technische universität dortmund

IT in Produktion und Logistik

Fachwissenschaftliche Projektarbeit

Validierung eines Datentransformationsprozesses von einer Datenquelle in eine andere Datenquelle

Felix Steinfurth
Matrikelnummer: 177508
Studiengang: Maschinenbau BA

ausgegeben am:
27.04.2018
eingereicht am:
27.10.2018

Erstprüfer: **Prof. Dr. Ing. Markus Rabe**

Zweitprüfer: **Astrid Klüter, M. Sc.**

Inhalt

1 Einleitung

Im Zuge der Digitalisierung spielen Online-Versandhändler eine immer größer werdende Rolle. Diese ersetzen nicht nur die klassischen Warenhäuser, sondern zunehmend auch Supermärkte. Dadurch steigen der logistische Aufwand und die Anforderungen an Distributionssysteme. Um auf die Anforderungen zu reagieren, werden Analysemethoden zur Entscheidungsunterstützung hinzugezogen. Dabei haben sich Simulationsverfahren etabliert (Rabe et al. 2017). Mit Hilfe der Simulation sollen die Distributionssysteme abgebildet und Möglichkeit zur Verbesserung der Prozesse und Strukturen gegeben werden. Erhobene Daten aus dem zu analysierenden System geben ein genaues Bild der Lieferkette wieder und ermöglichen somit eine detailgetreue Simulation. Da ein Großteil der Simulationsmodelle anhand von Daten erstellt wird und diese für die jeweilige Software in einem speziellen Format benötigt werden, ist ein Datentransformationsprozess ein unverzichtbarer Bestandteil jeder Datenbeschaffung (Laroque et al. 2013, S. 218). Eine Überprüfung des Transformationsprozesses mittels geeigneter Methoden ist dabei unumgänglich, da die Qualität der Daten direkten Einfluss auf die Simulationsergebnisse und damit auf den Erfolg einer Simulationsstudie hat (Wenzel et al. 2008, S. 119).

Am Fachgebiet IT in Produktion und Logistik wird an einem Simulationsprojekt geforscht, welches verschiedene Szenarien des Warenflusses eines Online-Supermarkts untersucht und bewertet. Ein Datensatz dient als Grundlage für die Simulation einer Supply-Chain, die die Logistik abbildet. Die Daten wurden bereits durch einen Datentransformationsprozess spezialisiert und in eine Form überführt, die in einer Simulationssoftware verwendet werden kann.

Ziel dieser fachwissenschaftlichen Ausarbeitung ist eine Validierung des Datentransformationsprozesses, der mit Hilfe des Simulationsschnittstelle SimChain erfolgte. Die Validierung soll sicherstellen, dass sich Modell und Realität in bestimmten Aspekten möglichst ähnlich verhalten (VDI 3633 - Blatt 1). Zudem wird untersucht, welche Defizite die Datenquelle aufweist. Hier wird veranschaulicht, ob das Simulationsmodell Lücken oder Fehler aufweist und wie sich bereits getroffene Annahmen auf die Ergebnisse auswirken. Die Ergebnisse der Validierung sollen Aufschluss geben, ob der aufbereitete Datensatz für die Simulation verwendet werden kann oder der Transformationsprozess beziehungsweise die Datengrundlage überarbeitet werden muss.

Dafür wird im ersten Kapitel zunächst die Thematik begrifflich in den Kontext der Simulation eingeordnet und anhand eines Vorgehensmodells der Ablauf der Validierung erklärt. Da es in der Literatur eine Vielzahl von Validierungsmethoden gibt, werden unterschiedliche Techniken betrachtet und geeignete für die Validierung ausgewählt. Des Weiteren wird das Simulationstool SimChain vorgestellt, um das Datenformat anhand der Funktionsweise anschaulich zu erklären. Der zu validierende Datentransformationsprozess ist nicht dokumentiert, sodass die Validierung ausschließlich auf Basis der gegebenen Daten erfolgt. Dazu werden diese in Kapitel 3 – „Datengrundlage der Validierung“ – detailliert beschrieben. Die eigentliche Validierung findet in Kapitel 4 statt. Die Prüfung der Daten lässt Rückschlüsse auf den Transformationsprozess zu und damit ebenfalls auf getroffene Annahmen, die in 4.1 aufgeführt werden. Daraufhin werden Prozess und Annahmen kritisch hinterfragt, sowie Verbesserungsvorschläge genannt, falls Fehler aufgedeckt werden. Anschließend wird die Datenquelle für die Simulationsstudie als Ergebnis der Validierung bewertet und Aussage darüber getroffen, ob und gegebenenfalls durch welche Korrekturen diese

zur Weiterverwendung geeignet ist. Zum Schluss werden alle Resultate der Projektarbeit im letzten Kapitel zusammengefasst.

2 Einordnung in den Kontext der Simulation und Vorgehen

Die Validierung des Datentransformationsprozesses geschieht vor dem Hintergrund einer Simulationsstudie in Bezug auf das Supply-Chain-Management eines Online-Supermarktes. Da die Simulation nach einem Vorgehensmodell erfolgt, soll auch der zu untersuchende Transformationsprozess einem Modell zugeordnet und die Validierung anhand dieses Modells durchgeführt werden. Dabei wird lediglich auf das Vorgehensmodell der Arbeitsgemeinschaft Simulation (ASIM) eingegangen, da hier insbesondere die Datentransformation und Validierung beschrieben sind und die Daten separat von den Modellierungsphasen betrachtet werden können (Rabe et al. 2008). Im Allgemeinen dient das Vorgehensmodell hauptsächlich zur Orientierung und es werden ausschließlich Begriffe, Methoden und Strukturen übernommen, die für das Thema dieser Arbeit relevant sind. In den nachfolgenden Ausführungen wird zuerst das Vorgehensmodell vorgestellt, dabei Grundbegriffe erläutert und dem vorliegenden Sachverhalt zugeordnet. Anschließend werden Validierungstechniken aus der Literatur diskutiert und auf Eignung in diesem Fall untersucht. Weiterhin wird die Verfahrensweise der Validierung mit der ausgewählten Technik beschrieben. Das Kapitel endet mit der Vorstellung des Simulationstools SimChain, um Struktur und Format der gegebenen Daten zu erklären.

2.1 ASIM-Vorgehensmodell

In der Literatur existiert eine Vielzahl an Vorgehensmodellen zu Simulationsstudien. Darunter befindet sich das Vorgehensmodell der Arbeitsgemeinschaft Simulation (ASIM), das durch eine klare Begriffsabgrenzung und Phasengliederung charakterisiert ist.

2.1.1 Struktur und Begriffe des Vorgehensmodells

„Ein Vorgehensmodell beschreibt im Wesentlichen den Weg von der Aufgabenspezifikation über ein Konzept und die Umsetzung des Modells mit einem Simulationswerkzeug bis zur Erzeugung von Ergebnissen“ (Rabe et al. 2017, S. 142). Das ASIM-Vorgehensmodell beschreibt diese Phasen und teilt sie in fünf Modellierungsphasen und zwei Phasen der Datenbehandlung ein, die gesondert betrachtet werden. Hierbei wird zwischen der „Datenbeschaffung“ und der „Datenaufbereitung“ unterschieden. Eine anschauliche Darstellung der Zusammenhänge der Phasen bietet Abbildung 1.

Bei der Datenbeschaffung werden Daten bereitgestellt, die durch Fachexperten möglichst unbearbeitet und direkt aus der Datenquelle bezogen werden. Diese Daten werden als Ergebnis der Phase „Rohdaten“ genannt. In der Phase Datenaufbereitung werden die Rohdaten in eine Form überführt, die das ausführbare Modell verwenden kann. Anschließend erfolgt eine Validierung der Daten, bei der die Eignung für die gegebene Aufgabenstellung bestimmt wird (Rabe et al. 2008, S. 52).

Die Rohdaten sowie die aufbereiteten Daten sind hier von besonderer Relevanz, da sie die Datengrundlage dieser Ausarbeitung darstellen. Der zu untersuchende Datentransformationsprozess findet sich in der Phase der Datenaufbereitung wieder. Die Schritte des ASIM-Vorgehensmodells können jedoch nicht vollständig durchlaufen werden, da die Phasenergebnisse im Gegensatz zu der eigentlichen Phase bereits gegeben sind

und die Datenaufbereitung selbst erst abgeleitet werden muss. Deshalb findet die Validierung anhand der Rohdaten und aufbereiteten Daten statt, ohne Bezug zur Datenaufbereitung.

