: Clusteranalyse mit k-means, mixed measures, 150 max runs, 1 000 Optimierungsschritten und k = 4**

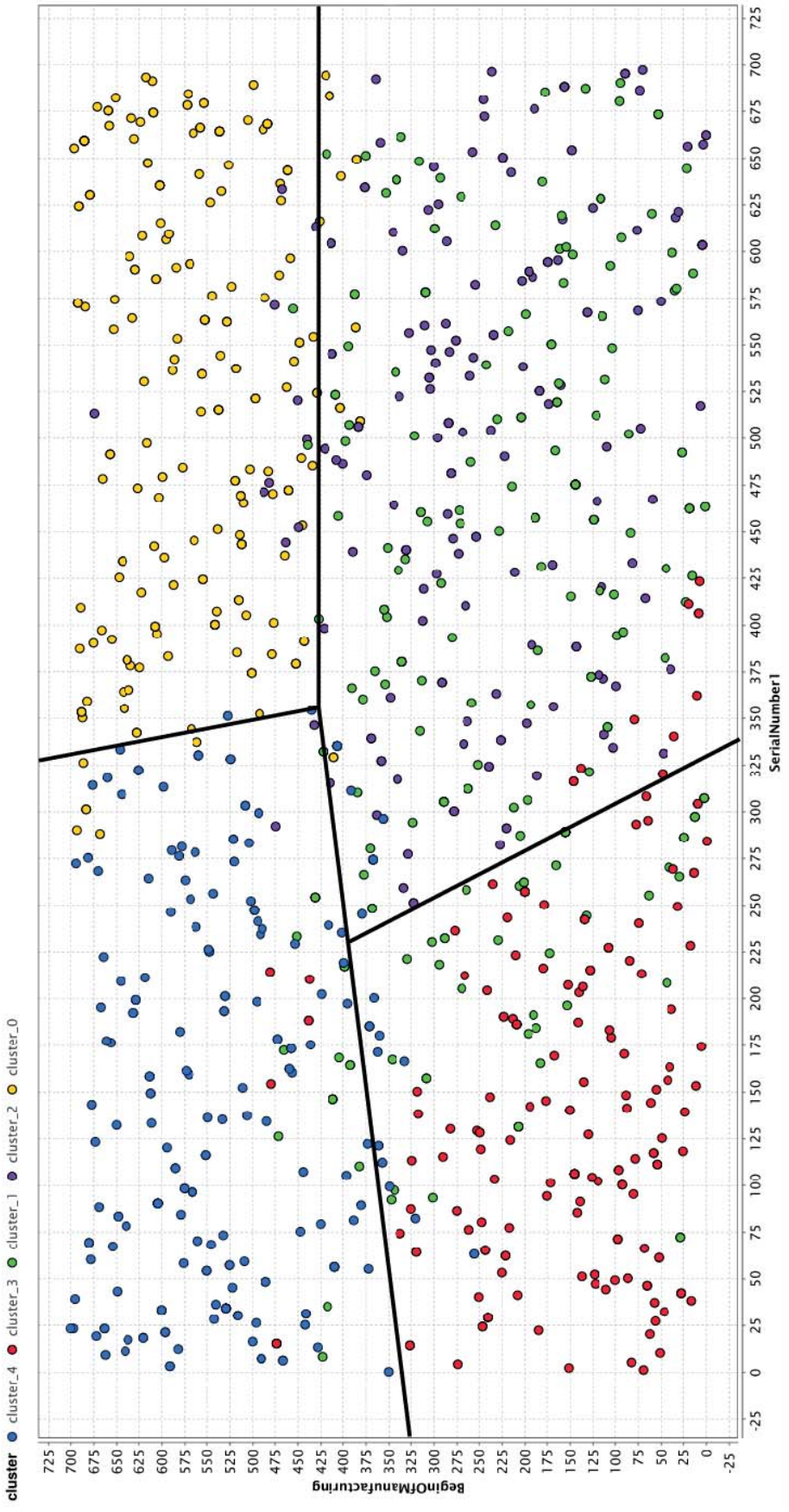



Abbildung 6.6: Clusteranalyse mit k-means, mixed measures, 150 max runs, 1 000 Optimierungsschritten und $k = 5$

#### 6.2.2.2 Assoziationsanalyse

Das Entdecken von Wirkzusammenhängen erfordert den Einsatz geeigneter Verfahren. Um interessante Wirkzusammenhänge zu entdecken, wurde als Data-Mining-Verfahren ein Regellerner in RapidMiner verwendet (Data-Mining-Algorithmus: RapidMiner Associtaion Rule). Um den Regellerner einsetzen zu können, erfolgte eine fachliche und technische Kodierung der Attribute. Tabelle 6.6 zeigt die ausgewählten Attribute der Assoziationsanalyse mit ihrem zugehörigen Datentyp nach Tabelle A.1 und gestattet eine Übersicht über den Anteil der technischen und fachlichen Kodierungen in der durchgeführten Assoziationsanalyse.

**Tabelle 6.6: Kodierung Assoziationslerner**

| Attribut | Datentyp | Beschreibung | Kodierungsart |
|---|---|---|---|
| GUID | String | Seriennummer des Werkstücks | fachlich |
| PlantID | Integer | Produktionsstandort | fachlich |
| LineID | Integer | Produktionslinie | technisch |
| ProductID | String | Produktnummer | fachlich |
| Day | String | Wochentag der Produktion | technisch |
| Shift | String | Arbeitsschicht nach Schichtfolge | fachlich |
| EndOfManufacturing | Datetime | Abschlusszeitpunkt Fertigung | technisch |
| Result | String | Statusmeldung | fachlich |
| NmbOfRepairs | Integer | Anzahl der Reparaturvorgänge | technisch |
| WorkSequence | Integer | Produktionsschritt | fachlich |
| WorkplaceID | Integer | Bearbeitungsstation | fachlich |
| Class | String | Automatische oder manuelle Bearbeitung | technisch |
| Value | Double | Messwert | technisch |
| Unit | String | Messwerteinheit | technisch |

Die fachliche Kodierung war für eine Vielzahl von Attributen unumgänglich, da eine rein technische Kodierung, wie z. B. die Discretize-Funktion (vgl. Abschnitt 4.4.4), keinen Ansatzpunkt für die Vielzahl von Arbeitsstationen oder die unterschiedlichen Produktnummern in den Daten lieferte. Die fachliche Kodierung konnte nur unter Einsatz von Kontextwissen durchgeführt werden, da sich die notwendi-

gen Kodierungsinformationen nicht aus den vorliegenden Daten erschlossen haben (vgl. Abschnitt 4.4.3). Hierzu wurde in einer Reihe von Treffen mit der Unternehmensfachseite festgelegt, wie potentiell notwendige Kodierungen durchgeführt werden können. Um die fachliche Kodierung in der praktischen Umsetzung von der Datenanreicherung aus Abschnitt 6.2.1.2 zu unterscheiden, muss hier der Fokus auf das anzuwendende Data-Mining-Verfahren gelegt werden. Daher wurden in der Folge nur Kodierungen durchgeführt, die für den Einsatz des RapidMiner-Assoziationslerners notwendig waren. Im Anschluss an die fachliche Kodierung erfolgte die technische Kodierung. Hauptaufgabe der technischen Kodierung war die Attributsumwandlung mittels des RapidMiner-Vorverarbeitungsoperators Discretize, um die Anzahl der Attributsausprägungen zu reduzieren (vgl. Abschnitt 4.4.4).

In Abhängigkeit des gewählten Gruppierungskriteriums wurden unterschiedliche technische Kodierungen durchgeführt (vgl. Abschnitt 6.2.1.2). Tabelle 6.7 zeigt beispielhaft die technische Kodierung des Attributs „Messergebnis“ auf einem gruppierten Datenbestand. Das Attribut „Zeilennummer“ fungiert als künstlicher Primärschlüssel, das Attribut „Testergebnis“ gibt an, ob die durchgeführte Prüfung erfolgreich war und die Produktionslinie repräsentiert eine der fünf Linien im Unternehmen. Das Attribut „Teilschicht“ wurde in der Vorverarbeitung unter Einbeziehung des unternehmensinternen Wissens über die Schichtsysteme, dem Kontextwissen, kodiert (vgl. Abschnitt 6.2.1.2). Das Attribut „Wochentag“ wurde ebenfalls im Rahmen der zuvor durchgeführten Vorverarbeitung mittels einer SQL-Routine auf Basis der existierenden Zeitstempel erzeugt und ersetzt in der Assoziationsanalyse die Originalzeitstempel der Unternehmensdaten (vgl. hierzu auch Tabelle 6.4). Die technische Kodierung des Attributs „Messergebnis“ wurde nur im Rahmen der durchgeführten Assoziationsanalyse verwendet und zählt somit zu den verfahrensspezifischen Kodierungen. Durch die Diskretisierungsfunktion in RapidMiner erfolgt eine äquidistante Aufteilung der Messergebnisse (vgl. Funktionen in Abschnitt 4.4.4). Das Ergebnis sind verschiedene Bereiche mit zugehörigen Intervallgrenzen, die von RapidMiner mit dem Schlüsselwort range und aufsteigender Nummerierung kodiert werden. Zur besseren Darstellbarkeit wurden die Intervallgrenzen auf drei Nachkommastellen gerundet. Dies hat zur Folge, dass die angegebenen Intervalle in den Experimenten in diesem Kapitel nicht disjunkt scheinen. Dies wird deutlich, wenn man die Intervallgrenzen der Bereiche range11 in der Zeilennummer 6 und range10 in der Zeilennummer 7 der Tabelle 6.7 vergleicht. Der Wert 173,5 ist Anfangs- bzw. Endpunkt beider Intervalle und folglich in ihrer Schnittmenge enthalten. In dieser und nachfolgenden Darstellungen wurde jedoch auf eine künstliche Anpassung der Intervallgrenzen verzichtet, da dies für die Evaluierung des Vorgehensmodells nicht notwendig ist.

Da gleiche Attributsausprägungen teilweise unterschiedliche Bedeutung hatten, gestaltete sich die technische Kodierung aufwendig. Die Erklärung hierfür liegt in den Attributsausprägungen und der kontextabhängigen Bedeutung identischer Attributsausprägungen. Die bereits durchgeführten Experimente basierten auf normalisierten Datenbanken, sodass identische Attributsausprägungen die gleiche fachliche

**Tabelle 6.7: Gruppierung nach Messeinheit = Milliampere, Discretize by Size und Size = 340**

| Zeilen-nummer | Messergebnis | Wochen-tag | Teilschicht | Test-ergebnis | Produk-tionslinie |
|---|---|---|---|---|---|
| 1 | range4 [2,185 - 10,475] | Montag | nachmittags | PASS | 3 |
| 2 | range8 [10,515 - 114,5] | Montag | nachmittags | PASS | 3 |
| 3 | range6 [10,485 - 10,495] | Montag | nachmittags | PASS | 3 |
| 4 | range9 [114,5 - 139] | Mittwoch | nachmittags | PASS | 3 |
| 5 | range1 [-∞ - 2,005] | Dienstag | morgens | PASS | 2 |
| 6 | range11 [173,5 - ∞] | Dienstag | nachts | PASS | 2 |
| 7 | range10 [139 - 173,5] | Mittwoch | nachmittags | PASS | 2 |
| 8 | range7 [10,495 - 10,515] | Mittwoch | nachmittags | PASS | 3 |
| 9 | range10 [139 - 173,5] | Mittwoch | nachts | PASS | 3 |
| 10 | range10 [139 - 173,5] | Mittwoch | nachts | PASS | 4 |

Bedeutung aufwiesen (vgl. Abschnitt 4.3.2). Dies hatte zur Folge, dass ein technisches Kodierungsverfahren pro Attribut bestimmt werden konnte und die Kodierung sowohl auf dem Gesamtdatenbestand als auch auf den gruppierten Daten erfolgen konnte. Im Anwendungsfeld 1 wiesen jedoch einzelne Gruppen spezifische Attributsausprägungen mit individueller Bedeutung auf. So konnte beispielsweise die Attributsausprägung „1“ je nach Gruppe die Bedeutung eines Boolean-Wertes oder des Messwertes 1 haben. Demzufolge mussten gruppenabhängige Kodierungsverfahren für einzelne Attribute bestimmt werden. Selbst wenn identische Kodierungsverfahren zum Einsatz kamen, war in vielen Fällen eine gruppenspezifische Verfahrensparametrierung notwendig, da die möglichen Attributsausprägungen im Wesentlichen von der entsprechenden Gruppe abhängig waren. Tabelle 6.8 zeigt eine technische Kodierung mit spezifischer Diskretisierungsfunktion und Parametrierung, die von dem bereits vorgestellten Beispiel in Tabelle 6.7 abweicht.

**Tabelle 6.8: Gruppierung nach Messeinheit = Milliohm, Discretize by Frequency und Rangeanzahl = 10**

| Zeilennummer | Messergebnis | Wochentag | Teilschicht | Testergebnis | Produktionslinie |
|---|---|---|---|---|---|
| 1 | range9 [155,35 - 163,8] | Montag | nachmittags | PASS | 3 |
| 2 | range10 [163,8 - ∞] | Mittwoch | nachmittags | PASS | 3 |
| 3 | range2 [119,55 - 125,55] | Mittwoch | nachmittags | PASS | 2 |
| 4 | range8 [149,05 - 155,35] | Dienstag | morgens | PASS | 2 |
| 5 | range2 [119,55 - 125,55] | Mittwoch | nachmittags | PASS | 2 |
| 6 | range9 [155,35 - 163,8] | Mittwoch | nachts | PASS | 3 |
| 7 | range5 [134,45 - 139,05] | Dienstag | morgens | PASS | 3 |
| 8 | range9 [155,35 - 163,8] | Samstag | nachts | PASS | 2 |
| 9 | range9 [155,35 - 163,8] | Mittwoch | nachmittags | PASS | 2 |
| 10 | range6 [139,05 - 144,9] | Mittwoch | nachmittags | PASS | 4 |

Die Erklärung für die gruppenspezifische Kodierung ist, dass im Gegensatz zu den vorherigen Experimenten Tabellen des hier diskutierten Datenbestands nicht normalisiert wurden (vgl. Abschnitt 6.2.1.2). Alle Attribute, auf die eine gruppenspezifische Kodierung angewandt wurde, stammen ursprünglich aus nicht-normalisierten Tabellen der zugrundeliegenden SC-Datenbank. Die SC-Datenbank sowie deren Normalisierungsgrad korrespondieren mit dem Aspekt der Datenqualität, der in den Eingangskriterien der MESC abgefragt wird (vgl. Abschnitt 4.1.2). In der praktischen Anwendung zeigte sich, dass die gestellten Fragen zum Thema Datenqualität in SC-Projekten weiter spezifiziert werden sollten, um eine ausreichende Informationsgrundlage für die Wissensentdeckung zu bilden.

Im Anschluss an die Kodierung erfolgte die Anwendung der zuvor ausgewählten Data-Mining-Verfahren (vgl. Phase 5 in Tabelle 3.3). Bei der praktischen Durchführung der MESC zur Entdeckung von Wirkzusammenhängen gab es verschiedene anwendungsfeldbezogene Herausforderungen. So mussten sinnvolle Gruppierungen für das Entdecken der Regeln bestimmt werden (vgl. Abschnitt 6.2.1.2) und in Abhängigkeit der gewählten Gruppen die Parametrierung der Modelle durchgeführt werden. Für eine spezifische Beschreibung der möglichen Gruppierungen im Anwendungsfeld 1 wird auf die im Rahmen dieser Arbeit entstandene Vorarbeit von Klein (2017) verwiesen. Tabelle 6.9 zeigt einen beispielhaften Ausschnitt der vorverarbeiteten Daten. Im unmittelbaren Vergleich zur gewählten Kodierung für die Clusteranalyse (vgl. Tabelle 6.5) bestätigt sich, dass die Kodierung im Bereich der SC verfahrensabhängig ist. In der Folge wird die Zuordnung der Kodierung zur Phase 4 der MESC als zielführend betrachtet. Für den Einsatz des Regellerners wurden im Minimum 1 700 künstliche Attribute durch die notwendigen Kodierungen erzeugt. Dies verdeutlicht die Komplexität der Kodierung im Umfeld der MESC-Datenbanken und zeigt, dass das Kontextwissen ein entscheidender Faktor bei zielführenden Kodierungen von Attributen ist (vgl. Tabelle 3.1).

**Tabelle 6.9: Beispielhafte Kodierung der Daten für RapidMiner Associtaion Rule**

| **ParamId = 02** | **ParamId = 04** | **LLimit = 0** | **ULimit = 0** | **GUID = C8F** | **Result = false** |
|---|---|---|---|---|---|
| false | true | false | true | false | false |
| false | true | false | true | false | true |
| false | true | false | true | false | true |
| false | true | false | true | false | false |
| false | true | false | true | false | true |

Im Anschluss erfolgte die MESC-Phase 5, in der die Data-Mining-Verfahren angewendet wurden (vgl. Phase 5 in Tabelle 3.3). In dieser Phase wurde die Fachseite intensiv eingebunden, denn aufgrund der Vielzahl von möglichen Modellen mussten diejenigen bestimmt werden, die für das Unternehmen die meiste Aussagekraft besaßen. Da bei den Regellernern die Anzahl der gefundenen Regeln in unmittelbarem Zusammenhang zu der Verfahrensparametrierung steht, wurde im ersten Schritt ein Vorauswahl von gruppierten Datenbeständen und zugehörigen Regeln festgelegt. Die Notwendigkeit der Vorauswahl wird deutlich, wenn man berücksichtigt, dass auf einer zufälligen Stichprobe von 100 000 Entitäten (Parametrierung: 1 600 Attribute, Support $>$ 0,8, Confidence $>$ 0,9) 27 000 Regeln erlernt wurden. Insbesondere den unterschiedlichen Gruppierungsmöglichkeiten der Unternehmensdatenbestände kommt in der Verfahrensgruppe der Regellerner eine besondere Bedeutung zu. Betrachtet man die Fragestellung des Unternehmens aus

Abschnitt 6.2.1, so werden Zusammenhänge zwischen selten auftretenden Fertigungsereignissen gesucht. Dies ist ein Unterschied zu dem klassischen Ansatz der Assoziationsanalyse, in dem häufig vorkommende Regeln von Interesse sind. Für das Anwendungsfeld 1 stellten sich in der Folge zwei grundsätzliche Experimentieransätze dar. Zum einen wurden auch selten vorkommende Attributsausprägungen in die Suchmenge für die Regellerner integriert (RapidMiner Frequent Item Sets mit niedrigem Support) oder alternativ die Berücksichtigung bestimmter Attributsausprägungen über die Parametrierung erzwungen. Dies implizierte, dass auch die selten vorkommenden fehlerhaften Prüfungen in den entdeckten Regeln aufzufinden waren. Einen anderen Experimentieransatz verfolgte die Durchführung der Assoziationsanalyse auf spezifischen Subgruppen. Hier wurden beispielsweise die Transaktionen, deren Attributsausprägungen für die Fertigungsprozesse nicht „pass“ sind, gruppiert und als Basis für die Regellerner verwendet. Tabelle 6.10 zeigt einen Ausschnitt der Regeln, die als Muster am Ende der MESC-Phase 5 der Fachseite zur Abstimmung vorgelegt wurden.

**Tabelle 6.10: Assoziationsregeln mit Support > 0,3, Confidence > 0,7**

| **Prämisse** | **Konklusion** | **Support** | **Confidence** |
|---|---|---|---|
| ParamDescription = PDES0000024 | NmbOfRepairs = range1 [0] | 0,758 | 0,957 |
| Result = PASS | NmbOfRepairs = range1 [0] | 0,801 | 0,960 |
| ResultCode = 2 | NmbOfRepairs = range2 [1-2] | 0,361 | 0,818 |
| Produktionslinie = 4 | NmbOfRepairs = range2 [1-2] | 0,322 | 0,824 |
| Teilschicht = Nachts | NmbOfRepairs = range2 [1-2] | 0,305 | 0,740 |

In den beispielhaften Regeln sind Attributsausprägungen für Prozesskennziffern (ParamDescription), Reperaturanzahlen (NmbOfRepairs), Teilschichten, Produktionslinien und Kennziffern für Fertigungsprozesse (ResultCode) enthalten. Als Grundlage für die Wissensentdeckung diente eine 100 000-Stichprobe, in der nur Transaktionen enthalten waren, deren Fertigungsprozess Fehler aufwies. Auf Ausgangsbasis der entdeckten Regeln konnten nun Hypothesen zu unterschiedlichen Wirkzusammenhängen in der Fertigung aufgestellt werden. In Tabelle 6.10 ist beispielsweise an der vorletzten Regel ersichtlich, dass zwischen der Produktionslinie 4 und einer Nachbesserung innerhalb der Fertigung ein Zusammenhang bestehen könnte. Die letzte Regel dieser Tabelle legt wiederum nah, dass die Nachtschicht diejenige Teilschicht ist, bei der es am häufigsten zu Nachbesserungen in der Fertigung kommt. Dies Ergebnis war von besonderem Interesse, da die Clusteranalyse die Vermutung nahelegte, dass keine der Schichten eine überproportionale Anzahl von fehlerhaften Fertigungsprozessen aufwies (vgl. Abbildung 6.4). Die Erklärung für die in erster Näherung widersprüchlichen Ergebnisse liegt in den verwendeten Attributen und ihrer Kodierung. Für die Assoziationsanalyse wur-

de die Anzahl der Reparaturen berücksichtigt, die Clusteranalyse verwendet nur die Information, ob die Produktion Fehler aufwies oder nicht. Die Kombination der beiden Ergebnisse führt demnach zu der Hypothese, dass auch in den anderen Schichten fehlerhafte Prozesse in vergleichbarer Größenordnung vorliegen. Die Ergebnisse der beiden Data-Mining-Verfahren begründen sich darin, dass es auch Fertigungsprozesse gibt, die eine größere Anzahl an Nachbesserungen benötigen sowie fehlerhafte Fertigungsprozesse, für die keine Nachbesserung in den Transaktionsdaten verzeichnet wurde. Der Domainbereich des Attributs „NmbOfRepairs“ weist als Maximalwert 8 Reparaturen auf. Demzufolge verteilen sich die Prozesse mit höheren Reparaturanzahlen sowie die fehlerhaften Prozesse ohne Nachbesserung auf die anderen beiden Teilschichten. In nachfolgenden Experimenten konnte diese Hypothese validiert werden. Des Weiteren zeigte sich, dass die Wochentage keinen Einfluss auf ein vermehrtes Auftreten von fehlerhaften Fertigungsprozessen hatten. Auch bei sehr geringem Support konnten keine Regeln entdeckt werden, die den Rückschluss auf einen interessanten Zusammenhang zu einem spezifischen Wochentag gestatteten. Die Regeln wurden mit der Fachseite diskutiert und zwei interessante Ausgangspunkte für eine erneute Durchführungen der Assoziationsanalyse festgelegt. Zum einen sollten die zugrundeliegenden Datenbestände nach Produktionslinien gruppiert werden, um zu untersuchen, ob es Regeln gibt, die spezifisch für die Produktionslinie 4 sind. Hiervon erhoffte sich das Unternehmen Anhaltspunkte für das gehäufte Auftreten von fehlerhaften Fertigungsprozessen in der betroffenen Linie. Zum anderen sollte das Auftreten der mehrmaligen Reparaturen innerhalb eines Fertigungsprozesses spezifiziert werden. Hierzu sollten nur diejenigen Transaktionen für die Assoziationsanalyse genutzt werden, deren Attributsausprägung für „NmbOfRepairs“ $> 2$ sind. Diese erneute Wissensentdeckung basiert auf den Ergebnis der zuvor durchgeführten und hier dokumentierten Wissensentdeckung und bedeutet eine erneute Durchführung der MESC. Die in diesem Abschnitt aufgezeigten Überlegungen zur Nutzung des entdeckten Wissens für weitere Data-Mining-Verfahren sind dem Schritt 7.2 der MESC zuzuordnen (vgl. Tabelle 3.3). Die erneute Durchführung würde jedoch eine wiederholte problemspezifische Durchführung der einzelnen MESC-Phasen beinhalten, da sich die Fragestellung der Wissensentdeckung im Anwendungsfeld 1 konkretisiert hat. Dies betrifft beispielsweise die Kodierung des Attributs „NmbOfRepairs“. Da die MESC mit Schritt 7.2 endet, ist die erneute Durchführung des Vorgehensmodells nicht mehr Teil dieser Forschungsarbeit. Für eine Betrachtung von weiterführenden Regeln in Bezug auf die Algorithmenparametrierung und die Gruppenbildung für das vorgestellte Anwendungsfeld wird jedoch auf die im Rahmen dieser Forschungsarbeit entstandenen Zuarbeiten vom Fachgebiet ITPL verwiesen (Klein 2017).

