

Kurzzusammenfassung

In dieser Arbeit wird ein Tool zur Clusteranalyse für zivile Verkehrsflugzeuge entwickelt und angewendet. Mit einer Clusteranalyse kann die Komplexität von Analysen zur zukünftigen Entwicklung von Flugzeugflotten und der zivilen Luftfahrt, die sich durch die große Anzahl verschiedener Flugzeugtypen ergibt, reduziert werden. Als Grundlage für die Entwicklung des Clusteranalyse-Tools werden zunächst bestehende Klassifikationen von Flugzeugen und Automobilen untersucht. Anschließend werden verschiedene Clusteranalyseverfahren vorgestellt und bewertet und ein Tool zur Clusteranalyse für zivile Verkehrsflugzeuge auf Grundlage des Medoid-Verfahrens entwickelt. Das Tool wird zur Identifikation von Cluster-Lösungen mit dem Ziel der Vereinfachung eines Flugplans aus der OAG-Datenbank angewendet. Im Rahmen der Vereinfachung wird jeder Flugzeugtyp durch den Repräsentanten seines Clusters ersetzt. Obwohl in den zugrundeliegenden Daten keine klare Clusterstruktur gefunden wird, werden die verschiedenen Flugzeugtypen durch die Cluster hinsichtlich zentraler charakteristischer Merkmale sehr gut separiert. Die Clusteranalyse kann daher als geeignete Methode zur Reduktion der durch die große Anzahl unterschiedlicher Flugzeugtypen bedingten Komplexität angesehen werden.

Abstract

In this thesis, a cluster analysis tool for civil aircraft types is developed and applied. The application of a cluster analysis is a way to reduce the complexity of analyses with respect to the future development of aircraft fleets and aviation in general. As a basis of the development of the cluster analysis tool, existing classification schemes for aircraft and passenger cars are analyzed at first. Subsequently, various algorithms for cluster analysis are presented and assessed taking their suitability for the application into account. A cluster analysis tool is developed based on the selected k-medoids algorithm. The tool is afterwards applied to simplify flight schedules from the OAG Database. For this purpose, each aircraft type is substituted by the representative aircraft of its cluster. Although the result does not show a clear cluster structure, the aircraft types appear to be properly separated regarding their characteristic attributes. The cluster analysis may therefore be considered as a suitable method in order to reduce the complexity arising from the large number of different aircraft types.

Inhaltsverzeichnis

Kurzzusammenfassung.....	2
Abstract.....	2
Inhaltsverzeichnis	3
Abbildungsverzeichnis	6
Tabellenverzeichnis	7
Abkürzungsverzeichnis	8
1 Einleitung	9
1.1 Problemstellung der Arbeit.....	9
1.2 Ziel der Arbeit	10
1.3 Inhalt und Aufbau der Arbeit.....	10
2 Bestehende Methoden zur Klassifikation ziviler Verkehrsflugzeuge und Automobile	12
2.1 Klassifikation von Flugzeugen	12
2.1.1 Klassifikation von Flugzeugen in der Wissenschaft	12
2.1.2 Klassifikation von Flugzeugen in der Luftfahrtpraxis	13
2.2 Klassifikation von Automobilen.....	15
2.3 Fazit und Implikationen für das Clustering in dieser Arbeit	16
3 Clusteranalyseverfahren	19
3.1 Einführung in die Clusteranalyse	19
3.2 Anforderungen an das Clusteranalyseverfahren.....	19
3.2.1 Muss-Anforderungen	20
3.2.2 Soll-Anforderungen.....	21
3.3 Auswahl eines Clusteranalyseverfahrens	21
3.3.1 Unvollständige, deterministische und probabilistische Clusteranalyseverfahren	22
3.3.1.1 Charakteristika der Verfahren	23
3.3.1.2 Vorauswahl geeigneter Verfahren.....	24
3.3.2 Überlappungsfreie deterministische Clusteranalyseverfahren	24
3.3.2.1 Funktionsweise der Verfahren	25
3.3.2.2 Anwendungsvoraussetzungen	27

3.3.2.3	Vorauswahl geeigneter Verfahren.....	29
3.3.3	K-Means-Verfahren und Medoid-Verfahren	32
3.3.3.1	Erklärung der Verfahren	33
3.3.3.2	Auswahl eines Verfahrens.....	34
4	Implementierung des Clusteranalyseverfahrens.....	37
4.1	Integration von Daten aus den verwendeten Datenbanken	37
4.2	Verarbeitung der Daten	38
4.2.1	Unähnlichkeitsmaß.....	38
4.2.2	Skalenniveau der Variablen	39
4.2.3	Umgang mit fehlenden Werten	41
4.2.4	Standardisierung.....	41
4.3	Struktur des Clusteranalysetools und Ablauf einer Clusteranalyse.....	41
4.4	Bestimmung der Anzahl der Cluster	46
4.4.1	Relativer Abfall der Summe der Unähnlichkeiten	46
4.4.2	Average Silhouette Width	47
5	Anwendung der Clusteranalyse auf zivile Flugzeugmuster	49
5.1	Spezifikation der Clusteranalyse	49
5.1.1	Clustering-Parameter.....	50
5.1.1.1	Festlegung der Clustering-Parameter	50
5.1.1.2	Berechnung der Clustering-Parameter	51
5.1.2	Verwendete Flugzeugtypen.....	53
5.2	Ergebnisse der Clusteranalyse	56
5.2.1	Anzahl der Cluster.....	56
5.2.2	7-Cluster-Lösung.....	58
5.2.3	17-Cluster-Lösung.....	61
5.3	Kritische Betrachtung des Clusteranalyse-Tools und der Ergebnisse.....	64
5.3.1	Erfüllung der an das Clusteranalyseverfahren gestellten Anforderungen.....	65
5.3.2	Eignung der Clusteranalyse zur Lösung der Problemstellung dieser Arbeit .	66
6	Zusammenfassung und Ausblick.....	69
	Literaturverzeichnis	72

Anlagen.....	73
Anlage 1 (Zuordnung der Flugzeuge aus OAG- und BADA-Datenbank)	74
Anlage 2 (Eidesstattliche Versicherung)	81

Abbildungsverzeichnis

Abbildung 3-1: Übersicht über die behandelten Clusteranalyseverfahren	22
Abbildung 4-1: Flussdiagramm Clustering	43
Abbildung 4-2: Flussdiagramm zur Ermittlung einer Clusterstruktur (Medoid-Verfahren)	45
Abbildung 4-3: Abnahme der Summe der Unähnlichkeiten mit steigender Anzahl der Cluster	47
Abbildung 5-1: Box Plot der verfügbaren Anzahl der Sitzplätze für 7 Cluster.....	59
Abbildung 5-2: Box Plot der verfügbaren Frachttonnen für 7 Cluster	59
Abbildung 5-3: Box Plot der verfügbaren Nutzlast für 7 Cluster.....	59
Abbildung 5-4: Box Plot mit der durchschnittlichen Flugstrecke für 7 Cluster.....	60
Abbildung 5-5: Box Plot mit der Reisefluggeschwindigkeit aus der OAG-Datenbank für 7 Cluster	60
Abbildung 5-6: Box Plot des spezifischen Kraftstoffverbrauchs für 7 Cluster	61
Abbildung 5-7: Box Plot der verfügbaren Anzahl der Sitzplätze für 17 Cluster.....	62
Abbildung 5-8: Box Plot der verfügbaren Frachttonnen für 17 Cluster.....	62
Abbildung 5-9: Box Plot der verfügbaren Nutzlast für 17 Cluster.....	63
Abbildung 5-10: Box Plot mit der durchschnittlichen Flugstrecke für 17 Cluster.....	63
Abbildung 5-11: Box Plot mit der Reisefluggeschwindigkeit aus der OAG-Datenbank für 17 Cluster	64
Abbildung 5-12: Box Plot des spezifischen Kraftstoffverbrauchs für 17 Cluster	64

Tabellenverzeichnis

Tabelle 2-1:	Zusammenfassung der Erkenntnisse aus der Analyse der Methoden zur Klassifikation ziviler Verkehrsflugzeuge und Automobile.....	17
Tabelle 3-1:	Muss-Anforderungen an das Clusteranalyseverfahren	20
Tabelle 3-2:	Soll-Anforderungen an das Clusteranalyseverfahren.....	21
Tabelle 3-3:	Anwendungsvoraussetzungen der Clusteranalyseverfahren	28
Tabelle 3-4:	Übersicht über die deterministischen, überlappungsfreien Clusteranalyseverfahren	30
Tabelle 3-5:	Vor- und Nachteile des K-Means-Verfahrens.....	35
Tabelle 3-6:	Vor- und Nachteile des Medoid-Verfahrens	36
Tabelle 4-1:	Interpretation der Average Silhouette Width	48
Tabelle 5-1:	Bei der Anwendung der Clusteranalyse verwendete OAG Specific Aircraft	53
Tabelle 5-2:	Bei der Anwendung der Clusteranalyse nicht berücksichtigte OAG Specific Aircraft	56
Tabelle 5-3:	Bewertung der Anzahl der Cluster	57

Abkürzungsverzeichnis

ACAS	Aircraft Analytical System
ACRISS	Association of Car Rental Industry Systems Standards
BADA	Base of Aircraft Data
FAA	Federal Aviation Administration
ICAO	International Civil Aviation Organization
MPV	Multi-Purpose Vehicle
OAG	Official Airline Guide
SUV	Sport Utility Vehicle
TAS	True Airspeed
VBA	Visual Basic for Applications

1 Einleitung

1.1 Problemstellung der Arbeit

Die kommerzielle Luftfahrt steht aktuell und in Zukunft vor großen Herausforderungen. Geprägt werden diese Herausforderungen beispielsweise durch den Bedarf nach einer Begrenzung von Schadstoff- und Lärmemissionen, die begrenzte Verfügbarkeit fossiler Energieträger sowie durch Kapazitätsprobleme an Flughäfen. Angesichts dieser Situation haben Analysen und Vorhersagen zu Veränderungen im Bereich der Luftfahrt, beispielsweise zur zukünftigen Entwicklung von Flugzeugflotten, eine hohe Relevanz für die Luftfahrtforschung.

Allerdings sind Analysen zur Vorhersage der zukünftigen Entwicklung von Flugzeugflotten durch die große Zahl von Flugzeugmodellen sehr komplex. Bereits im Jahr 2002 waren weltweit mehr als 300 verschiedene zivile Flugzeugmodelle und -konfigurationen im Einsatz [Eyers et al. 2004, S. 22].

Die Komplexität von Betrachtungen der zukünftigen Entwicklung der kommerziellen Luftfahrt wird durch zwei weitere Aspekte gesteigert. Der erste dieser Aspekte ist die große Anzahl der Einflussfaktoren auf die Entwicklung der kommerziellen Luftfahrt. Zu diesen Einflussfaktoren gehören beispielsweise die weltwirtschaftliche Entwicklung (Wirtschaftswachstum, Schwerpunktregionen, Grad der Globalisierung), die Entwicklung regulatorischer Rahmenbedingungen für die Luftfahrt (Steuern, Schadstoffemissionen, Lärm), die technologische Entwicklung und Weiterentwicklung von Flugzeugen (Energieeffizienz, Transportkapazität, Lärm), die Entwicklung der Kundenwünsche (Kosten, Komfort, Nachhaltigkeit), die Kosten und die Verfügbarkeit von Treibstoffen und anderen Rohstoffen oder die Kapazität von Flughäfen.

Der andere maßgebliche Aspekt ist die Komplexität der Wirkungszusammenhänge zwischen diesen Einflussfaktoren und der tatsächlichen zukünftigen Entwicklung des Luftfahrtmarktes und der Flugzeugflotten. Viele der Einflussfaktoren beeinflussen sich gegenseitig und ihre Vorhersage unterliegt großen Unsicherheiten. Zudem unterscheiden sich ihre Ausprägungen zum Teil erheblich zwischen verschiedenen Regionen der Welt. Außerdem sind die Wirkungen der Einflussfaktoren auf die zukünftige Entwicklung oft nur schwer prognostizierbar, vor allem im Zusammenspiel vieler Einflussfaktoren.

Zur Reduktion und Beherrschung dieser Komplexität gibt es verschiedene Ansätze. Im Bereich der Einflussfaktoren und deren Wirkmechanismen auf die Entwicklung der Luftfahrt setzt beispielsweise die Szenariotechnik an. Mit Zukunftsszenarien wird auf Basis der Identifikation und Auswahl der wichtigsten Einflussfaktoren und der Modellierung der Wirkungszusammenhänge zwischen diesen Einflussfaktoren eine überschaubare Anzahl plausibler Zukünfte abgeleitet. Jedes Szenario ist im Kern durch die Ausprägungen der

ausgewählten Einflussfaktoren gekennzeichnet. Aus gemeinsamen Eigenschaften verschiedener Szenarien können Vorhersagen über zukünftige Entwicklungen getroffen werden, die aufgrund der Gültigkeit in verschiedenen plausiblen Zukunftsszenarien besonders robust sind.

Ein anderer Ansatz, der auch in Kombination mit der Szenariotechnik angewendet werden kann, ist die Reduktion der Komplexität, die durch die große Anzahl verschiedener Flugzeugmodelle entsteht. Dieser Ansatz wird in dieser Arbeit mit dem Clustering, also der Gruppierung ziviler Flugzeugmuster verfolgt. Es werden aus mehreren Flugzeugtypen bestehende Cluster gebildet. Die zu einem Cluster gehörenden Flugzeugtypen sind sich untereinander bezüglich bestimmter, zuvor festgelegter Kriterien ähnlich. Die gebildeten Cluster können in anschließenden Betrachtungen jeweils stellvertretend für mehrere unterschiedliche Flugzeugmodelle verwendet werden.

1.2 Ziel der Arbeit

Ziel dieser Arbeit ist die Entwicklung eines Tools zur Durchführung von Clusteranalysen für zivile Flugzeugmuster. Die Ergebnisse der mit dem zu entwickelnden Tool durchgeführten Clusteranalysen sollen geeignet sein, die Komplexität von Flottenbetrachtungen durch eine Verringerung der Anzahl der betrachteten Flugzeugmodelle zu reduzieren. Das Clusteranalyse-Tool soll nach der Entwicklung ebenfalls im Rahmen dieser Arbeit erstmals angewendet werden. Die konkrete Problemstellung für die Anwendung ist die Vereinfachung eines Flugplans aus der OAG-Datenbank.

1.3 Inhalt und Aufbau der Arbeit

Um Erkenntnisse für die Entwicklung des Clusteranalyse-Tools zu gewinnen, werden in Kapitel 2 zunächst bestehende Systeme zur Klassifikation von Flugzeugen untersucht. Der Fokus liegt dabei auf der Systematik der verschiedenen Klassifikationen. Außerdem wird der Zusammenhang zwischen der jeweils mithilfe der Klassifikation zu lösenden Problemstellung und dem gewählten Ansatz zur Klassifikation untersucht. Da die systematische Klassifizierung von Flugzeugtypen bisher nur in relativ wenigen Fällen angewendet wird, wird zusätzlich die Systematik einiger Klassifikationen von Automobilen betrachtet.

In Kapitel 3 werden die allgemeinen theoretischen Grundlagen des Clustering behandelt und ein Clusteranalyseverfahren für das Clusteranalyse-Tool ausgewählt. Dafür werden zunächst die Anforderungen an ein für die Problemstellung dieser Arbeit geeignetes Clusteranalyseverfahren definiert. Anschließend werden verschiedene Clusteranalyseverfahren vorgestellt und anhand der Anforderungen hinsichtlich ihrer Eignung für diese Arbeit bewertet. Schließlich wird aus den vorgestellten Methoden eine besonders geeignete Methode ausgewählt.

In Kapitel 4 wird die Entwicklung des Clusteranalyse-Tools für zivile Flugzeugmuster beschrieben. Dabei werden insbesondere die verwendeten Datenbanken und die genaue Spezifikation des Clusteranalyseverfahrens hinsichtlich der Verarbeitung der Daten betrachtet. Außerdem werden die Struktur und der Aufbau des Clusteranalyse-Tools erklärt.

In Kapitel 5 wird das entwickelte Tool zur Clusteranalyse angewendet. Die konkrete Problemstellung für diese Anwendung ist die Vereinfachung eines Flugplans mithilfe einer Clusteranalyse für die enthaltenen Flugzeugtypen. Zunächst werden die für die Problemstellung relevanten Clustering-Parameter identifiziert und die Datenbasis für die Analyse erklärt. Es folgen die Auswertung der Ergebnisse sowie eine kritische Betrachtung des entwickelten Clusteranalyse-Tools und der Eignung der Clusteranalyse für die Lösung der Problemstellung dieser Arbeit.

Kapitel 6 fasst die Ergebnisse der Arbeit zusammen und gibt Anregungen für mögliche Weiterentwicklungen.

2 Bestehende Methoden zur Klassifikation ziviler Verkehrsflugzeuge und Automobile

In diesem Kapitel werden bestehende Klassifikationen ziviler Verkehrsflugzeuge aus der Wissenschaft und der Luftfahrtpraxis vorgestellt. Ziel des Kapitels ist die Gewinnung von Erkenntnissen für das Clustering von Flugzeugen in dieser Arbeit. Dazu werden Hintergrund, Zielsetzung, Methodik und Charakteristika ausgewählter bestehender Klassifikationen vorgestellt. Um neben den bestehenden Einteilungen von Flugzeugen weitere Erkenntnisse zur Methodik zu gewinnen, werden auch zwei bestehende Einteilungen von Automobilen betrachtet.