Die **Validierung** ist die Prüfung auf die Güte der zu untersuchenden Daten. Sie beurteilt die Genauigkeit der Darstellung (Balci 2003) beziehungsweise stellt nach VDI 3633 - Blatt 1 eine hinreichende Kongruenz von (Daten-)Modell und realem System sicher. Im ASIM-Vorgehensmodell wird die Validierung in der Verifikation und Validierung (V&V) zusammengefasst. Die V&V beschreibt ein systematisches Vorgehen mit Hilfe etablierter Validierungstechniken unter Betrachtung verschiedener Kriterien. Lediglich einige Techniken sind für die Überprüfung der Datenaufbereitung beziehungsweise des Datentransformationsprozesses geeignet. Diese Techniken werden im folgenden Kapitel vorgestellt.

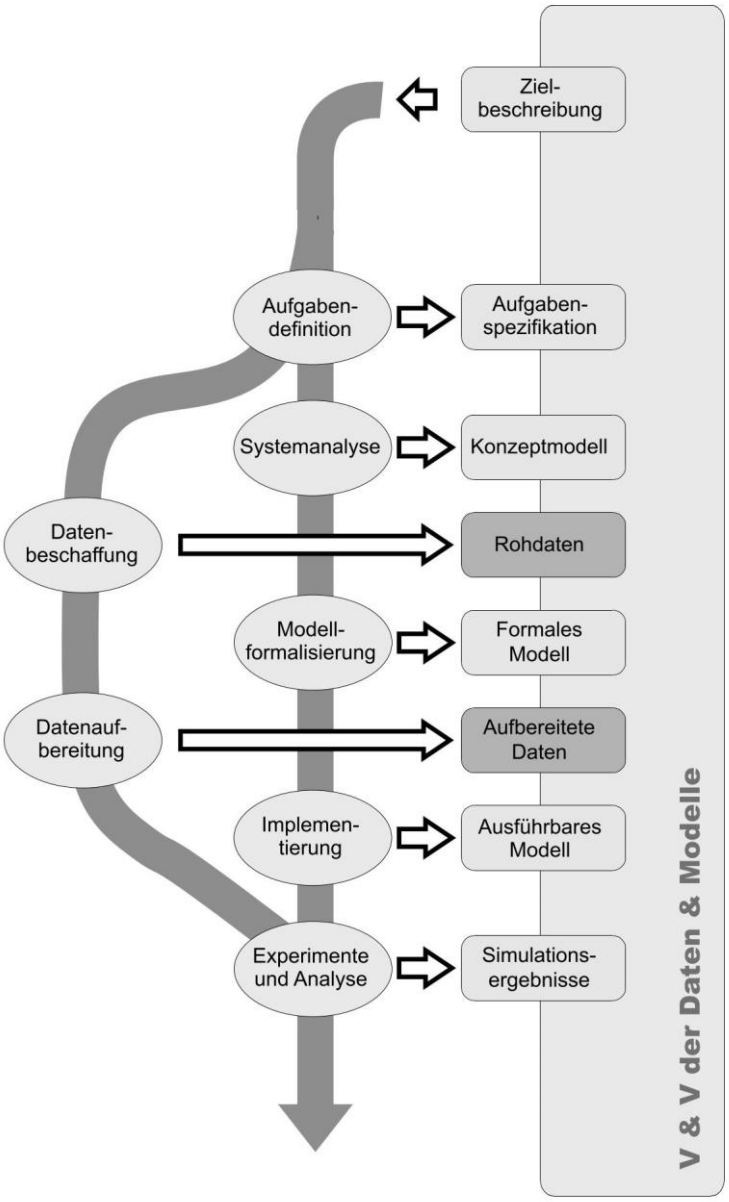

Abbildung 1: Vorgehensmodell bei der Simulation mit V&V (Rabe et al. 2008)

2.1.2 Geeignete Validierungstechniken

Rabe et al. (2008, S. 96) nennen 20 Techniken, von denen sechs für die Validierung der Rohdaten, und neun für die Validierung der aufbereiteten Daten generell geeignet sind. Abbildung 2 bietet dazu eine Übersicht über verwendbare Techniken in den einzelnen Phasenergebnissen. Im Folgenden werden diese erläutert und anschließend bezüglich ihrer Anwendbarkeit bewertet, um eine oder mehrere Techniken in Kombination für die Validierung des Datentransformationsprozesses auszuwählen.

Abbildung 2: Verwendbarkeit von V&V-Techniken im Verlauf der Simulationsstudie (Rabe et al. 2008, S. 113)

- Begutachtung: Die Begutachtung (engl. „Review“) setzt eine Beteiligung der Auftraggeber und Auftragnehmer der Simulationsstudie voraus und überprüft die Übereinstimmung der Studie mit den festgelegten Zielen (Balci 1994). Da die Technik unabhängig von der Simulationsstudie durchgeführt wird und die Projektbeteiligten fehlen, ist sie hier ungeeignet.
- Dimensionstest: Der Dimensionstest (engl. „Dimensional Consistency Test“) konzentriert sich auf Berechnungen und Formeln im Modell, die Fehler aufweisen können. Durch Nachrechnen der Dimension über die Einheiten werden die Fehler aufgedeckt (Rabe et al. 2008, S. 98). Der Test ist nur bedingt anwendbar, da nur wenige Berechnungen im Datensatz vorhanden sind.
- Monitoring: Mit Monitoring werden Momentaufnahmen oder Zeitverläufe des Simulationslaufs dargestellt und Werte von Zustandsgrößen und Variablen angezeigt. Dabei werden die Werte auf

Konsistenz geprüft (Rabe et al. 2008, S. 101). Diese Technik ist nicht durchführbar, da sie die Simulationsstudie einbindet.

- Schreibtischtest: Der Schreibtischtest (engl. „Desk Checking“) wird vor allem bei der eigenen Arbeit angewandt und prüft diese durch gründliches Durchgehen auf Vollständigkeit, Korrektheit, Konsistenz und Eindeutigkeit. Die Ausführung durch eine andere Person ist möglich und wird bevorzugt, da eigene Fehler oft übersehen werden (Balci 1994, S. 130). Aufgrund der Übertragbarkeit auf die gegebenen Daten ist die Methode geeignet.
- Sensitivitätsanalyse: Bei der Sensitivitätsanalyse (engl. „Sensitivity Analysis“) werden systematische Änderungen an Werten der Eingabeparameter vorgenommen und die Auswirkungen auf das Modellverhalten beobachtet. Unerwartete Auswirkungen decken eventuelle Inkonsistenzen auf (Balci 1994, S. 167). Der Einsatz dieser Technik benötigt jedoch Kenntnisse über das reale System, die nicht vorhanden sind.
- Statistische Techniken: Mit Hilfe von statistischen Techniken (engl. „Statistical Techniques“) werden Ausgabedaten des Modells mit dem Verhalten des entsprechenden realen Systems verglichen und bewertet. Dabei werden unter anderem angenommene Verteilungen kontrolliert (Rabe et al. 2008, S. 103). Diese Techniken sind geeignet und kombinierbar mit anderen Methoden.
- Strukturiertes Durchgehen: Ein Team, bestehend aus mehreren Projektbeteiligten, geht strukturiert (engl. „Walkthroughs“) die Dokumente und Phasenergebnisse durch und untersucht diese nach Fehlern (Balci 1994, S. 133). Die Anwendung dieser Technik ist nicht möglich, da die Validierung unabhängig von der Simulationsstudie geschieht.
- Turing-Test: Beim Turing-Test erhalten Fachexperten die Ausgabedaten des realen Systems und die Ausgabedaten des Modells unter ansonsten gleichen Rahmenbedingungen. Die Experten sollen nun zwischen beiden differenzieren. Die Erklärungen, wie sie die Daten voneinander unterscheiden können, enthalten hilfreiche Informationen, um das Modell zu verbessern (Balci 1994, S. 168). Die Technik erfordert Fachexperten mit Kenntnissen über das reale System, die nicht verfügbar sind.
- Validierung im Dialog: Bei der Validierung im Dialog (engl. „Face Validation“) wird das Modell von einem Team mit Kenntnissen über das reale System diskutiert und beurteilt, ob es plausibel ist (Balci 1994). Die Technik ist nicht geeignet, da sie ebenfalls die Simulationsstudie einbindet und weitere Personen benötigt werden.