Des Weiteren wurde versucht, Eigenschaften von gruppierten Daten über ein Regelmodell zu repräsentieren. Dies ist prinzipiell auf SC-Daten möglich, allerdings sind die resultierenden Regelmodelle im Anwendungsfeld 1 sehr einfach, da algorithmisch bedingt nur bestimmte Instanzen einer Klasse betrachtet werden und die komplexen fachlichen Zusammenhänge sich demzufolge nicht in den Regeln wider-

spiegeln. So wurde beispielsweise auf einer Stichprobe von 100 000 Datensätzen und mit dem RapidMiner-Rule-Induction-Operator (Algorithmus Repeated Incremental Pruning to Produce Error Reduction, RIPPER) und dem Fertigungsstatus als Klassifikator in den Experimenten ein Modell erlernt (vgl. Regelmodell 6.1), das jedoch für die Wissensentdeckung des Unternehmens nicht zielführend war. Dies lässt sich fachlich darin begründen, dass sich aus dem erlernten Modell keine Wirkzusammenhänge oder gar Handlungsanleitungen für das Unternehmen ableiten lassen, sondern lediglich ein einfacher Zusammenhang zwischen Transaktionsattributen in die Modellbildung eingeflossen ist.

**Regelmodell 6.1: Beispielhaftes Regelmodell auf Anwendungsfeld 1**

```
1 if ResultCode = range1 [0 - 0.500] then PASS (2529 / 14)
2 else FAIL (0 / 412)
3 correct: 2941 out of 2955 training examples
```

Abschließend kann festgestellt werden, dass das Entdecken von Wirkzusammenhängen auf den Unternehmensdaten durch die Anwendung der MESC möglich ist. Aufgrund der Vielzahl von gefundenen Regeln muss das Unternehmen jedoch in der nächsten Phase definieren, welche Muster von Interesse für das Datencockpit sind (vgl. interestingness measures in Abschnitt 2.3.2).

### 6.2.3 Weiterverarbeitung von Ergebnissen und Prozessbewertung

Die Weiterverarbeitung der MESC besteht aus den Schritten Auswahl sowie Darstellungsüberführung der entdeckten Muster (vgl. Phase 6 in Tabelle 3.3). Da die im Anwendungsfeld 1 gewonnen Erkenntnisse zur zukünftigen Konzipierung eines Datencockpits im Unternehmen dienen (vgl. Abschnitt 6.2.1.1), ist es essentiell, die Möglichkeiten nicht frühzeitig zu beschränken. Aus diesem Grund wurde die Auswahl der Muster in der durchgeführten Wissensentdeckung nicht weiter eingeschränkt. Da von der Unternehmensseite noch keine konkreten Anforderungen an die Auswahl der Muster gestellt wurden, wäre eine künstliche Einschränkung für die spätere Konzeption eine potentielle Risikoquelle. Je nach Konzept und Einbindung der Verfahren in eine Unternehmenslösung kann jedoch eine Einschränkung der Muster notwendig werden. Dies wird deutlich, wenn man die Vielzahl der Regeln betrachtet, die mittels Assoziationsanalysen im Anwendungsfeld 1 gefunden werden können (vgl. Abschnitt 6.2.2.2). Die erfolgreiche Darstellungsüberführung der Muster, der zweite Schritt dieser Phase, ist ebenfalls an das Vorliegen konkreter Anforderungen geknüpft. Die bisherigen Anforderungen des Unternehmens an die Wissensentdeckung beinhalten die Konzeption einer IT-Lösung in Form eines Datencockpits. Die Aufgabe der unternehmensspezifischen Musterdarstellung kann demnach innerhalb des zu entwickelnden Datencockpits umgesetzt werden. Aus der Anforderung der Datencockpitkonzeptionierung lässt sich in der Folge ableiten, dass im Anwendungsfeld 1 die Darstellungsüberführung in eine explizite und

von Menschen zu interpretierende Darstellungsform nicht zwingend notwendig ist (vgl. Abschnitt 4.6.2). Da die verwendeten Data-Mining-Modelle und entdeckten Muster aus der Software RapidMiner in XML-Dateien exportiert werden können, könnten verschiedene Interaktionskonzepte zwischen der Software zur Wissensentdeckung (vgl. Tabelle 6.3) und dem unternehmenseigenen Datencockpit realisiert werden. Ein mögliches Szenario wäre eine Schnittstellenimplementierung, mit der die Ergebnisse der Wissensentdeckung in das Datencockpit zu importieren wären. Das Exportieren der Muster in XML-Dateien bietet zudem den Vorteil, dass auch eine der MESC nachgelagerte Überführung in eine explizite Darstellungsform möglich ist. Aufgrund dieser Vorgehensweise müssen für die Konzeptionsphase des Datencockpits keine Einschränkungen getroffen werden und die endgültige Darstellungsüberführung der Muster kann von der Durchführung der MESC entkoppelt werden.

Die Bewertung des Data-Mining-Prozesses bildet die Abschlussphase der MESC (vgl. Phase 7 in Tabelle 3.3). Im Anwendungsfeld 1 bestand die wesentliche Aufgabe im Abschluss der V&V-Maßnahmen, die im folgenden Abschnitt ausführlich diskutiert werden. Zudem erfolgte die Projektdokumentation, die neben dieser Forschungsarbeit aus den Zuarbeiten von Klein (2017) und Li (2017) sowie einer Reihe von internen Datenmodellen, vorverarbeiten Datenbeständen, RapidMiner-Prozessdokumentationen und Mustern besteht. In dem Phasenschritt der Qualitätskontrolle wurde festgestellt, dass die Durchführung der MESC die Basis für ein Datencockpit-Konzept im Unternehmen darstellt. Durch den Einsatz der ausgewählten Data-Mining-Verfahren konnten Muster entdeckt werden, die als konzeptionelle Wissensbausteine für das Prozessmonitoring im Datencockpit fungieren.

Die Datenqualität stellte im Anwendungsfeld 1 ein erhebliches Problem dar, das schon bei der Auswahl und Vorverarbeitung der Daten zu zeitlichen Verzögerungen führte (vgl. z. B. Abschnitt 6.2.2.2). Die Datenqualität wurde in Phase 1 der MESC mit der beteiligten IT und der Fachseite diskutiert und als ausreichend eingestuft. Ein Grund für die Fehlbeurteilung liegt in der Interdisziplinarität der Fragestellung (vgl. auch Abschnitt 6.2.1.1). Die Fachseite und die IT auf Unternehmensseite waren mit den Anforderungen im Bereich KDD nicht vertraut und die KDD-Experten der Forschungsseite kannten zum damaligen Zeitpunkt weder die konkrete Datenlage noch die zugrundeliegenden Spezifikationen. Als weiterer Grund können die spezifischen Attribute angeführt werden, die als Grundlage für die Wissensentdeckung dienten. In den Diskussionen zum Thema Datenqualität stellte sich heraus, dass die Attribute, die nach der Vorverarbeitung als Grundlage für die Data-Mining-Verfahren dienten, in den bisherigen Produktionsanalysen nicht genutzt wurden. Hierzu zählten insbesondere logische Relationen zwischen Zeitstempeln und Eigenschaften der Unternehmensproduktion wie das Schichtmodell. Dieser Aspekt zeigt, dass nicht nur die entdeckten Muster neues Wissen für das Unternehmen beinhalten, sondern darüber hinaus die Anwendung der MESC

auch die Identifikation von neuen Wissensquellen für das Unternehmen unterstützen konnte.

Tabelle 6.11 gestattet einen beispielhaften Überblick über Probleme im Bereich der Datenqualität. Diese Probleme führten zur Wiederholung einzelner MESC-Phasen und der Gesamtaufwand für die Durchführung der Wissensentdeckung erhöhte sich.

**Tabelle 6.11: Beispielhafte Datenqualitätsprobleme im Anwendungsfeld 1**

| Qualitätsproblem | Problemerläuterung | Datenbeispiel |
|---|---|---|
| Fehlerhafter Datensatz | Fehlerhafte Attributsausprägung aufgrund falscher Zuordnungsfunktionen | Falsche Ausprägung in Schichtattribut in Bezug auf die Fertigungszeitstempel der Entitäten |
| | Produktionsfremde Einträge | Dummydatensätze, die angeben, wie Attribute befüllt werden sollen |
| Unvollständiger oder fehlender Datensatz | Erforderliche Attribute sind nicht befüllt | Fehlende Schlüsselattribute |
| | Fehlende Datensätze innerhalb einer Produktion | Teilschritte einer Fertigung fehlen |
| Unbrauchbarer Datenbankeintrag | Unverständliche Beschreibungen | Chinesische Zeichen als Messergebnis |
| | Fehlende Beschreibungen für Teilprozesse | |
| Nicht normalisierte Tabelle | Komplexe Strukturen innerhalb einer Tabelle (z. B. Hierarchie) | Datensatz wird über einen Schlüssel mit einem anderen Datensatz derselben Tabelle verbunden |
| | Attributsausprägungen können von vorherigen Datensätzen abhängig sein | Eine nicht bestandene Messung kann dazu führen, dass alle folgenden Messergebnisse des Werkstücks ebenfalls als nicht bestanden gekennzeichnet werden, obwohl ihre Messergebnisse im Toleranzbereich liegen |

**Tabelle 6.11: Beispielhafte Datenqualitätsprobleme im Anwendungsfeld 1 (Fortsetzung)**

| Qualitätsproblem | Problemerläuterung | Datenbeispiel |
|---|---|---|
| Inhomogene Attributsausprägung | Inhomogene Attributsausprägungen in verschiedenen Tabellen | Sowohl Widerstandsmesswerte als auch Temperaturmesswerte werden über ein Attribut ausgewiesen |
| Nicht spezifikationskonforme Datenbank | Unterschiedliche Datenbankstrukturen in verteilten Systemen | Einige Tabellen existieren nur an bestimmten Standorten oder als identisch deklarierte Tabellen weisen unterschiedliche Attribute auf |
| Redundante Entitäten | Redundante Entitäten in einer Vielzahl von Tabellen | Doppelte Entitäten in Tabellen, die per Tabellenaufbau mehre Unique-Attribute besitzen |

Die Datenqualitätsproblematik ist jedoch nicht spezifisch für die Wissensentdeckung in SC-Datenbanken. Vielmehr finden sich vergleichbare Schwierigkeiten oftmals in Unternehmen, die erst damit beginnen, ihre verfügbaren Unternehmensdatenbestände zu nutzen. In der abschließenden Bewertung des Anwendungsfelds 1 wurde dem Unternehmen empfohlen, die Datenbasis für das Datencockpit basierend auf den Erkenntnissen der MESC-Durchführung einzuschränken. Die Attribute der eingeschränkten Datenbasis sollten nachfolgend grundlegend bereinigt werden und ein kontinuierlicher Qualitätssicherungsprozess der Daten initiiert werden. Der Aspekt der kontinuierlichen Datenqualitätssicherung ist insbesondere dann essentiell, wenn die MESC für wiederkehrende Aufgaben im Bereich der Wissensentdeckung Einsatz finden soll. In weiterführenden Forschungsarbeiten könnte darüber hinaus geprüft werden, ob ein standardisierter Fragenkatalog zur Datenqualität Einsatz finden kann. Der Fragenkatalog könnte ein zielführendes Hilfsmittel zur Beurteilung der Datenqualität für die Wissensentdeckung in SCs sein. Als Gegenargument muss aufgeführt werden, dass die benötigte Eingangsqualität sogar innerhalb des KDD variiert. Zudem kann ein entwickelter Fragebogen mitunter keine Hilfestellung bieten, wenn, wie in dem hier diskutierten Anwendungsfeld, das Wissen zur Beantwortung der Fragen zu Projektbeginn nicht existiert.

### 6.2.4 Verifikation und Validierung der durchgeführten Phasen

In diesem Abschnitt erfolgt die Dokumentation der Evaluierung der entwickelten V&V der MESC (vgl. Argumentation in der Einleitung zu Abschnitt 6.2). Die

Evaluierung umfasst sowohl die Anordnung und die Aufgabenspektren der einzelnen V&V-Phasen (vgl. Tabelle 3.4) als auch den exemplarischen Einsatz von V&V-Techniken in MESC (vgl. Tabelle A.4).

Die V&V wurde zum jeweiligen Ende einer MESC-Phase durchgeführt und fungierte als Phasenabschluss. In Abschnitt 3.6 wurde angeführt, dass die Überprüfung von Ergebnissen auf Phasenebene eine angemessene Granularitätsstufe ist, da die meisten Ergebnisse erst zum Ende einer Phase überprüfbar sind. In der praktischen Durchführung zeigte sich, dass die eingesetzten V&V-Techniken der MESC zudem oftmals sehr zeitintensiv waren und das Wissen der Unternehmensexperten forderten. In der Folge unterstützt der Zeitaufwand die Konzeptionsentscheidung, auf Phasenebene zu arbeiten, denn eine Prüfung von Schrittergebnissen wäre aufgrund der fachseitigen Ressourcen praktisch nicht umsetzbar gewesen.

Die Entscheidung, die Phasenergebnisse sowohl intrinsisch als auch gegen die vorherigen Phasenergebnisse zu prüfen (vgl. Abbildung 3.4), hat sich als zielführend herausgestellt. Beide Prüfphasen konnten im Anwendungsfeld 1 Fehler in der Wissensentdeckung aufzeigen. Als Beispiel für die intrinsische Prüfung soll das Vorgehen in Prüfschritt 4,4 (vgl. Tabelle 3.4), der die Prüfung auf geeignete Auswahl von Data-Mining-Verfahren beinhaltet, erläutert werden. Dieser Schritt wurde ohne Beteiligung der Fachseite durchgeführt, da hierfür insbesondere umfassendes Wissen über mögliche Data-Mining-Verfahren notwendig war und das unternehmensbezogene Kontextwissen keinen Mehrwert für die verwendete V&V-Technik darstellte. Zum Einsatz kam die V&V-Technik „Test von Teilmodellen“ (vgl. Tabelle A.4), die sich im konkreten Fall als die Modellanwendungen auf kleinen Datenstichproben darstellte. Auf diesen Stichproben wurden unterschiedliche Clusterverfahren sowie verschiedene Ähnlichkeitsmaße angewandt und die Ergebnisse verglichen. Die Clusterverfahren, die beispielsweise die Mindestanforderungen an die Performance in der Modellerstellung nicht erfüllen oder keine ausreichend justierbaren Ähnlichkeitsmaße in RapidMiner besitzen, wurden in dieser Prüfung verworfen. Folglich wurden theoretisch ausgewählte aber in der praktischen Umsetzung auf den Datenbestand ungeeignete Verfahren ausgeschlossen.

Die Prüfung gegen die Phasenergebnisse der vorherigen Phase kann am Prüfschritt 2,1 (vgl. Tabelle 3.4) nachvollzogen werden. Dieser soll validieren, ob die ausgewählten Datenquellen und ihre Datenbestände für das Erreichen der zuvor festgelegten Zielbedingung geeignet sind. Für die Unternehmensdaten wurde eine Mischung aus Dokumentenprüfung und Inspektion (vgl. Tabelle A.4) festgelegt. Hierzu wurden zuerst die extrahierten Daten aus technischer Sicht überprüft, um beispielsweise sicherzustellen, dass die exportierten Datenbestände gültige Dateiformate aufweisen oder mandatorische Attribute wie IDs Attributsausprägungen besitzen. Im nächsten Schritt wurde ein Treffen mit der Fachseite durchgeführt, um die extrahierten Tabellen gegen die vorhandene Tabellenspezifikation zu prüfen. Hierbei konnte festgestellt werden, dass die Scandaten (vgl. Abbildung 6.1) zu wenige Entitäten beinhalteten. Die Tabelle beinhaltete nur ca. 11 000 verwendete Materialien. Demzufolge konnte eine Vielzahl der tatsächlich verwendeten

Materialien aus der Fertigungstabelle nicht zugeordnet werden. Die Daten wären somit für die ausgewählten Ziele der Wissensentdeckung nicht verwertbar gewesen, da sowohl das Entdecken von Wirkzusammenhängen wie auch die Clusteranalyse bei unvollständigen Daten in dieser Größenordnung nicht durchführbar sind. Da die Attribute der Tabelle Scandaten mit den Spezifikationen übereinstimmten, war der Fehler bei der ersten technischen Prüfung nicht entdeckt worden. In einer anschließenden Fehlersuche mit dem entsprechenden IT-Zulieferer konnte die Fehlerursache identifiziert und behoben werden. Es handelte sich um eine Beschränkung des Ausgabevolumens in den Skripten der ETL-Prozesse, die durch vorherige Exportaufgaben bedingt waren. In der Folge fehlten ca. 80 % der Entitäten. Nach der Fehlerbehebung erfolgte eine erneute Prüfung der Phasenergebnisse (vgl. Abschnitt 3.6).

An dem Prüfschritt 2,1 kann ebenfalls demonstriert werden, dass das Konzept des Dreiecksmodells mit der Unterscheidung zwischen intrinsischer Prüfung und Prüfung gegen die Vorphasen zielführend in MESC ist. Dies begründet sich darin, dass die durchgeführte intrinsische Prüfung in Schritt 2,2 nicht den expliziten Fokus der Zielerfüllungsüberprüfung hat und folglich diese Fehlerart in der Prüfung nicht zu identifizieren war. Dennoch konnten auch in in der intrinsischen Prüfung 2,2 Fehler in den Datenbeständen aufgedeckt werden. Es war ersichtlich, dass für zwei der IT-Systeme nicht nur die zuvor identifizierten Tabellen ausgewählt wurden, sondern vielmehr die vollständige Datenbank exportiert wurde. Dies hätte in den nachfolgenden Phasen zu einer nicht handhabbaren Verlängerung der Vorverarbeitungszeiten geführt. Als Folge der vollständigen Exports enthielten auch die ausgewählten Tabellen Transaktionen aus Zeiträumen, die für die Wissensentdeckung nicht festgelegt wurden. Dies hätte in der Anwendung der Data-Mining-Verfahren zu fehlerhaften Mustern geführt, da sich die Attributsausprägungen der relevanten Tabellen über die Zeit verändert haben. Zu Beginn der Produktion wurden beispielsweise bestimmte Attribute nicht protokolliert, die erst mit der Weiterentwicklung der Produktion hinzugekommen sind. Die Fehlerbehebung bestand in der Auswahl der für die Wissensentdeckung identifizierten Tabellen und Entitäten mit Zeitstempeln aus dem relevanten Zeitraum mittels SQL-Skripten. Dieses Beispiel verdeutlicht auch, warum eine Granularitätsstufe für die Prüfungen, die mehrere Phasen kumuliert betrachtet, in MESC keine Konzeptionsoption war. Dies begründet sich darin, dass in der Wissensentdeckung in SC-Datenbanken eine späte Fehleridentifikation zur Erschwerung der Fehlersuche und zu höheren Kosten in der Fehlerbehebung führt. Dies ist vergleichbar mit einer Fehleridentifikation in späten Phasen der Softwareentwicklung oder der Simulation. Im diskutierten Anwendungsfeld hätte der identifizierte Fehler aus Schritt 2,1 zu einer unrealistischen Datengruppe als Basis für die Data-Mining-Verfahren geführt. Als Konsequenz wäre bei der Mustervalidierung ein Rücksprung zur MESC-Phase 2 notwendig geworden und die Phasen hätten erneut durchgeführt werden müssen. In Anbetracht der diskutierten Zeitdauer für die Durchführung hätte diese Fehleridentifikation in Phase 5 zu einer Gesamtverlängerung des Projektes um ca. 80 % geführt (vgl. Abbildung 3.3).

Neben den hier diskutierten Fallbeispielen aus der praktischen V&V der MESC kamen weitere Techniken zum Einsatz. Die Tabelle 6.12 zeigt eine Auswahl der eingesetzten V&V-Techniken und gibt eine Übersicht über wesentliche Aktivitäten in einzelnen Schritten. Der Fokus dieser Tabelle liegt auf datenspezifischen Prüfaktivitäten im Rahmen der MESC. Die aufgelisteten Aktivitäten zeigen konkrete Realisierungen der allgemeinen V&V-Techniken in den einzelnen Prüfschritten, die spezifisch für den Bereich der Wissensentdeckung in SC-Datenbanken sind. Allgemeine Aktivitäten wie z.B. das Prüfen gegen die Spezifikationsdokumente, die ebenfalls in MESC enthalten sind, stehen nicht im Vordergrund der Dokumentation. Die Beschreibung der einzelnen Prüfschritte kann der Tabelle 3.4 entnommen werden. Die Vorgehensweisen der eingesetzten V&V-Techniken können aus der Literatur in Tabelle A.4 erschlossen werden. Für eine systematisierte Gegenüberstellung von V&V-Techniken sowie die detaillierte Beschreibung der Validierung von Wirkzusammenhängen im Anwendungsfeld 1 wird auf die im Rahmen dieser Forschungsarbeit am ITPL entstandene Arbeit von Klein (2017) hingewiesen.