In Abschnitt 2.1 werden die bestehenden Klassifikationen von Flugzeugen und in Abschnitt 2.2 Segmentierungen von Automobilen vorgestellt. Abschnitt 2.3 fasst die gewonnenen Erkenntnisse für das Clustering von Flugzeugen im Rahmen dieser Arbeit zusammen.

2.1 Klassifikation von Flugzeugen

2.1.1 Klassifikation von Flugzeugen in der Wissenschaft

In einer Studie über die Entwicklung der Lärmbelastung durch Flugzeuge an Flughäfen in der Europäischen Union hat ANOTEC Consulting (2003) die in Europa zugelassenen Flugzeuge in sieben generische Klassen eingeteilt. Basis für die Einteilung war die Anzahl der Sitzplätze in der vom Flugzeughersteller angegebenen Standardkonfiguration. Die kleinste generische Klasse umfasst in dieser Einteilung alle Flugzeuge mit bis zu 80 Sitzplätzen, die größte Klasse alle Flugzeuge mit mehr als 500 Sitzplätzen [ANOTEC Consulting 2003, S. I-32 - I-34].

Die generischen Klassen werden in dieser Studie verwendet, um die Vorhersage des zukünftigen Luftverkehrswachstums gemessen als Anzahl der Flugbewegungen zu vereinfachen. Hierzu erfolgt keine Betrachtung auf der Ebene jedes einzelnen Flugzeugtyps, sondern auf der Ebene der generischen Klassen. Zusammen mit der Flottenzusammensetzung zum Zeitpunkt der Erstellung der Studie und der Vorhersage der Entwicklung der Flottenzusammensetzung in der Zukunft wird damit die Anzahl der Flugbewegungen für verschiedene Flugzeugtypen vorhergesagt. Diese dient als Basis für die Berechnung der Lärmbelastung an Flughäfen [ANOTEC Consulting 2003, S. I-42 - I-44].

Eine detailliertere Einteilung von Flugzeugtypen in verschiedene Klassen haben Eyers et al. (2004) in ihrer Studie zur Untersuchung und Vorhersage der Emissionsbilanz der globalen Luftfahrt vorgenommen. Sie haben eine Flotte von 300 Flugzeugtypen durch die Auswahl von 40 repräsentativen Flugzeugtypen vereinfacht abgebildet. Dazu haben sie aus der Gesamtheit der 300 Flugzeugtypen eine Untermenge von 40 Flugzeugtypen identifiziert,

welche jeden Flugzeugtyp aus der Gesamtheit hinsichtlich Leistungsvermögen, Treibstoffverbrauch und Emissionen ausreichend gut repräsentieren können. [Eyers et al. 2004, S. 21-22].

Zunächst haben Eyers et al. die Flugzeugtypen in vier Hauptkategorien eingeteilt. Den überwiegenden Teil der gesamten Flotte macht die Hauptkategorie der großen Jets mit mehr als 100 Sitzplätzen aus. In die Hauptkategorie der Regional-Jets fallen Kurzstreckenflugzeuge mit Turbofan- oder Turbojet-Triebwerken und bis zu 100 Sitzplätzen. Die Gruppe der Turboprop-Flugzeuge umfasst alle Flugzeuge, die durch zwei oder mehr Turboprop-Triebwerke angetrieben werden. Die letzte Hauptkategorie, die Business-Jets beinhalten kleine Turbofan- oder Turboprop-Flugzeuge mit weniger als etwa 20 Sitzplätzen [Eyers et al. 2004, S. 22].

Anschließend haben Eyers et al. innerhalb der Hauptkategorien vier Unterkategorien gebildet und jeden Flugzeugtyp in eine Unterkategorie eingeordnet. Dafür haben sie die vier Kriterien Anzahl der Sitzplätze, verwendete Triebwerkstechnologien, maximales Startgewicht und grundlegende Konfiguration des Flugzeugs (Anzahl der Triebwerke etc.) verwendet. Aus jeder Unterkategorie wurde anschließend jener Flugzeugtyp als repräsentativer Vertreter für diese Unterkategorie ausgewählt, von dem global die größte Anzahl von Flugzeugen in Betrieb war [Eyers et al. 2004, S. 23-28].

Die Motivation für die Identifikation der 40 repräsentativen Flugzeugtypen war die Vereinfachung der anschließenden Ermittlung von Treibstoffverbrauch und Emissionen. Eine Betrachtung ohne diese Vereinfachung wäre aufgrund der großen Anzahl der Flugzeugtypen deutlich aufwändiger. Zudem waren die verwendeten Daten zum Treibstoffverbrauch nicht für alle Flugzeugtypen verfügbar. Die Emissionen der zivilen Luftfahrt wurden entsprechend zwar auf Basis verschiedener Datenbanken über Flugbewegungen mit Informationen zum jeweiligen Flugzeugtyp ermittelt, jedoch wurde für die Berechnung der Emissionen für die Flüge nicht von den tatsächlichen Flugzeugtypen, sondern von den festgelegten repräsentativen Flugzeugen mit ihren jeweiligen repräsentativen Triebwerken ausgegangen [Eyers et al. 2004, S. 11, 21-25, 39].

2.1.2 Klassifikation von Flugzeugen in der Luftfahrtpraxis

Neben der Gruppierung in wissenschaftlichen Studien werden Flugzeuge auch in der Luftfahrtpraxis für verschiedene Zwecke in Gruppen zusammengefasst. In diesem Abschnitt werden die Einteilung von Flugzeugen durch Fluggesellschaften sowie die Einteilung nach Wirbelschleppenkategorien durch Luftfahrtbehörden vorgestellt.

Umgangssprachlich ist die Einteilung von Flugzeugen in Kurz-, Mittel- und Langstreckenflugzeuge weit verbreitet. Die Begriffe Kurzstrecken-, Mittelstrecken- und Langstreckenflugzeug werden jedoch von den Flugzeugherstellern nicht definiert und haben keine sonstige allgemeingültige Definition. Jede Fluggesellschaft verwendet zudem eine eigene Ein-

teilung ihrer Flotte. In diese werden die eingesetzten Flugzeuge entsprechend dem jeweiligen Einsatzzweck im Unternehmen eingeordnet.

Die Deutsche Lufthansa teilt ihre Flotte beispielsweise in die Interkontinental-Flotte, die Kontinental-Flotte, die Regional-Flotte und die Cargo-Flotte ein. Zur Interkontinental-Flotte gehören die Airbus-Modelle A330, A340 und A380 sowie die Boeing 747. Die Kontinental-Flotte besteht aus der Airbus A320-Familie, sowie der Boeing 737 [Deutsche Lufthansa AG 2012]. Air France-KLM verwendet eine ähnliche Einteilung und unterscheidet zwischen Langstrecken-Flotte, Mittelstrecken-Flotte, Regional-Flotte und Fracht-Flotte. Die Einteilung der Flugzeugtypen in die Kategorie der Langstrecken-Flotte entspricht der Interkontinental-Flotte der Lufthansa, beinhaltet jedoch zusätzlich die Boeing 777 und die MD-11, die nicht zur Flotte der Lufthansa gehören. Die Mittelstrecken-Flotte von Air France-KLM besteht wie die Kontinental-Flotte der Lufthansa aus der Airbus A320-Familie und der Boeing 737 [Air France-KLM S.A. 2012].

Im Gegensatz zu Lufthansa und Air France-KLM unterscheidet die amerikanische United Continental Holdings lediglich zwischen „Mainline Aircraft“ und „Regional Aircraft“. Zu den „Mainline Aircraft“ gehören jene Flugzeugmodelle, die bei der Lufthansa und Air France-KLM zu den Interkontinental- und Kontinental-Flotten, beziehungsweise zur Langstrecken- und Mittelstrecken-Flotte gehören, sowie die Boeing 767 und 757, die zusätzlich zur Flotte von United Airlines gehören. Zu den „Regional Aircraft“ werden Flugzeugtypen gezählt, die bei Lufthansa und Air France-KLM typischerweise zur Regional-Flotte gehören [United Continental Holdings 2012; Deutsche Lufthansa AG 2012; Air France-KLM S.A. 2012].

Zum Zwecke der Festlegung der minimalen Abstände verschiedener Flugzeuge zur Vermeidung gefährlicher Turbulenzen durch die Wirbelschleppen vorausfliegender Flugzeuge teilt die amerikanische Luftsicherheitsbehörde „Federal Aviation Administration“ (FAA) die Flugzeuge in verschiedene Gewichtskategorien ein. In die kleinste Kategorie „Small“ fallen alle Flugzeuge mit einem maximalen Startgewicht von bis zu 41.000 Pfund (ca. 18,6 Tonnen). In die Klasse „Large“ werden Flugzeuge mit einem maximalen Startgewicht von mehr als 41.000 Pfund aber weniger als 300.000 Pfund (ca. 136,1 Tonnen) eingeteilt. Die Kategorie „Heavy“ umfasst Flugzeuge mit einem maximalen Startgewicht von mindestens 300.000 Pfund [Federal Aviation Administration 2011a, S. 5-5-3, Appendix A-1].

Für den Airbus A380-800 gilt in den USA derzeit eine temporäre Ergänzung der eigentlichen Richtlinien, die die spezielle Kennzeichnung mit der Kategorie „Super“ beinhaltet [Federal Aviation Administration 2011b, S. 1].

Eine sehr ähnliche Einteilung bezüglich der Vermeidung von Unfällen durch Wirbelschleppen mit den Kategorien „Light“, „Medium“, „Heavy“ und „Super“, jedoch mit teilweise abweichenden Gewichtsgrenzen, verwendet die International Civil Aviation Organization (deutsch: Internationale Zivilluftfahrtorganisation, ICAO) [International Ci-

vil Aviation Organization 2006, S. A-1; International Civil Aviation Organization 2007, S. 4-11].

Sowohl die Einteilung durch die FAA, also auch jene durch die ICAO dienen der Reduktion der Komplexität. Statt für jede Kombination zweier Flugzeugtypen Mindestabstände je nach vorausfliegendem Flugzeug festzulegen, ist bei diesen Einteilungen nur eine überschaubare Anzahl an Mindestabständen festzulegen. Durch individuelle Mindestabstände könnte bei Wahrung eines konstanten Mindestmaßes an Sicherheit durch eine teilweise engere Staffelung der Flugzeuge zwar die Kapazität von Flughäfen erhöht werden. Allerdings müsste eine sehr große Zahl von Mindestabständen aufwändig ermittelt werden; für 200 verschiedene Flugzeugtypen ergäben sich bereits 40.000 Kombinationen aus voraus- und hinterherfliegendem Flugzeug. Eine solch große Zahl von Mindestabständen wäre beispielsweise für Fluglotsen und Piloten kaum handhabbar.

Die Einteilung in Gruppen erfolgt auf Basis der Ausprägung einer einzigen Variablen, dem maximalen Startgewicht. Dieses hat einen ausreichend großen Einfluss auf die Intensität der vom vorausfliegenden Flugzeug gebildeten Wirbelschleppen sowie die Resistenz des hinterherfliegenden Flugzeugs gegenüber Wirbelschleppen, dass bei einer maximalen Vereinfachung die nötige Aussagekraft der Einteilung erhalten bleibt.

2.2 Klassifikation von Automobilen

Der Informations- und Datenanbieter IHS Global Insight verwendet für die Automobilindustrie zwei verschiedene Arten von Segmentierungen. Die sogenannten hausinternen Segmentierungen sind universelle Segmentierungen, die für verschiedene wiederkehrende Zwecke verwendet werden. Allgemein dienen sie vor allem der Analyse und Prognose der Entwicklung verschiedener Fahrzeugklassen. Neben den allgemeinen, hausinternen Segmentierungen gibt es kundenspezifische Segmentierungen. Diese werden individuell jeweils für eine spezifische Problemstellung des Kunden von IHS Global Insight erstellt.

Die hausinternen Segmentierungen dienen der Einteilung konkurrierender Fahrzeuge in Gruppen, deren zugehörige Fahrzeuge von potentiellen Käufern als mögliche Kaufobjekte wahrgenommen werden. Es werden also solche Fahrzeuge in Gruppen zusammengefasst, die aus Käufersicht eine möglichst hohe Ähnlichkeit besitzen. Die wichtigsten Zielgruppen dieser Segmentierungen sind Automobilhersteller und Automobilzulieferer. Für sie dienen die Segmentierungen der Analyse der Entwicklung der Markt- und Wettbewerbssituation in den relevanten Segmenten. [IHS Global Insight 2009, S. 408].

Eine der hausinternen Segmentierungen von IHS Global Insight verwendet die Größe der Fahrzeuge sowie die Positionierung im Markt, insbesondere hinsichtlich des Preises, als grundlegende Kriterien. Diese grundlegenden Kriterien werden unter anderem durch die Außen- und Innenabmessungen der Fahrzeuge, den Preis, Wettbewerbsmodelle und die

Marke repräsentiert. Für die Segmentierung werden die Fahrzeuge zunächst nach der grundsätzlichen Bauart in Fahrzeuge „klassischer Bauart“, Multi-Purpose Vehicles (MPV), Sport Utility Vehicles (SUV) und einige weitere Kategorien eingeteilt. Diese Kategorien werden anschließend nach verschiedenen Kriterien in die einzelnen Segmente unterteilt. Eine besondere Rolle spielen dabei Gesamtlänge und Preis des Fahrzeugs. Die Kriterien und deren Gewichtung unterscheiden sich jedoch zwischen den verschiedenen Kategorien. Zudem sind für bestimmte Segmente Besonderheiten definiert, um alle Fahrzeuge sinnvoll in die Segmente einzuordnen [IHS Global Insight 2009, S. 408-413].

Die kundenspezifischen Segmentierungen resultieren aus der Tatsache, dass die Kunden von IHS Global Insight (beispielsweise Automobilhersteller) für spezifische Fragestellungen maßgeschneiderte Segmentierungen des Marktes benötigen. Zu diesen Fragestellungen gehören beispielsweise die Vorbereitung der Entscheidung über den Eintritt in ein neues Marktsegment mit einem neuen Modell, die Analyse der Positionierung in einem bestimmten Segment im Vergleich zu den Wettbewerbern oder die Analyse der Wettbewerbssituation in verschiedenen Marktsegmenten. Je nach Zielsetzung sind die kundenspezifischen Segmentierungen entweder global einheitlich oder unterscheiden sich für verschiedene Regionen mit verschiedenen Marktcharakteristika.

Eine andere Einteilung von Automobilen nehmen Mietwagenunternehmen vor. Der Verband von Mietwagen- und anderen Reiseunternehmen ACRISS („Association of Car Rental Industry Systems Standards“) ist für diese international einheitliche Einteilung in verschiedene Fahrzeugkategorien verantwortlich. Die Einteilung der Fahrzeuge anhand von fünf verschiedenen Kriterien wird in einem vierstelligen Buchstabencode eindeutig abgebildet. Die Kriterien sind die Kategorie des Fahrzeugs (zum Beispiel Kleinwagen, Kompaktklasse, Mittelklasse, Luxusklasse), der Typ des Fahrzeugs (Bauart, zum Beispiel Kombi, Cabriolet, Geländewagen), die Art des Getriebes (Schaltgetriebe oder Automatikgetriebe), die Kraftstoffart und das Vorhandensein oder das Nicht-Vorhandensein einer Klimaanlage [ACRISS - Association of Car Rental Industry Systems Standards 2011].

2.3 Fazit und Implikationen für das Clustering in dieser Arbeit

Tabelle 2-1 fasst die Erkenntnisse aus der vorangegangenen Analyse der bestehenden Methoden zur Klassifikation ziviler Verkehrsflugzeuge und Automobile zusammen.

Tabelle 2-1: Zusammenfassung der Erkenntnisse aus der Analyse der Methoden zur Klassifikation ziviler Verkehrsflugzeuge und Automobile

Bereich	Erkenntnisse
Klassifikation von Flugzeugen	<ul style="list-style-type: none">• Methoden zur Klassifikation sind kaum verbreitet• Bestehende Klassifikationen sind stark zweckorientiert; meist wird dazu lediglich nach einem oder wenigen Kriterien gruppiert• Es wurde keine Anwendung der Clusteranalyse gefunden
Klassifikation von Automobilen	<ul style="list-style-type: none">• Es gibt ein einheitliches Klassifikationssystem für einige Standardanwendungen, für die meisten Anwendungen werden jedoch Segmentierungen für die spezifische Problemstellung erstellt• Es wurde keine Anwendung der Clusteranalyse gefunden

Zu den Einteilungen von Flugzeugtypen in verschiedene Gruppen lässt sich zusammenfassend feststellen, dass diese insgesamt bisher wenig verbreitet sind. In der wissenschaftlichen Literatur finden sich ebenso wie in der Luftfahrtpraxis nur relativ wenige verwendete Systeme zur Klassifikation von Flugzeugen. Die existierenden Klassifikationen sind stark auf ihren jeweiligen Zweck, zum Beispiel die Reduktion der Komplexität von Emissionsbetrachtungen auf Flottenebene, ausgerichtet. Für diese speziellen Zwecke reicht meist eine einfache Gruppierung nach einem oder wenigen Kriterien aus. Ein echtes Clustering auf Basis eines Clusteranalyseverfahrens konnte in der wissenschaftlichen Literatur und in praktischen Anwendungen nicht gefunden werden. Bei den Fluggesellschaften unterscheidet sich die Einteilung beispielsweise nach geografischen Gegebenheiten. Die amerikanische United Continental Holdings unterscheidet nicht zwischen Interkontinental-Flotte und Kontinental-Flotte. Entfernungen, die ausgehend von Europa in der Regel interkontinentale Strecken entsprechen, können in Nordamerika auch kontinentale Strecken sein.