Für die Validierung des Datentransformationsprozesses wird der Schreibtischtest ausgewählt, da er durch seine allgemeine Herangehensweise am besten geeignet ist, um einen Großteil der Fehler und Annahmen zu erfassen. Des Weiteren werden mit Hilfe von statistischen Tests die aufbereiteten Daten auf Plausibilität und Vollständigkeit geprüft. Beide Tests lassen sich miteinander kombinieren und ergänzen sich auf unterschiedlichen Ebenen. Andere Techniken sind nicht durchführbar oder eignen sich nicht für den vorliegenden Sachverhalt.

2.1.3 Vorgehensweise der Validierung

Die zu untersuchende Datentransformation entspricht der Datenaufbereitung des ASIM-Modells. Da in der Simulationsstudie die Dokumentation der Datenaufbereitung nicht vorliegt, kann sie lediglich hergeleitet und geschätzt werden. Informationsgrundlage sind die Phasenergebnisse vor und nach der Datenaufbereitung, dementsprechend die Rohdaten und die aufbereiteten Daten. Rabe et al. (2008, S. 181) erläutern die Vorgehensmethodik als Teil der Verifikation und Validierung, indem die aufbereiteten Daten in sich

(intrinsische Prüfung) und gegen die Rohdaten geprüft werden. Die intrinsische Prüfung untersucht die Daten auf Vollständigkeit, Konsistenz (Widerspruchsfreiheit), Genauigkeit und Plausibilität (Nachvollziehbarkeit). Die Prüfung gegen die Rohdaten erfolgt nach den gleichen Kriterien. Der Abgleich deckt vorhandene Unterschiede auf, welche Rückschlüsse auf getroffene Annahmen zulassen oder Fehler aufspüren. Diese Methodik wird sowohl mit Hilfe des Schreibtischtests als auch mit Hilfe der statistischen Tests durchgeführt.

2.2 Simulationstool SimChain

SimChain ist ein Simulationswerkzeug der Firma SimPlan, das in Kombination mit der Simulationssoftware Plant Simulation eingesetzt wird. Das Tool arbeitet als Schnittstelle für die Simulationssoftware und unterstützt bei der Erstellung und Anpassung von Supply-Chain-Modellen. SimChain generiert die aufbereiteten Daten auf Grundlage von definierten Parametern und Einstellungen. Weiterhin besteht die Möglichkeit, bestimmte Daten aus externen Datenquellen zu importieren.

Mit SimChain lässt sich eine virtuelle Lieferkette erstellen, deren Zusammensetzung über Parameter bestimmt wird. Beispielsweise können Orte („Locations“), Zeiten („Calender“), Artikel („SKU“), Transporte („Carrier“, „Means of Transport“), Beeinträchtigungen („Information Delay“, „Forecast Error“), Wege („Transport Relation“, „Sourcing Routes“) und weitere Komponenten angelegt und definiert werden (Melchior 2014). Eine der Hauptfunktionen ist das Generieren von Kundennachfrage. Diese wird entweder anhand von Vorgabewerten und Zufallsvariablen generiert oder aus Datenbanken mit realen Daten importiert (Fechteler und Gutenschwager 2014, S. 26).

Das Tool speichert alle Eingaben und generierten Daten im Excel-Spreadsheet-Format (xlsv). Außerdem ist es möglich, die Daten als einfach strukturierte Textdatei (csv-Format) für einen Simulationslauf mit Plant Simulation zu exportieren.

3 Datengrundlage des Transformationsprozesses

Es liegen zusammengetragene Daten über Logistik, Kundschaft und Geschäftszahlen des Online-Supermarkts als sogenannte Rohdaten und durch das Simulationstool SimChain erstellte und verarbeitete Daten als Aufbereitete Daten vor. Nicht bekannt ist, auf welche Weise und nach welchen Gesichtspunkten die Datenbeschaffung und die Datenaufbereitung erfolgten. Nachfolgend wird vor allem auf inhaltliche Aspekte und weniger auf das Datenformat eingegangen, da die Aufbereitung bezüglich der Modellparameter fehleranfälliger als die Formatierung ist. Letztere wird zudem durch die Simulationsschnittstelle SimChain gestützt, die Eingaben und getroffene Einstellungen in vorbestimmten Formaten speichert.

3.1 Rohdaten

Die Rohdaten bestehen aus den zwei Excel-Spreadsheet-Dateien (xlsv) „Economic Model“ und „2014 Results“. Das „Economic Model“ beinhaltet sieben Arbeitsmappen:

- *Inputs*: Hier finden sich zuoberst Kategorisierungen. Die Bevölkerungsdichte („Population Density“) und die sozioökonomischen Gruppen („Socio-Economic Group“) sind in fünf („Very Low“ bis „Very High“) beziehungsweise vier Kategorien („AB“ bis „DE“) eingeteilt. Darunter werden Variablen von Geschwindigkeiten, Zeiten und Kosten definiert. Eine Tabelle listet 16 Distributionszentren (Distribution Center, „DC“) und die jeweilige Postleitzahl des Bezirks („Postcode District“) auf. Daneben werden 105 belieferte Gebiete und Postleitzahlen angegeben. Kosten der Lieferwagen und Lastkraftwagen sind den restlichen Tabellen zu entnehmen.
- *Calculations*: Im Gegensatz zu den *Inputs* liegt hier der Fokus auf der Berechnung von aussagefähigen oder vergleichbaren Zahlen. Jeder Bezirk wird dazu in weitere Sektoren eingeteilt und zu jedem Sektor („Postcode sector“) werden detaillierte Angaben über Einwohner, Fläche und sozioökonomische Gruppen gemacht. Aus diesen werden Zahlen über Bestellungen, Lieferzeiten und –wege, Kosten, und Gewinne abgeleitet. Diese Tabelle ist mit 8035 Zeilen und dementsprechend 8035 Sektoren am umfangreichsten.
- *Stem mileage*: Diese Arbeitsmappe über die Laufleistung der Kundenlieferungen enthält drei Tabellen. Die erste berechnet die Entfernungen zwischen den Sektoren und den 16 Distributionszentren anhand ihrer Koordinaten. Als Endergebnis der Tabelle werden die geringste Entfernung und das entsprechende nächste Distributionszentrum für jeden Sektor dargestellt. Die zweite Tabelle berechnet spaltenweise und überschlägt die Kosten pro Bestellung für jedes Distributionszentrum. Da für jeden Sektor bereits das nächste Distributionszentrum bestimmt wurde, führt Tabelle drei die Lieferkosten an die Endkunden pro Bestellung für jeden Sektor auf.
- *Trunking*: Hier wird wie in der Arbeitsmappe „Stem mileage“ vorgegangen, betrachtet werden jedoch die Laufleistungen der Lastkraftwagen und die Distanz zwischen Produktionsstätte („Hub“) und Distributionszentren. Daraus wurden Lieferkosten an die Distributionszentren pro Bestellung für jeden Sektor berechnet.
- *Analysis*: In dieser Arbeitsmappe sind Zahlen über Bestellungen, Kosten und Gewinnspannen zusammengerechnet und nach Bevölkerungsdichten, sozioökonomische Gruppen und Gebieten

aufgeschlüsselt. Zu jedem Gebiet werden Bestellzahl, Umsatz, durchschnittliche Bestellsumme, Kosten, Gewinn und weitere Kennzahlen angegeben.

- *Population Density*: Hier ist ein Balkendiagramm dargestellt, das den Zusammenhang zwischen Bevölkerungsdichte, Umsatz und Kosten der Auftragsabwicklung aufzeigt.
- *Socio-Economic Group*: Dies ist ebenfalls ein Balkendiagramm, welches den Zusammenhang zwischen sozioökonomischer Gruppe, durchschnittliche Warenkorbgröße und Gewinn abbildet.

Die Datei „2014 Results“ enthält Tabellen über die Entwicklung von Bestellungen, Lieferungen, Umsatz und weitere ökonomische Kennzahlen. Hierbei werden die Jahre 2011 bis 2014 beziehungsweise 2013 bis 2014 verglichen und es wird eine Prognose für 2019 aufgestellt.

3.2 Aufbereitete Daten

Die aufbereiteten Daten bestehen aus vier xlsv-Dateien und 30 csv-Dateien. Die Daten wurden durch das Simulationstool SimChain erstellt und erhalten neben für das Datenmodell relevante Informationen auch Konfigurationsinformationen. Auf diese Daten wird nicht eingegangen, da sie nur für SimChain von Bedeutung sind.