**Tabelle 6.12: Einsatz von V&V-Techniken im Anwendungsfeld 1**

| **Phase** | **Prüf-schritte** | **Exemplarisch eingesetzte V&V-Technik** | **Wesentliche Aktivitäten** |
|---|---|---|---|
| Aufgaben-definition | 1,1 | Validierung im Dialog | Vielzahl von Treffen zwischen verschiedenen Stakeholdern dienten zur Überprüfung der schriftlich dokumentierten Ziele |
| Auswahl der relevanten Datenbestände | 2,2 | Begutachtung | Ausgewählte Datenbestände wurden sowohl in fachseitigen Treffen als auch gegen die vorliegenden Dokumente geprüft |
| | 2,1 | Dokumenten-prüfung und Inspektion | Dokumentenprüfung erfolgte gegen den spezifizierten Exportzustand der Tabellen und wurde durch Inspektionen ergänzt |
| Datenaufbereitung | 3,3 | Strukturiertes Durchgehen und Dokumenten-prüfung | Prüfung der einzelnen transformierten Attribute im direkten Dialog mit den Beteiligten des Forschungsbereichs. Operationen, die Kontextwissen benötigen, wurden mittels der zur Verfügung gestellten fachseitigen Spezifikationen geprüft |

**Tabelle 6.12: Einsatz von V&V-Techniken im Anwendungsfeld 1 (Fortsetzung)**

| Phase | Prüfschritte | Exemplarisch eingesetzte V&V-Technik | Wesentliche Aktivitäten |
|---|---|---|---|
| | 3,2 | Schreibtischtest | Menge der transformierten Attribute wurde in Äquivalenzklassen aufgeteilt (z. B. Aggregation auf Zeitstempeln) und aus den Äquivalenzklassen Stichproben für die manuelle Überprüfung der Transformationsoperationen gebildet |
| | 3,1 | Audit | Fachseitige Audits in Bezug auf aufbereitete Daten (z. B. Übergabe von Datenmodellen) |
| Vorbereitung des Data-Mining-Verfahrens | 4,4 | Test von Teilmodellen | Test welche Data-Mining-Verfahren ausgeschlossen wurden, nur beteiligte Forscher ohne Fachseite |
| | 4,3 | Test von Teilmodellen | Test der Forschungsbeteiligten ohne Fachseitenbeteiligung mit dem Fokus, anhand der Anzahl sowie Qualität der gefundenen Muster auf gruppierten Datenbeständen, die entsprechenden Vorbereitungsverfahren wie z. B. Transformationen zu prüfen |
| | 4,2 | Test von Teilmodellen | Wurde zusammen mit Prüfschritt 4,3 getestet, da eine Trennung beider Bereiche für die vorliegenden Datenbestände nur schwer zu realisieren war |
| | 4.1 | Audit | Besprechung erster Ergebnisstrukturen (z. B. einfache Regeln auf gruppierten Daten) aus den Tests von 4,2 und 4,3 |

**Tabelle 6.12: Einsatz von V&V-Techniken im Anwendungsfeld 1 (Fortsetzung)**

| Phase | Prüfschritte | Exemplarisch eingesetzte V&V-Technik | Wesentliche Aktivitäten |
|---|---|---|---|
| Anwendung der Data-Mining-Verfahren | 5,5 | Kreuzvalidierung | Güte und Generalisierungsfähigkeit verschiedener Experimentkonzepte und Data-Mining-Verfahren wurden mittels Kreuzvalidierung geprüft |
| | 5,4 | Review | Ohne Beteiligung der Fachseite, reine Prüfung der technischen und fachlichen Konfiguration der angewendeten Verfahren in RapidMiner |
| | 5,3 | Kreuzvalidierung | Implizit bei Prüfschritt 5,5 vertestet |
| | 5,2 | Validierung im Dialog | Mehrere Sitzungen mit Fachseitenbeteiligung, in denen Muster und die zugrundeliegenden, selektierten Daten mittels Kontextwissens validiert wurden |
| | 5,1 | Validierung im Dialog und Dokumentenprüfung | Ergebnisdokumentation von Unternehmen wurde geprüft sowie mögliche Interpretation der Ergebnisse mit der Fachseite diskutiert |
| Weiterverarbeitung der Data-Mining-Ergebnisse | 6,6 | - | Entfiel, da die Muster nicht weiterverarbeitet wurden (vgl. Abschnitt 6.2.3) |
| | 6,5 | Visualisierung und Review | Die Ergebnisse wurden visuell aufbereitet und mit der Fachseite diskutiert |
| | 6,4 | Visualisierung und Review | Prüfung erfolgte implizit in Schritt 5,4, Trennung war im Anwendungsfeld 1 nicht möglich |

**Tabelle 6.12: Einsatz von V&V-Techniken im Anwendungsfeld 1 (Fortsetzung)**

| Phase | Prüfschritte | Exemplarisch eingesetzte V&V-Technik | Wesentliche Aktivitäten |
|---|---|---|---|
| | 6,3 | Validierung im Dialog | Validierung, ob die Muster für das Unternehmen zu interpretieren sind, erfolgte mit der Fachseite im direkten Dialog |
| | 6,2 | Validierung im Dialog | Prüfung erfolgte implizit in Schritt 6,3, Trennung war im Anwendungsfeld 1 nicht möglich |
| | 6,1 | Validierung im Dialog | Validierung, ob die Muster für das Unternehmen das gesuchte Wissen darstellen, erfolgte mit der Fachseite im direkten Dialog |
| Bewertung des Data-Mining-Prozesses | 7,7 | Dokumentenprüfung | Prüfung der erstellten Dokumente durch Unternehmen und Forschungsbeteiligte |
| | 7,5 - 7,1 | Dokumentenprüfung | Aufgrund der Zeitersparnis wurden diese Schritte in 7,7 als Prüfinstanz integriert |

## 6.3 Anwendungsfeld 2: Datengenerierung mittels Plant Simulation

In diesem Anwendungsfeld wird ein DES-Modell zur Generierung von Transaktionsdaten diskutiert und aufgezeigt, wie aus den generierten Daten Wissen gewonnen werden kann. Der Schwerpunkt der Betrachtung liegt auf der Prozessabfolge von Simulation, Wissensentdeckung und simulationsunterstützter Validierung. Daher wird zuerst der Aufbau des Simulationsmodells und die Struktur der generierten Daten beschrieben. Im Anschluss werden ausgewählte Experimente für die Datengenerierung vorgestellt und die Resultate dargelegt. Hierbei soll insbesondere exemplarisch dokumentiert werden, welche Vorverarbeitungsschritte im Kontext der möglichen Datenaggregationsstufen durchzuführen sind (vgl. Tabelle 5.2). Im zweiten Abschnitt wird die Anwendbarkeit der simulationsunterstützten Validierung (vgl. Abschnitt 5.2.1) innerhalb der MESC mittels ausgewählter Ex-

perimente belegt und demonstriert, wie die grundlegenden Konzeptionsschritte in einer praktischen Anwendung umgesetzt werden können.

### 6.3.1 Aufbau des Simulationsmodells

Die erforderliche Modellierung und Simulation wurde mit dem Werkzeug Plant Simulation 12 der Siemens AG (vgl. Abschnitt 2.4) durchgeführt. Das Modell wurde am Fachgebiet ITPL im Rahmen dieser Dissertation erstellt. Die grundlegenden Konzepte und die technische Umsetzung wurden unter Zuarbeit von Arndt (2014) und Baydar (2016) entwickelt.

Abbildung 6.7 zeigt das Grundmodell in Plant Simulation. Das Simulationsmodell simuliert die operative Ebene einer SC und besteht aus Lieferanten, Transportmitteln (Flugzeug, LKW, Schiff), Umschlagspunkten bzw. Zwischenlager und einem Endabnehmer. Die Bausteine Lieferant 1-6 sind die Quellen der beweglichen Elemente (BE). In diesen werden die Produkte aus der Produktliste in zufälliger Reihenfolge erstellt. Die Produktliste beinhaltet in der Grundkonfiguration fünf verschiedene Produkte mit einer Mindest- und Höchstbestellmenge. Da nicht jeder Lieferant jedes Produkt liefert, gibt es Einschränkungen in der Produktzuordnung (z. B. Lieferant1 kann nur Produkte 1, 3 und 5 liefern). Im Methodenobjekt „Attributerzeugung“ werden den Produkten zufällige Volumina und Gewichte zugewiesen. Abhängig von dem Standort des Lieferanten, der Liefermenge, dem Liefervolumen und dem Gesamtgewicht der Lieferung wird für die BEs eine Lieferart bestimmt. Entsprechend der zugewiesenen Lieferart werden die erstellten Produkte an die Einzelstationen „Flugzeug1-3“, „LKW1-4“ oder „Schiff1-3“ weitergegeben. Die unterschiedlichen Transportmittel Flugzeug, LKW und Schiff haben spezifische Transportzeiten.

Die Stationen Flugzeug und Schiff liefern die Produkte an die Lager „UmschlagspunktFlugzeug“ bzw. „UmschlagspunktSchiff“. Dort haben sie eine Mindestverweildauer von zwei Stunden bevor sie, falls möglich, zu „Transport_Final“ oder aber zu „Zwischenlager1“ bzw. „Zwischenlager2“ transportiert werden. Die Stationen LKW liefern, falls möglich, direkt zu „Transport_Final“ oder aber zu „Zwischenlager1“ bzw. „Zwischenlager2“. Im „Zwischenlager1“ und „Zwischenlager2“ haben die Produkte in der Grundkonfiguration eine Mindestverweildauer von einem Tag. Von dort werden die Produkte, falls möglich, zu „Transport_Final“ oder aber zu „Transport_Ab_Zwischenlager“ geliefert. Bevor die Produkte letztlich bei der Senke „Endabnehmer“ ankommen, wird zu der Lieferzeit noch eine stochastische Verspätung hinzugerechnet. Die angekommenen BEs werden in einer Liste dokumentiert. In dieser Liste sind Informationen über die Produktart, die Lieferart, die Menge, das Volumen, aber auch die Zeiten gespeichert, an denen die BEs beispielsweise im Zwischenlager eingegangen und ausgegangen sind (Trace-Daten). Diese Liste bildet die Grundlage der generierten Datenbestände.

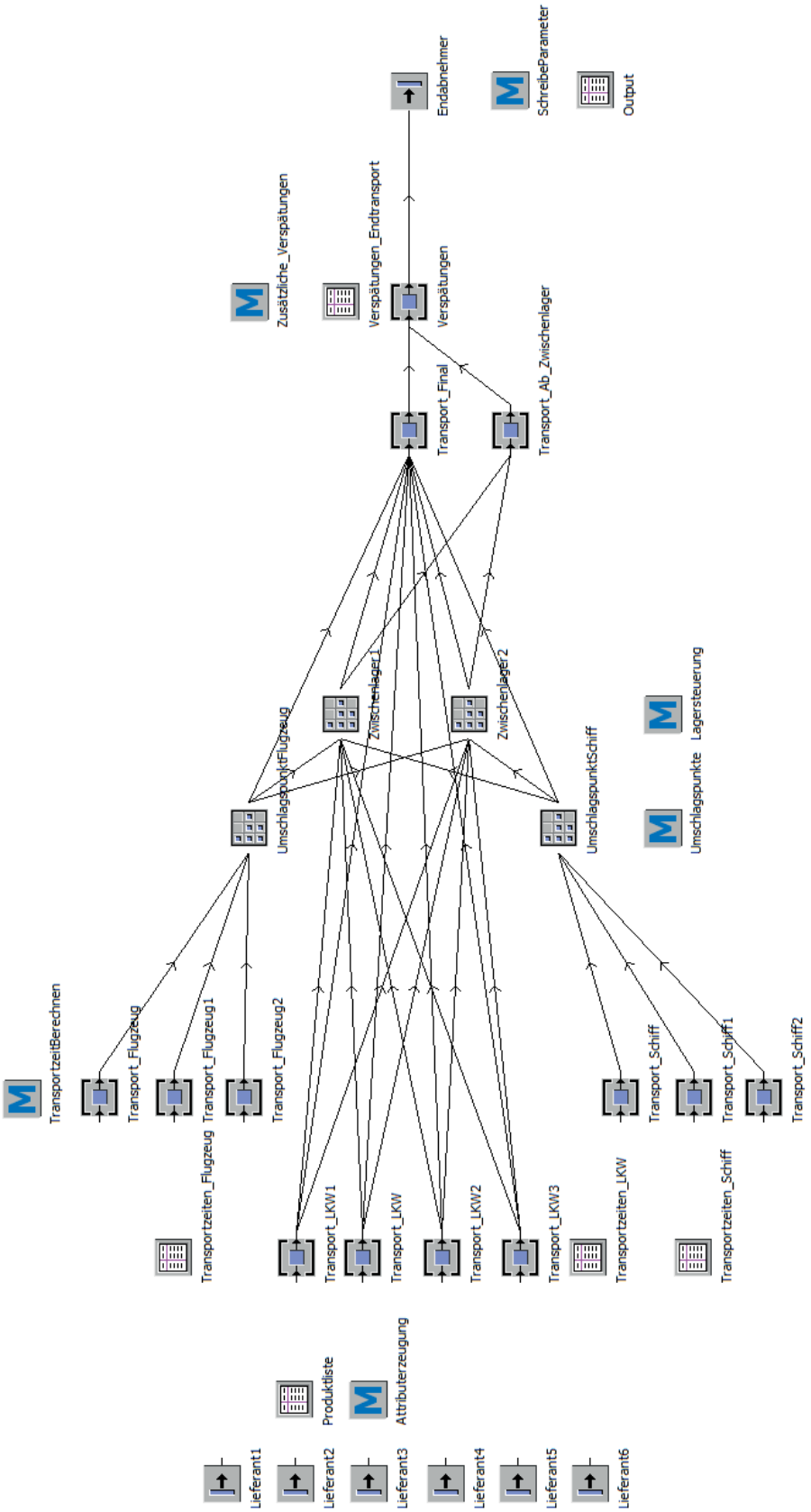



**Abbildung 6.7: Plant Simulation Grundmodell der SC**

Die dargelegte Beschreibung des Modellaufbaus fungiert als ein Konzeptmodell. Um die Datensicht zu berücksichtigen, wird das Konzeptmodell für die MESC-Analysen mit Schwerpunkt Data Farming und simulationsunterstützter Validierung um ein weiteres Beschreibungsmittel ergänzt. Da der Schwerpunkt des hier vorgestellten Simulationsmodells die Modellausgabe in Form der generierten Daten ist, wird in Tabelle 6.13 das in dieser Arbeit entwickelte Datenblatt des generierten Datenbestands als ergänzendes Beschreibungsmittel aufgeführt.

Jeder Simulationslauf generiert einen Datenbestand, der je nach Simulationsparametrierung und Simulationszeitrahmen eine gewisse Anzahl an Datensätzen generiert. Diese Datensätze werden in RapidMiner importiert (vgl. Abschnitt 2.4) und dort mit verschiedenen Vorverarbeitungsmethoden und Data-Mining-Verfahren bearbeitet und analysiert. Abbildung 6.8 zeigt den technischen Grundaufbau, der als Basis für alle weiterführenden Experimente dient. Die Datengenerierung erfolgt in Plant Simulation durch die Ausführung des SC-Modells. Der Datenexport kann im CSV-Format (vgl. Abschnitt 4.3.1) oder direkt in eine relationale Datenbank erfolgen. Die relationalen Datenbanken werden in der weiteren Diskussion nicht näher differenziert, da im Rahmen der Arbeit sowohl MySQL wie auch Microsoft SQL-Varianten zum Einsatz gekommen sind. Die CSV-Datei kann im Anschluss an den Simulationslauf bei Bedarf in eine relationale Datenbank importiert werden, bzw. die Tabellen einer relationalen Datenbank können in eine CSV-Datei exportiert werden. In RapidMiner erfolgt dann die Datenweiterverarbeitung und das Data Mining. Hierbei kann RapidMiner sowohl die relationale Datenbank als auch die CSV-Datei als Datenquelle verwenden.

**Tabelle 6.13: Beschreibung des generierten Datenbestands**

| **Kriterien** | **Beschreibung** |
|---|---|
| **Fachliche Kriterien** | |
| Branche | • Zulieferer Automobilbranche |
| Fachliche Beschreibung | • Lieferdaten von 6 Lieferanten an 20 Standorten<br>• Zeitraum vom 01.01.2015 bis zum 01.01.2025<br>• Zwischenlager und Umschlagsplätze für 10 Transportmittel<br>• Systemgrenzen von Warenausgang Lieferant bis Wareneingang Produzent |
| Exemplarische Attributsbeschreibung | • Bestellnummer: ID-Nummer der Bestellung<br>• ProductCode: ID-Nummer des bestellten Produkts |

**Tabelle 6.13: Beschreibung des generierten Datenbestands (Fortsetzung)**

| Kriterien | Beschreibung |
|---|---|
| | • Lieferweg: Beschreibung des Transportmittels<br>• Eingelagert: Datum und Zeitpunkt der Einlagerung in eines der Zwischenlager<br>• Gewicht: Gesamtgewicht einer Bestellung<br>• Volumen: Gesamtvolumen einer Bestellung |
| **Technische Kriterien** | |
| Quelle | • Plant Simulation SC-Modell |
| Umfang | • Variable Simulationsdauer mit 700 Datensätzen pro Jahr<br>• 15 Attribute in einer Tabelle |
| Technische Attributsbeschreibung | Datentypen nach Tabelle A.1:<br>• ProduktID: Zeichenkette<br>• Zielort: Ganzzahl<br>• Volumen: Ganzzahl<br>• Transportweg: Zeichenkette<br>• Ankunft: Datetime |
| **Untersuchungskriterien** | |
| Analysenotizen | • Bestellungen bestehen nur aus einer Produktgruppe<br>• n:m Beziehung zwischen Lieferanten und Produkten<br>• Wechsel des Transportmittels nur über Umschlagplatz möglich |
| Primärfrage | • Ist es möglich, SC-Konstellationen für Verspätungen in den generierten Daten mittels Data-Mining-Verfahren zu entdecken? |

Die beiden unterschiedlichen Exportoptionen wurden realisiert, um die Vorverarbeitung der generierten Daten (vgl. Aggregationsstufen in Abschnitt 5.1.2.1) sowohl direkt auf der Datenbank mittels geeigneter Skriptsprachen als auch in RapidMiner durchzuführen. Tabelle 6.14 ordnet die eingesetzten Werkzeuge den Phasen der MESC zu. Hierbei ist die Aufgabendefinition in Phase 1 ein manueller Prozess ohne Nutzung von Plant Simulation oder RapidMiner. Die Phase 2 beinhaltet die Datenbeschaffung und Datenauswahl und bedient sich einem Plant-Simulation-Modell zur Datengenerierung (vgl. Abbildung 6.7). Hierbei erstreckt

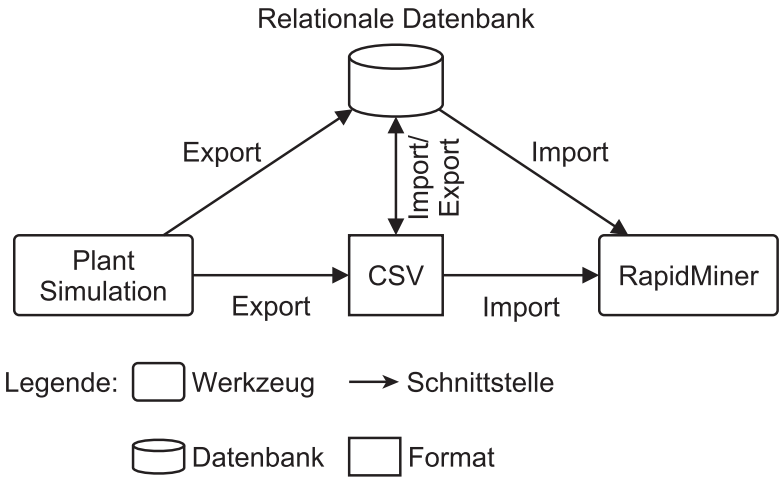



**Abbildung 6.8: Technischer Grundaufbau der Verbindung von Simulation und Wissensentdeckung**

sich die Phase bis zu der Ebene der relationalen Datenbank bzw. der CSV-Datei, da die relevanten Daten für die Wissensentdeckung unter Umständen erst auf dieser Ebene extrahiert werden. Die Phasen 3 und 4, die die Datenvorverarbeitung beinhalten, umfassen sowohl die Datenbank bzw. Datei als auch RapidMiner, da die Vorverarbeitung in beiden Bereichen durchgeführt werden kann. Die Anwendung der Data-Mining-Verfahren, Phase 5, wiederum erfolgt exklusiv unter Zuhilfenahme von RapidMiner. Die Weiterverarbeitung der Data-Mining-Ergebnisse in Phase 6 kann innerhalb von RapidMiner erfolgen (z. B. Darstellungstransformation) aber auch außerhalb durch die Anwendung geeigneter externer Werkzeuge. Im Rahmen der durchgeführten Experimente wurden die Muster beispielsweise zurück in die relationale Datenbank exportiert, um auf der Datenbank die Auswahl der einzelnen, für das SCM interessanten Muster durchzuführen (vgl. Interestingness in Abschnitt 4.6.1). Die Bewertung des Data-Mining-Prozesses in Phase 7 erfolgt abschließend als manueller Prozess ohne Werkzeugunterstützung und wurde daher außerhalb von RapidMiner verortet.

Die Evaluierung der einzelnen MESC-Phasen kann in den Experimenten unter Abschnitt 6.2 nachvollzogen werden und wird demzufolge in diesem Abschnitt nicht erneut aufgegriffen. Der für die in diesem Abschnitt diskutierten Experimente relevante Teil demonstriert Experimente zur Mustervalidierung und stellt die Konzepte der simulationsunterstützten Validierung aus Abschnitt 5.2 im praktischen Einsatz vor. Als Data-Mining-Verfahren wurde ein Regellerner angewandt, um Wirkzusammenhänge in den Transaktionsdaten aufzuzeigen. Hierzu wurden in den Experimenten mittels FPGrowth aus den generierten Daten Frequent Item Sets erstellt und auf diese ein Regellerner (RapidMiner Association Rule) angewandt (vgl. FPGrowth in Tabelle 2.9). In der in diesem Abschnitt durchgeführten Validierung wurde der Begriff Wirkzusammenhänge verwendet. Dieser Begriff impliziert eine manuelle Validierung z. B. durch Experten (vgl. hierzu auch Abschnitt 2.2.3.2). Sofern die manuellen Komponenten für den Experimentablauf und die Ergebnisbe-

**Tabelle 6.14: Einsatz von Werkzeugen in MESC**

| Phase | Plant Simulation | Datenbank | CSV | RapidMiner |
|---|---|---|---|---|
| 1 | ○ | ○ | ○ | ○ |
| 2 | ● | ● | ● | ○ |
| 3 | ○ | ● | ● | ● |
| 4 | ○ | ● | ● | ● |
| 5 | ○ | ○ | ○ | ● |
| 6 | ○ | ○ | ○ | ● |
| 7 | ○ | ○ | ○ | ○ |

**Legende:**
● Einsatz möglich ○ kein Einsatz

wertung irrelevant waren, sind diese nicht in den Experimentdokumentation dieses Abschnitts aufgeführt. Zusätzliche Experimente mit verschiedenen Mustern auf generierten Daten werden darüber hinaus in der im Rahmen dieser Forschungsarbeit entstandenen Arbeit zu Clusterverfahren von Su (2016) aufgezeigt.

### 6.3.2 Statistische Versuchsplanung in der Datengenerierung

In diesem Abschnitt wird demonstriert, wie die Transaktionsdatengenererierung in Plant Simulation mit dem Simulationsmodell durchgeführt wurde. Für die Motivation zur Datengenerierung mittels Data Farming wird auf Kapitel 5 verwiesen. Ein Teil der hier vorgestellten Experimente beruht auf Ergebnissen aus Vorarbeiten im Bereich Data Farming, die am ITPL durchgeführt wurden (Baydar 2016). In Ergänzung zu den theoretischen Ausführungen zur statistischen Versuchsplanung aus Abschnitt 2.3.3 sowie den Überlegungen zur Experimentvorbereitung aus Abschnitt 5.1.1 erfolgt hier nun die konkrete Umsetzung der Transaktionsdatengenerierung mittels des entwickelten Modells in Plant Simulation (vgl. Abbildung 6.7). Zu Beginn wurden für das gegebene Modell die Einflussgrößen bestimmt und aus diesen die Faktoren für die Versuchsplanung festgelegt (vgl. hierzu auch das Datenblatt in Tabelle 6.13). Tabelle 6.15 gibt eine Übersicht über die identifizierten Einflussfaktoren sowie die Anzahl ihrer möglichen Ausprägungen und ordnet jedem Faktor eine Position in der Einflussmatrix zu. Hierbei haben aktive Einflussfaktoren einen starken Einfluss auf die anderen Größen und passive Einflussfaktoren nur eine geringe Einflussstärke auf diese. Für die Verfahren zur Bestimmung der Faktoren und der Einflussmatrix wird auf die Vorarbeiten am Fachgebiet ITPL von Zimmermann (2016) verwiesen.