Auch im Bereich der Automobile gibt es keine universell anwendbare Segmentierung. Bei IHS Global Insight wird für einige Standardanwendungen zwar eine einheitliche Einteilung verwendet, für spezifische Fragestellungen werden jedoch jeweils eigene Segmentierungen entwickelt. Die Kriterien für jede dieser Segmentierungen werden von der jeweils zu lösenden Problemstellung abgeleitet.

In dieser Arbeit wird eine Einteilung entwickelt, die für verschiedene Anwendungen geeignet ist. Dafür sind zwei grundsätzliche Ansätze vorstellbar: Die Entwicklung einer universellen Einteilung unter Berücksichtigung aller für die verschiedenen Anwendungen relevanten Kriterien oder ein System, mit dem je nach Anwendung automatisch eine geeignete Einteilung vorgenommen werden kann. Eine fixe Einteilung in verschiedene Grup-

pen mithilfe eines umfassenden Katalogs von Kriterien reicht für die Problemstellung dieser Arbeit jedoch nicht aus. Daher wird in dieser Arbeit die Clusteranalyse weiter verfolgt, die es ermöglicht, mit relativ geringem Aufwand anwendungsspezifische Klassifikationen zu entwickeln.

3 Clusteranalyseverfahren

In diesem Kapitel werden Clusteranalyseverfahren vorgestellt und ein für die Problemstellung in dieser Arbeit geeignetes Clusteranalyseverfahren systematisch ausgewählt. Dazu wird in Abschnitt 3.1 zunächst eine Einführung in den Begriff der Clusteranalyse gegeben. In Abschnitt 3.2 werden die Anforderungen an ein Clusteranalyseverfahren für die Problemstellung dieser Arbeit entwickelt. In Abschnitt 3.3 werden die verschiedenen Arten von Clusteranalyseverfahren vorgestellt und anhand der entwickelten Anforderungen ein geeignetes Clusteranalyseverfahren ausgewählt.

3.1 Einführung in die Clusteranalyse

Die Clusteranalyse stellt eine Gruppe multivariater Analyseverfahren dar, mit denen ähnliche Objekte gruppiert werden. Die so entstandenen Gruppen von Objekten werden Cluster genannt. Ein Cluster sollte in sich möglichst homogen sein, das heißt die Objekte des Clusters sollten sich möglichst ähnlich sein. Im Gegensatz dazu sollten die verschiedenen Cluster möglichst heterogen, also unähnlich zueinander sein [Bacher et al. 2010, S. 15-18; Backhaus et al. 2011, S. 19, 397].

Man unterscheidet zwischen objekt- und variablenorientierten Clusteranalyseverfahren. Objektorientierte Clusteranalyseverfahren dienen der Ermittlung von Clustern verschiedener Objekte, die durch die Ausprägungen ihrer Merkmale gekennzeichnet sind. Bei einer variablenorientierten Clusteranalyse werden hingegen Cluster verschiedener, ähnlicher Merkmale (Variablen) der Objekte identifiziert [Bacher et al. 2010, S. 15-22]. Am Beispiel von Flugzeugen wäre das Ergebnis einer objektorientierten Clusteranalyse eine Reihe von Clustern von Flugzeugen, die sich in bestimmten Merkmalen, beispielsweise Anzahl der Sitzplätze, Flügelspannweite und Anzahl der Triebwerke ähnlich sind. Bei einer variablenorientierten Clusteranalyse könnten beispielsweise die Anzahl der Sitzplätze, das maximale Startgewicht und die Flügelspannweite ein Cluster bilden, weil zwischen diesen Merkmalen für die betrachteten Objekte eine hohe Korrelation besteht. Im Rahmen dieser Arbeit wird eine objektorientierte Clusteranalyse durchgeführt.

3.2 Anforderungen an das Clusteranalyseverfahren

In diesem Abschnitt werden die Anforderungen an ein Clusteranalyseverfahren für Flugzeuge im Rahmen dieser Arbeit entwickelt. Sie ergeben sich aus den spezifischen Aspekten der in dieser Arbeit behandelten Problemstellung sowie aus der geplanten Anwendung der entwickelten Methodik in der Forschung am Lehrstuhl für Luftfahrtsysteme der Technischen Universität München (siehe Abschnitte 1.1 und 2.3).

3.2.1 Muss-Anforderungen

Tabelle 3-1 zeigt die Bezeichnungen und Beschreibungen der durch das Clusteranalyse zu erfüllenden Muss-Anforderungen.

Tabelle 3-1: Muss-Anforderungen an das Clusteranalyseverfahren

Bezeichnung	Beschreibung
Objektorientierte Clusteranalyse	Das Verfahren muss für eine objektorientierte Clusteranalyse geeignet sein, weil Cluster von Objekten (Flugzeugtypen) und nicht von Ausprägungen von Variablen ermittelt werden sollen.
Eindeutige Zuordnung	Das Ergebnis des Verfahrens muss eine eindeutige Zuordnung von Flugzeugtypen zu Clustern sein; dabei muss jeder Flugzeugtyp genau einem Cluster angehören.
Unbegrenzte Parameteranzahl	Das Verfahren muss mit einer beliebig großen Anzahl von Clustering-Parametern (als Variablen verwendete Merkmale der Flugzeuge) anwendbar sein.
Flexibilität hinsichtlich Parametern	Das Verfahren muss mit verschiedenen Clustering-Parametern anwendbar sein; die Clustering-Parameter können also je nach Anwendungszweck des Clustering frei ausgewählt werden.
Flexibilität hinsichtlich Flugzeugtypen	Das Verfahren muss auf verschiedene Flugzeugflotten anwendbar sein; es muss also robust gegenüber der Streichung von Flugzeugen aus der Betrachtung, der Hinzunahme weiterer Flugzeuge oder der Betrachtung einer Teilgesamtheit einer Flotte sein.
Funktionsfähigkeit mit Datenmatrix	Das Verfahren muss auf Basis einer vorgegebenen Datenmatrix funktionieren, das heißt es liegen Informationen zu den relevanten Variablen zu jedem der betrachteten Flugzeuge vor; eine Ähnlichkeits- oder Unähnlichkeitsmatrix der verschiedenen Flugzeuge liegt zu Beginn nicht vor.
VBA	Das Verfahren muss als Algorithmus in Visual Basic for Applications (VBA) implementiert werden können.

3.2.2 Soll-Anforderungen

Tabelle 3-2 zeigt die Bezeichnungen und Beschreibungen der Anforderungen, die durch das Clusteranalyseverfahren möglichst erfüllt werden sollen.

Tabelle 3-2: Soll-Anforderungen an das Clusteranalyseverfahren

Bezeichnung	Beschreibung
Anzahl der Cluster	Das Verfahren soll eine sinnvolle Anzahl von Clustern ermitteln auf Basis eines vorgegebenen Bereichs für diese Anzahl durch den Nutzer.
Anschauliche Darstellung	Das Ergebnis der Clusteranalyse soll möglichst anschaulich dargestellt werden können.
Abgrenzung der Cluster	Die Aspekte, in denen sich die identifizierten Cluster voneinander unterscheiden, sollen dargestellt werden können.
Einfache Anwendung	Die Durchführung einer Clusteranalyse soll ohne detaillierte Kenntnis der Clusteranalyseverfahren und mit möglichst geringem Arbeitsaufwand möglich sein.
Zuverlässigkeit	Das Verfahren soll zuverlässig gute oder im besten Fall optimale Ergebnisse liefern; der Einfluss der Daten auf die Qualität der Ergebnisse soll so gering wie möglich sein.
Reproduzierbarkeit der Ergebnisse	Das Ergebnis soll reproduzierbar sein, also nicht vom Zufall abhängen.

3.3 Auswahl eines Clusteranalyseverfahrens

In diesem Abschnitt werden unterschiedliche Clusteranalyseverfahren vorgestellt. Abbildung 3-1 zeigt eine Übersicht über die behandelten Verfahren. In mehreren Schritten werden jeweils die für diese Arbeit am besten geeigneten Verfahren für die weitere, detailliertere Betrachtung ausgewählt, bis am Ende die Entscheidung für ein Clusteranalyseverfahren getroffen wird.

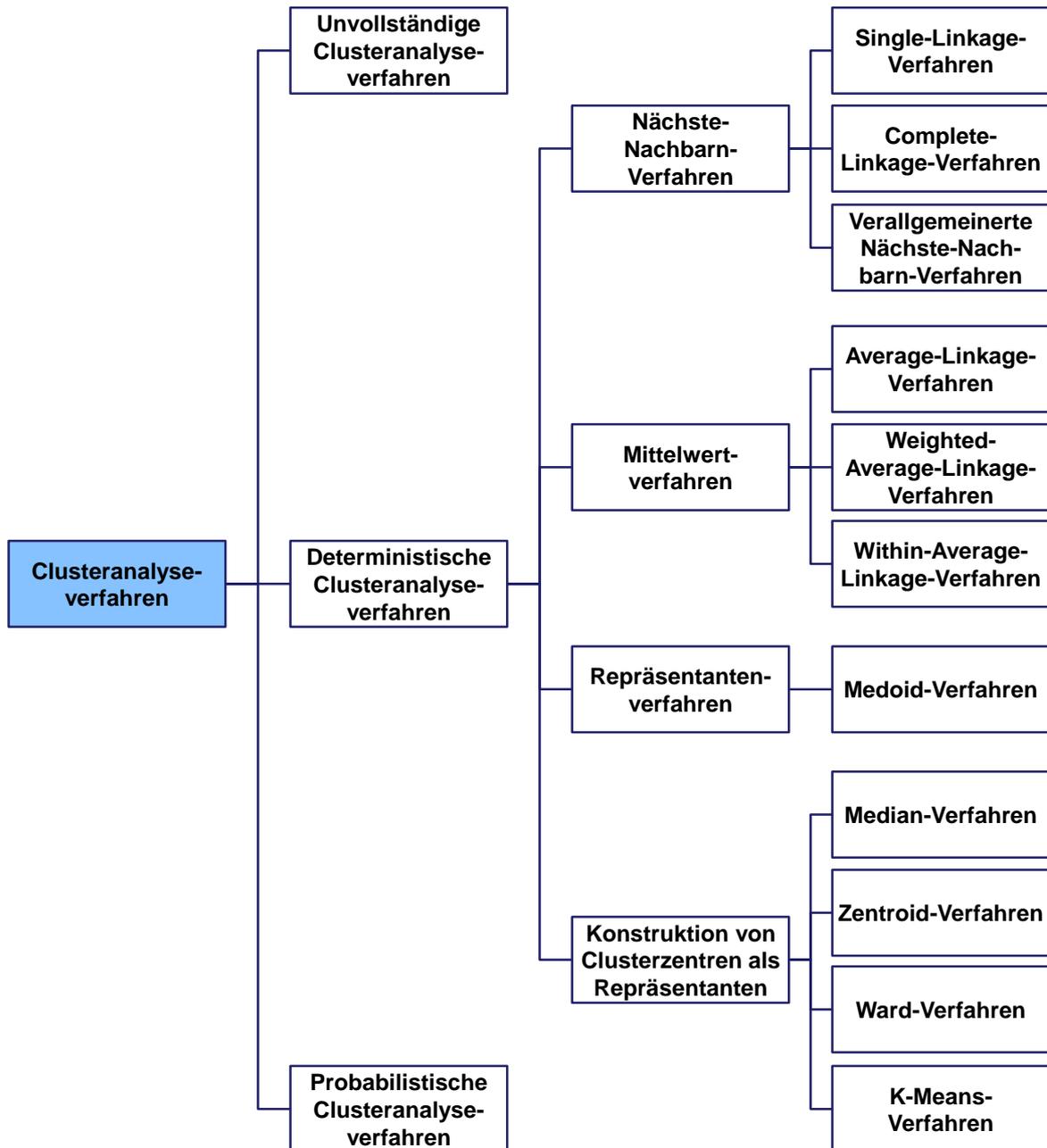


Abbildung 3-1: Übersicht über die behandelten Clusteranalyseverfahren

3.3.1 Unvollständige, deterministische und probabilistische Clusteranalyseverfahren

In diesem Abschnitt werden die prinzipiellen Funktionsweisen der verschiedenen grundlegenden Arten von Clusteranalyseverfahren erklärt und die geeigneten Verfahren für die weitere Betrachtung ausgewählt.

3.3.1.1 Charakteristika der Verfahren

Bacher et al. (2010) unterscheiden auf Basis der Zuordnung der Objekte zu den Clustern zwischen drei Arten von Verfahren zur Clusteranalyse: Die unvollständigen, die deterministischen und die probabilistischen Clusteranalyseverfahren.

Unvollständige Clusteranalyseverfahren

Unvollständige Clusteranalyseverfahren sind dadurch gekennzeichnet, dass die Zuordnung der Objekte zu den Clustern nicht durch das Verfahren, sondern durch den Anwender erfolgt. Unvollständige Clusteranalyseverfahren dienen lediglich der räumlichen Darstellung der Objekte, um dem Anwender deren Zuordnung zu Clustern zu erleichtern [Bacher et al. 2010, S. 18-19].

Deterministische Clusteranalyseverfahren

Mit deterministischen Clusteranalyseverfahren wird jedes Objekt automatisch einem oder, je nach Verfahren, möglicherweise mehreren Clustern zugeordnet. Verfahren, bei denen jedes Objekt nur einem Cluster zugeordnet wird, nennt man überlappungsfreie, nicht überlappende oder disjunktive Clusteranalyseverfahren. Verfahren, bei denen jedes Objekt auch mehreren Clustern angehören kann, nennt man überlappende Clusteranalyseverfahren [Bacher et al. 2010, S. 147].

Man unterscheidet bei den deterministischen Clusteranalyseverfahren weiter zwischen hierarchischen Verfahren und partitionierenden Verfahren. Die hierarchischen Verfahren sind durch eine schrittweise Bildung der Cluster gekennzeichnet. Bei den hierarchisch-agglomerativen Verfahren stellt als Ausgangspunkt jedes Objekt ein eigenes Cluster dar. Die Anzahl der Cluster wird davon ausgehend durch schrittweise Zusammenfassung der jeweils ähnlichsten Cluster mit jedem Schritt um Eins reduziert. Bei den hierarchisch-divisiven Verfahren werden ausgehend von einem einzigen, alle Objekte enthaltenden Cluster durch Zerteilung mehrere Cluster gebildet. Mit jedem Schritt steigt die Zahl der Cluster um Eins. Bei den partitionierenden Verfahren wird im Gegensatz zu den hierarchischen Verfahren von Beginn an eine Clusterzahl vorgegeben. Ausgehend von einer Anfangszuordnung der Objekte zur vorgegebenen Clusteranzahl werden die Cluster durch Verschieben der Objekte zwischen den Clustern schrittweise optimiert [Bacher et al. 2010, S. 19; Backhaus et al. 2011, S. 418].

Probabilistische Clusteranalyseverfahren

Mit probabilistischen Clusteranalyseverfahren werden die einzelnen Objekte nicht eindeutig einem oder mehreren Clustern zugeordnet. Stattdessen ist das Ergebnis des Clustering eine Reihe von Wahrscheinlichkeiten für die Zugehörigkeit von Objekten zu Clustern. Für jede Objekt-Cluster-Kombination wird die Wahrscheinlichkeit der Zugehörigkeit dieses

Objekts zu diesem Cluster angegeben. Die probabilistischen Clusteranalyseverfahren stellen damit eine Verallgemeinerung der deterministischen Clusteranalyseverfahren dar. Bei deterministischen Clusteranalyseverfahren werden die Objekte eindeutig bestimmten Clustern zugeordnet (bei probabilistischen Verfahren entspräche dies einer Wahrscheinlichkeit von 1) und anderen Clustern nicht zugeordnet (entspräche einer Wahrscheinlichkeit von 0) [Bacher et al. 2010, S. 19-20].

3.3.1.2 Vorauswahl geeigneter Verfahren

Die unvollständigen Clusteranalyseverfahren sind maximal für das Clustering nach drei verschiedenen Merkmalen geeignet, um eine maximal dreidimensionale Darstellung zu erhalten [Bacher et al. 2010, S. 19]. Mehr als drei Dimensionen sind nur schwer in einer Darstellung abzubilden und sind durch den Benutzer nur schwer vorstellbar. Eine Anforderung an das in dieser Arbeit verwendete Clusteranalyseverfahren ist jedoch, dass theoretisch beliebig viele, in jedem Fall aber mehr als drei verschiedene Merkmale für die Clusterbildung herangezogen werden können. Die unvollständigen Clusteranalyseverfahren werden daher nicht weiter betrachtet.

Eine weitere Anforderung ist, dass das Clusteranalyseverfahren eine eindeutige Zuordnung jedes Flugzeugs zu genau einem Cluster vornimmt. Bei den vorgestellten probabilistischen Verfahren werden die Flugzeuge durch die Verwendung von Wahrscheinlichkeiten jedoch nicht eindeutig Clustern zugeordnet. Daher werden die probabilistischen Verfahren ebenfalls nicht weiter betrachtet. Von den verbleibenden deterministischen Verfahren müssen die überlappenden deterministischen Clusteranalyseverfahren aufgrund der Anforderung der Eindeutigkeit der Zuordnung ebenfalls ausgeschlossen werden. Bei ihnen können die Objekte entgegen der Anforderungen jeweils gleichzeitig mehreren Clustern angehören.