Als Erstes wird die Datei „Basic Tables“ mit folgenden Arbeitsmappen betrachtet:

- *Continents* und *Countries*: Hierbei werden sieben Kontinente und 207 Länder aufgelistet, mit einem Primärschlüssel versehen und die Länder den Kontinenten zugeordnet.
- *Locations*: Die Tabelle besteht aus 266 Zeilen von Kunden und ihren Standorten, versehen mit Koordinaten, Sektor-PLZ, Stadt und Länderkennzeichnung. Alle Kunden sind aus London, UK.
- *Calender*: Hier werden Tage aufgezählt, an denen kein Versand stattfindet. Einziger Eintrag ist der 25. Dezember.
- *SKU*: In dieser Arbeitsmappe wird nur eine Bestandseinheit genannt, beschrieben als Paket („box“). Das Gewicht beträgt 1 und die Kohlenstoffdioxidemissionen wurden auf 0 festgelegt.
- *Carrier*: Hier ist ein Lager- und Transportträger spezifiziert, beispielsweise eine Palette. Sie ist 1€ wert, 1kg schwer und hat die Maße 604mm×414mm×231mm. Bei diesem Objekt gibt es ebenfalls keine Emissionen.
- *MeansOfTransport*: Als Transportmittel ist ein Van aufgeführt. Das maximale Gewicht der Zuladung beträgt 750kg, die Durchschnittsgeschwindigkeit 31km/h und die Kohlenstoffdioxidemissionen 0.
- *CarrierOnMeansOfTransport*: Ein Van kann mit bis zu 80 Transportträgern beladen werden.
- *Information_Delay*: Die Arbeitsmappe beschreibt die Verzögerungen im Informationsfluss abhängig von der Bestellmethode. Dabei ergeben sich zum Beispiel bei einer Bestellung per elektronischem Datenaustausch 0 Tage Verzögerung, per Telefon oder Fax 0 bis 2 Tage.
- *ForecastError*: Hier sind Abweichungen von der prognostizierten Nachfrage mit Wahrscheinlichkeiten pro prognostizierte Periode verknüpft.

Die Datei „Configuration Tables“ ist die umfangreichste der Excel-Spreadsheet-Dateien. Diese gibt Auskunft über Informationen in folgenden Arbeitsmappen:

- *Customer*: Hierbei werden, wie bereits in Tabelle *Locations*, 266 Kunden aus den einzelnen Sektoren aufgeführt.
- *Sites*: Die Arbeitsmappe beschreibt zwei Versandstandorte, von denen Waren versendet werden, anhand ihrer Sektor-PLZ.
- *Hubs_PlainSupplier*: Knotenpunkte, die als Zulieferer agieren, werden hier aufgeführt. Die Tabelle enthält keine Einträge.
- *sku_demand_external*: Die umfangreichste Tabelle dieser Datei enthält 2480 Zeilen über Kundenbestellungen vom 01.01.2014 bis zum 30.12.2014. Angegeben sind die Kundenstandorte, Bestellungsdatum, Lieferdatum, Menge an Artikeln und je eine Identifikationsnummer der Bestellung. Die Daten dieser Tabelle über die Nachfrage entstammen nicht einem Datengeneratoren, sondern wurden aus einer externen Datenquelle importiert.
- *Route*: Hier werden zwei Routen aufgeführt, die von den beiden Versandstandorten ausgehen. Weiterhin beträgt die maximale Tourenlänge 60000 km mit maximal 250 Knotenpunkten, die Transportkosten pro Kilometer 1,64 €. Andere Transportkosten bezogen auf Transportmittel und SKU werden nicht berücksichtigt.
- *Routecycle*: An jedem Tag der Woche werden pro Distributionszentrum 343 Transportmittel eingesetzt.

In der weniger umfangreichen xlsv-Datei „CustomerLatLon“ wird jeder Sektor mit einer geographischen Koordinate als Dezimalzahl verknüpft. Dabei besitzt der jeweilige Breitengrad 13 Nachkommastellen und der Längengrad 16 Nachkommastellen.

Die Datei „SitesLatLon“ ist ähnlich aufgebaut wie „CustomerLatLon“, jedoch werden ausschließlich die Koordinaten der zwei Distributionszentren angegeben. Hierbei besitzt der Breitengrad sechs und der Längengrad neun Nachkommastellen.

Die csv-Dateien haben unterschiedlich viele Einträge, welche jedoch demselben Schema entsprechen:

> Postleitzahl des Sektors („*PostcodeSector*“), Name des Bezirks („*Spoke*“), Versanddatum („*DayDate*“), Lieferungsdatum („*DayDelivery*“), Anzahl Artikel („*NumItems*“), Volumen („*Volume*“), Gewicht („*Weight*)“, Tag des Jahres („*DayOfYear*“).

Dabei bildet jede Datei in etwa ein Jahr ab. Zwischen den Dateien variiert die Anzahl der Tage zwischen 363 und 364, abhängig vom letzten Versand des Jahres, der entweder am 30.12. oder 31.12. erfolgt. Am 25.12. findet kein Versand statt, jedoch werden die versandten Güter zugestellt.

4 Ergebnisse der Validierung

Im Folgenden werden alle Ergebnisse der Validierung präsentiert und anschließend bewertet. Nach der Prüfung des Datensatzes erfolgt sowohl eine Bewertung des Modells in Bezug auf die aufbereiteten Daten, als auch eine abschließende Bewertung der Datenquelle auf Eignung zur Weiterverwendung in der Simulationsstudie.

4.1 Beschreibung des Datentransformationsprozesses durch Validierung

Die Validierung mit Hilfe der ausgewählten Techniken wird in die Prüfung der aufbereiteten Daten in sich und gegen die Rohdaten unterteilt. Dabei kommen jeweils der Schreibtischtest und statistische Techniken zum Einsatz, die in Kapitel 2.1.2 beschrieben und für die vorhandene Prüfung ausgewählt wurden.

4.1.1 Ergebnisse der intrinsischen Prüfung

Neben der allgemeinen Vorgehensweise bei der Durchführung des Schreibtischtest wird im Folgenden die spezielle angewandte Vorgehensweise erläutert. Hier wird zuerst jede Tabelle aus den Rohdaten und den aufbereiten Daten einzeln betrachtet. Der Fokus liegt auf den Kriterien Vollständigkeit, Plausibilität und Korrektheit, sodass auf fehlende Parameter, Wertebereiche und korrekte Einheiten geprüft wird. Auffällige Parameter werden dabei vermerkt. Daraufhin sind die Datensätze zusammenhängend zu untersuchen, um Beziehungen zwischen den Tabellen herzustellen und anhand dessen Inkonsistenzen aufzuzeigen. Hierbei ergibt sich ein gedankliches Modell, das mit Hilfe eines Schaubilds festgehalten werden kann. Abbildung 3 zeigt ein Schaubild, das einen Ausschnitt des gesamten Modells aus den aufbereiteten Daten darstellt und ausschließlich der Übersichtlichkeit dient.

Als Resultat des Schreibtischtest zeigt sich, dass die aufbereiteten Daten größtenteils widerspruchsfrei und plausibel sind. Einige Aspekte bleiben für die Simulationsstudie unberücksichtigt, zum Beispiel werden Parameter wie Emissionen nicht einberechnet. Des Weiteren werden verschiedene Komponenten nur in einer Ausprägung definiert, um das Modell zu vereinfachen. So wird beispielsweise nur eine Bestandseinheit „box“ genannt und nicht zwischen unterschiedlichen Produkten unterschieden. Da ein genauerer Detailierungsgrad nicht erforderlich ist, ist die Vereinfachung zulässig und aus Laufzeitgründen sogar erwünscht. Des Weiteren werden Bezirkssektoren auf einen einzelnen Punkt reduziert, der durch Koordinaten beschrieben wird. Da dieser nur zur Berechnung der Transportstrecken benötigt wird, ist eine Vereinfachung ebenfalls zulässig.