**Tabelle 6.15: Faktoren des SC-Modells in Plant Simulation (Grundkonfiguration)**

| Faktor | Beschreibung | Anzahl Ausprägungen | Zuordnung nach Einflussmatrix |
|---|---|---|---|
| ProduktID | Produkte der SC | 100 | aktiv |
| Mengenangabe | Mengenangabe pro Produkt | 1 000 | aktiv |
| Zulieferverfügbarkeit | Verfügbarkeit der Produkte beim jeweiligen Lieferanten | 2 | aktiv |
| Zwischenlager | Transportweg über Zwischenlager | 2 | aktiv |
| Abfahrtzeiten | Abfahrtzeiten auf Stunden gerundet | 24 | aktiv |
| Produktgewicht | Produktgewicht | 500 | passiv |
| Produktvolumen | Produktvolumen | 800 | passiv |
| Transportmittel | Transportmittel der SC | 10 | passiv |

Das Ziel der initialen Transaktionsdatengenerierung ist das Erzeugen einer Datenlandschaft, die in ihrer Ausprägung mit den Echtdaten des realen Systems vergleichbar ist. Die generierten Daten unterscheiden sich daher grundsätzlich von den generierten Daten, die in der simulationsunterstützten Validierung erzeugt werden. Während die Transaktionsdaten zur Anwendung der MESC über Simulationseingabegrößenmerkmale verfügen müssen, die die Wirklichkeit geeignet nachbilden, muss das für die Transaktionsdaten in der Validierungsphase nicht zutreffen. Es ist im Allgemeinen nicht möglich, in vertretbarer Zeit alle theoretisch möglichen Ausprägungen der Transaktionsdaten zu generieren. Dies wäre zudem ein Widerspruch zur realitätsnahen Abbildung, da eine vollständige Abdeckung von Ausprägungskombinationen im Regelfall keine Eigenschaft der Echtdaten in der SC ist. Aus diesen Gründen konzentrieren sich die hier diskutierten Experimente auf die Eingabegrößen. Die Auswirkungen der Eingabegrößen auf die Trace-Ausgabegrößen werden im Rahmen dieser Arbeit nicht mit statistischen Verfahren untersucht.

Die initiale Generierung der Transaktionsdaten erfolgt über eine festgelegte Anzahl an Replikationen. Die notwendige Replikationsanzahl hängt im Wesentlichen von den Zielvorgaben des zu erzeugenden Datenumfangs sowie der Konfiguration des Simulationsmodells ab (vgl. Abschnitt 5.1.1). Die Replikationen, die den Seed bei konstanten Stellgrößen variieren, erlauben das Erzeugen einer realitätsnahen Datenlandschaft. Insbesondere wurde durch die Verwendung von konstanten Stellgrößen widersprüchliche Konstellationen innerhalb eines Datenbestands verhindert. Sollten Simulationsexperimente (vgl. auch Simulationsexperimente in Abschnitt 2.3.3), die beispielsweise in ihren Konfigurationen ein Zwischenlager zum

einen begünstigen und zum anderen ausschließen, einen Datenbestand erzeugen, so sind potentiell zu entdeckende Wirkzusammenhänge fachlich ohne Aussagekraft. Replikationen für die Datengenerierung zu nutzen entspricht darüber hinaus einem der erprobten Standard-Szenarien im Data-Farming (vgl. Sanchez 2014 in Abschnitt 2.3.4).

Tabelle 6.16 zeigt einen Ausschnitt aus der Konfiguration der Simulationsstellgrößen, die für das Data Farming genutzt wurden. Die fachinhaltliche Bedeutung der Stellgrößen besteht in der Zuordnung von Lieferanten und Produkten, d. h. es wird festgelegt, welcher Lieferant in der SC welches Produkt liefern kann.

Da das Verhalten der beobachteten Trace-Ausgabegrößen durch den Zufall mitbestimmt wird (stochastische Simulation), muss die Aussagekraft der erzeugten Daten mittels statistischer Methoden verifiziert werden. Um zu prüfen, ob die generierten Daten statistisch signifikant sind (vgl. Abschnitt 5.1.1), wurden mehrere Replikationen mit festgelegter Parameterkonfiguration durchgeführt und die generierten Datenbestände verglichen. Der Simulationsumfang der generierten Datenbestände wurde auf 100 000 Transaktionen festgelegt. Tabelle 6.17 zeigt beispielhaft die ersten sechs Bestellmengen der ProduktID = 1 für variierende Seeds. Zusätzlich wurde der Erwartungswert und die Varianz des entsprechenden Transaktionsmerkmals für die generierten Datenbestände ermittelt und in der Tabelle angegeben. Die Anzahl der Replikationen wurde auf 6 festgelegt. Um die Replikationsanzahl zu bestimmen, diente die Rule-of-Thumb-Methode als Anhaltswert. Diese Regel schlägt 3-5 Replikationen für ein Simulationsmodell vor (Hoad et al. 2007). Erwartungsgemäß sind Erwartungswert und Varianz für die verschiedenen Seeds in mathematisch vergleichbaren Größenordnungen. Die Wurzel aus der Varianz, die Standardabweichung, ergibt im Mittel den Wert 5,5 für das untersuchte Transaktionsmerkmal. Bereits an den dargestellten Werten wird ersichtlich, dass keine Werte angenommen wurden, die sich mit einer höheren Wahrscheinlichkeit als $\leq 5\,\%$ um mehr als das Doppelte der Standardabweichung vom Erwartungswert unterscheiden. Das Doppelte der Standardabweichung entspricht in diesem Fall dem Wert 10 und der Erwartungswert liegt im Mittel bei 59. Um die getroffene Aussage zu widerlegen, müssten in der Folge vermehrt Werte $< 49$ oder $> 69$ für das dargestellte Transaktionsmerkmal generiert worden sein. In den generierten Datenbeständen traf die Regel bezüglich der $\leq 5\,\%$ Abweichung auf kein Merkmal der Transaktionsdaten zu. Basierend auf dem durchgeführten Überlegungen darf folglich das generierte Simulationsergebnis als statistisch signifikant eingestuft werden.

Da jedoch der Fokus der hier durchgeführten Simulation das Erzeugen von ausreichend großen Datenbeständen ist, sind die statistischen Überlegungen nicht weiter zielführend. Vielmehr muss in Anlehnung an Abschnitt 5.1.2 festgelegt werden, wie viele Simulationsläufe notwendig sind, um einen ausreichenden Datenbestand zu generieren. Da es sich bei dem SC-Modell um ein nicht-terminierendes System handelt, ist nur ein Simulationslauf für die Transaktionsdatengenerierung notwendig. Der Umfang der erzeugten Daten wurde auf 100 000 Transaktionen festgelegt.

**Tabelle 6.16: Auszug der Stellgrößen-Ausgangskonfiguration des Plant Simulation SC-Modells**

| ProduktID | Gewicht | Volumen | Produktname | Produkt-klasse | Bestellung | | Lieferant | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | minimal | maximal | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 10 | 20 | Hauptrahmen | a | 100 | 200 | false | false | false | true | true | true |
| 2 | 100 | 80 | Hauptverkleidung | b | 1 | 10 | true | true | true | true | true | true |
| 3 | 1 | 1 | Befestigungen | a | 200 | 500 | true | false | true | false | true | true |
| 4 | 2 | 5 | Kette | a | 50 | 70 | true | true | true | true | true | true |
| 5 | 5 | 2 | Kolben | b | 10 | 20 | false | true | false | true | false | false |
| 6 | 10 | 10 | Zylinderkopf | a | 250 | 300 | false | false | true | false | false | true |
| 7 | 7 | 8 | Kupplung | b | 120 | 400 | true | false | true | false | true | false |
| 8 | 8 | 6 | Gabelbrücke | a | 75 | 150 | false | false | false | true | true | true |
| 9 | 2 | 3 | Ventile | b | 5 | 10 | false | false | false | true | true | true |
| 10 | 5 | 3 | Vorderradbremse | b | 200 | 300 | true | true | true | true | true | true |
| 11 | 5 | 9 | Pleuel | b | 50 | 70 | true | false | true | false | true | true |
| 12 | 7 | 5 | Federbein | a | 1 | 20 | true | true | true | true | true | true |
| 13 | 20 | 15 | Kurbelwelle | a | 250 | 325 | false | true | false | true | false | false |
| 14 | 10 | 8 | Auspuffsystem | a | 120 | 200 | false | false | true | false | false | true |
| 15 | 6 | 45 | Hinterer Hilfsrahmen | b | 10 | 20 | true | false | true | false | true | false |

Die generierten Transaktionsdaten wurden im Anschluss mittels MESC untersucht und verschiedene Data-Mining-Verfahren angewandt. Tabelle 6.18 zeigt einen Auszug aus den generierten Transaktionsdaten. Dieser Datenbestand wurde bereits um das aggregierte Attribut „Verspätung“ angereichert (vgl. Abschnitt 5.1.2.1).

**Tabelle 6.17: Generierte Transaktionen bei variierendem Seed**

| Produktmenge | Seed 1 | Seed 2 | Seed 3 | Seed 4 | Seed 5 | Seed 6 |
|---|---|---|---|---|---|---|
| Bestellung 1 | 50 | 61 | 58 | 67 | 64 | 58 |
| Bestellung 2 | 58 | 62 | 60 | 60 | 62 | 60 |
| Bestellung 3 | 69 | 57 | 64 | 60 | 56 | 64 |
| Bestellung 4 | 53 | 62 | 59 | 67 | 61 | 59 |
| Bestellung 5 | 64 | 58 | 53 | 51 | 57 | 53 |
| Bestellung 6 | 68 | 61 | 63 | 56 | 61 | 63 |
| Erwartungswert | 59,63 | 59,40 | 59,34 | 59,66 | 59,39 | 59,31 |
| Varianz | 33,94 | 34,67 | 32,99 | 33,4 | 34,66 | 33,18 |

**Tabelle 6.18: Auszug aus den generierten Transaktionsdaten (siehe auch Datenblatt in Tabelle 6.13)**

| ProduktID | Menge | Lieferdatum | Zwischenlager | Verspätung |
|---|---|---|---|---|
| 4 | 50 | 07.01.2015 16:18 | Zwischenlager1 | ja |
| 5 | 58 | 08.01.2015 06:30 | Zwischenlager2 | nein |
| 4 | 69 | 11.01.2015 17:07 | Zwischenlager1 | nein |
| 6 | 53 | 08.01.2015 20:49 | nein | nein |
| 7 | 64 | 08.01.2015 22:29 | nein | nein |

In Bezug auf das Einsparpotential der MESC bei geeigneter Konzeption der Trace-Ausgabegrößen (vgl. Tabelle 5.2) wurden in diesem Experiment der MESC-Schritt 3.1 „Überführung in ein Standarddatenformat“ und der MESC-Schritt 3.2 „Gruppierung“ eingespart. Die Ausgabe der Plant Simulation Daten erfolgt in Excel, was bereits dem in Abschnitt 4.3.1 empfohlenen Tabellenformat entspricht. Die Gruppierung erfolgte hier anhand von Transportmitteln, um eine Vergleichbarkeit der Transportzeiten sicherzustellen. Des Weiteren wurden mit der fachlichen Kodierung der Verspätung als Merkmal der Transaktionsdaten ebenfalls Arbeitsschritte, die dem MESC-Schritt 4.3 „Fachliche Kodierung“ zuzuordnen sind, eingespart.

Aufgrund der Auswahl des Data-Mining-Verfahrens in Bezug auf die Datenbeschaffenheit waren als Vorverarbeitungsschritte der MESC noch Teile der Schrit-

te 4.3 und 4.4 „Technische Kodierung“ durchzuführen. Hierbei wurde die technische und fachliche Kodierung der Attribute in einem iterativen Prozess schrittweise angepasst, um geeignete Datenaggregationsstufen für die Entdeckung von Wirkzusammenhängen festzulegen. Dieser Vorgang kann an den Zeitstempeln der Transaktionsdaten nachvollzogen werden (vgl. Tabelle 6.18). Die Zeitstempel, aus deren Kombination sich aggregierte Daten wie die Verspätung ergeben, sind für die eigentliche Wissensentdeckung ungeeignet, da zu viele Einzelwerte vorliegen. In mehreren Versuchen wurden unterschiedliche Diskretisierungen und Attributsumwandlungen mittels Kontextwissens erprobt, wobei sich im Resultat eine mittlere Aggregationsstufe als zielführend herausgestellt hat. Hierbei wurden die Zeitstempel auf Wochentage gemappt, da vermutet wurde, dass unterschiedliche Wochentage einen Einfluss auf die verschiedenen Transportmittel haben. Eine feinere Granulierung des Zeitstempelmappings (Wochentag plus vormittags (06.00 - 11.59) oder nachmittags (12.00 - 17.59) oder nachts (18.00 - 5.59)) führte zu keinen neuen Erkenntnissen.

Abbildung 6.9 zeigt das Ergebnis der angewandten Data-Mining-Verfahren auf dem gruppierten Datenbestand „Transportmittel = Flugzeug“. In der Abbildung wurde als initiale Darstellungsart ein Graph gewählt (vgl. Abschnitt 4.4.1), der drei Wirkzusammenhänge darstellt, die im Kontext der Verspätung entdeckt wurden. In der Darstellung gibt die erste Zahl hinter einer Regel den Support an. Der Support steht für die relative Häufigkeit der Transaktionen, in denen die Regel anwendbar ist. Der zweite Wert in der Darstellung ist die Confidence, die für die relative Häufigkeit der Beispiele steht, in denen die Regel auch gültig ist. Sie gibt den Prozentanteil der Transaktionen an, für die gilt, wenn die Item Sets der Prämisse enthalten sind, sind auch die Item Sets der Konklusion enthalten. Das Attribut Verspätung stellt die Konklusion dar und die verbleibenden Attribute mit ihren Attributsausprägungen und Kombinationen bilden die Prämissen. Tabelle 6.19 zeigt die Wirkzusammenhänge aus Abbildung 6.9 in der Darstellungsform eines Regelwerks. Diese Darstellungsform wird für die weitere Diskussion zugrunde gelegt, da Veränderungen in Regeln sowie ihren statistischen Parametern leicht ersichtlich sind.

Diese dargestellten Regeln ergeben sich aus der in den Experimenten gewählten und trainierten Parametrierung der Verfahren. Hierbei ist insbesondere wichtig, dass die Verspätung = true als Element der Konklusion festgelegt wurde. Somit entfallen Regeln, die Wirkzusammenhänge aufzeigen, die keinen Bezug in der Konklusion zur Verspätung aufweisen. Zusätzlich wurden nur Regeln betrachtet, deren Confidence über 0,8 liegt. Regeln zu Verspätungen, deren Confidence geringer ist, wurden zwar in den Experimenten betrachtet, sind aber aufgrund der Übersichtlichkeit der Ergebnisse hier nicht weiter diskutiert. Zusätzlich wurde der Support auf einen kleinen Wert gesetzt, um auch seltene Kombinationen von Transaktionsdaten zu inkludieren. In den Regeln ist Z_Lagerdauer als Lagerdauer in Tagen, Anzahl als Liefermengeneinheit pro Produkt und Anbieter sowie Dauer als Gesamtlieferzeit vom Lieferant zum Zielort enthalten. Die Kommawerte in den At-

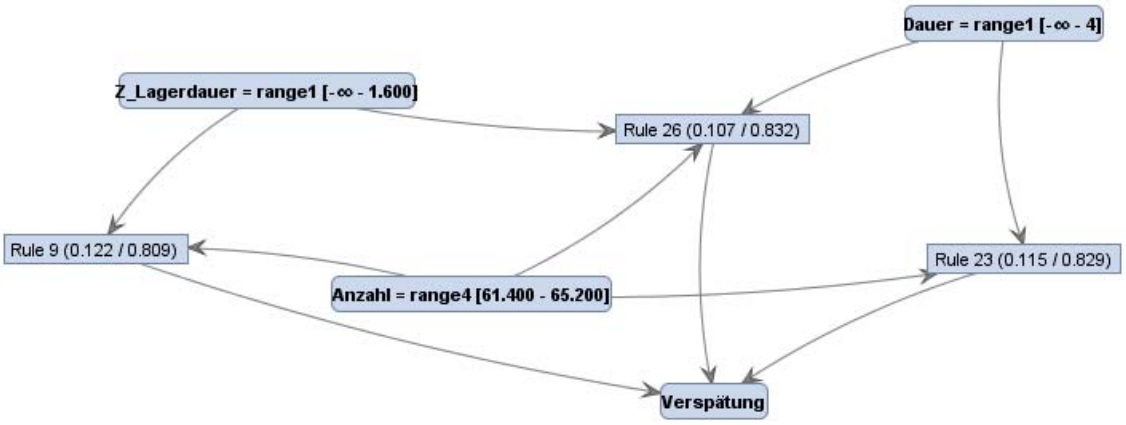



**Abbildung 6.9: Regeln als Ergebnis der Assoziationsanalyse auf den generierten Transaktionsdaten**

**Tabelle 6.19: Assoziationsregeln mit Support > 0,1, Confidence > 0,8 und Frequent Item Set „Verspätung = true“ in der Konklusion**

| Nr. | Prämisse | Konklusion | Support | Confidence |
|---|---|---|---|---|
| 9 | Z_Lagerdauer = range1 [-∞ – 1,600], Anzahl = range4 [61,400 – 65,200] | Verspätung | 0,122 | 0,809 |
| 23 | Dauer = range1 [-∞ – 4], Anzahl = range4 [61,400 – 65,200] | Verspätung | 0,115 | 0,829 |
| 26 | Z_Lagerdauer = range1 [-∞ – 1,600], Dauer = range1 [-∞ – 4], Anzahl = range4 [61,400 – 65,200] | Verspätung | 0,107 | 0,832 |

tributsausprägungen sind durch die eingesetzten Diskretisierungparameter in der technischen Kodierung gegeben (vgl. Abschnitt 4.4.4). Alle Regeln wurden in der Datengruppe des Transportmittels Flugzeug gefunden. Des Weiteren zeigt sich mit der Reduzierung der Confidence auf 0,7 eine vermehrtes Auftreten der Regel Produktgruppe 4. Die genaue Parametrierung der relevanten Vorverarbeitungsschritte und Data-Mining-Verfahren kann im Anhang der Tabelle B.7 entnommen werden.

Das vorgestellte Regelwerk wurde im Rahmen der MESC-Anwendung mittels Kreuzvalidierung, einer Standardmethode des KDD zur Mustervalidierung (vgl. Literatur zur Validierung in Abschnitt 2.3.2), überprüft. Das Vorgehen für die MESC wurde in Abschnitt 4.5.2 diskutiert. Der nachfolgende Abschnitt greift die entwickelten Verfahren der simulationsunterstützten Validierung auf und zeigt, wie die Regeln zur Verspätung mittels DES validiert werden können.

### 6.3.3 Statistische Versuchsplanung in der simulationsunterstützten Validierung

Bei der simulationsunterstützten Validierung nimmt die Parametereinstellung des Simulationsmodells einen wichtigen Stellenwert ein. Die Parametrierung des Simulationsmodells wird benötigt, um die Hypothesen zu dem entdeckten Muster in geeigneter Weise auf das Modell zu übertragen. Die Übertragung durch Veränderung von Stellgrößen wurde in ihren theoretischen Grundlagen in Abschnitt 5.2.2 vorgestellt. Der aktuelle Abschnitt demonstriert nun die praktische Umsetzbarkeit der gezielten Veränderung der Datenbasis zur Mustervalidierung.

Das zu validierende Muster vom Subtyp Regel aus Tabelle 6.19 bildet die Ausgangslage der Experimente zur simulationsunterstützten Validierung. Zu Beginn wurden verschiedene Hypothesen zu den gefundenen Regeln aufgestellt. Die folgende Aufzählung stellt eine Auswahl der Hypothesen dar:

1. Wenn die Produktgruppe eine Auswirkung auf die Verspätung hat, muss ein vermehrtes Auftreten der Produktgruppe 4 die Auftrittswahrscheinlichkeit für das Muster erhöhen.
2. Wenn die reine Bestellmenge eine Auswirkung auf die Verspätung hat, muss sich bei separater Anpassung der Bestellmenge in allen Produktgruppen die Auftrittswahrscheinlichkeit für das Muster erhöhen.
3. Wenn die Verweildauer in einem ausgezeichneten Zwischenlager Einfluss auf die Verspätung hat, muss bei Umgehung des Zwischenlagers die Auftrittswahrscheinlichkeit für das Muster verringert werden.
4. Wenn die Lieferdauer eines Transportmittels keinen Einfluss auf die Verspätung hat, muss bei Anpassung der Lieferdauer für Flugzeuge die Auftrittswahrscheinlichkeit für das Muster in der relevanten Datengruppe konstant sein.

Die ausgewählten Hypothesen stellen nur eine Teilmenge der möglichen Hypothesen dar. Ausschlaggebend für die Arbeit mit Teilmengen ist, dass zu jeder SC-Regel eine Vielzahl von möglichen Hypothesen aufgestellt werden kann. Hier sind insbesondere sämtliche Formen der Negierung von Aussagen aufzuführen sowie die Kombination von mehreren Hypothesen. Die Hypothese 1 kann beispielsweise auch als „Wenn die Produktgruppe keine Auswirkung auf die Verspätung hat, darf ein vermehrtes Auftreten der Produktgruppe 4 die Auftrittswahrscheinlichkeit für das Muster nicht verändern“ formuliert werden. Eine Kombination von Hypothesen wäre beispielsweise „Wenn die Produktgruppe und die Bestellmenge eine Auswirkung auf die Verspätung haben, muss ein vermehrtes Auftreten der Produktgruppe 4 und die Variation der Bestellmenge die Auftrittswahrscheinlichkeit für das Muster erhöhen“. Ein Überprüfen des vollständigen Hypothesenraums zu einem Muster ist nicht zielführend und oftmals praktisch auch nicht umsetzbar. In den für die praktische Demonstration ausgewählten Hypothesen wurden gezielt nur Teilaussa-

gen der Regeln integriert, mit dem Ziel, so zusätzlich zu validieren, ob tatsächlich nur die Kombination der einzelnen Merkmalausprägungen die Verspätung bedingt.

Die ausgewählten Hypothesen wurden als Ausgangspunkt für die Stellgrößenveränderung im Simulationsmodell genutzt. Die Veränderung des Simulationsmodells und der daraus resultierenden Datenlandschaft wurde durch Veränderung der Einflussfaktoren herbeigeführt (vgl. Tabelle 6.15). Die Veränderung der Einflussfaktoren muss für jede Hypothese einzeln durchgeführt werden. Aus den vier aufgeführten Hypothesen ergaben sich folgende Veränderungen der Faktoren im Simulationsmodell:

- Für Experiment 1 wurde die Bestellhäufigkeit für Produktgruppe 4 erhöht und zusätzlich eingestellt, dass jeder Lieferant Produktgruppe 4 liefern kann.
- Für Experiment 2 wurde die Bestellmenge aller Produktgruppen auf eine identische Bestellmenge (50-70 Bestellmenge in Tabelle 6.16) festgelegt.
- Für Experiment 3 wurden die Transportwege über die ausgezeichneten Zwischenlager eliminiert.
- Für Experiment 4 wurden die Transportzeiten des Transportmittels Flugzeug so verändert, dass diese in den Größenordnungen mit der Datengruppe „LKW“ und „Auto“ vergleichbar waren.

Auf den genierten Datenbeständen wurde erneut der RapidMiner-Assoziationslerner mit der Parametrierung des Data Farmings (vgl. Tabelle B.7 im Anhang) angewandt. Die unveränderte Parametrierung stellt einen wesentlichen Bestandteil der Experimente dar (vgl. hierzu auch Steuerungsablauf in Abbildung 5.6). Tabelle 6.20 zeigt die Wirkzusammenhänge, die mittels MESC im Rahmen von Experiment 1 erzeugt wurden. Die Nummerierung (Nr.) in allen Experimenten entspricht den Regelnummern nach absteigendem Support in RapidMiner.

**Tabelle 6.20: Assoziationsregeln mit Support > 0,1 und Confidence > 0,8 aus Experiment 1**

| Nr. | Prämisse | Konklusion | Support | Confidence |
|---|---|---|---|---|
| 9 | Z_Lagerdauer = range1 [-∞ – 1,600], Anzahl = range4 [61,400 – 65,200] | Verspätung | 0,321 | 0,802 |
| 23 | Dauer = range1 [-∞ – 4], Anzahl = range4 [61,400 – 65,200] | Verspätung | 0,211 | 0,828 |
| 26 | Z_Lagerdauer = range1 [-∞ – 1,600], Dauer = range1 [-∞ – 4], Anzahl = range4 [61,400 – 65,200] | Verspätung | 0,145 | 0,831 |

Die gefundenen Regeln zeigen, dass sich die Hypothese bestätigt hat. Obwohl die Confidence relativ konstant geblieben ist, hat sich durch die künstliche Erhöhung von Produktgruppe 4 in der SC der Support für die relevanten Regeln erhöht. Der gestiegene Support bedeutet, dass in Experiment 1 wesentlich mehr Transaktionen enthalten sind, auf die diese Regeln anwendbar sind.