Im folgenden Abschnitt werden die überlappungsfreien deterministischen Clusteranalyseverfahren näher betrachtet.

3.3.2 Überlappungsfreie deterministische Clusteranalyseverfahren

In diesem Abschnitt werden von den vorgestellten Arten von Clusteranalyseverfahren die überlappungsfreien deterministischen Clusteranalyseverfahren, welche die Anforderungen dieser Arbeit an ein Clusteranalyseverfahren erfüllen, näher behandelt. Zunächst werden die Funktionsprinzipien dieser Verfahren vorgestellt (Abschnitt 3.3.2.1). Anschließend werden die jeweiligen Anwendungsvoraussetzungen behandelt (Abschnitt 3.3.2.2) und die für diese Arbeit am besten geeigneten Verfahren für die weitere Betrachtung ausgewählt (Abschnitt 3.3.2.3).

3.3.2.1 Funktionsweise der Verfahren

In diesem Abschnitt werden die Funktionsprinzipien der überlappungsfreien deterministischen Clusteranalyseverfahren vorgestellt. Bei diesen Verfahren unterscheidet man vier Gruppen: Die Nächste-Nachbarn-Verfahren, die Mittelwertverfahren, die Repräsentanten-Verfahren und die Verfahren zur Konstruktion von Clusterzentren als Repräsentanten.

Nächste-Nachbarn-Verfahren

Bei den Nächste-Nachbarn-Verfahren werden die Cluster auf Basis der Nächste-Nachbar-Beziehungen zwischen den einzelnen Objekten gebildet. Zwei Objekte sind nächste Nachbarn, wenn ihre Ähnlichkeit einen bestimmten Schwellenwert überschreitet oder ihre Unähnlichkeit einen bestimmten Schwellenwert unterschreitet. Jedes Objekt kann theoretisch also zwischen null und beliebig viele nächste Nachbarn haben. Man unterscheidet drei verschiedene Nächste-Nachbarn-Verfahren: Die Complete-Linkage-Verfahren, die verallgemeinerten Nächste-Nachbarn-Verfahren und die Single-Linkage-Verfahren [Bacher et al. 2010, S. 148-149].

Bei den Complete-Linkage-Verfahren müssen alle Objekte eines Clusters nächste Nachbarn zueinander sein. Dadurch werden bei diesem Verfahren sehr hohe Anforderungen an die Homogenität der Cluster gestellt. Im Gegensatz dazu stellen die Single-Linkage-Verfahren nur sehr geringe Anforderungen an die Homogenität der Cluster. Bei ihnen wird gefordert, dass jedes Objekt in einem Cluster mindestens einen nächsten Nachbarn hat. Dies ermöglicht kettenartige Cluster, in denen sich im Extremfall nur die direkt aneinander angrenzenden Objekte (analog direkt aneinander angrenzenden Kettengliedern) ähnlich sind, viele Objektpaare im Cluster jedoch nur eine sehr geringe Ähnlichkeit aufweisen [Bacher et al. 2010, S. 148-149].

Die verallgemeinerten Nächste-Nachbarn-Verfahren erlauben durch die Veränderung von Parametern eine Änderung der Strenge der Anforderungen an die Homogenität der ermittelten Cluster. Die Strenge der Anforderungen kann damit zwischen den Extremen des Complete-Linkage-Verfahrens und des Single-Linkage-Verfahrens eingestellt werden. Ein Ansatz der verallgemeinerten Nächste-Nachbarn-Verfahren fordert, dass jedes Objekt eines Clusters eine festgelegte Mindestzahl nächster Nachbarn hat. Der Unterschied zum Complete-Linkage-Verfahren besteht darin, dass nicht alle Objekte in einem Cluster nächste Nachbarn sein müssen. Stattdessen muss jedes Objekt eine bestimmte Mindestanzahl nächster Nachbarn in seinem Cluster haben. Nach einem anderen Ansatz muss jedes Objekt in seinem Cluster zumindest einen k -ten Nachbarn haben. Dazu wird für jedes Objekt die Rangfolge der Nachbarn nach ihrer Ähnlichkeit ermittelt. Der erste Nachbar ist der Nachbar mit der größten Ähnlichkeit, der zweite Nachbar derjenige mit der zweitgrößten Ähnlichkeit und der k -te Nachbar derjenige mit der k -größten Ähnlichkeit. Es wird in diesem Ansatz gefordert, dass für jedes Objekt in einem Cluster zumindest eines der diesem

Objekt ähnlichsten Objekte ebenfalls zu diesem Cluster gehört [Bacher et al. 2010, S. 148-149, 259-263].

Mittelwertverfahren

Bei den Mittelwertverfahren werden die Cluster auf Basis der durchschnittlichen paarweisen Ähnlichkeiten oder Unähnlichkeiten der Objekte in einem Cluster sowie zwischen verschiedenen Clustern gebildet. Die durchschnittliche Ähnlichkeit muss dabei ein bestimmtes Maß überschreiten. Der Unterschied zu den Nächste-Nachbarn-Verfahren besteht darin, dass nicht die Ähnlichkeit der nächsten Nachbarn entscheidend für die Clusterbildung ist, sondern die durchschnittlichen Ähnlichkeiten der Elemente in einem Cluster [Bacher et al. 2010, S. 149-150, 264-265].

Man unterscheidet bei den Mittelwertverfahren zwischen Average-Linkage-, Weighted-Average-Linkage-, Within-Average-Linkage-Verfahren. Diese unterscheiden sich vor allem in der Art und Weise der Berechnung der Ähnlichkeiten oder Unähnlichkeiten der Objekte innerhalb eines Clusters sowie zwischen verschiedenen Clustern [Bacher et al. 2010, S. 149-150, 264-265].

Repräsentanten-Verfahren

Die Repräsentanten-Verfahren sind dadurch gekennzeichnet, dass jedes der ermittelten Cluster durch eines seiner Objekte repräsentiert wird. Dazu werden aus der Gesamtheit der Objekte zunächst jene identifiziert, die eine hohe Anzahl nächster Nachbarn haben und daher gut als Repräsentanten geeignet sind. Anhand der Nutzervorgabe über die gewünschte Anzahl der Cluster werden die Repräsentanten festgelegt. Anschließend werden den festgelegten Repräsentanten ähnliche Objekte zugeordnet. Es gibt überlappungsfreie und überlappende Repräsentanten-Verfahren; hier werden aufgrund der Anforderungen nur die überlappungsfreien Verfahren betrachtet [Bacher et al. 2010, S. 150, 277-279, 282-283].

Verfahren zur Konstruktion von Clusterzentren als Repräsentanten

Die Verfahren zur Konstruktion von Clusterzentren als Repräsentanten unterscheiden sich von den Repräsentanten-Verfahren in der Hinsicht, dass nicht einzelne Objekte als Repräsentanten für die Cluster ausgewählt werden, sondern durch die Mittelwerte der zu einem Cluster gehörenden Objekte Clusterzentren als hypothetische Repräsentanten gebildet werden. Zu den Verfahren gehören das Median-, das Zentroid-, das Ward- und die K-Means-Verfahren. Bei den Median- und Zentroid-Verfahren werden die Cluster so bestimmt, dass die Clusterzentren maximale Entfernungen voneinander haben. Beim Ward-Verfahren und den K-Means-Verfahren wird dagegen die Streuung der Clusterzentren maximiert [Bacher et al. 2010, S. 150].

3.3.2.2 Anwendungsvoraussetzungen

Die vier im vorangegangenen Abschnitt vorgestellten Gruppen von deterministischen Clusteranalyseverfahren haben teilweise einige grundlegende Anwendungsvoraussetzun-

gen. Diese vier grundlegenden Anwendungsvoraussetzungen und ihr Zutreffen auf die vier Gruppen von deterministischen Clusteranalyseverfahren werden in diesem Abschnitt erläutert.

Eine Clusteranalyse kann grundsätzlich auf zwei verschiedenen Formen von Daten aufbauen: Zum einen auf einer Ähnlichkeits- oder Unähnlichkeitsmatrix, und zum anderen auf einer Datenmatrix. Eine Ähnlichkeitsmatrix enthält die paarweise Ähnlichkeit der zu untersuchenden Objekte, eine Unähnlichkeitsmatrix enthält entsprechend die paarweise Unähnlichkeit der Objekte. Eine Datenmatrix stellt eine weniger aggregierte Form der Daten dar. Sie enthält die Ausprägungen aller für die Clusteranalyse relevanten Variablen aller zu untersuchenden Objekte. Aus den (Roh)-Daten einer Datenmatrix kann eine Ähnlichkeits- oder Unähnlichkeitsmatrix ermittelt werden. Aus einer Ähnlichkeits- oder Unähnlichkeitsmatrix kann jedoch keine Datenmatrix ermittelt werden. Die erste der vier grundlegenden Anwendungsvoraussetzungen der Clusteranalyseverfahren ist das Vorliegen einer Datenmatrix. In diesem Fall werden die Rohdaten für die Clusteranalyse benötigt und eine bloße Ähnlichkeits- oder Unähnlichkeitsmatrix reicht nicht aus [Bacher et al. 2010, S. 150-152].

Sofern eine Datenmatrix verwendet wird, müssen die relevanten Variablen unter Umständen in quantitativer Form vorliegen, um Mittelwertbildungen der Variablen zu ermöglichen. Für einige Verfahren ist zudem die Verwendung eines bestimmten Ähnlichkeits- / Unähnlichkeitsmaßes vorgeschrieben. Dieses Maß bestimmt, wie aus der gegebenenfalls verwendeten Datenmatrix die Ähnlichkeits- oder Unähnlichkeitsmatrix berechnet wird. Die vierte grundlegende Anwendungsvoraussetzung ist die Notwendigkeit der Verwendung eines metrischen Ähnlichkeits- / Unähnlichkeitsmaßes [Bacher et al. 2010, S. 150-152].

Tabelle 3-3 zeigt, welche vier grundlegenden Voraussetzungen für die Anwendung der verschiedenen überlappungsfreien Clusteranalyseverfahren erfüllt werden müssen.

Tabelle 3-3: Anwendungsvoraussetzungen der Clusteranalyseverfahren
[nach Bacher et al. 2010, S. 150-154]

	Nächste-Nachbarn-Verfahren	Mittelwertverfahren	Repräsentanten-Verfahren	Clusterzentren als Repräsentanten
Ist eine Datenmatrix erforderlich?	Nein	Nein	Nein	Ja
Muss eine verwendete Datenmatrix quantitative Variablen beinhalten?	Nein	Nein	Nein	Ja
Ist ein bestimmtes Ähnlichkeitsmaß / Unähnlichkeitsmaß erforderlich?	Nein	Nein	Nein	Ja
Ist ein metrisches Ähnlichkeitsmaß / Unähnlichkeitsmaß erforderlich?	Nein	Ja	Nein	Nein

Für die Nächste-Nachbarn-Verfahren und die Repräsentanten-Verfahren muss keine der vier genannten grundlegenden Anforderungen erfüllt werden [Bacher et al. 2010, S. 150-152].

Für die Mittelwertverfahren ist aufgrund der nötigen Mittelwertbildung der Ähnlichkeiten oder Unähnlichkeiten ein metrisches Ähnlichkeits- / Unähnlichkeitsmaß erforderlich [Bacher et al. 2010, S. 150-152].

Für die Verfahren zur Konstruktion von Clusterzentren als Repräsentanten muss die quadrierte euklidische Distanz als Ähnlichkeits- / Unähnlichkeitsmaß verwendet werden. Für beispielhafte Clusteranalysen liefern die entsprechenden Verfahren auch mit anderen Ähnlichkeits- / Unähnlichkeitsmaßen die gleichen oder ähnliche Ergebnisse. Dies ist jedoch nicht für alle Clusteranalysen gegeben. Die Anwendungsvoraussetzung der Verwendung der quadrierten euklidischen Distanz als Ähnlichkeits- / Unähnlichkeitsmaß wird daher in dieser Arbeit als gegeben angenommen. Die quadrierte euklidische Distanz ist kein metrisches Unähnlichkeitsmaß, da sich durch das Quadrat eine doppelt große Unähnlichkeit zweier Objekte nicht zwangsläufig in einem doppelten Unähnlichkeitsmaß widerspiegelt [Bacher et al. 2010, S. 150-153].

Alle vier Anwendungsvoraussetzungen können mit der Problemstellung und Ausgangssituation dieser Arbeit erfüllt werden. Die nötigen Daten können in eine Datenmatrix überführt werden. Die relevanten Variablen sind überwiegend quantitativer Natur. Einzelne nominale und ordinale Variablen können bei Bedarf jedoch mithilfe von Dummy-Variablen wie quantitative Variablen behandelt werden [Bacher et al. 2010, S. 336-337]. Bezüglich des verwendeten Unähnlichkeitsmaßes gibt es in dieser Arbeit a priori keine Einschränkungen.

3.3.2.3 Vorauswahl geeigneter Verfahren

In den vorangegangenen Abschnitten werden die verschiedenen Gruppen von deterministischen Clusteranalyseverfahren und die jeweiligen grundlegenden Anwendungsvoraussetzungen vorgestellt. Tabelle 3-4 gibt einen zusammenfassenden Überblick über die Funktionsprinzipien, sowie die Vor- und Nachteile der verschiedenen überlappungsfreien deterministischen Clusteranalyseverfahren. Auf Basis dieser Vorstellung und der wichtigsten Vor- und Nachteile der Verfahren für die Anwendung im Rahmen dieser Arbeit werden in diesem Abschnitt die geeigneten Verfahren weiter eingegrenzt. Dazu spielt in diesem Schritt vor allem die Zuverlässigkeit des Verfahrens im Hinblick auf die Erzielung möglichst optimaler Ergebnisse eine Rolle. Aus verschiedenen Gruppen miteinander vergleichbarer Verfahren soll zudem jeweils nur das am besten geeignete Verfahren weiter betrachtet werden.

Tabelle 3-4: Übersicht über die deterministischen, überlappungsfreien Clusteranalyseverfahren

Verfahren	Funktionsprinzip	Vorteile	Nachteile
Nächste-Nachbarn-Verfahren	<ul style="list-style-type: none"> • Grundlage: Nächste-Nachbarn-Beziehung (Überschreitung eines Schwellenwertes für die Ähnlichkeit) • Complete-Linkage-Verfahren: Alle Objekte des Clusters sind nächste Nachbarn • Single-Linkage-Verfahren: Jedes Objekt hat mindestens einen nächsten Nachbarn in seinem Cluster • Verallgemeinerte-Nächste-Nachbarn-Verfahren: Parametrische Variation zwischen Complete-Linkage- und Single-Linkage-Verfahren 	<ul style="list-style-type: none"> • Einfache und schnelle Verfahren • Geringe Anwendungsvoraussetzungen 	<ul style="list-style-type: none"> • Single- und Complete-Linkage-Verfahren sind hierarchische Verfahren • Zu hohe (Complete-Linkage-Verfahren), bzw. zu geringe (Single-Linkage-Verfahren) Homogenitätsanforderung
Mittelwertverfahren	<ul style="list-style-type: none"> • Bildung der Cluster auf Basis der durchschnittlichen Ähnlichkeit der enthaltenen Objekte 	<ul style="list-style-type: none"> • Einfache und schnelle Verfahren • Geringe Anwendungsvoraussetzungen 	<ul style="list-style-type: none"> • Hierarchische Verfahren • Teilweise Auftreten von Inversionen (Verschmelzungen ähnlicherer Cluster als im vorherigen Schritt)
Repräsentanten-Verfahren	<ul style="list-style-type: none"> • Objekte mit besonders vielen nächsten Nachbarn werden zu Repräsentanten • Andere Objekte werden dem jeweils nächsten Repräsentanten zugeordnet 	<ul style="list-style-type: none"> • Cluster können durch die Repräsentanten veranschaulicht werden • Vorgabe der gewünschten Anzahl der Cluster 	<ul style="list-style-type: none"> • Viele Repräsentanten-Verfahren erfordern Vorgabe von (nicht formal begründbaren) Schwellenwerten für die Homogenität der Cluster / Heterogenität zwischen den Clustern • Viele Repräsentanten-Verfahren liefern überlappende Clusterstrukturen oder es verbleiben unklassifizierte Objekte
Verfahren zur Konstruktion von Clusterzentren als Repräsentanten	<ul style="list-style-type: none"> • Es werden hypothetische Repräsentanten aus den Mittelwerten der Cluster gebildet • Die Cluster werden durch Minimierung der Entfernungen der Objekte von den Clusterzentren bzw. Maximierung der Clusterzentren voneinander gebildet 	<ul style="list-style-type: none"> • Cluster können durch die hypothetischen Repräsentanten veranschaulicht werden • Vorgabe der gewünschten Anzahl der Cluster 	<ul style="list-style-type: none"> • Median-, Zentroid- und Ward-Verfahren sind hierarchische Verfahren • Beim Median- und Zentroid-Verfahren sind Inversionen möglich

Die Single-Linkage- und Complete-Linkage-Verfahren haben den gravierenden Nachteil einer zu geringen beziehungsweise einer zu hohen Homogenitätsanforderung. Ihre Anwendung ist in der Regel nur dann zu empfehlen, wenn keine Datenmatrix vorliegt oder die Anwendungsvoraussetzungen der anderen Verfahren nicht erfüllt sind. Im Fall dieser Arbeit liegt jedoch eine Datenmatrix vor und die Anwendungsvoraussetzungen der anderen Verfahren können grundsätzlich erfüllt werden [Bacher et al. 2010, S. 162-155, 274-275].