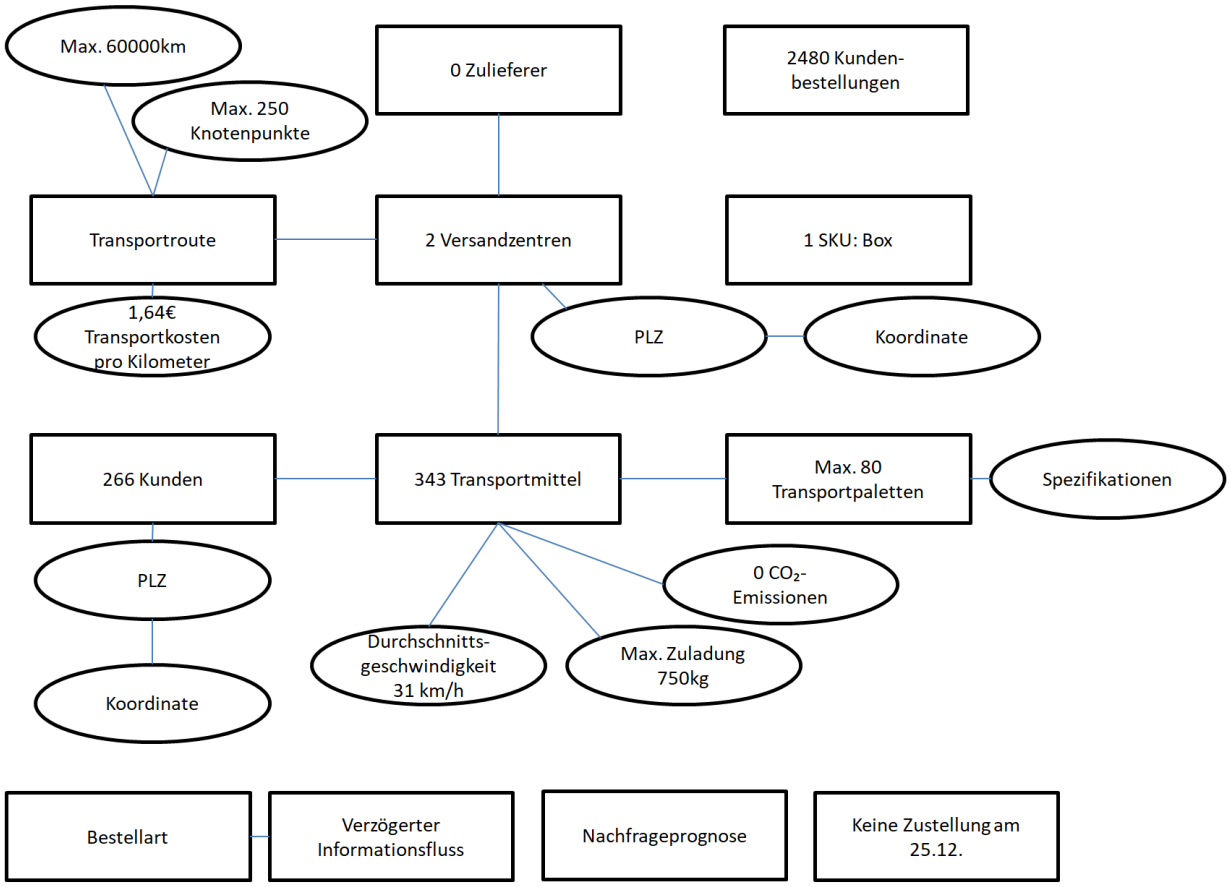

Abbildung 3: Zusammenhänge zwischen Modellkomponenten in den aufbereiteten Daten

In der Tabelle *Calender* wird zwar der 25. Dezember als freier Tag deklariert, jedoch gilt der Feiertag nur für das Versandzentrum und nicht für die Zusteller. Das wird sowohl in *sku_demand_external* als auch in den csv-Dateien ersichtlich, da in den Tabellen Bestellungen mit Bestell- und Lieferdatum aufgelistet sind. Die Zustellung erfolgt in dem Modell immer einen Tag nach der Bestellung. Daher sind zwar Einträge mit Bestelldatum vom 24.12. und Lieferdatum am 25.12. vorhanden, jedoch gibt es keine Bestellungen mit Bestelldatum vom 25.12. und Lieferdatum am 26.12. Daraus ergeben sich zwar keine Inkonsistenzen, jedoch kann man in der Realität davon ausgehen, dass Bestellungen an jedem Tag aufgegeben werden können. Liegt die Bestellung an einem Feiertag, verzögert sich nur die Zustellung um einen weiteren Tag.

Die Tabelle *sku_demand_external* ist von besonderer Bedeutung, da sie Aussagen über den Datentransformationsprozess liefert. Laut der technischen Dokumentation von SimChain (Fechteler und Gutenschwager 2014, S. 71) wird die Nachfrage nicht durch Zufallszahlen generiert, falls diese Tabelle mit Daten gefüllt ist. Anderenfalls würden Daten aus dem Zufallsgenerator in der Tabelle *SKU_Demand* festgehalten werden, die im vorliegenden Datensatz allerdings keine Einträge enthält. Daher sind die Kundenbestellungen aus *sku_demand_external* aus einer externen Datenquelle importiert. Da die Datenquelle nicht dokumentiert ist und eine Generierung der Daten außerhalb von SimChain nicht ausgeschlossen werden kann, ist eine Überprüfung der Kundennachfrage aus der Tabelle unerlässlich. Hierbei werden mittels statistischer Tests die Verteilung und Mittelwerte der Kundenbestellungen auf Konsistenz und Plausibilität untersucht. Als statistische Größe wird die durchschnittliche Bestellmenge

gewählt, da sie die aussagekräftigste Größe der Tabelle ist. Werden die pro Bezirkssektor berechneten Werte absteigend nach der Bestellmenge sortiert, ergibt sich die in Abbildung 4 dargestellte Verteilung.

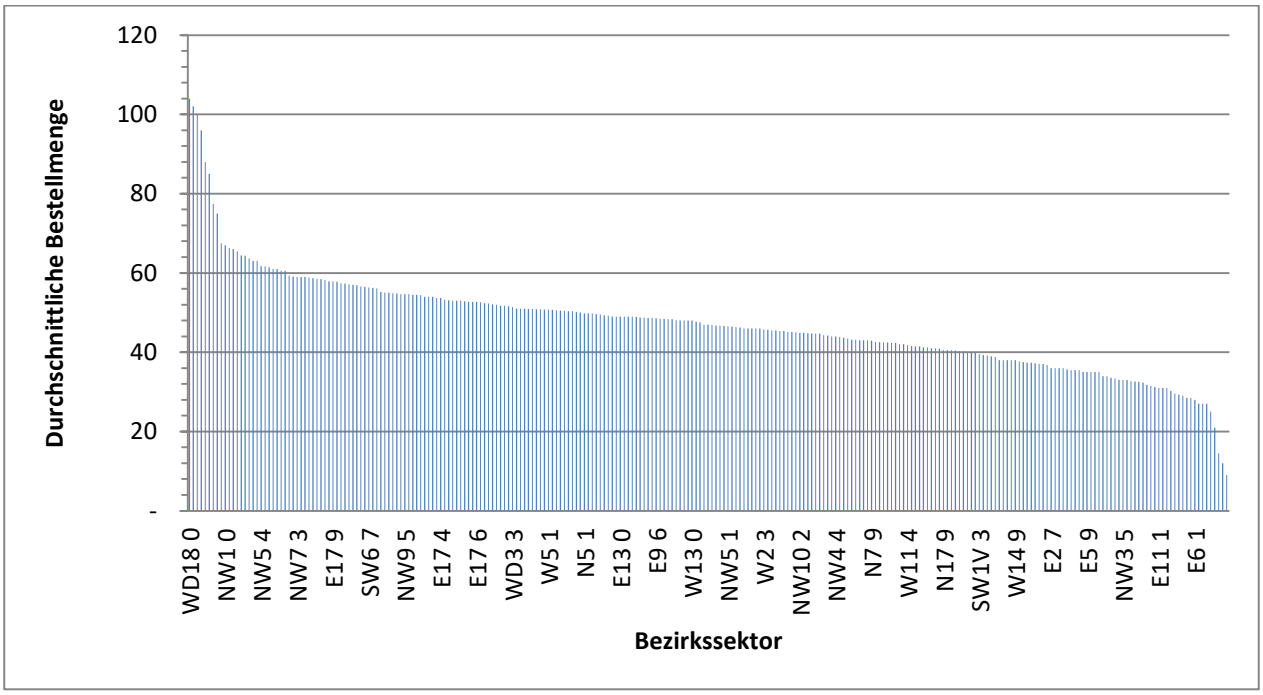

Abbildung 4: Durchschnittliche Bestellmenge der 266 Sektoren

Der Mittelwert dieser Verteilung liegt bei 48,8 Bestellungen pro Sektor. Die Ungleichverteilung ist plausibel, da eine sozio-ökonomische Ungleichheit in der Kundschaft angenommen werden kann, die sich auf das Bestellverhalten auswirkt. Im folgenden Kapitel wird darauf Bezug genommen und untersucht, ob umsatzstarke Sektoren aus den Rohdaten mit den Sektoren übereinstimmen, die in den aufbereiteten Daten viele Bestellungen aufweisen. Allerdings muss hinzugefügt werden, dass ca. 14% der Sektoren (37 von 261) nur eine Bestellung verzeichnet haben, sodass die Verteilung aufgrund des kleinen Datensatzes nur marginal aussagekräftig ist.