Im Anschluss wurde das Experiment 2 mit gleicher Parameterkonfiguration durchgeführt. Hier müsste das Ergebnis weiter untersucht werden, da die Aussage nicht eindeutig ist. Tabelle 6.21 zeigt den Wirkzusammenhang, der mittels MESC im Rahmen von Experiment 2 erzeugt wurde. Das Ergebnis legt nahe, dass die Einlagerdauer in einem expliziten Zwischenlager nun ausschlaggebend für die Verspätung ist. Die Regeln 9 und 26 aus Tabelle 6.20 sind nun nicht mehr gültig, da die Bestellmenge in der Attributsausprägung von „Anzahl“ nun kein ausschlaggebender Faktor mehr ist. Hier hat sich die Hypothese, dass die reine Bestellmenge ausschlaggebend ist, nicht bestätigt. Vielmehr legt die gefundene Regel 5 aus Tabelle 6.20 nah, die Auswirkung der Zwischenlager in weiteren Experimenten zu untersuchen.

**Tabelle 6.21: Assoziationsregeln mit Support > 0,1 und Confidence > 0,8 aus Experiment 2**

| Nr. | Prämisse | Konklusion | Support | Confidence |
|---|---|---|---|---|
| 5 | Z_Lagerdauer = range1 [-∞ – 2,200] | Verspätung | 0,230 | 0,809 |

Das Experiment 3 beschäftigt sich mit der Auswirkung der Zwischenlager auf die gefundene Verspätung. Zu diesem Zweck wurde eine Hypothese formuliert, die besagt, dass sich die Auftrittswahrscheinlichkeit für das Muster verringert, wenn spezifische Zwischenlager in der SC vermieden werden. Diese Hypothese wurde bestätigt, da unter den 12 gefundenen Regeln aus diesem Experiment keine Regel enthalten war, die einen Rückschluss auf eine Verspätung gestattet. Um zu prüfen, in wie weit sich nun mögliche Auftrittswahrscheinlichkeiten verringert haben, wurden in einem weiteren Experimentzyklus der Support und die Confidence reduziert. Dies verfolgt das Ziel, auch seltenere Regeln in den generierten Datenbeständen zu finden. Die ersten drei Regeln, die einen Rückschluss auf die Verspätung gestatten, sind in Tabelle 6.22 dargestellt. Es wird fachlich angemerkt, dass diese Regeln nicht interessant sind (vgl. interestingness measures in Abschnitt 2.3.2), da der Unterschied in den Größenordnungen zwischen Confidencewert und Support im Vergleich zu den vorangestellten Experimenten (z. B. Experiment in Tabelle 6.20) nur marginal ist. Dies begründet sich darin, dass ein Confidencewert in einer vergleichbaren Größenordnung wie ein Supportwert nur den Support widerspiegelt.

Die Regeln 38-40 aus Tabelle 6.22 stellen somit keine Aussage über den Zusammenhang zwischen den Transaktionsattributen dar. Es wird dennoch deutlich, dass auch in den selteneren Verspätungsregeln spezifische Zwischenlager keine Rolle

**Tabelle 6.22: Assoziationsregeln mit Support > 1 und Confidence > 0,2 aus Experiment 3**

| Nr. | Prämisse | Konklusion | Support | Confidence |
|---|---|---|---|---|
| 38 | Dauer = range1 [-∞ – 7] | Verspätung | 0,206 | 0,284 |
| 39 | Anzahl = range1 [-∞ – 100,200] | Verspätung | 0,166 | 0,275 |
| 40 | Dauer = range1 [-∞ – 7], Anzahl = range1 [-∞ – 100,200] | Verspätung | 0,157 | 0,271 |

spielen. Im Rückschluss ist ersichtlich, dass die Vermeidung von spezifischen Zwischenlagern dazu führt, dass die Verspätungsregeln nur noch selten und mit keinem direktem Zusammenhang zu bestimmten Zwischenlagern aufzufinden sind. Dieses Experiment unterstützt die zu Beginn aufgestellte Hypothese und dient zur Validierung der Aussage, dass ein spezifisches Zwischenlager Anteil an den auftretenden Verspätungen hat.

In Experiment 4 wurde die Lieferdauer des Transportmittels Flugzeug angepasst, um zu prüfen, ob die Lieferzeit in der Datengruppe Flugzeug einen Einfluss auf die gefundenen Regeln hat. Mit dieser Anpassung fanden sich in der spezifischen Datengruppe keine Verspätungen mehr, die einen Bezug zur Lieferzeit hatten. Somit konnte diese Hypothese bestätigt werden. Wie allerdings schon in Experiment 2 gesehen, reichen oftmals die Ergebnisse aus nur einem Simulationsexperiment nicht aus, um eine Hypothese anzunehmen oder zu widerlegen. Hier ist es ratsam, ausgehend von den letzten Experimentergebnissen, neue ergänzende Hypothesen zu entwickeln und weitere Simulationen durchzuführen. Für Experiment 4 konnte zwar ein Zusammenhang mit der Lieferzeit nachgewiesen werden, aber es lässt sich nicht ausschließen, dass weitere Spezifika in den Flugzeugtransaktionen zu Verspätungen führen. In einem weiteren Versuch wurde der Assoziationsregellerner auf die Menge aller Transaktionen angewandt, um zu prüfen, ob es noch Regeln mit Verspätungen gibt, die einen spezifischen Bezug zu einem Transportmittel aufweisen. In dem Experiment wurde sukzessive der Support verringert, um auch seltenere Regeln zu finden. Tabelle 6.23 zeigt die zwei Regeln, die noch einen Bezug zu spezifischen Flugzeugstrecken aufweisen und als Konklusion die Verspätung beinhalten.

**Tabelle 6.23: Assoziationsregeln mit Support > 0,2 und Confidence > 0,7 aus Experiment 4**

| Nr. | Prämisse | Konklusion | Support | Confidence |
|---|---|---|---|---|
| 119 | Lieferweg = Flugzeug 0 | Verspätung | 0,330 | 0,703 |
| 169 | Lieferweg = Flugzeug 1 | Verspätung | 0,221 | 0,713 |

In nachfolgenden Experimenten kann mit dem hier vorgestellten Ansatz validiert werden, welche weiteren SC-Komponenten gegebenenfalls in Kombination mit der Verspätung auf den beiden Transportstrecken zu identifizieren wären. Es können auch weitere Hypothesen aufgestellt werden, indem beispielsweise die beiden Flugstrecken für den nächsten Simulationslauf gesperrt werden. Das Prinzip wurde in Experiment 3 mit der Vermeidung eines spezifischen Zwischenlagers demonstriert.

## 6.4 Zusammenfassung der Evaluierungsergebnisse

Bei der Evaluierung im Anwendungsfeld 1 wurde festgestellt, dass sowohl die Schritte und die Schrittreihenfolge als auch die zugehörigen Phasen und deren Abfolge in der MESC zielführend sind. In Abschnitt 6.2.1.2 wurde detailliert ausgeführt, dass die Reihenfolge der Phasen und die enthaltenen Schritte eine bedarfsgerechte Datenauswahl ermöglichen. Des Weiteren konnte gezeigt werden, dass spezifische Schritte wie die Gruppierung von Daten im SC-Umfeld von praktischer Bedeutung sind (vgl. Abschnitt 6.2.2.1). Die bereitgestellte Datenqualität stellt in allen Phasen eine Herausforderung dar, was jedoch nicht als explizites Problem der MESC interpretiert werden darf (vgl. Abschnitt 6.2.3).

Im Bezug auf das Dreiecksmodell konnte gezeigt werden, dass aufgrund der zeitlichen Randbedingungen eine Durchführung der V&V auf Phasenebene eine praxisnahe Konzeptionsentscheidung darstellt (vgl. Abschnitt 6.2.4). Die Durchführung der MESC im Anwendungsfeld 1 hat zudem gezeigt, dass trotz vorgegebener Struktur des Dreiecksmodells bestimmte Entscheidungen im V&V-Umfeld projektspezifisch sind. In dem hier diskutierten Anwendungsfeld entfiel ein Prüfschritt und verschiedene vom Konzept separat angelegte Prüfschritte wurden in einen Prüfschritt kumuliert (vgl. Tabelle 6.12). Die Kumulation erfolgte jedoch nur bei Prüfschritten, die in direktem zeitlichem oder inhaltlichem Bezug standen und die dieselbe V&V-Technik nutzten. Zudem zeigte sich, dass die phasenbezogene Fehleridentifikation im Anwendungsfeld 1 das ausschlaggebende Kriterium war, um Fehler in der Data-Mining-Modellbildung von fehlerhaften SC-Datenbeständen zu unterscheiden (vgl. Diskussion zur Datenqualität in Abschnitt 6.2.3). Die Ausführung der MESC war durch die Wahlmöglichkeiten bei einer Vielzahl der eingesetzten V&V-Techniken gekennzeichnet. Die Wahl wurde im Anwendungsfeld 1 von den Projektrestriktionen, z. B. Anzahl der Treffen oder eingesetzte Ressourcen, dominiert. Es wird aber explizit darauf hingewiesen, dass es sich nur um mögliche V&V-Techniken handelt und bei einer erneuten Durchführung der MESC auch alternative Techniken Einsatz finden könnten. Als Orientierung bei der Durchführung der MESC wird auf die Tabelle A.4 verwiesen.

Es kann konstatiert werden, dass das Konzept des Dreiecksmodells ein notwendiger Bestandteil des hier entwickelten KDD-Vorgehensmodells ist. Der phasenbegleitende Einsatz von V&V-Techniken ist in der Modellausführung erforderlich, denn nur ein phasenbezogenes, zeitnahes Identifizieren von Fehlern ermög-

licht eine präzise Fehlerlokalisierung und schnelle Fehlerbehebung. Die verfügbaren V&V-Techniken aus Simulation und Softwareentwicklung konnten im Anwendungsfeld 1 nutzbringenden Einsatz finden.

Die Evaluierung des Anwendungsfelds 2 hat gezeigt, dass die Generation von Transaktionsdaten für die Wissensentdeckung ein sinnvolles Methodenelement darstellt, da diese die Echtdaten der SC ergänzen oder ersetzen kann (vgl. Abschnitt 6.3.2). Die so generierten Transaktionsdaten wurden im Anschluss mittels MESC auf mögliche Muster untersucht und diese dann mittels des Methodenelements der simulationsunterstützten Validierung geprüft. Bei diesem Evaluierungsschritt zeigt sich, dass das Expertenwissen, das zur Formulierung der Hypothesen notwendig ist (vgl. Abschnitt 6.3.3), ein notwendige Vorbedingung für die erfolgreiche Durchführung dieses Methodenelements ist. Es wurde des Weiteren aufgezeigt, dass dieser Schritt sowohl das Kontextwissen über die SC zur Formulierung der Hypothesen benötigt als auch Kenntnisse des Simulationsmodells zur bedarfsgerechten Datengenerierung voraussetzt. Aus diesem Grund ist der Einsatz der entwickelten Methodenelemente in der Praxis nur bei geeigneten Randbedingungen, wie z.B. kritischen Entscheidungen, zu empfehlen (vgl. hierzu auch Abschnitt 5.3).

Die umfassende Untersuchung von möglichen Aggregationsebenen der Trace-Ausgabegrößen war aufgrund des spezifischen Simulationsmodells nur exemplarisch möglich (vgl. auch Zielsetzung im Abschnitt 6.1). Hier müssen zukünftige Anwendungen der entwickelten Methode zeigen, ob sich weitere, über die theoretischen Überlegungen aus Abschnitt 5.1.2 hinausgehende, Prinzipien ableiten lassen.

Es kann konstatiert werden, dass die vorgestellten Verfahren zur Generierung von Transaktionsdaten und zur simulationsunterstützten Validierung in der Praxis einsetzbar sind. Das Aufstellen von Hypothesen hat sich in diesen Experimenten als iterativer Prozess dargestellt, der zwar den Einsatz von Experten zur Hypothesenformulierung fordert, aber im Gegenzug auch für Laien interpretierbare Ergebnisse in Form von existierenden, nicht existierenden oder sich verändernden Mustern liefert.

Des Weiteren konnte dargelegt werden, dass der in dieser Arbeit entwickelte Musterbegriff eine detaillierte Beschreibung einzelner Sachverhalte ermöglicht. Insbesondere die explizite Benennung von Subtypen (vgl. Abschnitt 4.4) erleichterte die Dokumentation der Experimente (vgl. z. B. Abschnitt 6.3.3). Im Rückschluss wird der Musterbegriff als zielführend für die Beschreibung von Sachverhalten in der MESC angesehen. Eine darüberhinausgehende Aussage kann im Rahmen dieser Arbeit nicht getroffen werden (vgl. auch Abschnitt 6.1).

Zusammenfassend lässt sich feststellen, dass die Evaluierung der einzelnen Methodenelemente (vgl. Abbildung 5.7) gezeigt hat, dass die Methode zur Wissensentdeckung in SC-Datenbanken geeignet ist.

# 7 Zusammenfassung und Ausblick

Dieses Kapitel dient der Zusammenfassung der vorliegenden Arbeit und eröffnet einen Ausblick auf weitere, sich aus dieser Arbeit ergebende Forschungsthemen. Die Wissensentdeckung ist ein notwendiger Schritt, um die Unsicherheiten in der SC aufgrund der steigenden Komplexität von SCs und deren Datenbeständen zu kontrollieren. Unter der Vielzahl von Methodenelementen zur Wissensentdeckung in großen Datenbeständen sind die Vorgehensmodelle des KDD ein etablierter Ansatz. Es gibt jedoch Probleme in der Anwendung existierender Vorgehensmodelle im Umfeld der SC. Die Hintergründe der Probleme und die sich daraus ergebenden Fragestellungen wurden im Kapitel 2 dargestellt und in vier Forschungsfragen zusammengefasst. Hieraus ergab sich das Kernziel der Arbeit, ein Vorgehensmodell zur Wissensentdeckung in SC-Datenbanken zu entwickeln. Das zu entwickelnde Modell sollte einen für die SC geeigneten Musterbegriff beinhalten, Ansätze zur Datengenerierung aufnehmen und eine modellbegleitende Validierung integrieren.

Basierend auf den hergeleiteten Charakteristika der SC wurden Anforderungen an ein Vorgehensmodell aufgestellt und aus den existierenden KDD-Modellen ein geeignetes Referenzmodell ausgewählt. Unter Nutzung des Referenzmodells sowie Validierungsmodellen aus der Simulation wurde im Anschluss eine eigene Methode zur Wissensentdeckung in SC-Datenbanken entwickelt, die Methode zur Musterextraktion in SCs (MESC) (vgl. Kapitel 3). Diese Methode verfügt aufgrund der Integration des Validierungsmodells „Dreiecksmodell“ nun über eine modellbegleitende Validierungsfunktionalität (vgl. Forschungsfrage 3 in Abschnitt 2.4.2). Im Anschluss wurden die Phasen der MESC erläutert und ein methodisches Vorgehen für ausgewählte Schritte diskutiert (vgl. Kapitel 4). Hierbei wurden der strukturelle Aufbau der MESC und Anweisungen für die praktische Durchführung dargelegt (vgl. Forschungsfrage 2 in Abschnitt 2.4.2). Im Zuge der einzelnen Phasenuntersuchungen wurde der Musterbegriff formal definiert; dies gestattet die Verwendung einer einheitlichen Terminologie in der Phasendokumentation der MESC (vgl. Abschnitt 4.4.1). Der so geschaffene Begriff berücksichtigt die verschiedenen Muster, deren Klassen und Subklassen und ermöglicht die Trennung des Musterbegriffs von seinen Darstellungsarten. Insbesondere das Kontextwissen, das in den vorherrschenden Musterbegriffen unberücksichtigt geblieben ist, wurde in die geschaffenen Definitionen integriert, um die Verwendung im SC-Kontext zu ermöglichen (vgl. Forschungsfrage 1 in Abschnitt 2.4.2). Im Anschluss an die vorliegende Forschungsarbeit muss die fortwährende Anwendung des geschaffenen Musterbegriffs im Projekteinsatz zeigen, ob die gewählten Musterklassen und Subklassen alle Anwendungsfelder der SC-Wissensentdeckung abdecken. In diesem Rahmen ist eine weitere Unterteilung der geschaffenen Begrifflichkeiten denkbar, um spezifische Musterarten, beispielsweise in der SC-Simulation, strukturell zu erfassen.

Des Weiteren wurde festgestellt, dass die Validierung der Muster ein zeitintensiver Prozess ist, der Kontextwissen benötigt. Im Rahmen der nachfolgenden Untersu-

chungen wurde aufgezeigt, dass die Datenbereitstellung den Einsatz von MESC im SC-Umfeld begrenzen kann. Aus diesen Gründen wurde im Anschluss diskutiert, welchen methodischen Anteil die Simulation zur Wissensentdeckung in SC-Datenbanken beisteuern kann. Hierbei wurden die Aggregationsstufen der Daten diskutiert und erstmals die Methode des Data Farmings auf die Generierung von Transaktionsdaten angewandt (vgl. Forschungsfrage 4 in Abschnitt 2.4.2). Die Überlegungen zum Data Farming bildeten die Basis für die Entwicklung einer simulationsunterstützten Validierung, die im Anschluss vorgestellt wurde. Die Grundidee des Verfahrens beruht auf der Validierung von Mustern durch das Bilden von Hypothesen sowie der hypothesengetriebenen Veränderung der Datenbasis für die Wissensentdeckung. Im Rahmen der Konzeptdarstellung wurden Potentiale und Anwendungsmöglichkeiten der simulationsunterstützten Validierung unter dem Aspekt der Praxistauglichkeit diskutiert (vgl. Abschnitt 5.2.3). Abschließend wurde das Data Farming und die simulationsunterstützte Validierung der Muster in die MESC integriert.

Für die Validierungsphase wurden in Ergänzung zu den Phasen-Experimenten aus Kapitel 4 zwei Anwendungsfelder definiert. Im ersten Anwendungsfeld wurde die vollständige Wissensentdeckung mittels MESC auf einem großen Industriedatenbestand durchgeführt (vgl. Abschnitt 6.2). Bei der Durchführung der einzelnen Phasen hat sich gezeigt, dass die Ausarbeitungen zu den einzelnen Phasen in der Praxis tragfähig sind. Das zweite Anwendungsfeld thematisiert die Datengenerierung und simulationsunterstützte Validierung in MESC anhand eines SC-Simulationsmodells (vgl. Abschnitt 6.3). Aus den Experimenten konnte der Schluss gezogen werden, dass die Simulation sowohl zum Zweck der Datengenerierung als auch für die Mustervalidierung gewinnbringend ist.

Die Experimente bestätigten auch die These, dass die Konzeption der simulationsunterstützten Validierung ein projektindividueller Prozess ist und somit der effiziente Einsatz nicht uneingeschränkt zu empfehlen ist. Es zeigte sich zudem, dass der Aufbau von Simulationsmodellen von komplexen SCs für den reinen Zweck der Mustervalidierung sehr zeitintensiv ist und der Einsatz der simulationsunterstützten Validierung in der Praxis nur bei bereits existierenden SC-Simulationsmodellen oder bei kritischen Anwendungen empfohlen werden kann.

Wie bereits im Abschnitt 2.4 dargestellt, wurden in dieser Arbeit keine spezifischen Datenstrukturen oder Konzepte aus dem Bereich Big Data untersucht. In diesem Kontext würden verschiedene Aspekte der Modellentwicklung eine Neubewertung erfordern. Beispielsweise müssten die Charakteristika von SC-Datenbanken und die daraus abgeleiteten Anforderungen um Aspekte der Big-Data-Bestände erweitert werden (vgl. Abschnitt 3.2 und 3.3). Hinsichtlich der Kernaspekte der SC-Datenbestände in dieser Arbeit ergeben sich durch den Big-Data-Aspekt mögliche Veränderungen. Dies wird insbesondere an dem geplanten Unternehmensnutzen von Big Data im Bereich Ergänzung von Entscheidungsgrundlagen sowie dem herausragenden Stellwert der Transaktionsdaten in Bezug auf das Datenvolumen

von Unternehmen deutlich (Bitkom 2014). Im Anbetracht dieser Entwicklung kann über eine zukünftige Adaption der MESC nachgedacht werden.

Auch die Validierungsmöglichkeiten durch die Simulation bieten weiteres Entwicklungspotential im KDD. In diesem Themenfeld ist zukünftig zu prüfen, welche Schritte und Phasen in den KDD-Vorgehensmodellen ebenfalls von der Simulation profitieren könnten. Des Weiteren wurde durch das Einbeziehen der Simulation in das Forschungsfeld des KDD aufgezeigt, dass die Aggregationsstufen der Daten zum Einsparen von Bearbeitungsschritten im KDD bzw. sogar zum Entfallen von vollständigen Vorgehensmodellphasen führen können (vgl. Abschnitt 5.1.2.1). Eine Weiterentwicklung dieser Konzepte kann zukünftig zu einem „Lean-KDD“ führen, welches die zeitintensive Vorverarbeitungsphase der Wissensentdeckung möglichst schlank und effizient gestaltet. Die Einbindung von Methoden und Verfahrensweisen aus der Disziplin des Lean Managements in das Konzept der Wissensentdeckung sind jedoch noch nicht hinreichend untersucht. Eine Potentialanalyse der Lean-Management-Verfahrensweise im Hinblick auf das KDD ist somit in weiterführenden Arbeiten denkbar.

Im Hinblick auf den MESC-Schritt der Qualitätskontrolle (vgl. Abschnitt 4.7.1) gibt es alternative Möglichkeiten, die die Optimierung im Nachgang der eigentlichen Wissensentdeckung vermeiden. Es handelt sich hierbei um die agilen Methoden, die aus der Softwareentwicklung bekannt sind. Der Stand der Technik hat gezeigt, dass in den KDD-Vorgehensmodellen die Qualitätskontrolle kaum Berücksichtigung findet sowie eine Integration agiler Techniken nicht Teil der aktuellen Forschung ist (vgl. Abschnitt 2.3.1), sodass eine Untersuchung empfehlenswert ist. Allerdings birgt die Integration agiler Techniken auch viele Risiken für die Wissensentdeckung und geht einher mit einer Vielzahl von Unternehmensanforderungen, wie beispielsweise dem Reifegrad von Prozessen, der Dokumentationsstrukturen oder der Projektmanagementkultur. Aufgrund der praktischen Perspektive dieser Forschungsarbeit wurde kein agiles Vorgehen hinsichtlich einer Qualitätskontrolle in MESC untersucht. Es wird aber explizit auf den Forschungsbedarf in diesem Bereich hingewiesen.

Die vorliegende Arbeit hat Definitionen und Konzepte im Bereich der Muster, der Methoden zur Wissensentdeckung sowie der Integration von Simulation in das Themenfeld des KDD aufgezeigt. Des Weiteren konnte dargelegt werden, dass Vorverarbeitungsmethoden und zielführendes Einbinden von Kontextwissen essentiell im Bereich der SC-Wissensentdeckung sind. Einige der in dieser Arbeit aufgestellten Konzepte sollten in zukünftigen Forschungsarbeiten auf ihre Generalisierbarkeit geprüft werden. In dieser Arbeit ist die Integration des Dreiecksmodells, das eine modellbegleitende V&V im Bereich KDD gestattet, hervorzuheben. Im Rahmen dieser Arbeit wurde das Dreiecksmodell in ein spezifisches Vorgehensmodell integriert, um dieses zu komplettieren. Ob diese Methode auch auf andere KDD-Vorgehensmodelle übertragbar ist, wurde nicht untersucht. Hier wäre ein generelles Konzept zur Erweiterung existierender KDD-Vorgehensmodelle denkbar,

das insbesondere die Anpassung des Dreiecksmodells auf die spezifischen Phasen der einzelnen Modelle beschreibt.