Zudem gehören die Single-Linkage- und Complete-Linkage-Verfahren stets zu den hierarchisch-agglomerativen Verfahren. Die hierarchisch-agglomerativen und die hierarchisch-divisiven Verfahren haben jedoch den schwerwiegenden Nachteil, dass einmal getroffene Cluster-Entscheidungen nicht wieder rückgängig gemacht werden können. Im Fall der hierarchisch-agglomerativen Verfahren verbleiben nach der Verschmelzung zweier Cluster im Laufe des Verfahrens alle zuvor in diesen beiden Clustern enthaltenen Objekte anschließend im gleichen Cluster. Im Fall der hierarchisch-divisiven Verfahren können die Objekte eines geteilten Clusters nicht wieder Bestandteile des gleichen Clusters werden. Die hierarchischen Clusteranalyseverfahren führen daher nicht immer zu optimalen Ergebnissen. Die Single-Linkage- und Complete-Linkage-Verfahren werden daher von der weiteren Betrachtung ausgeschlossen.

Die verallgemeinerten-Nächste-Nachbarn-Verfahren sind aus der Gruppe der Nächste-Nachbarn-Verfahren am besten für die Clusteranalyse im Rahmen dieser Arbeit geeignet. Jedoch empfehlen Bacher et al. (2010) die Anwendung der Nächste-Nachbarn-Verfahren für eine objektorientierte Clusteranalyse auf Basis einer Datenmatrix generell nur, wenn die Voraussetzungen des K-Means- und des Ward-Verfahrens nicht erfüllt sind [Bacher et al. 2010, S. 153]. Diese Verfahren können die zusätzlichen Informationen, die bei Verfügbarkeit einer Datenmatrix vorliegen, besser nutzen und können potentiell bessere Ergebnisse liefern. Aus der Gruppe der Nächste-Nachbarn-Verfahren wird daher im Folgenden zunächst kein Verfahren weiter betrachtet. Ein partitionierendes, verallgemeinertes-Nächste-Nachbarn-Verfahren würde die Anforderungen dieser Arbeit an ein Clusteranalyseverfahren jedoch erfüllen und stellt eine mögliche Alternative gegenüber a priori besser geeignet erscheinenden Verfahren dar.

Von den Mittelwertverfahren ist in der Regel das Weighted-Average-Linkage-Verfahren dem Average-Linkage-Verfahren aufgrund der besseren Interpretierbarkeit und dem Within-Average-Linkage-Verfahren aufgrund der Vermeidung von Inversionen vorzuziehen. Das Auftreten einer Inversion ist durch die Verschmelzung zweier Cluster gekennzeichnet, die sich einander ähnlicher sind als die beiden im vorhergehenden Schritt verschmolzenen Cluster. Cluster, die sich besonders ähnlich sind, sollten jedoch vor Clustern, die sich weniger ähnlich sind, verschmolzen werden [Bacher et al. 2010, S. 268, 275].

Die Mittelwertverfahren gehören ebenso wie die Single-Linkage- und Complete-Linkage-Verfahren zu den hierarchisch-agglomerativen Verfahren und haben den Nachteil, dass

Fehlentscheidungen des Algorithmus nicht korrigiert werden können. Ebenso wie die Nächste-Nachbarn-Verfahren werden die Mittelwertverfahren von Bacher et al. (2010) für eine objektorientierte Clusteranalyse auf Basis einer Datenmatrix nur empfohlen, wenn die Voraussetzungen des K-Means- und des Ward-Verfahrens nicht erfüllt sind. Die Mittelwertverfahren werden aus diesen Gründen nicht weiter betrachtet [Bacher et al. 2010, S. 153, 274-275].

Ein bedeutender Nachteil vieler Algorithmen aus der Gruppe der Repräsentanten-Verfahren ist die Notwendigkeit der Vorgabe von Schwellenwerten für die Homogenität innerhalb der Cluster und die Heterogenität zwischen den Clustern. Für die Höhe dieser Schwellenwerte gibt es bisher keine formale Begründung, vielmehr werden geeignete Werte in der Regel durch den Nutzer durch Ausprobieren bestimmt. Dies ist nicht mit der Anforderung kompatibel, dass das Verfahren mit geringem Arbeitsaufwand und ohne detaillierte Kenntnisse des Clusteranalyseverfahrens möglich sein soll. Außerdem werden mit vielen Repräsentanten-Verfahren überlappende Clusterstrukturen ermittelt oder nicht alle Objekte auch Clustern zugeordnet. Beides verstößt gegen die Anforderungen an ein Clusteranalyseverfahren im Rahmen dieser Arbeit [Bacher et al. 2010, S. 277-283].

Im Folgenden wird aus der Gruppe der Repräsentanten-Verfahren nur das Medoid-Verfahren nach Kaufman/Rousseeuw betrachtet. Es liefert eine überlappungsfreie Clusterstruktur, bestehend aus der vom Nutzer vorgegebenen Anzahl von Clustern. Dabei werden alle Objekte einem Cluster zugeordnet. Eine Vorgabe von Schwellenwerten ist nicht nötig [Kaufman/Rousseeuw 1990, S. 102-104].

Aus der Gruppe der Verfahren zur Konstruktion von Clusterzentren als Repräsentanten sind in der Regel das Ward-Verfahren und das K-Means-Verfahren dem Zentroid- und dem Median-Verfahren vorzuziehen, da bei den letztgenannten Verfahren ebenfalls Inversionen auftreten können. Sowohl das Zentroid- und das Median-Verfahren, als auch das Ward-Verfahren haben zudem den Nachteil, zu den hierarchisch-agglomerativen Verfahren zu gehören. Folglich wird aus dieser Gruppe im Folgenden nur noch das K-Means-Verfahren betrachtet [Bacher et al. 2010, S. 153, 268, 274-275, 285-295].

3.3.3 K-Means-Verfahren und Medoid-Verfahren

Im letzten Abschnitt wurden die überlappungsfreien deterministischen Clusteranalyseverfahren vorgestellt und das K-Means-Verfahren und das Medoid-Verfahren als für diese Arbeit am besten geeignete Verfahren für die weitere Betrachtung ausgewählt. In diesem Abschnitt werden diese beiden Verfahren detailliert vorgestellt (Abschnitt 3.3.3.1) und auf Basis der Betrachtung der jeweiligen Vor- und Nachteile eines der beiden Verfahren für die Clusteranalyse in dieser Arbeit ausgewählt (Abschnitt 3.3.3.2).

3.3.3.1 Erklärung der Verfahren

In diesem Abschnitt wird zunächst das K-Means-Verfahren und anschließend das Medoid-Verfahren ausführlich vorgestellt.

K-Means-Verfahren

Beim K-Means-Verfahren wird zunächst die gewünschte Anzahl der Cluster vorgegeben. Zu Beginn des Verfahrens werden die Objekte zufällig auf die festgelegte Anzahl der Cluster aufgeteilt.¹ Anschließend werden die anfänglichen Clusterzentren als Mittelwerte der Ausprägungen aller Variablen der Objekte in einem Cluster berechnet [Bacher et al. 2010, S. 299].

Im nächsten Schritt wird die Zuordnung der Objekte zu den Clustern geändert, indem jedes Objekt dem nächstgelegenen Cluster zugeordnet wird. Die Entfernung wird dabei mithilfe der quadrierten euklidischen Distanz gemessen. Diese ist die Summe der quadrierten Entfernungen zwischen Objekt und Clusterzentrum für alle verwendeten Variablen. Sie entspricht im zwei- und dreidimensionalen Fall der quadrierten Entfernung zwischen zwei Punkten. Die Berechnung der Clusterzentren und die Änderung der Zuordnung der Objekte zu den Clustern werden so oft wiederholt, bis keine Änderungen der Zuordnung mehr auftreten und damit zumindest ein lokales Optimum der Clusterzuordnung erreicht ist. Mit jeder dieser Iterationen wird die Streuungsquadratsumme über alle Cluster minimiert. Die Streuungsquadratsumme ist die Summe der quadrierten Abweichungen der Variablenausprägungen der Objekte von ihrem jeweiligen Clusterzentrum über alle verwendeten Variablen und Objekte [Bacher et al. 2010, S. 299-300].

Medoid-Verfahren

Das Medoid-Verfahren funktioniert auf Basis einer Unähnlichkeitsmatrix. Bei anfänglichem Vorliegen einer Datenmatrix muss aus dieser zunächst die Unähnlichkeitsmatrix ermittelt werden. Außerdem muss vor dem Start des Verfahrens die gewünschte Anzahl der Cluster vorgegeben werden [Kaufman/Rousseeuw 1990, S. 68, 102-104].

Das Medoid-Verfahren verläuft in zwei Hauptschritten. Im ersten Hauptschritt erfolgt die Auswahl der Repräsentanten, die beim Medoid-Verfahren Medoide genannt werden. Der zweite Hauptschritt ist die eigentliche Bildung der Cluster durch die Zuordnung aller Objekte, die nicht als Medoide bestimmt wurden zum jeweils nächsten Medoid [Kaufman/Rousseeuw 1990, S. 68, 102-104].

Die Festlegung der Medoide wiederum erfolgt in zwei Teilschritten: Im ersten Teilschritt wird eine Ausgangskonfiguration für die Auswahl der Medoide ermittelt. Die Anzahl der

¹ Es gibt Weiterentwicklungen des K-Means-Verfahrens, bei denen die Startkonfiguration nicht zufällig festgelegt wird.

Medoide in der Ausgangskonfiguration entspricht der vorab festgelegten, gewünschten Anzahl der Cluster. Als erstes Medoid wird jenes Objekt ausgewählt, für das die Summe der Unähnlichkeiten zu allen anderen Objekten minimal ist. Die restlichen Medoide für die Ausgangskonfiguration werden schrittweise unter Minimierung der Summe der Unähnlichkeiten der nicht ausgewählten Objekte zum jeweils nächsten Medoid ausgewählt. Die Festlegung eines weiteren Medoids bewirkt, dass sich die Zuordnung der Objekte zu den Medoiden nach dem Kriterium der minimalen Unähnlichkeit ändert. Damit sinkt die Summe der Unähnlichkeiten der Objekte zum jeweils nächsten Medoid durch Festlegung eines zusätzlichen Medoids. Als zusätzliches Medoid wird immer dasjenige Objekt ausgewählt, für das die Summe aller Unähnlichkeiten am stärksten sinkt [Kaufman/Rousseeuw 1990, S. 102-103].

Der zweite Teilschritt zur Festlegung der Medoide ist der paarweise Austausch der mit der Ausgangskonfiguration als Medoide ausgewählten Objekte. Ziel der Austausche ist die weitere Minimierung der Summe der Unähnlichkeiten aller nicht ausgewählten Objekte zum jeweils nächsten Medoid. Bei einem Tausch wird ein Objekt zum Medoid, während ein bisheriges Medoid zum normalen Objekt wird. Es werden so lange paarweise Austausche vorgenommen, bis kein solcher Austausch unter Verringerung der Summe der Unähnlichkeiten mehr möglich ist [Kaufman/Rousseeuw 1990, S. 103-104].

3.3.3.2 Auswahl eines Verfahrens

Nach der detaillierten Vorstellung des K-Means-Verfahrens und des Medoid-Verfahrens im vorangegangenen Abschnitt wird in diesem Abschnitt eines der Verfahren für die Anwendung für das Clustering von Flugzeugen ausgewählt. Dazu werden zunächst die Vor- und Nachteile beider Verfahren vorgestellt. Anschließend werden diese Vor- und Nachteile im Hinblick auf die Problemstellung dieser Arbeit gegeneinander abgewogen und eines der Verfahren ausgewählt.

Gegenüber dem K-Means-Verfahren hat das Medoid-Verfahren den Vorteil einer höheren Anschaulichkeit des Ergebnisses, weil jedes Cluster durch ein tatsächlich existierendes Objekt repräsentiert wird. Allerdings wird ein Cluster durch sein Clusterzentrum im Allgemeinen besser repräsentiert als durch eines der Objekte des Clusters.

Ein Vorteil des K-Means-Verfahrens ist die hohe Verbreitung des Verfahrens [Tan et al. 2005, S. 497]. Daher gibt es zum K-Means-Verfahren relativ viele Studien zu Verbesserungen des Verfahrens und zur Interpretation der Ergebnisse.

Sowohl das K-Means-Verfahren als auch das Medoid-Verfahren haben den Nachteil, dass sie nicht immer das optimale Ergebnis liefern. In der Regel gibt es für die Zielfunktionen beider Algorithmen ein globales Optimum sowie viele lokale Optima. Beide Algorithmen führen jedoch nicht notwendigerweise zum globalen Optimum. Beim K-Means-Verfahren hängt das Ergebnis von der zufälligen Anfangskonfiguration ab. Durch eine häufige Wie-

derholung des gesamten Algorithmus mit unterschiedlichen Anfangskonfigurationen kann die Wahrscheinlichkeit, das globale Optimum zu erreichen, gesteigert werden. Im Gegensatz dazu ist das Ergebnis des Medoid-Verfahrens nicht zufallsabhängig und daher zuverlässig reproduzierbar [Bacher et al. 2010, S. 305; Kaufman/Rousseeuw 1990, S. 102-104].

Das K-Means-Verfahren ist empfindlich gegenüber Objekten, die sich in den Ausprägungen der verwendeten Parameter stark von allen anderen Objekten unterscheiden [Tan et al. 2005, S. 513]. Beim Medoid-Verfahren werden diese Ausreißer hingegen meist automatisch isoliert [Kaufman/Rousseeuw 1990, S. 117]. Durch die Verwendung der quadrierten euklidischen Distanz tendiert das K-Means-Verfahren zudem zur Bildung gleichgroßer, kugelförmiger Cluster, auch wenn die tatsächlichen Cluster keine Kugelform haben oder verschiedene Ausdehnungen haben [Tan et al. 2005, S. 513].

Tabelle 3-5 und Tabelle 3-6 fassen die Vor- und Nachteile des K-Means-Verfahrens und des Medoid-Verfahrens zusammen.

Tabelle 3-5: Vor- und Nachteile des K-Means-Verfahrens

Vorteile des K-Means-Verfahrens	Nachteile des K-Means-Verfahrens
<ul style="list-style-type: none"> • Hypothetisch gebildete Repräsentanten werden durch die Mittelwerte der Variablen aller Objekte eines Clusters gebildet und sind daher meist noch repräsentativer für das Cluster als das repräsentativste Cluster-Mitglied • Hohe Verbreitung des Verfahrens 	<ul style="list-style-type: none"> • Das Ergebnis ist abhängig von der zufälligen Startkonfiguration • Das Ergebnis ist nicht immer optimal; es kann ein lokales und nicht das globale Optimum sein • Tendenz zur Bildung gleichgroßer, kugelförmiger Cluster, auch wenn die tatsächlichen Cluster keine Kugelform oder verschiedene Ausdehnungen haben • Empfindlich gegenüber Ausreißern

Tabelle 3-6: Vor- und Nachteile des Medoid-Verfahrens

Vorteile des Medoid-Verfahrens	Nachteile des Medoid-Verfahrens
<ul style="list-style-type: none"> • Real existierende Repräsentanten sind anschaulicher als hypothetisch gebildete Repräsentanten • Es müssen keine Startwerte vorgegeben werden • Das Ergebnis ist im Allgemeinen unabhängig vom Zufall • Ausreißer in den Daten werden meist automatisch isoliert, indem sie ein eigenes Cluster bilden 	<ul style="list-style-type: none"> • Das Ergebnis ist nicht immer optimal; es kann ein lokales und nicht das globale Optimum sein • Die Cluster können durch einen hypothetischen Repräsentanten besser repräsentiert werden als durch einen real existierenden Repräsentanten

Die Auswahl eines Clusteranalyseverfahrens erfolgt auf Basis einer Abwägung der jeweiligen Vor- und Nachteile. Beide Verfahren haben den Nachteil, dass die Ergebnisse nicht immer dem globalen Optimum entsprechen. Die Tatsache, dass die Ergebnisse des K-Means-Verfahrens von der zufälligen Anfangskonfiguration abhängen ist eine Verletzung der Soll-Anforderung, dass die Ergebnisse reproduzierbar sein müssen. Dagegen ist der zweite Nachteil des Medoid-Verfahrens, dass die Cluster durch hypothetische Repräsentanten besser repräsentiert werden können als durch real existierende Objekte weniger gewichtig. Die damit einhergehende hohe Anschaulichkeit der Cluster ist von großem Vorteil. Außerdem können weiterführende Betrachtungen und Berechnungen auf genauen Daten des Repräsentanten basierend durchgeführt werden, ohne ein hypothetisches Flugzeug auf Basis eines Clusterzentrums entwerfen zu müssen. Ein weiterer Vorteil des Medoid-Verfahrens, der nicht durch Vorteile des K-Means-Verfahrens aufgewogen werden kann, ist die geringere Empfindlichkeit gegenüber Ausreißern in den Daten.

Für die Clusteranalyse in dieser Arbeit wird daher das Medoid-Verfahren verwendet.

4 Implementierung des Clusteranalyseverfahrens

Das im vorangegangenen Kapitel ausgewählte Medoid-Verfahren wurde in VBA in Kombination mit Microsoft Excel implementiert. In diesem Kapitel werden Aufbau, Ablauf und Funktionsweise des so entstandenen Clusteranalyse-Tools behandelt.