Weiterhin wurde der Datensatz zusammenhängend betrachtet, indem in der Realität voneinander abhängige Parameter ins Verhältnis gesetzt wurden. Dabei passen die Daten der Tabelle *Routecycle* nicht zu den Daten der Tabelle *sku_demand_external*. Die Zahl der Auslieferungsfahrzeuge muss in einem glaubhaften Verhältnis zur Anzahl Bestellungen stehen. Allerdings stehen in *Routecycle* 343 deklarierte Transportmittel pro Distributionszentrum, demnach insgesamt 686 Fahrzeuge jeden Tag, 2480 Bestellungen im Jahr gegenüber. Bei maximal 13 und durchschnittlich 5,25 Bestellungen pro Tag, die aus Enfield beziehungsweise maximal 6 und durchschnittlich 2,09 Bestellungen pro Tag, die aus Ruislip geliefert werden, wird der Großteil der Fahrzeugflotte nicht gebraucht werden. Ein bis zwei Transportmittel je Distributionszentrum sind vollkommen ausreichend.

Die Tabelle *Route* definiert die maximale Tourenlänge eines Fahrzeugs mit einer Strecke von 60000 km. Die in der Tabelle *MeansOfTransport* angegebene Durchschnittsgeschwindigkeit eines Vans beträgt 31 km/h, bei maximaler Tourenlänge wäre ein Van über 80 Tage unterwegs. Für eine Auslieferungsroute ist dieser Wert vollkommen unrealistisch.

Die auf bis zu 16 Nachkommastellen genaue Dezimalzahl der Ortskoordinaten von Kunden und Distributionszentren (Exceltabelle *CustomerLatLon* und *SitesLatLon*) ist ein typischer Modellierungsfehler (vgl. Wenzel und Bernhard 2008). Die 16. Nachkommastelle einer Koordinate entspricht einer Granularität von 11 Pikometern ($11*10^{-12}$ m). Hier liegt der Nutzen eines extrem hohen Detailierungsgrades in keinem Verhältnis zum Aufwand, vor allem vor dem Hintergrund, dass ein großflächiger Bezirkssektor mit diesen Koordinaten dargestellt ist.

4.1.2 Ergebnisse der Prüfung gegen die Rohdaten

Die Prüfung der aufbereiteten Daten gegen die Rohdaten erfolgt nach folgendem Schema: Mit Hilfe des Schreibtischtests werden die modellrelevanten Komponenten und Parameter aus dem Datensatz ermittelt und notiert. Dabei ist die Prüfung auf Modellrelevanz und Bezug zum simulierten Sachverhalt bei allen Daten erforderlich und wird grundsätzlich für jede Tabelle einzeln durchgeführt. Anschließend werden diese Daten in einem gedanklichen oder grafischen Modell festgehalten und den aufbereiteten Daten gegenübergestellt.

Die Gegenüberstellung deckt weitere Diskrepanzen auf. Vor allem fällt auf, dass die Rohdaten in der Tabelle *Calculations* weitaus mehr belieferte Bezirkssektoren aufweisen als die aufbereiteten Daten in der Tabelle *sku_demand_external*. Hier stehen 8035 Bezirkssektoren aus ganz Großbritannien nur 266 Bezirkssektoren aus London gegenüber. Die Tabelle *Inputs* enthält zudem Informationen über Bestellungen. Die Anzahl der Bestellungen insgesamt liegt bei über 8,6 Millionen, bei einer durchschnittlichen Anzahl von 1081 Bestellungen pro Bezirkssektor. Die Bestellzahlen der aufbereiteten Daten entsprechen mit insgesamt 2480 und durchschnittlich 3,875 Bestellungen pro Sektor nur einem Bruchteil der Zahlen aus dem realen System. Das entspricht eine um den Faktor 30 größere Anzahl an Bezirkssektoren und um den Faktor 3400 größere Anzahl an Bestellungen in den Rohdaten. Auffällig ist zudem, dass der Bezirkssektor „E20 1“ aus *sku_demand_external* in den Tabellen der Rohdaten überhaupt nicht existiert.

Die Tabelle *Inputs* liefert weitere Daten, die die aufbereiteten Daten nicht erwähnen. Da die Rohdaten Informationen über ein erheblich größeres Auslieferungsgebiet enthalten, sind 16 statt zwei Distributionszentren aufgeführt. Des Weiteren fehlen in den aufbereiteten Daten zwei Zulieferer („Hubs“), die alle Distributionszentren beliefern. Diese Zulieferer befinden sich am selben Standort wie die Distributionszentren 1 und 2.

Ein Vergleich der Bestellmengen pro Bezirkssektor zwischen Roh- und aufbereiteten Daten ist nicht möglich, da in den Rohdaten keine Angaben zu Bestellmengen enthalten sind. Dies wäre eine Möglichkeit, um die Verteilungen zu vergleichen und somit Schlüsse auf die Realitätsnähe der aufbereiteten Daten zuzulassen. Allerdings sind in den Rohdaten in der Tabelle *Calculations* Informationen zu den Jahresumsätzen und Anzahl Bestellungen der einzelnen Sektoren vorhanden. Für die folgende Überprüfung werden die Bezirkssektoren, die in der Tabelle *sku_demand_external* der aufbereiteten Daten gegeben sind, in der Tabelle *Calculations* der Rohdaten herausgesucht und die zugehörigen Jahresumsätze sowie Bestellzahlen bestimmt. Anschließend wird der Quotient dieser beiden Zahlen ermittelt, der Jahresumsatz geteilt durch Anzahl Bestellungen ergibt den durchschnittlichen Preis einer Bestellung, der durchschnittlichen Bestellsumme. Da die durchschnittliche Bestellsumme theoretisch mit der durchschnittlichen Bestellmenge korreliert (bei ausreichend großer Datenmenge), wird auch im Diagramm eine ähnliche Verteilung erwartet.

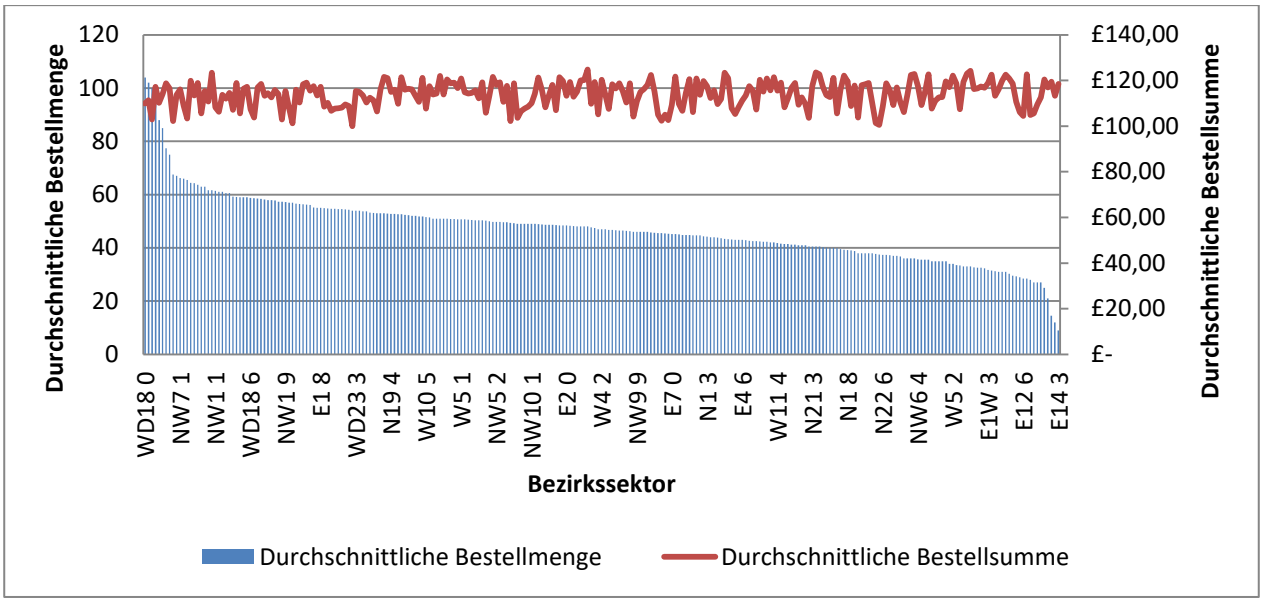

Abbildung 5: Zusammenhang zwischen Bestellmenge aus den aufbereiteten Daten und Umsatz aus den Rohdaten