Abschließend kann festgehalten werden, dass im Rahmen der vorliegenden Arbeit eine valide Methode mit begleitender V&V für die Wissensentdeckung in SC-Datenbanken entwickelt wurde. In diesem Kontext wurde ein neuer Musterbegriff geschaffen, Methoden zur Generierung von Transaktionsdaten aufgezeigt und erste Ansätze zur simulationsunterstützten Validierung als neue V&V-Technik eingeführt. Für das im Rahmen der Methode entwickelte Vorgehensmodell existiert im Hinblick auf die steigende Komplexität in der SC (Pfeiffer et al. 2013) ein bedeutsames Handlungsfeld, in dem die Anwendung möglich und gewinnbringend ist. Das entwickelte Vorgehensmodell bietet zudem Erweiterungspotential im Bereich Big Data und agiler Methoden. Einige der in dieser Arbeit entwickelten Konzepte bieten Ansätze zur Generalisierbarkeit für den Bereich des KDD und aufgezeigte Nebenthemen beinhalten Potential zur intensiven Auseinandersetzung. Aus diesen Gründen ist eine Anschlussforschung in den in diesem Kapitel aufgeführten Feldern motiviert.

# Literaturverzeichnis


Adriaans, P.; Zantinge, D.: Data mining. Boston: Addison Wesley Professional, 1996.

Alkharboush, N.; Li, Y.: A decision rule method for assessing the completeness and consistency of a data warehouse. In: Hoeber, O.; Li, Y.; Huang, X. J. (Hg.): Proceedings of the 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology. Bd. 3. Piscataway, NJ: IEEE, 2010, S. 265–268.

Alpar, P.; Niedereichholz, J.: Data Mining im praktischen Einsatz: Verfahren und Anwendungsfälle für Marketing, Vertrieb, Controlling und Kundenunterstützung. Wiesbaden: Vieweg+Teubner, 2000.

Anand, S. S.; Patrick, A. R.; Hughes, J. G.; Bell, D. A.: A data mining methodology for cross-sales. Knowledge-Based Systems 10 (1998) 7, S. 449–461.

Anane, R.; Younas, M.; Tsai, C.-F.; Chao, K.-M.: Agent-based transactional framework for the supply chain. In: Proceedings of the first international conference on machine learning and cybernetics. Piscataway, NJ: IEEE, 2002, S. 1956–1961.

Arndt, H.: Supply Chain Management: Optimierung logistischer Prozesse. 6. Aufl. Wiesbaden: Gabler, 2013.

Arndt, V.: Ereignisdiskrete Simulation einer Supply Chain zur Generierung von Transaktionsdaten. Dortmund: Technische Universität Dortmund, Fachgebiet IT in Produktion und Logistik, Masterarbeit, 2014.

Arnold, D.; Furmans, K.: Materialfluss in Logistiksystemen. 6. Aufl. Berlin: Springer, 2009.

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

Ayers, J. B.: Encyclopedia of supply chain management. Boca Raton: CRC, 2012.

Balci, O.: Verification, validation, and testing. In: Banks, J. (Hg.): Handbook of simulation. Hoboken: John Wiley & Sons, 1998, S. 335–393.

Banks, J.: Principles of simulation. In: Banks, J. (Hg.): Handbook of simulation. Hoboken: John Wiley & Sons, 1998, S. 3–30.

Baydar, E.: Data Farming Konzept in Tecnomatix Plant Simulation. Dortmund: Technische Universität Dortmund, Fachgebiet für IT in Produktion und Logistik, Bachelorarbeit, 2016.

Becker, M.; Wenning, B.-L.; Görg, C.; Gehrke, J. D.; Lorenz, M.; Herzog, O.: Agent-based and discrete event simulation of autonomous logistic processes. In: Borutzky, W.; Orsoni, A.; Zobel, R. (Hg.): Proceedings of the 20th European conference on modelling and simulation. Bonn: ECMS, 2006, S. 566–571.

Beckmann, H.: Prozessorientiertes Supply Chain Engineering: Strategien, Konzepte und Methoden zur modellbasierten Gestaltung. Wiesbaden: Gabler, 2012.

Beckmann, N.: Untersuchung des Einsatzes von Vorgehensmodellen des Knowledge Discovery in Databases für Bereiche der Logistik. Dortmund: Technische

Universität Dortmund, Fachgebiet IT in Produktion und Logistik, Masterarbeit, 2015.

Berry, M. J. A.; Linoff, G.: Mastering data mining: the art and science of customer relationship management. New York: John Wiley & Sons, 2000.

Bishop, C.: Pattern recognition and machine learning. New York: Springer, 2006.

Bissantz, N.; Hagedorn, J.: Data Mining (Datenmustererkennung). Wirtschaftsinformatik 51 (2009) 1, S. 139–144.

Bitkom: Potenziale und Einsatz von Big Data: Ergebnisse einer repräsentativen Befragung von Unternehmen in Deutschland. Bitkom - Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V. (Hg.). Bitkom: Berlin, 2014.

Blackstone, J. H.: APICS dictionary. 13. Aufl. Chicago: APICS The Association for Operations Management, 2010.

Bogon, T.; Timm, I. J.; Jessen, U.; Schmitz, M.; Wenzel, S.; Lattner, A. D.; Paraskevopoulos, D.; Spieckermann, S.: Towards assisted input and output data analysis in manufacturing simulation: The EDASim approach. In: Laroque, C.; Himmelspach, J.; Pasupathy, R.; Rose, O.; Uhrmacher, A. M. (Hg.): Proceedings of the 2012 winter simulation conference (WSC 2012). Piscataway, NJ: IEEE, 2012, S. 1–13.

Borade, A. B.; Sweeney, E.: Decision support system for vendor managed inventory supply chain: a case study. International Journal of Production Research 53 (2015) 16, S. 4789–4818.

Borchers, J.: A pattern approach to interaction design. New York: John Wiley & Sons, 2001.

Brade, D.: A generalized process for the verification and validation of models and simulation results. Neubiberg: Universität der Bundeswehr München, Fakultät für Informatik, Dissertation, 2003.

Brandstein, A. G.; Horne, G. E.: Data farming: a meta-technique for research in the 21st century. In: Hoffmann, F. G.; Horne, G. E. (Hg.): Maneuver warfare science. Quantico: United States Marine Corps Combat Development Command, 1998, S. 93–99.

Braun, C.; Hafner, M.; Wortmann, F.: Methodenkonstruktion als wissenschaflicher Erkenntnisansatz: Arbeitspapier. Universität St. Gallen (Hg.). Institut für Wirtschaftsinformatik: St. Gallen, 2004.

Brinkkemper, S.: Method engineering: engineering of information systems development methods and tools. Information and Software Technology 38 (1996) 4, S. 275–280.

Bronštejn, I. N.; Semendyayev, K. A.; Musiol, G.; Mühlig, H.: Handbook of mathematics. 6. Aufl. Berlin: Springer, 2015.

Bullinger, H.-J.; Spath, D.; Warnecke, H.-J.; Westkämper, E.: Handbuch Unternehmensorganisation: Strategien, Planung, Umsetzung. 3. Aufl. Berlin: Springer, 2009.

Cabena, P.; Hadjinian, P.; Stadler, R.; Verhees, J.; Zanasi, A.: Discovering data mining: from concept to implementation. Upper Saddle River, N.J.: Prentice Hall, 1998.

Chandra, C.; Grabis, J.: Supply chain configuration: concepts, solutions and applications. New York: Springer, 2007.

Che, D.; Safran, M.; Peng, Z.: From big data to big data mining: challenges, issues, and opportunities. In: Hong, B.; Meng, X.; Chen, L.; Winiwarter, W.; Song, W. (Hg.): Database systems for advanced applications. Bd. 7827. Berlin: Springer, 2013, S. 1–15.

Chen, P. P. S.: The entitiy-relationship model - toward a unified view of data. ACM Transactions on Database Systems 1 (1976) 1, S. 9–36.

Chopra, S.; Meindl, P.: Supply Chain Management: Strategie, Planung und Umsetzung. 5. Aufl. Hallbergmoos: Pearson, 2014.

Choy, K. L.; Lee, W. B.; Lo, V.: Design of a case based intelligent supplier relationship management system-the integration of supplier rating system and product coding system. Expert Systems with Applications 25 (2003) 1, S. 87–100.

Christopher, M.: Logistics and supply chain management: strategies for reducing cost and improving service. 2. Aufl. London: Financial Times, 1998.

Christopher, M.: Logistics & supply chain management. 4. Aufl. New York: Financial Times Prentice Hall, 2011.

Christopher, M.; Lee, H.: Mitigating supply chain risk through improved confidence. International Journal of Physical Distribution & Logistics Management 34 (2004) 5, S. 388–396.

Cios, K. J.: Medical data mining and knowledge discovery. Bd. 60. Heidelberg: Physica, 2001.

Cleve, J.; Lämmel, U.: Data mining. München: De Gruyter Oldenbourg, 2014.

Collier, K.; Medidi, M.; Sautter, D.: Visualization in the knowledge discovery process. In: Fayyad, U. M.; Grinstein, G. G.; Wierse, A. (Hg.): Information visualization in data mining and knowledge discovery. San Francisco: Morgan Kaufmann, 2002, S. 121–122.

Cooley, R.; Mobasher, B.; Srivastava, J.: Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems 5 (1999) 1, S. 5–32.

Cooper, M. C.; Lambert, D. M.; Pagh, J. D.: Supply chain management: more than a new name for logistics. The International Journal of Logistics Management 8 (1997) 1, S. 1–14.

Corsten, H.; Gössinger, R.: Einführung in das Supply Chain Management. 2. Aufl. München: Oldenbourg, 2008.

Dasu, T.; Johnson, T.: Exploratory data mining and data cleaning. Hoboken: John Wiley & Sons, 2003.

Dengel, A.: Semantische Technologien: Grundlagen – Konzepte – Anwendungen. Berlin: Spektrum, 2012.

DIN IEC 60050-351: Internationales Elektrotechnisches Wörterbuch - Teil 351: Leittechnik. Berlin: Beuth, 2014.

Draisbach, U.: Partitionierung zur effizienten Duplikaterkennung in relationalen Daten. Wiesbaden: Vieweg+Teubner, 2012.

Düsing, R.: Knowledge Discovery in Databases - Begriff, Forschungsgebiet, Prozess und System. In: Chamoni, P.; Gluchowski, P. (Hg.): Analytische Informationssysteme. Berlin: Springer, 2010, S. 281–306.

Dworatschek, S.: Grundlagen der Datenverarbeitung. 8. Aufl. Berlin: De Gruyter, 1989.

Dyer, R.: MySQL in a nutshell. Köln: O'Reilly, 2008.

European Commision: NESSI – Big Data White Paper: Big Data - A New World of Opportunities. European Commision (Hg.). December 2012.

Fahrmeir, L.; Künstler, R.; Pigeot, I.; Tutz, G.: Statistik: Der Weg zur Datenanalyse. 7. Aufl. Berlin: Springer, 2010.

Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.: The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM 39 (1996a) 11, S. 27–34.

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

Fayyad, U. M.; Uthurusamy, R. (Hg.): Knowledge discovery in databases: papers from the 1994 AAAI workshop. Menlo Park: AAAI, 1994.

Feigenbaum, E. A.; McCorduck, P.: The fifth generation: artificial intelligence and Japan's computer challenge to the world. Reading: Addison Wesley Longman, 1983.

Fleischmann, B.; Meyr, H.: Supply chain planning. In: Jürgens, D.; Grünert, T. (Hg.): Logistik Management. Wiesbaden: Vieweg+Teubner, 2001, S. 13–29.

Fockel, R.: Methoden des Data Mining im praktischen Einsatz. Nüsser, W.; Weigand, C. (Hg.). Fachhochschule der Wirtschaft: Paderborn, Shaker, 2009.

Frank, E.; Witten, I. H.: Making better use of global discretization. In: Bratko, I.; Dzeroski, S. (Hg.): Proceedings of 16th international conference on machine learning. San Francisco: Morgan Kaufmann, 1999, S. 115–123.

Frawley, W. J.; Piatetsky-Shapiro, G.; Matheus, C. J.: Knowledge discovery in databases: an overview. AI Magazine 13 (1992) 3, S. 57–70.

Freund, Y.; Schapire, R. E.: Large margin classification using the perceptron algorithm. Machine Learning 37 (1999) 3, S. 277–296.

Gabriel, R.; Gluchowski, P.; Pastwa, A.: Data Warehouse & Data Mining. Herdecke: W3L, 2009.

Gamma, E.; Helm, R.; Johnson, R.; Vlissides, J.: Design patterns: elements of reusable object-oriented software. Boston: Addison-Wesley, 1995.

García, S.; Luengo, J.; Herrera, F.: Data preprocessing in data mining. Bd. 72. Cham: Springer, 2015.

Geng, L.; Hamilton, H. J.: Interestingness measures for Data Mining: a survey. ACM Computing Surveys 38 (2006) 3.

Gerke, K.; Claus, A.; Mendling, J.: Process mining of RFID-based supply chains. In: Hofreiter, B.; Werthner, H. (Hg.): 11th IEEE conference on commerce and enterprise computing (CEC'09). Washington: IEEE, 2009, S. 285–292.

Ghadge, A.; Dani, S.; Chester, M.; Kalawsky, R.: A systems approach for modelling supply chain risks. Supply Chain Management: An International Journal 18 (2013) 5, S. 523–538.

Gibson, B. J.; Mentzer, J. T.; Cook, R. L.: Supply chain management: the pursuit of a consensus definition. Journal of Business Logistics 26 (2005) 2, S. 17–25.

Giese, A.: Differenziertes Performance Measurement in Supply Chains. Wiesbaden: Gabler, 2012.

Girard, A.; Pappas, G. J.: Verification using simulation. In: Hespanha, J. P.; Tiwari, A. (Hg.): Hybrid systems: computation and control. Bd. 3927. Berlin: Springer, 2006, S. 272–286.

Goldkuhl, G.; Lind, M.; Seigerroth, U.: Method integration: the need for a learning perspective. IEEE Proc., Softw. (IEEE Proceedings - Software) 145 (1998) 4, S. 113–118.

Guhl, P.: Erstellung eines konzeptuellen Datenbankschemas im Umfeld von Supply Chains. Dortmund: Technische Universität Dortmund, Fachgebiet IT in Produktion und Logistik, Masterarbeit, 2014.

Günther, H.-O.; Tempelmeier, H.: Produktion und Logistik. 9. Aufl. Berlin: Springer, 2011.

Gürez, E.: Zuordnung von Data Mining-Methoden zu problemspezifischen Fragestellungen von Supply Chain Management-Aufgaben. Dortmund: Technische Universität Dortmund, Fachgebiet IT in Produktion und Logistik, Bachelorarbeit, 2015.

Hanhijärvi, S.: Multiple hypothesis testing in pattern discovery. In: Elomaa, T.; Hollmén, J.; Mannila, H. (Hg.): Discovery science. Bd. 6926. Berlin: Springer, 2011, S. 122–134.

Hansen, H. R.; Neumann, G.: Wirtschaftsinformatik 2: Informationstechnik. 9. Aufl. Stuttgart: Lucius & Lucius, 2005.

Harland, C. M.: Supply chain management: relationships, chains and networks. British Journal of Management 7 (1996) s1, S. 63–80.

Harrell, C.; Ghosh, B. K.; Bowden, R.: Simulation using ProModel. 3. Aufl. New York: McGraw-Hill, 2012.

Harrison, A.; van Hoek, R. I.: Logistics management and strategy: competing through the supply chain. 3. Aufl. Harlow: Financial Times Prentice Hall, 2008.

Hastie, T.; Tibshirani, R.; Friedman, J. H.: The elements of statistical learning: data mining, inference, and prediction. 2. Aufl. New York: Springer, 2009.

Heisig, P.: Integration von Wissensmanagement in Geschäftsprozesse. Berlin: eureki, 2005.

Hellingrath, B.; Kuhn, A.: Supply Chain Management: Optimierte Zusammenarbeit in der Wertschöpfungskette. Berlin: Springer, 2002.

Hellingrath, B.; Laakmann, F.; Nayabi, K.: Auswahl und Einführung von SCM-Softwaresystemen. In: Beckmann, H. (Hg.): Supply Chain Management. Berlin: Springer, 2004, S. 99–122.

Hellström, D.; Johnsson, M.: Using discrete event simulation in supply chain planning. In: Solem, O. (Hg.): Proceedings of the 14th annual conference for Nordic researchers in logistics. Trondheim: NOFOMA, 2002, S. 13–28.

Hetzel, W.: The complete guide to software testing. Wellesley: QED Information Sciences, 1984.

Hippner, H.; Wilde, K. D.: Der Prozess des Data Mining im Marketing. In: Hippner, H.; Küsters, U.; Meyer, M.; Wilde, K. D. (Hg.): Handbuch Data Mining im Marketing. Wiesbaden: Vieweg, 2001, S. 21–94.

Hoad, K.; Robinson, S.; Davies, R.: Automating DES output analysis: how many replications to run. In: Henderson, S. G.; Biller, B.; Hsieh, M.-H.; Shortle, J.; Tew, J. D.; Barton, R. R. (Hg.): Proceedings of the 2007 Winter Simulation Conference. Washington, D.C.: IEEE, 2007.

Hult, G. T. M.; Ketchen, D. J.; Cavusgil, S. T.; Calantone, R. J.: Knowledge as a strategic resource in supply chains. Journal of Operations Management 24 (2006) 5, S. 458–475.

Jodin, D.; Kuhnt, S.; Wenzel, S.: Methodennutzungsmodell zur Informationsgewinnung in großen Netzen der Logistik. In: Buchholz, P.; Clausen, U. (Hg.): Große Netze der Logistik. Berlin: Springer, 2009, S. 1–18.

John, G. H.: Enhancements to the data mining process. Stanford: Stanford University, Department of Computer Science, Dissertation, 1997.

Jünemann, R.: Materialfluß und Logistik: Systemtechnische Grundlagen mit Praxisbeispielen. Berlin: Springer, 1989.

Kamble, S.; Desai, A.; Vartak, P.: Data mining and data warehousing for supply chain management. In: Proceedings of the international conference on communication, information & computing technology. Piscataway, NJ: IEEE, 2015, S. 1–6.

Kemppainen, K.; Vepsäläinen, A. P.: Trends in industrial supply chains and networks. International Journal of Physical Distribution & Logistics Management 33 (2003) 8, S. 701–719.

Kleijnen, J. P. C.; Sanchez, S. M.; Lucas, T. W.; Cioppa, T. M.: State-of-the-art review: a user's guide to the brave new world of designing simulation experiments. INFORMS Journal on Computing 17 (2005) 3, S. 263–289.

Klein, J.: Anwendung von Data Mining auf produktionslogistische Massendaten mit Schwerpunkt Verifikation und Validierung. Dortmund: Technische Universität Dortmund, Fachgebiet IT in Produktion und Logistik, Masterarbeit, 2017.

Klösgen, W.; Zytkow, J. M.: Knowledge discovery in databases terminology. In: Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. (Hg.): Advances in knowledge discovery and data mining. Menlo Park: AAAI, 1996, S. 573–592.

Knobloch, B.: Der Data-Mining-Ansatz zur Analyse betriebswirtschaftlicher Daten. Otto-Friedrich-Universität: Bamberg, Otto-Friedrich-Universität, 2000.

Knobloch, B.: A framework for organizational data analysis and organizational data mining. In: Nemati, H.; Barko, C. D. (Hg.): Organizational data mining. Hershey: IGI Global, 2003, S. 334–356.

Kohonen, T.: Self-organizing maps. 3. Aufl. Berlin: Springer, 2001.

Köster, C.: Vorgehen zur Berücksichtigung von Wissen zu Wirkzusammenhängen in Simulationsstudien für Supply Chains. Dortmund: Technische Universität Dortmund, Fachgebiet IT in Produktion und Logistik, Masterarbeit, 2015.

Krause, J.: PHP - Grundlagen und Lösungen: Webserver-Programmierung unter Windows und Linux. München: Hanser Fachbuch, 1999.

Krcmar, H.: Einführung in das Informationsmanagement. Berlin: Springer, 2011.

Kuhn, A.; Hellingrath, B.; Hinrichs, J.: Logistische Assistenzsysteme. In: ten Hompel, M. (Hg.): Software in der Logistik - Weltweit sichere Supply Chains. München: Huss, 2008, S. 20–26.

Küppers, B.: Data Mining in der Praxis: Ein Ansatz zur Nutzung der Potentiale von Data Mining im betrieblichen Umfeld. Frankfurt am Main: Peter Lang, 1999.

Kurgan, L. A.; Musilek, P.: A survey of knowledge discovery and data mining process models. The Knowledge Engineering Review 21 (2006) 1, S. 1–24.

Lambert, D. M.: Supply chain management: processes, partnerships, performance. 2. Aufl. Sarasota: Supply Chain Management Institute, 2005.

Lassmann, W.: Wirtschaftsinformatik: Nachschlagewerk für Studium und Praxis. Wiesbaden: Gabler, 2006.

Law, A. M.: Simulation modeling and analysis. 5. Aufl. New York: McGraw-Hill Education, 2014.

Le, T.; Phung, D.; Nguyen, K.; Venkatesh, S.: Fast one-class support vector machine for novelty detection. In: Cao, T.; Lim, E.-P.; Zhou, Z.-H.; Ho, T.-B.; Cheung, D.; Motoda, H. (Hg.): Advances in knowledge discovery and data mining. Bd. 9078. Cham: Springer, 2015, S. 189–200.

Lee, H. L.; Ng, S. M.: Introduction to the special issue on global supply chain management. Production and Operations Management 6 (1997) 3, S. 191–192.

Li, Y.: Anwendung von Data Mining auf produktionslogistischen Massendaten mit Schwerpunkt Datenvorverarbeitung. Dortmund: Technische Universität Dortmund,
Fachgebiet IT in Produktion und Logistik, Masterarbeit, 2017.

Lieber, D.; Erohin, O.; Deuse, J.: Wissensentdeckung im industriellen Kontext: Herausforderungen und Anwendungsbeispiele. Zeitschrift für wirtschaftlichen Fabrikbetrieb 108 (2013) 6, S. 388–393.

Liu, H.; Motoda, H.: Instance selection and construction for data mining. Boston: Springer US, 2001.

Lysons, K.; Farrington, B.: Purchasing and supply chain management. 7. Aufl. London: Financial Times Prentice Hall, 2005.

Maimon, O.; Rokach, L.: Data mining and knowledge discovery handbook. 2. Aufl. New York: Springer US, 2010.

Mariscal, G.; Marbán, Ó.; Fernández, C.: A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review 25 (2010) 2, S. 137–166.

März, L.; Krug, W.; Rose, O.; Weigert, G.: Simulation und Optimierung in Produktion und Logistik: Praxisorientierter Leitfaden mit Fallbeispielen. Berlin: Springer, 2011.

Mattfeld, D. C.; Vahrenkamp, R.: Logistiknetzwerke: Modelle für Standortwahl und Tourenplanung. 2. Aufl. Wiesbaden: Gabler, 2014.

Mayer, G.; Spieckermann, S.; Wenzel, S.: Steigerung der Produktivität in Simulationsstudien mit Assistenzwerkzeugen. Zeitschrift für wirtschaftlichen Fabrikbetrieb 107 (2012) 3, S. 174–177.

Meinhardt, I.; Sunarjo, M. F.; Marquardt, H.-G.: Bestimmung des stochastischen Zeitverhaltens in Supply Chains. In: Lasch, R.; Janker, C. G. (Hg.): Logistik Management. Wiesbaden: Deutscher Universitätsverlag, 2005, S. 123–133.

Mentzer, J. T.; DeWitt, W.; Keebler, J. S.; Min, S.; Nix, N. W.; Smith, C. D.; Zacharia, Z. G.: Defining supply chain management. Journal of Business Logistics 22 (2001) 2, S. 1–25.

Mertens, P.; Back, A.; Becker, J.; König, W.; Krallmann, H.; Rieger, B.; Scheer, A.-W.; Seibt, D.; Stahlknecht, P.; Strunz, H.; Thome, R.; Wedekind, H.: Lexikon der Wirtschaftsinformatik. 4. Aufl. Berlin: Springer, 2001.