In Abschnitt 4.1 wird zunächst die Integration der Daten aus den verwendeten Datenbanken erläutert. Vor der eigentlichen Implementierung mussten nach der Entscheidung für das Medoid-Verfahren zudem einige weitere Entscheidungen zur Verarbeitung der Daten durch das Clusteranalyse-Tool und damit zur genauen Ausgestaltung des Verfahrens getroffen werden. Die entsprechenden Entscheidungsmöglichkeiten sowie die jeweils getroffene Entscheidung werden in Abschnitt 4.2 behandelt. Abschnitt 4.3 behandelt die Struktur des Clusteranalysetools und den Ablauf einer Clusteranalyse. Dabei werden das Einlesen der Daten, die Verarbeitung der Daten in Vorbereitung der Bestimmung der Clusterstrukturen und letztlich die Ermittlung der verschiedenen Clusterstrukturen behandelt. Neben der eigentlichen Bestimmung der Clusterstruktur sind im entwickelten Clusteranalyse-Tool auch zwei Hilfsmittel zur Entscheidung für eine sinnvolle Anzahl der Cluster vorhanden. Diese werden in Abschnitt 4.4 erklärt.

4.1 Integration von Daten aus den verwendeten Datenbanken

In diesem Abschnitt werden die Anforderungen an das Format der Daten aus den vom Clusteranalyse-Tool verwendeten Datenbanken sowie die Zuordnung der Flugzeuge aus den verschiedenen Datenbanken behandelt.

Das entwickelte Clusteranalyse-Tool verfügt über Schnittstellen zur Verwendung von Daten aus drei verschiedenen Datenbanken. Es können flugzeugspezifische Daten aus der Official Airline Guide (OAG)-Datenbank, aus der Base of Aircraft Data (BADA)-Datenbank und der Aircraft Analytical System (ACAS)-Datenbank verwendet werden.

Die OAG-Datenbank liegt in Form einer Datenbank für Microsoft Access vor. Die für die Clusteranalyse benötigten Daten werden mithilfe einer Datenbankabfrage extrahiert und in eine Excel-Datei im XLSX-Format exportiert. Aus dieser Datei liest das Clusteranalyse-Tool die benötigten Informationen ein.

Die BADA-Datenbank liegt in Form einzelner Textdateien vor. Das Clusteranalyse-Tool liest die gesuchten Daten direkt aus diesen Textdateien. Verwendet werden dabei Informationen aus den „Operating Performance Files“ und den „Performance Table Files“. Eine Erweiterung auf weitere Dateitypen ist einfach möglich.

Die ACAS-Datenbank liegt in Form eines Programms mit eigener Benutzeroberfläche vor. Für die Clusteranalyse muss die Liste aller weltweit registrierten Flugzeuge in eine CSV-

Datei exportiert werden. Diese CSV-Datei wird vom Clusteranalyse-Tool später eingelesen.

Um Daten über Flugzeuge aus den verschiedenen Datenbanken gleichzeitig zu verwenden, ist aufgrund der unterschiedlichen Bezeichnung der Flugzeugtypen in den Datenbanken eine Zuordnung dieser Bezeichnungen der Flugzeugtypen erforderlich. Entsprechend der Zielsetzung der Arbeit soll die Cluster-Analyse auf die Flugzeugtypen der OAG Datenbank angewendet werden. Daher stellen diese Flugzeugtypen die Objekte der Clusteranalyse dar. Die Zuordnung der Flugzeuge aus der BADA-Datenbank zu den Flugzeugen aus der OAG-Datenbank erfolgt 1:1. Das heißt, einem Flugzeug aus der OAG-Datenbank wird genau ein Flugzeug aus der BADA-Datenbank zugeordnet, wenn ein entsprechendes Flugzeug in der BADA-Datenbank vorhanden ist.

Da die ACAS-Datenbank anders als OAG- und BADA-Datenbank keine Flugzeugtypen-Datenbank, sondern eine Datenbank einzelner registrierter Flugzeuge ist, findet bei der Zuordnung der Flugzeuge aus der ACAS-Datenbank zu den Flugzeugen aus der OAG-Datenbank eine 1:n Zuordnung statt. Das heißt, jedem Flugzeugtyp aus der OAG-Datenbank werden alle passenden Flugzeuge aus der ACAS-Datenbank zugeordnet. Um die entsprechenden Daten aus der ACAS-Datenbank zu gewinnen, wird der Mittelwert der entsprechenden Variablen über alle passenden Flugzeuge gebildet.

4.2 Verarbeitung der Daten

In diesem Abschnitt werden die Entscheidungen erläutert, die nach der Entscheidung für das Medoid-Verfahren für dessen genaue Ausgestaltung und Implementierung nötig waren. Dazu gehören die Entscheidung für eines der möglichen Unähnlichkeitsmaße (Abschnitt 4.2.1), die Behandlung von Variablen mit unterschiedlichem Skalenniveau (Abschnitt 4.2.2), der Umgang mit fehlenden Werten (Abschnitt 4.2.3) und die Standardisierung der Daten (Abschnitt 4.2.4).

4.2.1 Unähnlichkeitsmaß

Ein sehr verbreitetes Unähnlichkeitsmaß ist die euklidische Distanz. Die ebenfalls Verwendung findende quadrierte euklidische Distanz entspricht der Summe der quadrierten Differenzen aller Variablen zweier Objekte. Formel 4-1 zeigt die Berechnung der quadrierten euklidischen Distanz im Fall von drei Dimensionen (x, y und z).

$$\text{Quadrierte euklidische Distanz} = x^2 + y^2 + z^2 \quad (4-1)$$

Die euklidische Distanz ist die Quadratwurzel der quadrierten euklidischen Distanz. Im zwei- und dreidimensionalen Fall entspricht die euklidische Distanz dem Abstand der Ob-

jekte in der Ebene, beziehungsweise im Raum [Backhaus et al. 2011, S. 411]. Formel 4-2 zeigt die Berechnung der euklidischen Distanz im Fall von drei Dimensionen.

$$\text{Euklidische Distanz} = \sqrt{x^2 + y^2 + z^2} \quad (4-2)$$

Ein anderer Ansatz wird mit der Manhattan-Distanz verfolgt. Sie entspricht im zwei- oder dreidimensionalen Fall nicht dem Abstand oder der Luftlinie zwischen zwei Punkten, sondern der zwischen zwei Wolkenkratzern tatsächlich zurückzulegenden Entfernung unter der Annahme senkrecht aufeinander stehender Straßen. Die Manhattan-Distanz berechnet sich als Summe der absoluten Differenzen der einzelnen Variablen zweier Objekte [Kaufman/Rousseeuw 1990, S. 12].

Die euklidische Distanz und die Manhattan-Distanz sind die am praxisrelevantesten Spezialfälle der Minkowski-Metrik [Kaufman/Rousseeuw 1990, S. 12]. Beide haben jedoch den Nachteil, dass Korrelationen zwischen den Variablen in der Clusteranalyse indirekt zu einer stärkeren Gewichtung von solchen Unterschieden zwischen den Objekten führen, die sich in den Ausprägungen mehrerer (daher korrelierter) Variablen ausdrücken. Die Mahalanobis-Distanz korrigiert diesen Effekt. Allerdings stellt sie hohe Anforderungen an die verwendeten Daten. Sie darf vor allem nur dann verwendet werden, wenn die Variablen annähernd normalverteilt sind [Backhaus et al. 2011, S. 450; Tan et al. 2005, S. 81].

Aufgrund der Gewichtung von kleinen und großen Differenzen in verschiedenen Variablen und der hohen Anwendungsvoraussetzungen der Mahalanobis-Distanz wird in dieser Arbeit die euklidische Distanz verwendet. Die quadrierte euklidische Distanz, die Manhattan-Distanz und die Mahalanobis-Distanz sind im entwickelten Tool zur Clusteranalyse jedoch ebenfalls verfügbar.

4.2.2 Skalenniveau der Variablen

Man unterscheidet nach dem Skalenniveau vier verschiedene Arten von Variablen: Nominale Variablen, ordinale Variablen, intervallskalierte Variablen und verhältnisskalierte Variablen. Alle vorgestellten Unähnlichkeitsmaße erfordern mindestens ein intervallskaliertes Skalenniveau der Variablen [Backhaus et al. 2011, S. 401, 409-411]. Variablen mit anderen Skalenniveaus können jedoch in transformierter Form trotzdem verwendet werden [Bacher et al. 2010, S. 184-185; Kaufman/Rousseeuw 1990, S. 35-36].

Nominale Variablen

Nominale Variablen sind durch Bestimmtheit gekennzeichnet. Für sie ist die Verwendung von Gleichheit und Ungleichheit zulässig, jedoch kein Größenvergleich sowie keine Addition, Subtraktion, Multiplikation oder Division. Ein Beispiel für eine nominale Variable im Bereich von Flugzeugtypen ist der Triebwerkstyp. Die Variable kann je nach Einteilung

beispielsweise die Ausprägungen Turbofan, Turbojet, Turboprop und Piston Engine haben [Tan et al. 2005, S. 25-27].

Ordinale Variablen

Ordinale Variablen zeichnen sich neben ihrer Bestimmtheit durch das Vorliegen einer Ordnung aus. Daher ist neben Aussagen über Gleichheit oder Ungleichheit der Größenvergleich definiert. Es sind Aussagen der Art „Ausprägung A ist größer als Ausprägung B“ möglich. Ein Beispiel für eine ordinale Variable ist die Wirbelschleppenategorie eines Flugzeugs. Die verschiedenen Ausprägungen der Variable sind geordnet, jedoch sind keine Aussagen über den Abstand der Ausprägungen bekannt. Weiter mathematische Operationen sind daher nicht zulässig [Tan et al. 2005, S. 25-27].

Intervallskalierte und verhältnisskalierte Variablen

Für intervallskalierte Variablen sind zusätzlich Addition und Subtraktion zulässig, weil der Abstand zwischen verschiedenen Ausprägungen quantifizierbar ist. Weil der Nullpunkt nicht entsprechend gewählt ist, sind jedoch keine Multiplikation und Division zulässig. Beispiele für intervallskalierte Variablen sind das Kalenderdatum und die Temperatur in Grad Celsius [Bacher et al. 2010, S. 185; Tan et al. 2005, S. 25-27].

Für verhältnisskalierte Variablen ist auch die Verhältnisbildung sinnvoll. Die Operationen Multiplikation und Division sind zulässig. Die Anzahl der Sitzplätze oder der Triebwerke eines Flugzeugs sind verhältnisskalierte Variablen [Bacher et al. 2010, S. 185; Tan et al. 2005, S. 25-27].

Intervallskalierte Variablen und verhältnisskalierte Variablen bezeichnet man gemeinsam auch als quantitative Variablen [Bacher et al. 2010, S. 185; Tan et al. 2005, S. 25-27].

Da die vorgestellten Unähnlichkeitsmaße jeweils mindestens ein intervallskaliertes Skalenniveau der Variablen erfordern, ist für nominale und ordinale Variablen eine gesonderte Behandlung nötig [Backhaus et al. 2011, S. 401, 409-411].

Nominale Variablen werden in dieser Arbeit, wie üblich, in Dummy-Variablen überführt. Für jede Ausprägung einer nominalen Variablen wird eine Dummy-Variable eingeführt. Jede Dummy-Variable nimmt den Wert eins an, wenn die Ausprägung zutrifft und den Wert null, wenn die Ausprägung nicht zutrifft. Ordinale Variablen werden durch ihren Rang dargestellt und anschließend wie intervallskalierte Variablen behandelt. Alternativ könnten sie auch in Dummy-Variablen überführt werden. Dabei würden jedoch die Informationen über die Ordnung der Ausprägungen verloren gehen [Bacher et al. 2010, S. 184-185; Kaufman/Rousseeuw 1990, S. 35-36].

4.2.3 Umgang mit fehlenden Werten

Ein Problem bei der Berechnung der Unähnlichkeiten der verschiedenen Objekte können fehlende Werte einiger Variablen für bestimmte Objekte darstellen. In dieser Arbeit wird die folgende, von Kaufman/Rousseeuw (1990) und Tan et al. (2005) empfohlene, Lösung dieses Problems verwendet: Die Unähnlichkeiten für jedes Objektpaar werden nur basierend auf denjenigen Variablen berechnet, für die für beide Objekte die Werte verfügbar sind. Im Umkehrschluss werden Variablen, deren Werte für eines oder beide beteiligten Objekte nicht verfügbar sind, bei der Berechnung der Unähnlichkeit zwischen diesen beiden Objekten vernachlässigt. Diese Lösung ist zulässig, wenn relativ zur Anzahl der Variablen und Objekte nur wenige Werte fehlen [Kaufman/Rousseeuw 1990, S. 14-15; Tan et al. 2005, S. 40-41].

Bei Verwendung der Mahalanobis-Distanz ist diese Vorgehensweise nicht möglich. Stattdessen werden fehlende Werte einer Variable in diesem Fall durch den durchschnittlichen Wert dieser Variable, gemittelt über alle Flugzeuge, ersetzt.

4.2.4 Standardisierung

Damit jede Variable unabhängig von ihrem Wertebereich und ihrer Standardabweichung das gleiche Gewicht für die Bildung der Cluster hat, werden alle Variablen vor der eigentlichen Clusterbildung auf einen Erwartungswert von eins und eine Standardabweichung von null standardisiert.

Die Werte der durch Transformation aus den nominalen Variablen entstandenen Dummy-Variablen müssen zusätzlich angepasst werden. Der Grund dafür ist, dass die Präsenz der nominalen Variable in mehreren standardisierten Dummy-Variablen zu einem höheren Gewicht der entsprechenden nominalen Variable führen würde. Die Werte der Dummy-Variablen werden bei Verwendung der euklidischen Distanz, der quadrierten euklidischen Distanz oder der Mahalanobis-Distanz daher mit der Quadratwurzel aus 0,5 und bei Verwendung der Manhattan-Distanz mit 0,5 multipliziert [Bacher et al. 2010, S. 185].

4.3 Struktur des Clusteranalysetools und Ablauf einer Clusteranalyse

Das im Rahmen dieser Arbeit entwickelte Clusteranalyse-Tool wird über eine Bedienoberfläche in Microsoft Excel gesteuert. Für das Einlesen der Daten aus den verwendeten Datenbanken und die eigentliche Berechnung der gesuchten Clusterstrukturen kommen zwei separate Makros zum Einsatz. Die Ergebnisse (Clusterstrukturen) des Clustering sowie die Bewertung der Qualität der gefundenen Lösungen werden in tabellarischer und grafischer Form ebenfalls in Excel dargestellt.

Zunächst müssen in Vorbereitung einer Clusteranalyse für eine Flugzeugflotte die in den verwendeten Datenbanken enthaltenen Flugzeuge erfasst und zugeordnet werden. Die ent-

sprechenden Informationen müssen im Tabellenblatt „Aircrafts“ eingetragen werden. Anschließend können mit dem ersten der beiden Makros die benötigten Daten zu den im Tabellenblatt „Aircrafts“ enthaltenen Flugzeugen eingelesen werden. Das Einlesen der Daten wird mit einem Klick auf den Button „Read Data“ im Tabellenblatt „Input“ gestartet.

Die einzulesenden Daten sind im Quellcode des Makros spezifiziert und können dort geändert oder erweitert werden. Zunächst werden die benötigten Daten aus den verschiedenen Datenquellen eingelesen und in die Tabellenblätter „Datenmatrix“ und „Datenmatrix_Q“ geschrieben. In diesen Tabellenblättern werden die Ausprägungen aller verfügbaren Variablen für alle verfügbaren Flugzeugtypen gespeichert und dargestellt. Der Unterschied zwischen den beiden Tabellenblättern besteht in der Darstellung der Ausprägung ordinaler Variablen. Diese werden im Tabellenblatt „Datenmatrix“ mit ihren Bezeichnungen geführt, während sie im Tabellenblatt „Datenmatrix_Q“ für die Weiterverarbeitung durch Zahlenwerte ausgedrückt werden. Nach dem Einlesen der benötigten Daten werden jene Variablen berechnet, für die Daten aus mehreren Datenquellen benötigt werden. Diese Variablen werden ebenfalls in die Tabellenblätter mit den Datenmatrizen geschrieben.

Nach dem Einlesen der Daten sind diese in den beiden Datenmatrizen erhalten. Sie müssen nur im Falle einer Aktualisierung der Daten (beispielsweise aufgrund einer neuen Version einer Datenbank) oder einer Änderung der enthaltenen Flugzeuge im Tabellenblatt „Aircrafts“ oder einer Änderung der Zuordnung dieser Flugzeuge neu eingelesen werden.

Die eigentliche Clusteranalyse wird mit dem Button „Clustering“ im Tabellenblatt „Input“ gestartet. Der grundlegende Ablauf der Clusteranalyse ist in Abbildung 4-1 dargestellt. Zunächst werden die nötigen Eingaben des Nutzers eingelesen. Das sind die zu verwendenden Flugzeugtypen und Clustering-Parameter (Variablen), der Bereich der gewünschten Anzahl der Cluster (minimale und maximale Anzahl) und das zu verwendende Unähnlichkeitsmaß. Im nächsten Schritt werden die Datenmatrix aus dem Tabellenblatt „Datenmatrix_Q“ eingelesen und die nominalen Variablen durch Dummy-Variablen ersetzt. Im Falle der Verwendung der Mahalanobis-Distanz werden die Kovarianzmatrix der verwendeten Clustering-Parameter und deren Inverse berechnet. Aus den Daten sowie der Kovarianzmatrix wird anschließend die Unähnlichkeitsmatrix berechnet. Bei Verwendung eines anderen Unähnlichkeitsmaßes müssen die Daten lediglich standardisiert und aus den standardisierten Daten die Unähnlichkeitsmatrix ermittelt werden. Im nächsten Schritt folgt die eigentliche Berechnung der Clusterstrukturen aus der Unähnlichkeitsmatrix für alle gewünschten Zahlen von Clustern. Zum Schluss werden die Hilfsmaße zur Bestimmung einer geeigneten Anzahl der Cluster ermittelt und die Ergebnisse in Excel dargestellt.