Das Diagramm in Abbildung 5 zeigt keinen Zusammenhang zwischen der durchschnittlichen Bestellmenge aus den aufbereiteten Daten und der durchschnittlichen Bestellsumme aus den Rohdaten. Während die in Kapitel 4.1.1 aufgezeigte Verteilung der Bestellmenge durch eine abnehmende Funktionskurve beschrieben wird, liegt die durchschnittliche Bestellsumme nahezu gleichverteilt um den Mittelwert von 113,96 £ bei einer Standardabweichung von 5,68 £. Die erwartete Korrelation lässt sich nicht nachweisen, sodass hier ein weiteres Missverhältnis zwischen Roh- und aufbereiteten Daten vorliegt.

Von Interesse ist zudem, ob der kleine Datensatz der aufbereiteten Daten zu den Sektoren repräsentativ für die Rohdaten ist. Die Tabelle *Analysis* gibt dazu Auskunft über Bestellzahlen und durchschnittliche Bestellsummen; die Tabelle *Stem mileage* beinhaltet Informationen zu den Entfernungen und Transportwegen. Mit Hilfe dieser Werte ist ein aussagekräftiger Vergleich möglich. Auch hier kommen statistische Tests vor, die die Verteilungen untersuchen. Auf die Bestellzahlen wurde bereits in diesem Kapitel eingegangen, verglichen werden nun Mittelwert, Minimum und Maximum der durchschnittlichen Bestellsumme und des Transportwegs zum nächstgelegenen Distributionszentrum. Die Londoner Bezirkssektoren, auf die sich der aufbereitete Datensatz reduziert, haben einen mittleren Transportweg von 5,9 km und maximal 10,3 km, alle Sektoren hingegen im Mittel 34 km und maximal 239, 4 km. Das Minimum beträgt bei beiden Datensätzen 0 km. Die durchschnittliche Bestellsumme liegt in London nach den Daten bei 113,98 £, kleinste bei 99,91 £ und größte bei 124,84 £. Im gesamten vereinigten Königreich sind es im Mittel 111,13 £, minimal 86,45 £ und maximal 126,56 £. Die Abweichungen sind signifikant, sodass die Sektoren der aufbereiteten Daten als nicht repräsentativ gelten.

In den Rohdaten finden sich außerdem Zahlen zu den eingesetzten Lieferwagen. Laut Arbeitsmappe *Inputs* stellt ein Lieferwagen 163 Bestellungen pro Woche zu. Bei 8.684.000 jährlichen Bestellungen und 8476 Lieferungen pro Lieferwagen im Jahr, entspricht die Größe der Fahrzeugflotte insgesamt 1024 Fahrzeuge. Dabei liegt das Verhältnis Bestellungen pro Fahrzeug mit 8476 in den Rohdaten deutlich höher als 3,6 in den aufbereiteten Daten. Die Transportkosten unterscheiden sich ebenfalls. Laut Rohdaten betragen sie 0,35€ pro Kilometer anstatt 1,64€. Anders als in den aufbereiteten Daten werden weitere auf Fahrer und Bestellungen bezogene Kosten genannt.

4.2 Bewertung und Verbesserungsvorschläge für das Modell

Die aufbereiteten Daten sollten vor allem in Hinsicht auf Genauigkeit und Vollständigkeit überarbeitet werden. Die inhaltlich größte Abweichung von den Rohdaten ist der viel geringere Datensatz bezüglich Kunden(-standorten) und Bestellzahlen. Dass es sich um eine beabsichtigte Datenreduktion handelt, ist ausgeschlossen. Diese erlaubt „die Vernachlässigung der Daten, die nicht zu einer für die Aufgabenstellung relevanten Information beitragen“ (Wenzel und Bernhard 2008, S. 499). Da die Daten durchaus relevant sind, liegt ein solcher Fall nicht vor. Die Relevanz lässt sich an den Untersuchungen zur Repräsentativität zeigen, die einen deutlichen Unterschied zwischen Roh- und aufbereiteten Daten veranschaulichen.

Neben den 8035 Bezirkssektoren sollten auch 16 Distributionszentren und zwei Zulieferer mit den entsprechenden Standorten in das Modell übernommen werden. Der in der Realität nicht existierende Bezirkssektor „E20 1“ darf hingegen nicht im Modell vorkommen. Da Aufwand und Fehleranfälligkeit bei einer manuellen Eingabe aller Standorte ziemlich hoch ist, sollte überprüft werden, ob ein Import der, auf die wesentlichen Parameter reduzierten, Rohdaten durchführbar ist. In Kapitel 3.1 ist beschrieben, in welchen Tabellen die benötigten Informationen zu finden sind. Die Bestellzahlen pro Bezirkssektor sowie die Anzahl Fahrzeuge pro Distributionszentrum können direkt aus den Rohdaten übertragen werden, ansonsten wäre eine Generierung der Anzahl Bestellungen mit Hilfe des Mittelwerts ebenso denkbar. Aus den Tabellen der Rohdaten lassen sich zwar keine Bestellmengen entnehmen, die aufbereiteten Daten sind jedoch in dieser Hinsicht plausibel und können weiterverwendet werden. Um den Detailierungsgrad zu erhöhen, sollten die Kosten der Transportmittel, Fahrer und Bestellungen aufgegliedert und in das Modell eingefügt werden.

Alle genannten Parameter, deren Werte nicht plausibel sind, sollten im Datenmodell korrigiert werden. Das beinhaltet die Anzahl Transportmittel in *Routecylce* und die maximale Tourenlänge in der Tabelle *Route*. Eine Empfehlung für zu verwendende Werte wird hier nicht gegeben, da eine Datenreduktion in Bezug auf die Bestellzahlen dennoch möglich ist und viele Werte mit diesen in ein plausibles Verhältnis gesetzt werden müssen. Plausible Verhältnisse sollten aus dem Kapitel 4.1.2 entnommen werden, sie orientieren sich an den Rohdaten. Falls die Emissionen in der Simulation nicht betrachtet werden, müssen keine anderen Werte eingegeben werden. Eine Änderung der Granularität der Koordinaten ist nicht zwingend notwendig, da der Aufwand bereits in der Phase der Datenaufbereitung erfolgte. Ein möglicherweise größerer Rechenaufwand durch mehr Nachkommastellen ist vernachlässigbar.

4.3 Bewertung der Datenquelle des Simulationsmodells

Eine Durchführung der Simulation auf Grundlage der aufbereiteten Daten wird nicht empfohlen. Die erkannten Fehler beim Datentransformationsprozess sind erheblich und mindern die Glaubwürdigkeit der Simulationsergebnisse. Dies hat Auswirkungen auf die Akzeptanz der Simulationsstudie und ihrer Ergebnisse.

Sofern die aufbereiteten Daten jedoch in den beschriebenen Punkten korrigiert und die Verbesserungsvorschläge befolgt werden, steht einer Verwendung in der Simulation nichts entgegen. Das überarbeitete Model spiegelt dabei das Verhalten des realen Systems im geforderten Detailierungsgrad wieder. Durch die V&V wird außerdem Glaubwürdigkeit hergestellt, die wiederum die Akzeptanz der Ergebnisse fördert. Dies bedeutet nicht, dass das Datenmodell keine weiteren Fehler aufweist. Die Arbeit

und die Methoden der V&V weisen ausschließlich auf Fehler hin, können jedoch nicht die Abwesenheit von Fehlern beweisen. Eine wiederholte Prüfung mit Hilfe der V&V nach einer Überarbeitung ist unerlässlich um weitere Fehler aufzuspüren, die entweder nicht gefunden wurden oder durch die Überarbeitung entstehen.