Mertens, P.; Bodendorf, F.; König, W.; Picot, A.; Schumann, M.; Hess, T.: Grundzüge der Wirtschaftsinformatik. 11. Aufl. Berlin: Springer, 2012.

Messaoud, R. B.; Boussaid, O.; Rabaséda, S. L.: Mining association rules in OLAP cubes. In: Proceedings of the 1st international conference on innovations in information technology. Piscataway, NJ: IEEE, 2006, S. 1–5.

Meyr, H.; Stadtler, H.: Types of supply chains. In: Stadtler, H.; Kilger, C. (Hg.): Supply chain management and advanced planning. Berlin: Springer, 2005, S. 65–80.

Mierswa, I.; Wurst, M.; Klinkenberg, R.; Scholz, M.; Euler, T.: YALE: rapid prototyping for complex data mining tasks. In: Ungar, L.; Craven, M.; Gunopulos, D.; Eliassi-Rad, T. (Hg.): Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2006, S. 935–940.

Mohammadi, L.; Fazlollahtabar, H.; Mahdavil, I.; Tajdin, A.: Data mining on return items in a reverse supply chain. In: Proceedings of the 4th international conference on industrial engineering and operations management. Canton, MI: IEOM, 2014, S. 1467–1472.

Moody, D. L.; Kortink, M. A.: From enterprise models to dimensional models: a methodology for data warehouse and data mart design. In: Jeusfeld, M. A.; Shu, H.; Staudt, M.; Vossen, G. (Hg.): Proceedings of the 2nd international workshop on design and management of data warehouses. Stockholm: DMDW, 2000, S. 1–12.

Morik, K.; Klingspor, V.: Informatik kompakt: Eine grundlegende Einführung mit Java. Berlin: Springer, 2006.

Niemann, H.: Methoden der Mustererkennung. Frankfurt am Main: Akademische Verlagsgesellschaft, 1974.

North, K.: Wissensorientierte Unternehmensführung: Wertschöpfung durch Wissen. 5. Aufl. Lehrbuch. Wiesbaden: Gabler, 2011.

Nowitzky, J.: Partitionierungstechniken in Datenbanksystemen: Motivation und Überblick. Informatik-Spektrum 24 (2001) 6, S. 345–356.

Oedekoven, D.: Nutzenpotenziale harmonisierter Stammdaten in den Prozessen der Auftragsabwicklung von Auftragsfertigern. Aachen: Apprimus, 2011.

Otto, B.; Hüner, K. M.: Funktionsarchitektur für unternehmensweites Stammdatenmanagement. Universität St. Gallen (Hg.). Institut für Wirtschaftsinformatik: St. Gallen, 2009.

Parimala, N.; Pahwa, P.: Coalescing data marts. International Scholarly and Scientific Research & Innovation 11 (2008) 2, S. 3931–3936.

Parshutin, S.: Managing product life cycle with multiagent data mining system. In: Perner, P. (Hg.): Advances in data mining. applications and theoretical aspects. Berlin: Springer, 2010, S. 308–322.

Pelleg, D.; Moore, A.: X-means: extending K-means with efficient estimation of the number of clusters. In: Langley, P. (Hg.): Proceedings of the seventeenth international conference on machine learning. Burlington: Morgan Kaufmann, 2000, S. 727–734.

Petersohn, H.: Data Mining: Verfahren, Prozesse, Anwendungsarchitektur. München: Oldenbourg, 2005.

Pfeiffer, D.; Anwander, S.; Hellingrath, B.: A simulation approach to evaluate supply chain flexibility. In: Sprague, R. H. (Hg.): 46th Hawaii international conference on system sciences. Piscataway, NJ: IEEE, 2013, S. 1134–1143.

Piatetsky-Shapiro, P.: What main methodology are you using for your analytics, data mining, or data science projects? poll. 2014. URL: `http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html` (zuletzt geprüft am 06. 02. 2017).

Piazza, F.: Data Mining im Personalmanagement: Eine Analyse des Einsatzpotenzials zur Entscheidungsunterstützung. Wiesbaden: Gabler, 2010.

Piro, A.; Gebauer, M.: Definition von Datenarten zur konsistenten Kommunikation im Unternehmen. In: Hildebrand, K.; Gebauer, M.; Hinrichs, H.; Mielke, M. (Hg.): Daten- und Informationsqualität. Wiesbaden: Vieweg+Teubner, 2011, S. 143–156.

Probst, G.; Raub, S.; Romhardt, K.: Wissen managen: Wie Unternehmen ihre wertvollste Ressource optimal nutzen. 5. Aufl. Wiesbaden: Gabler, 2006.

Quinlan, J. R.: Induction of decision trees. Machine Learning 1 (1986) 1, S. 81–106.

Rabe, M.; Deininger, M.: State of art and research demands for simulation modeling of green supply chains. International Journal of Automation Technology 6 (2012) 3, S. 296–303.

Rabe, M.; Scheidler, A. A.: An approach for increasing the level of accuracy in supply chain simulation by using patterns on input data. In: Tolk, A.; Diallo, S. Y.; Ryzhov, I. O.; Yilmaz, L.; Buckley, S.; Miller, J. A. (Hg.): Proceedings of the 2014 winter simulation conference. Piscataway, NJ: IEEE, 2014, S. 1897–1906.

Rabe, M.; Scheidler, A. A.: Farming for Mining - Entscheidungsunterstützung mittels Simulation im Supply Chain Management. In: Rabe, M.; Clausen, U. (Hg.): Simulation in production and logistics 2015. Stuttgart: Fraunhofer, 2015, S. 671–679.

Rabe, M.; Spieckermann, S.; Wenzel, S.: Verifikation und Validierung für die Simulation in Produktion und Logistik: Vorgehensmodelle und Techniken. Berlin: Springer, 2008.

RapidMiner Inc.: RapidMiner, the industry's #1 open source predictive analytics platform. 2016. URL: `https://rapidminer.com/` (zuletzt geprüft am 06. 02. 2017).

Rashidi, L.; Rajasegarar, S.; Leckie, C.: An embedding scheme for detecting anomalous block structured graphs. In: Cao, T.; Lim, E.-P.; Zhou, Z.-H.; Ho, T.-B.; Cheung, D.; Motoda, H. (Hg.): Advances in knowledge discovery and data mining. Bd. 9078. Cham: Springer, 2015, S. 215–227.

Reiber, W.: Vom Fachexperten zum Wissensunternehmer: Wissenspotenziale stärker nutzen, die persönliche Wirksamkeit erhöhen. Wiesbaden: Gabler, 2013.

Reinartz, T.; Wirth, R.: The need for a task model for knowledge discovery in databases. In: Kodratoff, Y.; Nakhaeizadeh, G.; Taylor, C. (Hg.): Workshop notes statistics, machine learning, and knowledge discovery in databases. 1995, S. 19–24.

Riha, I. V.: Kosten- und leistungsoptimierter Betrieb kooperativer Logistiknetzwerke. In: Buchholz, P.; Clausen, U. (Hg.): Große Netze der Logistik. Berlin: Springer, 2009, S. 75–99.

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

Roy, R.: Industrial knowledge management: a micro-level approach. London: Springer, 2001.

Rubinstein, R. Y.: Optimization of computer simulation models with rare events. European Journal of Operational Research 99 (1997) 1, S. 89–112.

Rubinstein, R. Y.; Kroese, D. P.: Simulation and the Monte Carlo method. 2. Aufl. Hoboken: John Wiley & Sons, 2008.

Runkler, T. A.: Data Mining: Methoden und Algorithmen intelligenter Datenanalyse. Wiesbaden: Vieweg+Teubner, 2010.

Sanchez, S. M.: Simulation experiments: better data, not just big data. In: Tolk, A.; Diallo, S. Y.; Ryzhov, I. O.; Yilmaz, L.; Buckley, S.; Miller, J. A. (Hg.): Proceedings of the 2014 winter simulation conference. Piscataway, NJ: IEEE, 2014, S. 805–816.

Säuberlich, F.: KDD und Data Mining als Hilfsmittel zur Entscheidungsunterstützung. Frankfurt am Main: Peter Lang, 2000.

Schaffranietz, K.; Neumann, F.: Wissensgenerierung aus Datenbanken. In: Keuper, F.; Neumann, F. (Hg.): Wissens- und Informationsmanagement. Wiesbaden: Gabler, 2009, S. 149–177.

Schemm, J. W.: Zwischenbetriebliches Stammdatenmanagement: Lösungen für die Datensynchronisation zwischen Handel und Konsumgüterindustrie. Berlin: Springer, 2012.

Serdarasan, S.: A review of supply chain complexity drivers. Computers & Industrial Engineering 66 (2013) 3, S. 533–540.

Sharafi, A.: Knowledge Discovery in Databases: Eine Analyse des Änderungsmanagements in der Produktentwicklung. Wiesbaden: Gabler, 2013.

Siemens Industry Software GmbH: Siemens PLM Software. 2016. URL: https://www.plm.automation.siemens.com/ (zuletzt geprüft am 06. 02. 2017).

Smola, A. J.; Schölkopf, B.: A tutorial on support vector regression. Statistics and Computing 14 (2004) 3, S. 199–222.

Sowa, J. F.: Knowledge representation: Logical, philosophical, and computational foundations. Pacific Grove: Brooks Cole, 2000.

Steglich, M.; Feige, D.; Klaus, P.: Logistik-Entscheidungen: Modellbasierte Entscheidungsunterstützung in der Logistik mit LogisticsLab. 2. Aufl. München: De Gruyter Oldenbourg, 2016.

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

Stevens, G. C.: Integrating the supply chain. International Journal of Physical Distribution & Materials Management 19 (1989) 8, S. 3–8.

Stickel, E.; Groffmann, H.-D.; Rau, K.-H.: Gabler: Wirtschaftsinformatik Lexikon. Wiesbaden: Gabler, 1997.

Stolzle, W.; Otto, A.: Supply Chain Controlling in Theorie und Praxis: Aktuelle Konzepte und Unternehmensbeispiele. Wiesbaden: Gabler, 2003.

Su, W. Z.: Knowledge discovery in suppy chain transaction data by applying data farming. Dortmund: Technische Universität Dortmund, Fachgebiet IT in Produktion und Logistik, Masterarbeit, 2016.

Talia, D.; Trunfio, P.: Service-oriented distributed knowledge discovery. Boca Raton: CRC, 2013.

Tan, K. H.; Zhan, Y. Z.; Ji, G.; Ye, F.; Chang, C.: Harvesting big data to enhance supply chain innovation capabilities: an analytic infrastructure based on deduction graph. International Journal of Production Economics (2015) 165, S. 223–233.

ten Hompel, M.; Schmidt, T.: Warehouse-Management: Organisation und Steuerung von Lager- und Kommissioniersystemen. 3. Aufl. Berlin: Springer, 2008.

ten Hompel, M.; Heidenblut, V.: Taschenlexikon Logistik: Abkürzungen, Definitionen und Erläuterungen der wichtigsten Begriffe aus Materialfluss und Logistik. 3. Aufl. Berlin: Springer, 2011.

Terzi, S.; Cavalieri, S.: Simulation in the supply chain context: a survey. Computers in Industry 53 (2004) 1, S. 3–16.

Tiemeyer, E.: Handbuch IT-Management: Konzepte, Methoden, Lösungen und Arbeitshilfen für die Praxis. 2. Aufl. München: Carl Hanser, 2007.

Töpfer, A.: Lean Six Sigma: Erfolgreiche Kombination von Lean Management, Six Sigma und Design for Six Sigma. Berlin: Springer, 2009.

Turban, E.; Sharda R.; Delen, D.: Decision support and business intelligence systems. 9. Aufl. Upper Saddle River, N.J.: Prentice Hall, 2011.

Vahrenkamp, R.; Kotzab, H.: Logistik: Management und Strategien. 7. Aufl. München: Oldenbourg, 2012.

van der Aalst, W.: Process mining: discovery, conformance and enhancement of business processes. Berlin: Springer, 2011.

van der Aalst, W.; Weijters, A.: Process mining: a research agenda. Computers in Industry 53 (2004) 3, S. 231–244.

VDI-Richtlinie 3633 Blatt 1: Simulation von Logistik-, Materialfluss- und Produktionssystemen - Grundlagen. Beuth, 2014.

VDI-Richtlinie 3633 Blatt 3: Simulation von Logistik-, Materialfluß- und Produktionssystemen - Experimentplanung und -auswertung. Beuth, 1997.

Wagemann, A.: Wirkzusammenhänge beim Planparallelpolieren von Hochleistungskeramik. Dissertation. Aachen: Shaker, 1994.

Wang, R. Y.; Strong, D. M.: Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems 12 (1996) 4, S. 5–33.

Wannenwetsch, H.: Integrierte Materialwirtschaft und Logistik: Beschaffung, Logistik, Materialwirtschaft und Produktion. 4. Aufl. Berlin: Springer, 2010.

Wannenwetsch, H.; Nicolai, S.: E-Supply-Chain-Management: Grundlagen - Strategien - Praxisanwendungen. 2., überarb. und erw. Aufl. Wiesbaden: Gabler, 2004.

Weigel, N.: Datenqualitätsmanagement - Steigerung der Datenqualität mit Methode. In: Hildebrand, K.; Gebauer, M.; Hinrichs, H.; Mielke, M. (Hg.): Daten- und Informationsqualität. Wiesbaden: Vieweg+Teubner, 2011, S. 69–86.

Weiss, S. M.; Indurkhya, N.: Predictive data mining: a practical guide. Burlington: Morgan Kaufmann, 1998.

Wellbrock, W.: Innovative Supply-Chain-Management-Konzepte: Branchenübergreifende Bedarfsanalyse sowie Konzipierung eines Entwicklungsprozessmodells. Wiesbaden: Gabler, 2015.

Wenzel, S.; Abel, D.; Willmann, C.: Wissensarbeit in der Digitalen Fabrik - Der Zwiespalt zwischen Systematisierung und Kreativität. In: Spath, D. (Hg.): Wissensarbeit - Zwischen strengen Prozessen und kreativem Spielraum. Berlin: GITO, 2011, S. 251–276.

Wenzel, S.; Boyaci, P.; Jessen, U.: Simulation in production and logistics: trends, solutions and applications. In: Dangelmaier, W.; Blecken, A.; Delius, R.; Klöpfer, S. (Hg.): Advanced Manufacturing and Sustainable Logistics: 8th International Heinz Nixdorf Symposium, IHNS 2010, Paderborn, Germany, April 21-22, 2010. Proceedings. Berlin: Springer, 2010, S. 73–84.

Wenzel, S.; Weiß, M.; Collisi-Böhmer, S.; Pitsch, H.; Rose, O.: Qualitätskriterien für die Simulation in Produktion und Logistik: Planung und Durchführung von Simulationsstudien. Berlin: Springer, 2008.

Werner, H.: Supply Chain Management: Grundlagen, Strategien, Instrumente und Controlling. 4. Aufl. Wiesbaden: Gabler, 2010.

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

Wiendahl, H.-P.: Betriebsorganisation für Ingenieure. 5. Aufl. München: Carl Hanser, 2004.

Wilding, R.: The supply chain complexity triangle: uncertainty generation in the supply chain. International Journal of Physical Distribution & Logistics Management 28 (1998) 8, S. 599–616.

Winn, T.; Calder, P.: Is this a pattern? IEEE Software 19 (2002) 1, S. 59–66.

Wrobel, S.: Data Mining und Wissensentdeckung in Datenbanken. Künstliche Intelligenz 12 (1998) 1, S. 6–10.

Wrobel, S.; Wettschereck, D.; Verkamo, I.; Siebes, A.; Mannila, H.; Kwakkel, F.; Klösgen, W.: User interactivity in very large scale data mining. In: Dilger, W.; Schlosser, M.; Zeidler, J.; Ittner, A. (Hg.): Beiträge zum 9. Fachgruppentreffen Maschinelles Lernen der GI Fachgruppe 1.1.3. Chemnitzer Informatik-Berichte. Chemnitz: Technische Universität Chemnitz-Zwickau, 1996, S. 125–130.

Würthele, V. G.: Datenqualitätsmetrik für Informationsprozesse. Zürich: Technische
Hochschule Zürich, Dissertation, 2003.

Wyatt, J. L.; Petters, D. D.; Hogg, D.: From animals to robots and back: reflections on hard problems in the study of cognition: a collection in honour of Aaron Sloman. Bd. 22. Cham: Springer, 2014.

Yan, X.; Su, X. G.: Linear regression analysis: theory and computing. Singapore: World Scientific, 2009.

Ziegler, J.: Systematische Untersuchung von möglichen Datenkategorien in Supply Chains. Dortmund: Technische Universität Dortmund, Fachgebiet IT in Produktion und Logistik, Bachelorarbeit, 2015.

Zimmermann, R.: Statistische Versuchsplanung für Data Farming-Konzepte in Tecnomatix Plant Simulation. Dortmund: Technische Universität Dortmund, Fachgebiet IT in Produktion und Logistik, Bachelorarbeit, 2016.

# Abbildungsverzeichnis

# Tabellenverzeichnis

# Abkürzungsverzeichnis

| | |
|---|---|
| ASIM | Arbeitsgemeinschaft Simulation |
| BE | bewegliches Element |
| BO | Business-Objekt |
| Crisp-DM | Cross Industry Standard Process for Data Mining |
| CSV | Comma-separated values |
| DES | Discrete Event Simulation |
| DM | Data Mining |
| DMM | Data-Mining-Model |
| ERM | Entity-Relationship-Modellen |
| ETL | Extraktion-Transformation-Laden |
| GPU | Grafikprozessoren |
| GUID | Seriennummern der gefertigten Produkte |
| ID | Identifikatoren |
| ITPL | IT in Produktion und Logistik |
| IuK | Informations- und Kommunikationstechnologien |
| KDD | Knowledge Discovery in Databases |
| KNN | Künstliche Neuronale Netze |
| MESC | Methode zur Musterextraktion in SCs |
| RFID | Radio-frequency identification |
| RIPPER | Repeated Incremental Pruning to Produce Error Reduction |
| RM | Referenzmodell |
| SC | Supply Chain |
| SCM | Supply Chain Management |
| SCOR | Supply-Chain-Operations-Reference-Modells |
| SEMMA | Sample, Explore, Modify, Model and Assess |
| SOM | Self-Organizing Maps |
| SQL | Structured Query Language |
| V&V | Verifikation und Validierung |
| VDI | Verein Deutscher Ingenieure |
| XML | Extensible Markup Language |

# Symbolverzeichnis

| | |
|---|---|
| $D$ | Datenbank |
| $\mathcal{G}$ | Subtypklasse logisch-numerisch |
| $\mathcal{H}$ | Subtypklasse statistisch |
| $k$ | Parameter |
| $K(x)$ | Klassifikationsfunktion |
| $n, m, w$ | Zählvariablen |
| $\mathcal{P}$ | Subtypklasse elementar |
| $t_i$ | Transaktion |

# Anhang A: Tabellen zum Stand der Wissenschaft

**Tabelle A.1: Datentypen in Anlehnung an Dyer (2008) und Krause (1999)**

| **Bezeichnung** | **Wertebereich/ Genauigkeit** | **Speicher-bedarf** | **Anmerkungen** |
|---|---|---|---|
| **Numeric** | | | |
| **Bezeichnung** | **Wertebereich/ Genauigkeit** | **Speicher-bedarf** | **Anmerkungen** |
| TINYINT | $0 \ldots 255$ | 1 Byte | Ganzzahl |
| SMALLINT | $-2^{15} \ldots 2^{15} - 1$ | 2 Bytes | Ganzzahl |
| MEDIUMINT | $-2^{23} \ldots 2^{23} - 1$ | 3 Bytes | Ganzzahl |
| INT | $-2^{31} \ldots 2^{31} - 1$ | 4 Bytes | Ganzzahl |
| BIGINT | $-2^{63} \ldots 2^{63} - 1$ | 8 Bytes | Ganzzahl |
| DECIMAL[M[,D]] | Abhängig von Parametern M, D | • M+2 Bytes: $D > 0$<br>• M+1 Bytes: $D = 0$<br>• D+2 Bytes: $M < D$ | |
| FLOAT[M,D] | Wertebereich uneinheitlich | 4 Bytes | |
| DOUBLE[M,D] | $-1.798E + 308 \ldots 1.798E + 308$ | 8 Bytes | |
| REAL | siehe DOUBLE | 8 Bytes | |
| BIT[M] | von M abhängig | 1 Bit ... 8 Byte | Definiert eine M-Bit lange Zahl |
| BOOLEAN | Null, 0, 1 | 1 Byte | Ist ein TINYINT |
| SERIAL | $0 \ldots 2^{64} - 1$ | 8 Bytes | Wie BIGINT UNSIGNED NOT NULL |

**Tabelle A.1: Datentypen in Anlehnung an Dyer (2008) und Krause (1999) (Fortsetzung)**

| **Bezeichnung** | **Wertebereich/ Genauigkeit** | **Speicher-bedarf** | **Anmerkungen** |
|---|---|---|---|
| **Date and Time** | | | |
| **Bezeichnung** | **Wertebereich/ Genauigkeit** | **Speicher-bedarf** | **Anmerkungen** |
| DATE | 01.01.1000 ... 31.12.9999 | 3 Bytes | Format: yyyy.mm.dd |
| DATETIME | 01.01.1000 ... 31.12.9999 | 8 Bytes | Format: yyyy.mm.dd hh:mm:ss |
| TIMESPAMP | 01.01.1970 ... 19.01.2038 | 4 Bytes | Format: yyyy.mm.dd hh:mm:ss |
| TIME | -838:59:59 ... 838:59:59 | 3 Bytes | Format: hh:mm:ss |
| YEAR | 1901 ... 2155 | 1 Byte | Format: yyyy |
| **Strings** | | | |
| **Bezeichnung** | **Wertebereich/ Genauigkeit** | **Speicher-bedarf** | **Anmerkungen** |
| CHAR[M] | Zeichenkette bis Länge von 255 | M Bytes | M definiert die Zeichenkettenlänge |
| VARCHAR[M] | Zeichenkette bis Länge von 255 | L Bytes | L ist die tatsächliche Länge der Zeichenkette |
| TINYTEXT | $0 \ldots 255$ | L+1 Bytes | L:Textlänge |
| TEXT | $0 \ldots 2^{16} - 1$ | L+2 Bytes | L:Textlänge |
| MEDIUMTEXT | $0 \ldots 2^{24} - 1$ | L+3 Bytes | L:Textlänge |
| LONGTEXT | $0 \ldots 2^{32} - 1$ | L+4 Bytes | L:Textlänge |
| BINARY[M] | M von $0 \ldots 255$ | M Bytes | Binärer String mit Länge M |
| VARBINARY[M] | M von $0 \ldots 255$ | M Bytes | Ist Binary |
| TINYBLOB[M] | variable Daten | L+1 Bytes | L:Länge, wie TINYTEXT |
| BLOB | variable Daten | L+2 Bytes | L:Länge, wie TEXT |
| MEDIUMBLOB | variable Daten | L+3 Bytes | L:Länge, wie MEDIUMTEXT |

**Tabelle A.1: Datentypen in Anlehnung an Dyer (2008) und Krause (1999) (Fortsetzung)**

| Bezeichnung | Wertebereich/ Genauigkeit | Speicher-bedarf | Anmerkungen |
|---|---|---|---|
| LONGBLOB | variable Daten | L+4 Bytes | L:Länge, wie LONG-TEXT |
| ENUM | bis max. $2^{16}$ Einträge | 1 oder 2 Bytes | Liste für Werte |
| SET | bis max. 64 Einträge | 1-8 Bytes | Liste für Strings |