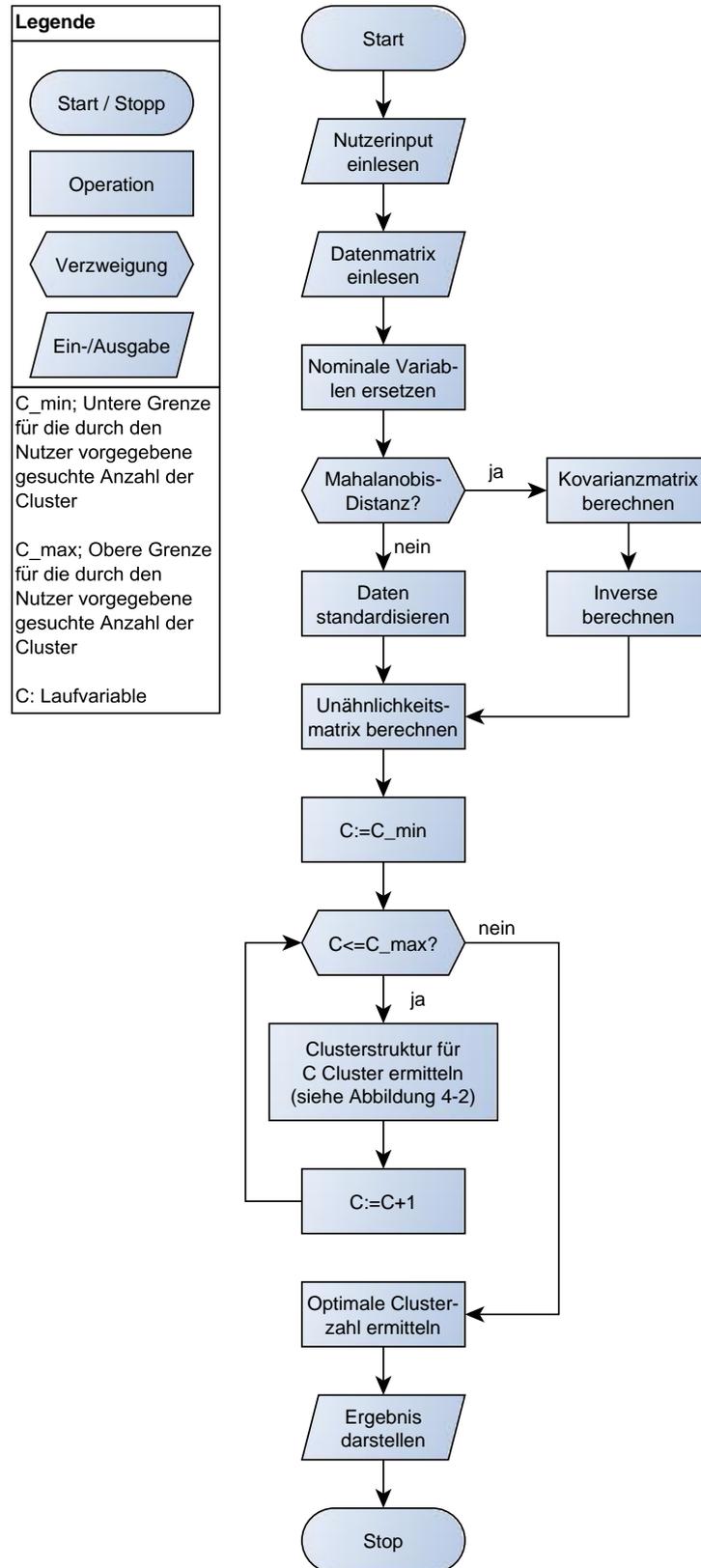


Abbildung 4-1: Flussdiagramm Clustering

Die Ermittlung einer Clusterstruktur für eine vorgegebene Anzahl von Clustern ist in einem separaten Flussdiagramm in Abbildung 4-2 dargestellt. Die dargestellte Prozedur greift auf die benötigten Daten aus dem in Abbildung 4-1 dargestellten Hauptprogramm zu. Zunächst wird entsprechend dem Ablauf des Medoid-Verfahrens das erste Medoid festgelegt. Dies ist jenes zu clusternde Objekt, für das die Summe der Unähnlichkeiten zu allen anderen Objekten minimal ist. Die Anzahl der bereits festgelegten Medoide wird anschließend auf eins gesetzt. Nun werden schrittweise so lange weitere Medoide bestimmt, bis die Anzahl der Medoide der gesuchten Anzahl der Cluster entspricht. Das Kriterium für die Festlegung dieser Medoide ist die Minimierung der Summe der Unähnlichkeiten zwischen jedem Objekt und dem jeweils nächsten Medoid. Wenn die Anzahl der festgelegten Medoide der gesuchten Clusteranzahl entspricht, ist die Anfangskonfiguration erreicht. Anschließend wird versucht, diese Anfangskonfiguration durch den Austausch der festgelegten Medoide zu verbessern. Ziel ist weiterhin die Minimierung der Summe der Unähnlichkeiten zwischen den Objekten und dem jeweils nächsten Medoid. Für die Identifikation eines möglichen Tauschs werden alle Möglichkeiten des Tauschs eines Medoids gegen ein Objekt, das kein Medoid ist, evaluiert. Wenn mehrere Tauschmöglichkeiten die Summe der Unähnlichkeiten reduzieren würden, wird derjenige Tausch durchgeführt, der zur größten Abnahme der Summe der Unähnlichkeiten führt. Nach einem Tausch werden erneut alle Möglichkeiten evaluiert und so lange Austausche durchgeführt, bis kein Tausch mit Reduktion der Summe der Unähnlichkeiten mehr möglich ist. Am Ende der Bestimmung der Clusterstruktur werden die Cluster durch Zuordnung jedes Objekts zum jeweils nächsten Medoid gebildet.

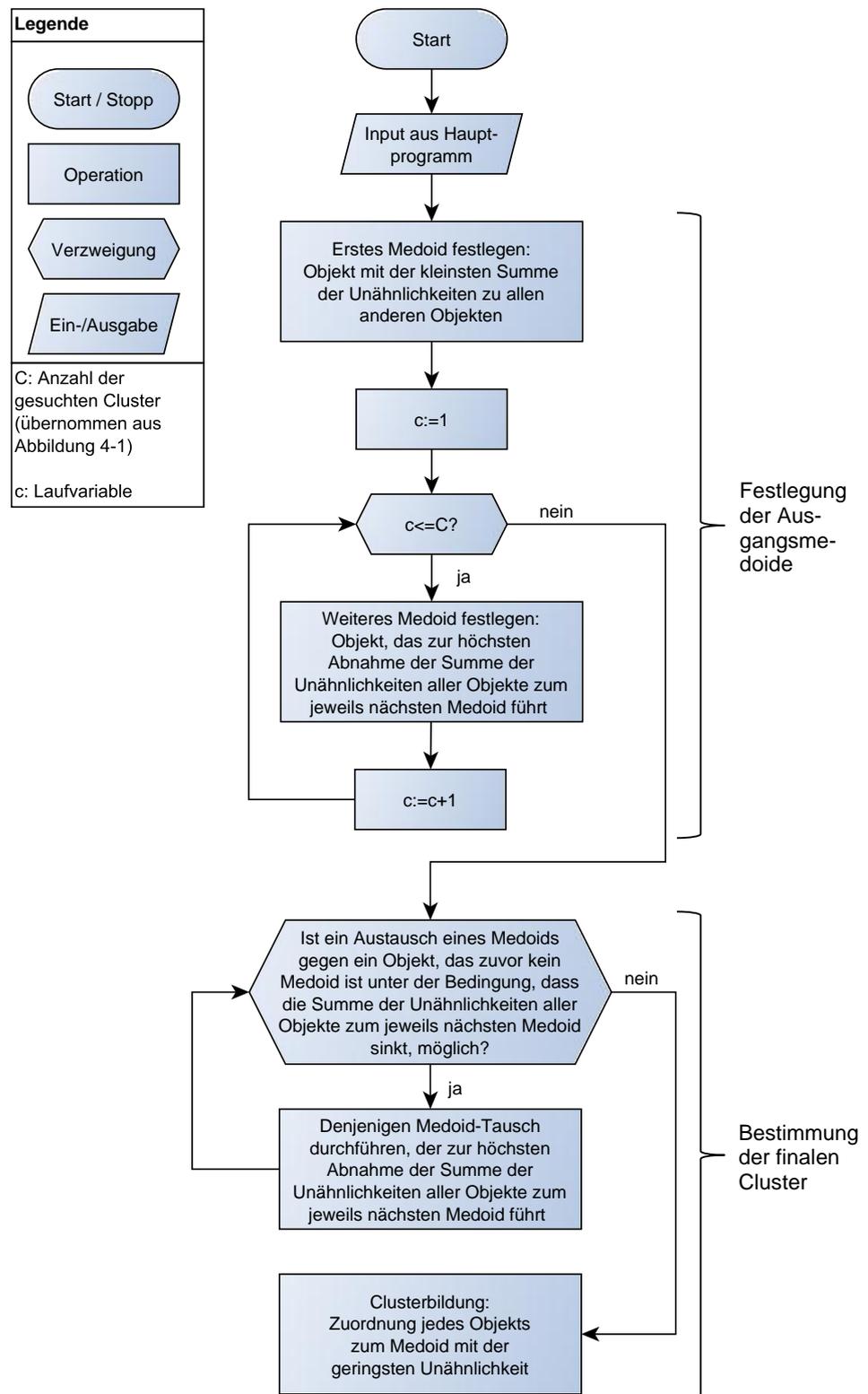


Abbildung 4-2: Flussdiagramm zur Ermittlung einer Clusterstruktur (Medoid-Verfahren)

4.4 Bestimmung der Anzahl der Cluster

Die Anzahl der Cluster hängt von den zugrundeliegenden Daten und den verwendeten Variablen zur Clusteranalyse ab. Das bedeutet, dass es entsprechend der verschiedenen möglichen Anzahlen der Cluster Lösungen der Clusteranalyse gibt, die die zugrundeliegende wahre Struktur der Daten besser erklären, und solche, die diese weniger gut erklären. Eine absolute Bewertung der Qualität einer Clusterstruktur ist eine der größten Herausforderungen der Clusteranalyse. Für die relative Bewertung der Qualität verschiedener Lösungen gibt es jedoch einige Hilfsmittel [Bacher et al. 2010, S. 305-309, 470-472].

Für das Medoid-Verfahren sind im entwickelten Clusteranalyse-Tool zwei verschiedene Hilfsmittel zur relativen Bewertung der Clusterstrukturen und damit zur Festlegung der Anzahl der Cluster verfügbar: Ein Hilfsmittel ist der relative Abfall der Summe der Unähnlichkeiten bei Erhöhung der Anzahl der Cluster; das andere Hilfsmittel ist die „Average Silhouette Width“ nach Kaufman/Rousseeuw (1990).

4.4.1 Relativer Abfall der Summe der Unähnlichkeiten

Beim Hilfsmittel des relativen Abfalls der Unähnlichkeiten wird als Grundlage der relativen Bewertung verschiedener Clusterstrukturen die Summe der Unähnlichkeit der verschiedenen Clusterstrukturen verwendet. Die direkte Verwendung der Summe der Unähnlichkeiten ist nicht sinnvoll, weil diese beim Medoid-Verfahren bei Erhöhung der Anzahl der Cluster immer sinkt. Eine gewisse Aussagekraft kommt jedoch der Stärke des Abfalls der Summe der Unähnlichkeiten bei einer Erhöhung der Anzahl der Cluster um 1 zu [Bacher et al. 2010, S. 299, 305-311]. Eine gute Clusterstruktur liegt dann vor, wenn der auf diese Clusterstruktur folgende (bei einer Erhöhung der Anzahl der Cluster um 1) relative Abfall der Summe der Unähnlichkeiten deutlich geringer ist als der vorhergehende relative Abfall der Unähnlichkeiten. Das Prinzip ist analog zur Identifikation von Knicken in einem Diagramm, in dem die Summe der Unähnlichkeiten über die Anzahl der Cluster aufgetragen ist. In der Regel sind diese Knicke im Diagramm jedoch nur schwer erkennbar [Bacher et al. 2010, S. 299, 305-311]. Abbildung 4-3 zeigt ein solches Diagramm. Es basiert auf den gleichen Daten wie Tabelle 5-3. Die im Vergleich relativ guten Clusterstrukturen sind hier jedoch deutlich schwerer erkennbar.

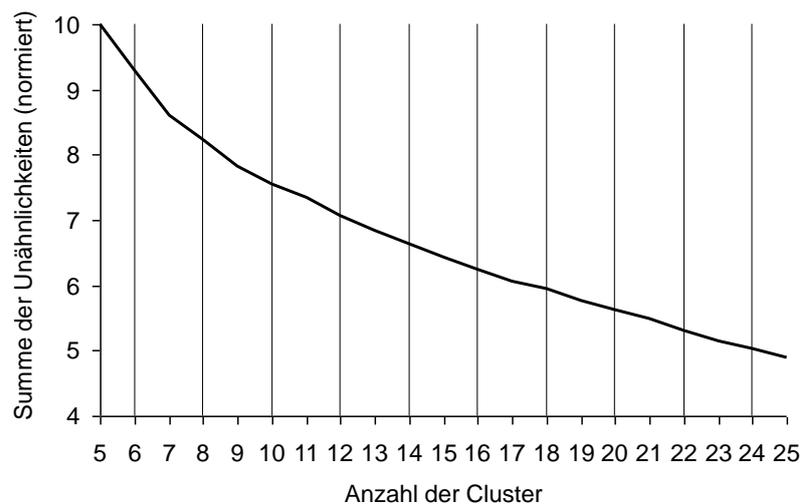


Abbildung 4-3: Abnahme der Summe der Unähnlichkeiten mit steigender Anzahl der Cluster

Das beschriebene Vorgehen wird von Bacher et al. (2010) unter anderem für das K-Means-Verfahren empfohlen. Im Fall des K-Means-Verfahrens wird dazu der relative Abfall der Streuungsquadratsumme in den Clustern betrachtet. Die Streuungsquadratsumme in den Clustern entspricht dabei der Summe der quadrierten euklidischen Distanzen zwischen den Objekten und ihren Clusterzentren. Sie ist gleichzeitig diejenige Größe, die bei der Clusteranalyse nach dem K-Means-Verfahren minimiert wird. Sie ist äquivalent zur Summe der Unähnlichkeiten beim Medoid-Verfahren, welche gleichzeitig die Zielgröße des Medoid-Verfahrens darstellt. Der Unterschied besteht darin, dass die Summe der Unähnlichkeiten sich auf die Unähnlichkeiten zwischen den Objekten und ihren Repräsentanten, und nicht auf die Unähnlichkeiten zwischen den Objekten und ihren Clusterzentren als hypothetische Repräsentanten bezieht. Bei Verwendung eines von der quadrierten euklidischen Distanz abweichenden Unähnlichkeitsmaßes wird dieses Unähnlichkeitsmaß auch für die Ermittlung der Auswahl der Anzahl der Cluster verwendet, um die Konsistenz zwischen der Ermittlung der optimalen Clusterstruktur und der besten Anzahl der Cluster zu wahren [Bacher et al. 2010, S. 299, 305-311].

4.4.2 Average Silhouette Width

Die „Average Silhouette Width“ ist der Durchschnitt einer Bewertung der Qualität der Clusterzuordnung („Silhouette Width“) jedes Objekts. Die Qualität der Zuordnung jedes Objekts zu seinem Cluster basiert auf der durchschnittlichen Unähnlichkeit des Objekts zu den anderen Objekten des gleichen Clusters sowie der durchschnittlichen Unähnlichkeit des Objekts zu den Objekten des nächstgelegenen Clusters. Als nächstgelegenes Cluster eines Objekts ist jenes Cluster definiert, für das die durchschnittliche Unähnlichkeit mit dem betrachteten Objekt minimal ist, unter der Bedingung, dass es sich nicht um das Clus-

ter handelt, zu dem das betrachtete Objekt selbst gehört [Kaufman/Rousseeuw 1990, S. 83-87].

Eine gute Clusterzuordnung ist dadurch gekennzeichnet, dass die durchschnittliche Unähnlichkeit des Objekts zu den Objekten des gleichen Clusters sehr gering ist im Verhältnis zur durchschnittlichen Unähnlichkeit zu den Objekten der anderen Cluster. Eine geringe durchschnittliche Unähnlichkeit innerhalb des Clusters deutet auf eine hohe Homogenität innerhalb des Clusters hin; eine hohe Unähnlichkeit mit den anderen Clustern auf eine hohe Heterogenität zwischen den Clustern. Die durchschnittliche „Silhouette Width“ aller Objekte stellt somit ein Maß für die Qualität der gefundenen Clusterstruktur dar [Kaufman/Rousseeuw 1990, S. 83-87].