Eine Alternative zur Überarbeitung der aufbereiteten Daten ist eine erneute Datenaufbereitung, hier bieten die Kapitel 4.1 und 4.2 eine Vielzahl an Hinweisen, welche Punkte zu beachten sind. Kapitel 3 stellt dazu eine umfangreiche Übersicht zum Sammeln der relevanten Rohdaten bereit.

5 Zusammenfassung

Ziel dieser Arbeit ist eine Validierung des Datentransformationsprozesses, der einen Rohdatensatz in eine aufbereitete Form zur Weiterverwendung in einer Simulation überführt hat. Dafür wird sich am ASIM-Vorgehensmodell orientiert, das verschiedene Phasen einer Simulationsstudie aufzählt. Die Phasen Datenbeschaffung und Datenaufbereitung werden differenziert und deren Phasenergebnisse Rohdaten und aufbereitete Daten als Begriffe eingeführt. Neben der Validierung, die im Vorgehensmodell in der Kombination Verifikation und Validierung (V&V) beschrieben wird, findet sich der Datentransformationsprozess in der Phase der Datenaufbereitung wieder. Da die Dokumentation der Phase nicht vorliegt, wird die V&V ausschließlich auf die Roh- und aufbereiteten Daten angewandt. Dazu werden Validierungstechniken vorgestellt. Als geeignete Techniken werden der Schreibtischtest und statistische Tests ausgewählt. Angelehnt an das ASIM-Vorgehensmodell erfolgt die Prüfung der aufbereiteten Daten zunächst intrinsisch, also in sich selber ohne Bezug zum realen System, und anschließend gegen die Rohdaten.

Der Datensatz wurde auf Grundlage der Rohdaten mit Hilfe vom Simulationstool SimChain aufbereitet, in dem die Daten einzeln manuell eingefügt, nach Vorgabewerten generiert oder importiert wurden. Die Daten sind in Form von Excel- und csv-Dateien vorhanden, wobei erstere alle Daten und Einstellung beinhalten und letztere eine Tabelle der Bestellungen und Lieferungen eines Jahres abbildet, die für die Weiterverwendung in der Simulationssoftware bestimmt ist.

Bei der intrinsischen Prüfung werden nicht nur die generellen Daten untersucht, sondern auch die Einstellungen, die das Modellverhalten vorgeben und begrenzen. Dabei ergeben sich einige Unstimmigkeiten zwischen den Modellparametern: Die Anzahl der Transportfahrzeuge steht in einem viel zu großen Verhältnis zu der Bestellzahl, die maximale Tourenlänge ist mit 60000 km überdimensioniert und der Detailierungsgrad der Koordinaten ist viel höher als benötigt. Ebenso wird die Verteilung der durchschnittlichen Bestellsummen pro Sektor betrachtet, die analysierte Ungleichverteilung wird als plausibel eingeschätzt. Die Verteilung kann jedoch auch durch den sehr kleinen Datensatz beeinflusst sein, beispielsweise haben 14% der Sektoren nur eine Bestellung verzeichnet.

Die Prüfung gegen die Rohdaten offenbart gravierende Unterschiede in der Größe des Datensatzes: Die Anzahl der belieferten Bezirkssektoren ist in den Rohdaten um den Faktor 30 größer, die Anzahl an Bestellungen um einen Faktor über 3000. Außerdem geben die aufbereiteten Daten nur zwei statt 16 Distributionszentren an. Zwei Zulieferer, die die Distributionszentren beliefern, fehlen ebenfalls in dem aufbereiteten Datenmodell. Um die beiden Datensätze in Hinsicht auf die Bestellzahlen zu untersuchen, wird das bereits in der intrinsischen Prüfung verwendete Diagramm hinzugezogen. Da in den Rohdaten keine Informationen über die Bestellmengen vorhanden sind, werden die Bestellmengen aus den aufbereiteten Daten der aus den Rohdaten berechneten Bestellsumme gegenübergestellt. Hier wird eine Korrelation erwartet, die der direkte Vergleich jedoch nicht nachweisen kann. Die unterschiedlichen Verteilungen sprechen gegen ein realitätsnahes Modellverhalten. Zwischen beiden Datensätzen unterscheiden sich zudem die Transportkosten in Höhe und Detailierung, sowie die Fahrzeugflottengröße im Verhältnis zu den Bestellungen erheblich.

Die These, dass es sich bei dem viel kleineren Datensatz der aufbereiteten Daten um eine beabsichtigte Datenreduktion handelt, wird ebenfalls verworfen. Der reduzierte Datensatz ist nicht repräsentativ für die Rohdaten.

Abschließend wird nicht empfohlen, die gegebenen aufbereiteten Daten in der Simulation weiterzuverwenden. Die V&V stellt erhebliche Fehler und vor allem Missverhältnisse zu den Rohdaten fest, die zu überarbeiten sind. Verbesserungsvorschläge und Hinweise bei der Datenaufbereitung sind in Kapitel 4.2 zu finden. Eine Verwendung in der Simulation wird erst nach einer anschließenden erneuten Prüfung mit Hilfe der V&V empfohlen. Die bereits ausgewählten Validierungstechniken können dazu eingesetzt werden.

Abkürzungsverzeichnis

ASIM	Arbeitsgemeinschaft Simulation
CSV	Comma-seperated values
DC	Distribution Center
DTP	Datentransformationsprozess
PLZ	Postleitzahl
SKU	Stock Keeping Unit
V&V	Verifikation und Validierung
XLSV	Excel Spreadsheet (Dateiformat)

Abbildungsverzeichnis

Literaturverzeichnis

Balci, Osman (1994): Validation, verification, and testing techniques throughout the life cycle of a simulation study. In: Jeffrey D. Tew (Hg.): 1994 Winter Simulation Conference proceedings. Lake Buena Vista, Florida, December 11 - 14, 1994. New York, NY: Assoc. for Computing Machinery.

Balci, Osman (2003): VERIFICATION, VALIDATION, AND CERTIFICATION OF MODELING AND SIMULATION APPLICATIONS. In: S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, eds. (Hg.): Proceedings of the 35th conference on Winter simulation driving innovation. Winter Simulation Conference 2003: Winter Simulation Conference (ACM Digital Library), S. 150–158.

Fechteler, Till; Gutenschwager, Kai (2014): SimChain Technical Documentation. Version 1.0: SimPlan AG.

Laroque, Christoph; Klaas, Alexander; Dangelmaier, Wilhelm (Hg.) (2013): Simulation in Produktion und Logistik 2013. [Entscheidungsunterstützung von der Planung bis zur Steuerung ; 15. ASIM Fachtagung] ; Paderborn, 09. - 11. Oktober 2013. Fachtagung Simulation in Produktion und Logistik; Gesellschaft für Informatik; ASIM-Fachtagung "Simulation in Produktion und Logistik". Paderborn: Heinz-Nixdorf-Inst. Univ. Paderborn (ASIM-Mitteilung, 147).

Melchior, Julia (2014): SimChain - Tutorial for Modeling via the User Interface. Online verfügbar unter https://www.simchain.net/images/Media/SimChain_Tutorial.pdf, zuletzt geprüft am 16.10.2018.

Rabe, Markus; Gutenschwager, Kai; Spieckermann, Sven; Wenzel, Sigrid (2017): Simulation in Produktion und Logistik. Grundlagen und Anwendungen. Berlin: Springer Vieweg.

Rabe, Markus; Spiekermann, Sven; Wenzel, Sigrid (2008): Verifikation und Validierung für die Simulation in Produktion und Logistik. Vorgehensmodelle und Techniken. Berlin, Heidelberg: Springer (VDI-Buch).

VDI 3633 - Blatt 1: Simulation von Logistik-, Materialfluss- und Produktionssystemen.

Wenzel, Sigrid; Bernhard, Jochen (2008): Definition und Modellierung von Systemlasten für die Simulation logistischer Systeme. In: Peter Nyhuis (Hg.): Beiträge zu einer Theorie der Logistik. Berlin, Heidelberg: Springer Berlin Heidelberg, S. 487–509.

Wenzel, Sigrid; Collisi-Böhmer, Simone; Pitsch, Holger; Rose, Oliver; Weiß, Matthias (2008): Qualitätskriterien für die Simulation in Produktion und Logistik. Planung und Durchführung von Simulationsstudien. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg (VDI-Buch).