**Tabelle A.2: Literaturbeispiele - Einsatz von Data-Mining-Verfahren im SCM nach Su (2016)**

| | Managementbereiche | Spezifische Aufgabe und Data-Mining-Einsatz |
|---|---|---|
| **SCOR-Prozesse** | | |
| Beschaffen | Lieferantenmanagement | Entscheidungsbäume: Lieferantenprobleme in Verbindung mit der Beurteilung der Wahrscheinlichkeit für jedes auftretende Ergebnis reduzieren (Choy et al. 2003) |
| Planen | Erweiterte Planung und Optimierung | Regressionsanalyse und statistische Methoden: Benötigte Mengeneinheiten prognostizieren (Lysons und Farrington 2005) |
| Herstellen | Produktionsintegration und -intelligenz | Assoziationsregeln: Grundursache der Produktfehler identifizieren, Fertigungskapazität optimieren, zustandsorientierte Instandhaltung ermöglichen (Turban et al. 2011) |
| Liefern | Transport Management System, Warehousemanagement | Genetische Algorithmen: Hypothesengenerierung für den lieferantengesteuerten Bestand unter unsicheren Nachfragesituationen (Borade und Sweeney 2015) |
| Rückliefern | Rückführendes Logistikmanagement | Clusteralgorithmen: Verbesserung der Prozessqualität mittels Kategorisierung der rückgelieferten Rohstoffe durch den k-Means Algorithmus (Mohammadi et al. 2014) |

**Tabelle A.2: Literaturbeispiele - Einsatz von Data-Mining-Verfahren im SCM nach Su (2016) (Fortsetzung)**

| | Managementbereiche | Spezifische Aufgabe und Data-Mining-Einsatz |
|---|---|---|
| **SC bezogene Prozesse** | | |
| Konstruktion und Design | Produktlebenszyklusmanagement | Multi-Agente-Data-Mining-Systeme: Entscheidungen für die Produktionsplanung aufgrund von historischen Nachfragewerten sowie dem Produktlebenszyklus treffen (Parshutin 2010) |
| Vertrieb, Marketing, Kundenservice | Customer Relationship, Kundendienst, Ersatzteilmanagement | Cluster-Algorithmen: Clusterung von Unternehmenskunden mittels demographischen Verteilung und Einkaufsverhalten (Turban et al. 2011) |

**Tabelle A.3: Bewertungsgrundlagen im Überblick nach Küppers (1999)**

| Kriterium | Problem | Bewertungsgrundlage |
|---|---|---|
| **Anwenderorientierte Kriterien** | | |
| Interessantheit | • Maschinelle Bewertung der Interessantheit<br>• Behandlung von Anomalien | • Fokussierungsmöglichkeit auf interessante Muster<br>• Einbindung von Hintergrundwissen |
| Interpretierfähigkeit/ Verständlichkeit | • Ergebnisse in verständlicher Form darstellen und Transparenz der Methode | • Methodische Transparenz |
| Autonomiegrad | • Spannungsfeld zwischen allgemeiner Verwendbarkeit und Autonomie | • Nicht anwendbar, da von der Parametrisierung abhängig |
| **Methodenorientierte Kriterien** | | |
| Charakterisierung von Unsicherheit | • Fehlende Unsicherheitsmaße zu jedem ermittelten Ergebnis | • Vorhandensein bzw. Integration statistischer Maßzahlen |
| Explizite und implizite Annahmen | • Anwendbarkeit von Methoden hängt stark von den zu analysierenden Daten ab | • Vorhandensein von Beschränkungen |
| Regularisierung (Über-/ Unteranpassung) | • Methodisches Modell passt zu gut bzw. schlecht auf das Problem | • Nicht anwendbar, da von der Parametrisierung abhängig |
| **Datenorientierte Kriterien** | | |
| Datendeformation | • Vorverarbeitung der Daten teilweise notwendig | • Notwendigkeit der Vorverarbeitung |
| Datenqualität | • Reale Datenbestände haben häufig eine mindere Datenqualität | • Empfindlichkeit auf mindere Datenqualität |
| Verarbeitbare Datenmenge/ Laufzeitverhalten | • Große Datenmengen sind teilweise nicht verarbeitbar | • Notwendigkeit der Performance-Verbesserung |

**Tabelle A.4: Beispielhafter Einsatz von möglichen V&V-Techniken in MESC**

| Phase | Prüfschritt | Exemplarische V&V-Techniken | Exemplarische Literatur |
|---|---|---|---|
| Aufgabendefinition | 1,1: Prüfung auf Vollständigkeit und Plausibilität der gegebenen Randbedingungen und festgelegten Zielbedingungen | • Validierung im Dialog<br>• Begutachtung<br>• Strukturiertes Durchgehen | Rabe, Spieckermann et al. (2008) |
| Auswahl der relevanten Datenbestände | 2,2: Prüfung und Relevanz der verwendeten Datenquellen und Datenbestände | • Validierung im Dialog<br>• Schreibtischtest<br>• Begutachtung | Rabe, Spieckermann et al. (2008) |
| | 2,1: Prüfung, ob die ausgewählten Datenquellen und Datenbestände für das Erreichen der Zielbedingung geeignet sind | • Dokumentenüberprüfung<br>• Inspektion<br>• Audit<br>• Test von Teilmodellen | Balci (1998) |
| Datenaufbereitung | 3,3: Prüfung, ob die vorliegenden Daten entsprechend transformiert worden sind | • Begutachtung<br>• Strukturiertes Durchgehen | Rabe, Spieckermann et al. (2008) |
| | 3,2: Prüfung der Datenverarbeitung gegen den relevanten Datenbestand | • Validierung im Dialog<br>• Schreibtischtest | Rabe, Spieckermann et al. (2008) |
| | 3,1: Prüfung, ob die aufbereiteten Daten für das Erreichen der Zielbedingung geeignet sind | • Dokumentenüberprüfung<br>• Inspektion<br>• Audit<br>• Test von Teilmodellen | Balci (1998) |
| Vorbereitung des Data-Mining-Verfahrens | 4,4: Prüfung auf geeignete Auswahl des Data-Mining-Verfahren | • Vergleich mit anderen Modellen<br>• Test von Teilmodellen | Rabe, Spieckermann et al. (2008) |

**Tabelle A.4: Beispielhafter Einsatz von möglichen V&V-Techniken in MESC (Fortsetzung)**

| Phase | Prüfschritt | Exemplarische V&V-Techniken | Exemplarische Literatur |
|---|---|---|---|
| | 4,3: Prüfung, ob die Datenaufbereitung als Vorbereitung für das Data-Mining-Verfahren ausreichend ist | • Begutachtung<br>• Strukturiertes Durchgehen | Rabe, Spiekermann et al. (2008) |
| | 4,2: Prüfung, ob die Auswahl der relevanten Daten den Anforderungen des ausgewählten Data-Mining-Verfahrens entspricht | • Validierung im Dialog<br>• Strukturiertes Durchgehen<br>• Test von Teilmodellen | Rabe, Spiekermann et al. (2008) |
| | 4,1: Prüfung, ob das Data-Mining-Verfahren für die Erfüllung der Aufgabenstellung geeignet ist | • Dokumentenüberprüfung<br>• Inspektion<br>• Audit<br>• Test von Teilmodellen | Balci (1998) |
| Anwendung der Data-Mining-Verfahren | 5,5: Prüfung auf richtige Anwendung des Data-Mining-Verfahrens | • Vergleich mit anderen Modellen<br>• Ursachen-Wirkungs-Graph<br>• DM-Modellbewertung mittels Testdaten (z. B. Kreuzvalidierung) | Rabe, Spiekermann et al. (2008) und Steinlein (2004) |
| | 5,4: Prüfung, ob das Data-Mining-Verfahren für ihre Anwendung zuvor richtig vorbereitet worden ist | • Begutachtung<br>• Schreibtischtest<br>• Strukturiertes Durchgehen | Rabe, Spiekermann et al. (2008) |

**Tabelle A.4: Beispielhafter Einsatz von möglichen V&V-Techniken in MESC (Fortsetzung)**

| Phase | Prüfschritt | Exemplarische V&V-Techniken | Exemplarische Literatur |
|---|---|---|---|
| | 5,3:Prüfung, ob die Datenaufbereitung für die Anwendung des Data-Mining-Verfahrens geeignet ist | • Dokumentenüberprüfung<br>• Inspektion<br>• Audit<br>• Test von Teilmodellen | Balci (1998) |
| | 5,2: Prüfung, ob die Datenselektion eine fachgerechte Anwendung des Data-Mining-Verfahrens ermöglicht | • Validierung im Dialog<br>• Visualisierung<br>• Strukturiertes Durchgehen | Collier et al. (2002) und Rabe, Spiekermann et al. (2008) |
| | 5,1: Prüfung, ob durch die Anwendung des Data-Mining-Verfahrens die Zielbedingungen erfüllt werden | • Dokumentenüberprüfung<br>• Inspektion<br>• Audit<br>• Test von Teilmodellen | Balci (1998) |
| Weiterverarbeitung der Data-Mining-Ergebnisse | 6,6: Prüfung auf ordnungsgemäße Weiterverarbeitung der Data-Mining-Ergebnisse | • Vergleich mit anderen Modellen<br>• Test von Teilmodellen<br>• Visualisierung | Balci (1998), Collier et al. (2002) und Rabe, Spiekermann et al. (2008) |
| | 6,5: Prüfung, ob die Anwendung der Data-Mining-Verfahren interpretierbare Daten als Resultat liefert | • Review<br>• Vergleich mit anderen Modellen<br>• Test von Teilmodellen | Rabe, Spiekermann et al. (2008) |
| | 6,4: Prüfung, ob das ausgewählte Data-Mining-Verfahren interpretierbare Daten liefern kann | • Review<br>• Test von Teilmodellen<br>• Visualisierung | Collier et al. (2002) und Rabe, Spiekermann et al. (2008) |

**Tabelle A.4: Beispielhafter Einsatz von möglichen V&V-Techniken in MESC (Fortsetzung)**

| Phase | Prüfschritt | Exemplarische V&V-Techniken | Exemplarische Literatur |
|---|---|---|---|
| | 6,3: Prüfung, ob die Daten für die Interpretation fachlich richtig aufbereitet worden sind | • Begutachtung<br>• Schreibtischtest<br>• Validierung im Dialog | Rabe, Spieckermann et al. (2008) |
| | 6,2: Prüfung, ob die Datenselektion für die Interpretation ausreichend ist, oder ob andere Daten für die Interpretation selektiert werden müssen | • Begutachtung<br>• Schreibtischtest<br>• Strukturiertes Durchgehen | Rabe, Spieckermann et al. (2008) |
| | 6,1: Prüfung, ob die durch Interpretation gewonnenen Erkenntnisse den vordefinierten Zielen der Aufgabenstellung genügen | • Dokumentenüberprüfung<br>• Inspektion<br>• Audit<br>• Test von Teilmodellen | Balci (1998) |
| Bewertung des Data-Mining-Prozesses | 7,5: Prüfung, ob die Anwendung des Data-Mining-Verfahrens genügend dokumentiert ist | • Dokumentenüberprüfung<br>• Schreibtischtest<br>• Strukturiertes Durchgehen | Balci (1998) und Rabe, Spieckermann et al. (2008) |
| | 7,4: Prüfung, ob die Dokumentation für den Auswahlprozess des Data-Mining-Verfahrens ausreichend ist | • Begutachtung<br>• Schreibtischtest<br>• Strukturiertes Durchgehen | Rabe, Spieckermann et al. (2008) |
| | 7,3: Prüfung, ob der Datenaufbereitungsprozess und die Prozessergebnisse ausreichend dokumentiert sind | • Begutachtung<br>• Schreibtischtest<br>• Strukturiertes Durchgehen | Rabe, Spieckermann et al. (2008) |

**Tabelle A.4: Beispielhafter Einsatz von möglichen V&V-Techniken in MESC (Fortsetzung)**

| Phase | Prüfschritt | Exemplarische V&V-Techniken | Exemplarische Literatur |
|---|---|---|---|
| | 7,2: Prüfung, ob die Dokumentation des Datenauswahlprozesses und der Prozessergebnisse ausreichend ist | • Begutachtung<br>• Schreibtischtest<br>• Strukturiertes Durchgehen | Rabe, Spieckermann et al. (2008) |
| | 7,1: Prüfung, ob die Aufgabenstellung unter Berücksichtigung von Randbedingungen und Zielkriterien ausreichend dokumentiert ist | • Strukturiertes Durchgehen<br>• Audit | Balci (1998) und Rabe, Spieckermann et al. (2008) |

# Anhang B: Datenblätter und Experimente

**Tabelle B.1: Beschreibung des ersten Datenbestands**

| Kriterien | Beschreibung |
|---|---|
| **Fachliche Kriterien** | |
| Branche | • Zulieferer Automobilbranche |
| Fachliche Beschreibung | • Lieferdaten von 197 Lieferanten an 23 Standorte<br>• Zeitraum vom 27.01.2009 bis zum 13.09.2012<br>• Bestelländerungen zusätzlich enthalten |
| Exemplarische Attributsbeschreibung | • OrderNumber: ID-Nummer der Bestellung<br>• StepCounter: Teilschritt einer Bestellung<br>• ProductCode: ID-Nummer des bestellten Produkts<br>• Shipment: Beschreibung des Transportwegs<br>• DateOfTransaction: Datum und Zeitpunkt der Erstellung der Transaktion |
| **Technische Kriterien** | |
| Quelle | • Data Warehouse |
| Umfang | • 1 570 121 Datensätze<br>• 67 Attribute in vier Tabellen |
| Technische Attributsbeschreibung | Datentypen nach Tabelle A.1:<br>• Bestellnummer: Zeichenkette<br>• Schrittzähler: Ganzzahl<br>• Produktcode: Zeichenkette<br>• Transportweg: Zeichenkette<br>• Transaktionsdatum: Datetime |

**Tabelle B.1: Beschreibung des ersten Datenbestands (Fortsetzung)**

| Kriterien | Beschreibung |
|---|---|
| **Untersuchungskriterien** | |
| Analysenotizen | • Produktcode besitzt innere Struktur<br>• Multilinguale Attributsausprägungen, daher doppelte Einträge mit identischem Inhalt<br>• Bestellungen bestehen aus mehreren Transaktionen |
| Primärfrage | • Ist es möglich Verspätungen in der SC aufgrund der gegebenen Daten vorherzusagen? |

**Tabelle B.2: Beschreibung des zweiten Datenbestands**

| Kriterien | Beschreibung |
|---|---|
| **Fachliche Kriterien** | |
| Branche | • Lebensmittelindustrie |
| Fachliche Beschreibung | • Lieferdaten<br>• Zeitraum vom 01.10.2014 bis zum 31.03.2015<br>• Mitarbeiter und Lagerpositionen<br>• Bestelländerungen zusätzlich enthalten |
| Exemplarische Attributsbeschreibung | • TA-Nummer: ID-Nummer der Bestellung<br>• TA-Position: Zähler des aktuellen Bestellschritts<br>• Mitarbeiter: Personalnummer des Bearbeiters<br>• Artikel: ID-Nummer des bestellten Produkts<br>• Bewegungsart: Beschreibung des Transportwegs<br>• Soll-Menge: Erwartete Liefermenge<br>• Ist-Menge: Tatsächliche Liefermenge |
| **Technische Kriterien** | |
| Quelle | • SAP-System |
| Umfang | • 109 579 Datensätze<br>• 15 Attribute in einer Tabelle |

**Tabelle B.2: Beschreibung des zweiten Datenbestands (Fortsetzung)**

| **Kriterien** | **Beschreibung** |
|---|---|
| Technische Attributsbeschreibung | Datentypen nach Tabelle A.1:<br>• Transaktionsnummer: Ganzzahl<br>• Transaktionsposition: Ganzzahl<br>• Mitarbeiter: Zeichenkette<br>• Artikel: Ganzzahl<br>• Bewegungsart: Ganzzahl<br>• Soll-Menge: Gleitkommazahl(DOUBLE)<br>• Ist-Menge: Gleitkommazahl(DOUBLE) |
| **Untersuchungskriterien** | |
| Analysenotizen | • Attributsausprägungen sind überwiegend Kürzel<br>• Zusammengesetzter Schlüssel aus Transaktionsnummer und Transaktionsposition |
| Primärfrage | • Nach welchen Merkmalen können Artikel gruppiert werden? |

**Tabelle B.3: Auszug aus Datenbestand B.1 mit Fabrikationsnummer**

| Transaktions-nummer | Fabrikationsnum-mer | Produktbeschrei-bung | Bestell-menge | Zielort | Lieferweg |
|---|---|---|---|---|---|
| 37 721 426 | 56 410 571AV | 7 886 048 | 5 | Italien 3 | OUR FORWARDER |
| 37 678 834 | 48 120 361AT | 7 886 058 | 10 | Italien 4 | OUR FORWARDER |
| 37 721 434 | 59 520 773BT | 7 886 062 | 5 | Italien 4 | OUR FORWARDER |
| 45 365 022 | 74 110 361A | 9 313 018 | 100 | Italien 2 | OUR FORWARDER |
| 45 920 982 | 18 920 101A | 9 310 530 | 280 | Italien 2 | BY CARRIER |
| 46 388 632 | 48 410 901A | 9 315 224 | 10 | Italien 1 | YOUR FORWARDER |
| 45 097 918 | 46 013 311A | 9 315 538 | 100 | Italien 1 | OUR FORWARDER |
| 46 389 098 | 49 421 431A | 9 315 644 | 240 | Italien 2 | YOUR FORWARDER |
| 63 066 892 | 59 820 991A | 13 051 426 | 192 | Thailand 1 | YOUR FORWARDER |
| 63 070 080 | 34 420 392A | 12 361 502 | 50 | Thailand 1 | OUR FORWARDER |
| 56 018 238 | 51 017 561A | 11 225 378 | 100 | Italien 2 | BY SEA |
| 65 854 764 | 57 112 751B | 13 603 774 | 70 | Italien 1 | BY TRUCK |
| 66 757 866 | 28 240 893A | 13 757 652 | 176 | Italien 1 | BY SEA |
| 66 757 942 | 47 110 252BA | 13 463 492 | 80 | Italien 3 | YOUR FORWARDER |
| 66 757 958 | 22 522 451E | 13 763 706 | 27 | Italien 1 | YOUR FORWARDER |
| 66 757 998 | 47 120 121D | 13 615 436 | 30 | Thailand 1 | YOUR FORWARDER |
| 66 758 018 | 82 510 422A | 13 291 468 | 100 | Italien 1 | BY TRUCK |
| 66 758 094 | 59 520 091AQ | 13 608 038 | 10 | Italien 2 | OUR FORWARDER |

**Tabelle B.4: Auszug aus Datenbestand B.1 mit aufgesplitteter Fabrikationsnummer**

| Transaktionsnummer | Seriennummer | Sorte | Produktbeschreibung | Bestellmenge | Zielort |
|---|---|---|---|---|---|
| 37 721 426 | 56 410 571 | AV | 7 886 048 | 5 | Italien 3 |
| 37 678 836 | 48 120 351 | AT | 7 886 056 | 10 | Italien 4 |
| 37 678 834 | 48 120 361 | AT | 7 886 058 | 10 | Italien 4 |
| 37 721 434 | 59 520 773 | BT | 7 886 062 | 5 | Italien 4 |
| 45 365 022 | 74 110 361 | A | 9 313 018 | 100 | Italien 2 |
| 45 920 982 | 18 920 101 | A | 9 310 530 | 280 | Italien 2 |
| 46 388 632 | 48 410 901 | A | 9 315 224 | 10 | Italien 1 |
| 45 097 918 | 46 013 311 | A | 9 315 538 | 100 | Italien 1 |
| 46 389 098 | 49 421 431 | A | 9 315 644 | 240 | Italien 2 |
| 63 066 892 | 59 820 991 | A | 13 051 426 | 192 | Thailand 1 |
| 63 070 080 | 34 420 392 | A | 12 361 502 | 50 | Thailand 1 |
| 56 018 238 | 51 017 561 | A | 11 225 378 | 100 | Italien 2 |
| 65 854 764 | 57 112 751 | B | 13 603 774 | 70 | Italien 1 |
| 66 757 866 | 28 240 893 | A | 13 757 652 | 176 | Italien 1 |
| 66 757 942 | 47 110 252 | BA | 13 463 492 | 80 | Italien 3 |
| 66 757 958 | 22 522 451 | E | 13 763 706 | 27 | Italien 1 |
| 66 757 998 | 47 120 121 | D | 13 615 436 | 30 | Thailand 1 |
| 66 758 018 | 82 510 422 | A | 13 291 468 | 100 | Italien 1 |
| 66 758 094 | 59 520 091 | AQ | 13 608 038 | 10 | Italien 2 |

In der aufgeführten Tabelle ist die Fabrikationsnummer in die Attribute „Seriennummer“ und „Sorte“ unterteilt worden.

**Tabelle B.5: Vergleich von Clusterverfahren auf dem Datenbestand B.1**

| Clusterverfahren | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Marker |
|---|---|---|---|---|---|---|---|---|---|---|---|
| k-Means | 34 280 | 25 974 | 5 154 | 18 652 | 201 | 93 | 730 | 8 973 | 3 567 | 2 376 | 2 |
| Expectation Maximization Clustering | 34 945 | 1 803 | 60 569 | 72 | 27 | 4 | 1 | 29 | 203 | 2 347 | 1 |
| Random Clustering | 9 871 | 10 050 | 9 936 | 9 927 | 9 994 | 10 039 | 10 172 | 10 021 | 9 962 | 10 028 | 8 |
| X-Means | 35 148 | 62 476 | 163 | 2 213 | - | - | - | - | - | - | 1 |
| Support Vector Clustering | 96 390* | 51 | 29 | 60 | 80 | 90 | 130 | 108 | 42 | 110 | 1 |

Die Experimente umfassen einen Datenbestand von 100 000 Entitäten und wurde mit k = 10 initiiert (sofern parametrierbar). Die Clusternummerierung wurde gemäß der größten Übereinstimmung der Cluster ausgewählt, damit die Nummerierung die Clusterinhalte und nicht die internen Algorithmusparameter widerspiegelt. Der Marker zeigt eine ausgewählte Entität und den ihr zugeordneten Cluster.

* Dieser Cluster wird als Noise klassifiziert.

**Tabelle B.6: Vergleich von Clusterverfahren auf dem Datenbestand B.2**

| Clusterverfahren | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Marker |
|---|---|---|---|---|---|---|---|---|---|---|---|
| k-Means | 7 405 | 12 933 | 12 100 | 7 652 | 14 623 | 11 817 | 11 228 | 1 935 | 8 770 | 11 537 | 9 |
| Expectation Maximization Clustering | 13 056 | 14 060 | 4 964 | 9 932 | 5 101 | 9 994 | 20 112 | 12 932 | 9 759 | 1 090 | 2 |
| Random Clustering | 9 898 | 10 023 | 9 980 | 9 946 | 10 004 | 9 907 | 10 119 | 10 088 | 10 015 | 10 020 | 3 |
| X-Means | 1 701 | 32 211 | 43 954 | 22 134 | - | - | - | - | - | - | 2 |
| Support Vector Clustering | 100 000* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Die Experimente umfassen einen Datenbestand von 100 000 Entitäten und wurde mit k = 10 initiiert (sofern parametrierbar). Die Clusternummerierung wurde gemäß der größten Übereinstimmung der Cluster ausgewählt, damit die Nummerierung die Clusterinhalte und nicht die internen Algorithmusparameter widerspiegelt. Der Marker zeigt eine ausgewählte Entität und den ihr zugeordneten Cluster.

* Dieser Cluster wird als Noise klassifiziert.

**Tabelle B.7: Parametrierung FPGrowth und Assoziationslerner in RapidMiner für Experimente zur Datengenerierung und simulationsunterstützten Validierung in Abschnitt 6.3**

| Verfahren | Parameter | Wert |
|---|---|---|
| Discretize by Binning | bins | 5 |
| FPGrowth | min number of item sets | 10 |
| FPGrowth | min support | 0,1 |
| FPGrowth | must contain | Verspätung = true |
| AssociationRules | criterion | confidence |
| AssociationRules | min confidence | 0,8 |