Basierend auf Erfahrungswerten liefern Kaufman und Rousseeuw auch eine Interpretation der „Average Silhouette Width“ in Form einer absoluten Bewertung der Qualität der gefundenen Clusterstruktur. Die entsprechenden Bewertungen sind in Tabelle 4-1 dargestellt [Kaufman/Rousseeuw 1990, S. 87-88].

Tabelle 4-1: Interpretation der Average Silhouette Width²

Average Silhouette Width	Vorgeschlagene Interpretation
0,71 - 1,00	Starke Clusterstruktur gefunden
0,51 - 0,70	Erkennbare Clusterstruktur gefunden
0,26 - 0,50	Schwache, möglicherweise nur zufallsbedingte Clusterstruktur gefunden
≤ 0,25	Keine erkennbare Clusterstruktur gefunden

² Nach Kaufman/Rousseeuw 1990, S. 88

5 Anwendung der Clusteranalyse auf zivile Flugzeugmuster

In diesem Kapitel wird die Anwendung der Clusteranalyse auf zivile Flugzeugmuster im Rahmen dieser Arbeit beschrieben. Die allgemeine Motivation dieser Arbeit und der Entwicklung eines Tools zur Clusteranalyse für Flugzeuge ist die Reduktion der Komplexität der Betrachtung der zukünftigen Entwicklung von Flugzeugflotten. In diesem Themenbereich gibt es jedoch vielfältige mögliche Anwendungsfälle einer Clusteranalyse. Für diese verschiedenen Anwendungsfälle kann das gleiche Clusteranalyse-Tool mit dem identischen Clusteranalyseverfahren verwendet werden.

Die Parameter, nach denen die Flugzeugtypen in Cluster eingeteilt werden, sowie die in der Clusteranalyse betrachteten Flugzeuge sollten jedoch spezifisch für die konkrete Problemstellung ausgewählt werden. Die Problemstellung, für welche die Clusteranalyse in dieser Arbeit durchgeführt werden soll, ist die Vereinfachung eines Flugplans hinsichtlich der Anzahl unterschiedlicher Flugzeugtypen. Dies wird durch die Bildung von Clustern der verschiedenen Flugzeugtypen und das anschließende Ersetzen jedes Flugzeugs im Flugplan durch den Repräsentanten seines Clusters bewerkstelligt. Im Ergebnis muss für auf dem Flugplan basierende Analysen nicht jeder Flugzeugtyp aufwändig modelliert werden. Stattdessen muss lediglich eine überschaubare Anzahl von repräsentativen Flugzeugen für die gebildeten Cluster modelliert werden. Um in den späteren Analysen keine verfälschten Ergebnisse zu erhalten, müssen die Transportleistung der vereinfachten Flugzeugflotte und deren Kraftstoffverbrauch möglichst genau jenen Werten der realen Flotte entsprechen. In Abschnitt 5.1 wird die Spezifikation der Clusteranalyse für diese konkrete Problemstellung entwickelt. Dazu gehören vor allem die Auswahl der Clustering-Parameter und die Auswahl der behandelten Flugzeuge.

In Abschnitt 5.2 werden die Ergebnisse der Anwendung der Clusteranalyse vorgestellt. Dazu werden auf Basis der Bewertung der Qualität unterschiedlicher Cluster-Lösungen zwei ausgewählte Clusterstrukturen präsentiert.

Eine kritische Betrachtung des Clusteranalyse-Tools erfolgt in Abschnitt 5.3. Dabei wird die Erfüllung der definierten Anforderungen überprüft und es werden basierend auf den Erkenntnissen aus der Anwendung der Clusteranalyse Schlüsse zur Eignung des Verfahrens für die Lösung der Problemstellung dieser Arbeit gezogen.

5.1 Spezifikation der Clusteranalyse

In diesem Abschnitt wird die Spezifikation der Clusteranalyse für die Vereinfachung eines Flugplans mithilfe einer Clusteranalyse entwickelt. Dazu werden in Abschnitt 5.1.1 die für diese Problemstellung nötigen Clustering-Parameter identifiziert und ihre Berechnung erklärt. In Abschnitt 5.1.2 wird ein Überblick über die für diese Clusteranalyse verwendeten Flugzeugtypen gegeben.

Für die vorliegende Clusteranalyse wird, wie in Abschnitt 4.2.1 erläutert, die euklidische Distanz als Unähnlichkeitsmaß verwendet. Sie wird für einen weiten Bereich der Anzahl der Cluster von 5 bis 25 durchgeführt.

5.1.1 Clustering-Parameter

In diesem Abschnitt werden die Parameter, die für das Clustering der Flugzeuge im Rahmen dieser Arbeit verwendet werden, behandelt. Dazu wird zunächst hergeleitet, welche Parameter entsprechend dem Anwendungsziel in dieser Arbeit erforderlich sind (Abschnitt 5.1.1.1). Anschließend wird die Berechnung der festgelegten Parameter erklärt (Abschnitt 5.1.1.2).

5.1.1.1 Festlegung der Clustering-Parameter

Im Folgenden wird die Festlegung der Clustering-Parameter für die Clusteranalyse zur Vereinfachung des Flugplans behandelt. Dazu wird zum einen hergeleitet, welche Parameter für die Vereinfachung des Flugplans nötig sind. Zum anderen wird die Berechnung der ausgewählten Parameter erläutert.

Im Rahmen der Vereinfachung des Flugplans sollen die Transportkapazität und der Kraftstoffverbrauch der für diesen Flugplan eingesetzten Flugzeugflotte möglichst nicht verändert werden. Der zu ersetzende Flugzeugtyp und der als Ersatz verwendete Flugzeugtyp müssen sich in ihren Charakteristika bezüglich der Transportkapazität und des Kraftstoffverbrauchs daher möglichst ähnlich sein.

Zur Abbildung der Transportkapazität eines Flugzeugs werden die Transportkapazitäten für Passagiere und Fracht separat betrachtet, weil unterschiedliche Transportgüter unterschiedlichen Dienstleistungen der Fluggesellschaften entsprechen.

Damit ein Flugzeug auf einer bestimmten Flugstrecke die gleiche Passagiertransportkapazität anbieten kann, müssen die Anzahl der Sitzplätze sowie die Reisefluggeschwindigkeit möglichst ähnlich sein. Insbesondere für Kurzstreckenflüge könnten auch Faktoren wie die Abfertigungszeit am Boden („turnaround time“) oder die durchschnittlichen Geschwindigkeiten im Steig- und Sinkflug relevant sein, weil der Reiseflug nur einen geringen Anteil an der gesamten Einsatzzeit des Flugzeugs für eine Flugstrecke hat. Aufgrund der für eine Clusteranalyse ratsamen Beschränkung auf die Verwendung der wichtigsten Faktoren sowie der Verfügbarkeit geeigneter Daten, werden für die Abbildung der Transportkapazität für Passagiere nur die Anzahl der Sitzplätze und die Reisefluggeschwindigkeit verwendet.

Die Betrachtung für die Frachttransportkapazität ist äquivalent zur Passagiertransportkapazität. Anstelle der verfügbaren Sitzplätze werden jedoch die verfügbaren Frachttonnen verwendet. Zur Abbildung der Transportkapazität für Fracht werden daher die verfügbaren Frachttonnen und die Reisefluggeschwindigkeit verwendet.

Um auch den Kraftstoffverbrauch dieser Flugzeuge möglichst gut annähern zu können, sind weitere Parameter nötig. Diese sind der spezifische Kraftstoffverbrauch bezogen auf die verfügbare Transportkapazität, die verfügbare Nutzlast, die Reisefluggeschwindigkeit und die typische Flugstrecke des Flugzeugs. Mit diesen Parametern kann der gesamte Kraftstoffverbrauch eines Flugzeugs auf einer bestimmten Strecke näherungsweise ermittelt und mit anderen Flugzeugen in Relation gesetzt werden.

Der spezifische Kraftstoffverbrauch dient der Abbildung der Kraftstoffeffizienz der verschiedenen Flugzeuge. Um eine Verzerrung durch die verfügbare Nutzlast und die Reisefluggeschwindigkeit zu vermeiden, wird der spezifische Kraftstoffverbrauch auf die verfügbare Transportkapazität bezogen. Eine ähnlich große verfügbare Nutzlast und Reisefluggeschwindigkeit sind ebenfalls wichtig, um beim Ersetzen eines Flugzeugs durch einen Repräsentanten einen möglichst ähnlichen Kraftstoffverbrauch für eine gegebene Flugstrecke zu erhalten. Die typische Flugstrecke ist wichtig, um auch den Anteil des Kraftstoffverbrauchs eines Flugzeugs im Steig- und Sinkflug abbilden zu können. Für Flugzeuge, die typischerweise auf Kurzstrecken eingesetzt werden, ist dieser Anteil beispielsweise deutlich höher als bei Langstreckenflugzeugen.

Da Flugzeuge mit Jetantrieb und Flugzeuge mit Turboprop-Antrieb jeweils eine grundlegend verschiedene Charakteristik des Kraftstoffverbrauchs haben, wird zusätzlich der Antriebstyp als Parameter verwendet.

5.1.1.2 Berechnung der Clustering-Parameter

In diesem Abschnitt wird die Berechnung der verwendeten Clustering-Parameter erläutert. Die Daten für die Clustering-Parameter stammen aus der OAG-Datenbank und der BADA-Datenbank. Im Clusteranalyse-Tool sind zwar Daten aus der ACAS-Datenbank verfügbar, diese werden jedoch für die Anwendung des Verfahrens im Rahmen dieser Arbeit nicht benötigt. Aus der OAG-Datenbank werden alle geplanten Flüge für den Monat Juni im Jahr 2008 verwendet. Es wird vereinfachend angenommen, dass der Juni hinsichtlich der angebotenen Flüge repräsentativ für das gesamte Jahr ist. Eventuelle Änderungen des Flugplans oder Flugausfälle, die in den Flugplandaten aus der OAG-Datenbank nicht enthalten sind, können nicht berücksichtigt werden. Von der BADA-Datenbank wird die Version 3.10 verwendet.

Anzahl der verfügbaren Sitzplätze

Die Anzahl der Sitzplätze eines Flugzeugs wird als durchschnittliche Anzahl der verfügbaren Sitzplätze eines Flugzeugs, bezogen auf die gesamte angebotene Flugstrecke eines Flugzeugs (Großkreisentfernung) im Betrachtungszeitraum berechnet. Aus der OAG-Datenbank wird dazu der nach der Großkreisentfernung (DISTKM) und der Anzahl der Flüge (FREQUENCY) gewichtete Durchschnitt der verfügbaren Sitzplätze (SEATS) eines Flugzeugs berechnet.

Verfügbare Frachttonnen

Die verfügbaren Frachttonnen eines Flugzeugs werden analog der Anzahl der Sitzplätze als durchschnittlich verfügbare Frachttonnen, bezogen auf die gesamte angebotene Flugstrecke eines Flugzeugs (Großkreisentfernung) im Betrachtungszeitraum berechnet. Aus der OAG-Datenbank wird dazu der nach der Großkreisentfernung (DISTKM) und der Anzahl der Flüge (FREQUENCY) gewichtete Durchschnitt der verfügbaren Frachttonnen (FREIGHTTONS) eines Flugzeugs berechnet.

Verfügbare Nutzlast

Die verfügbare Nutzlast setzt sich aus der verfügbaren Nutzlast für die Passagierbeförderung und den verfügbaren Frachttonnen zusammen. Sie wird unter der Annahme, dass ein Sitzplatz einer verfügbaren Nutzlast von 100 kg entspricht, aus der zuvor ermittelten Anzahl der Sitzplätze und der verfügbaren Frachttonnen ermittelt.

Reisefluggeschwindigkeit

Als Reisefluggeschwindigkeit wird die wahre Fluggeschwindigkeit („True Airspeed“, TAS) des Flugzeugs bei maximaler Flughöhe aus den „Performance Table Files“ der BADA-Datenbank verwendet.

Durchschnittliche Flugstrecke

Die durchschnittliche Flugstrecke eines Flugzeugs wird als arithmetischer Durchschnitt der Großkreisentfernung vom Startflughafen zum Zielflughafen für alle Flüge eines Flugzeugs berechnet. Eventuelle Zwischenstopps werden dabei nicht in Form einer Erhöhung der Entfernung berücksichtigt, weil diese nicht zu einer höheren Beförderungsleistung im Vergleich zu einem Direktflug auf der gleichen Strecke führen. Aus der OAG-Datenbank wird zur Berechnung der durchschnittlichen Flugstrecke der mit der Anzahl der Flüge (FREQUENCY) eines Flugplaneintrags gewichtete Durchschnitt der Großkreisentfernung (DISTKM) berechnet.

Spezifischer Kraftstoffverbrauch

Der spezifische Kraftstoffverbrauch ist der Kraftstoffverbrauch des jeweiligen Flugzeugs bezogen auf die angebotene Transportleistung. Er wird aus dem Kraftstofffluss (Masse des verbrauchten Kraftstoffs pro Zeit) durch Division durch die verfügbare Nutzlast und die Reisefluggeschwindigkeit ermittelt. Die Einheit des so berechneten spezifischen Kraftstoffverbrauchs ist Kilogramm Kraftstoff je verfügbare Tonne Nutzlast und Kilometer (Tonnenkilometer). Als Kraftstofffluss wird der „Fuel Flow“ bei maximaler Flughöhe und nominaler Flugzeugmasse aus den „Performance Table Files“ der BADA-Datenbank verwendet. Für die verfügbare Nutzlast und die Reisefluggeschwindigkeit werden die beiden entsprechenden, ebenfalls in diesem Abschnitt beschriebenen, Parameter herangezogen.

Verwendung findende quadrierte euklidische Distanz entspricht der Summe der quadrierten Differenzen aller Variablen zweier Objekte. Formel 5-1 zeigt die Berechnung des spezifischen Kraftstoffverbrauchs..

$$\text{Spez. Kraftstoffverbrauch} = \frac{\text{Kraftstofffluss}}{\text{Verfügbare Nutzlast} \times \text{Reisefluggeschw.}} \quad (5-1)$$

Antriebstyp

Mit der Variable Antriebstyp werden die Flugzeuge in Flugzeuge mit Jetantrieb und Turbopropantrieb unterteilt. Die entsprechende Information kommt aus der Variable „Category“ der OAG-Datenbank, in der zwischen „Wide bodied Jets“, „Narrow bodied Jets“ und „Turboprops“ unterschieden wird.

Neben den erläuterten Clustering-Parametern sind im Clusteranalyse-Tool weitere Parameter für andere Anwendungszwecke verfügbar. Dies sind die Reisefluggeschwindigkeit aus der OAG-Datenbank, das maximale Startgewicht („Maximum Takeoff Weight“, MTOW) zum einen aus der OAG-Datenbank und zum anderen aus der ACAS-Datenbank, der Kraftstofffluss, der zur Berechnung des spezifischen Kraftstoffverbrauchs verwendet wird, die Wirbelschleppenkategorie, die Anzahl der Triebwerke, die Flügelspannweite und das Verhältnis aus verfügbarer Nutzlast und maximalem Startgewicht. Weitere Parameter können jederzeit hinzugefügt werden.

5.1.2 Verwendete Flugzeugtypen

Für die Anwendung der Clusteranalyse werden insgesamt 185 Flugzeugtypen verwendet. Diese sind in Tabelle 5-1 dargestellt.

Tabelle 5-1: Bei der Anwendung der Clusteranalyse verwendete OAG Specific Aircraft

OAG Specific Aircraft	OAG Specific Aircraft Name	OAG Specific Aircraft	OAG Specific Aircraft Name
ND2	Aerospatiale (nord) 262	74F	Boeing 747 (Freighter)
ABF	Airbus A300 (Freighter)	74M	Boeing 747 (Mixed Configuration)
AB3	Airbus A300 Passenger	747	Boeing 747 (Passenger)
ABY	Airbus A300-600 Freighter	741	Boeing 747-100 (Passenger)
AB6	Airbus A300-600 Passenger	74X	Boeing 747-200 (Freighter)
AB4	Airbus A300B2 /B4 Passenger	742	Boeing 747-200 (Passenger)
ABX	Airbus A300B4 /A300C4 /A300F4	743	Boeing 747-300 /747-100 /200 Sud (Pax)
31F	Airbus A310 Freighter	74D	Boeing 747-300 /747-200 Sud (Mxd Config)
310	Airbus A310 Passenger	74E	Boeing 747-400 (Mixed Configuration)
312	Airbus A310-200 Passenger	744	Boeing 747-400 (Passenger)
31Y	Airbus A310-300 Freighter	74Y	Boeing 747-400F (Freighter)
313	Airbus A310-300 Passenger	74L	Boeing 747SP Passenger
318	Airbus A318	757	Boeing 757 (Passenger)

5 Anwendung der Clusteranalyse auf zivile Flugzeugmuster

OAG Specific Aircraft	OAG Specific Aircraft Name
32S	Airbus A318 /319 /320 /321
319	Airbus A319
320	Airbus A320
321	Airbus A321
330	Airbus A330
332	Airbus A330-200
333	Airbus A330-300
340	Airbus A340
342	Airbus A340-200
343	Airbus A340-300
345	Airbus A340-500
346	Airbus A340-600
380	Airbus A380-800 Passenger
ANF	Antonov AN-12
AN4	Antonov AN-24
AN6	Antonov AN-26 /30 /32
A28	Antonov An28 /Pzl Mielec M-28 Sky-truck
AN7	Antonov AN-72 /74
AT4	ATR 42-300 /320
AT5	ATR 42-500
AT7	ATR 72
ATR	ATR42 /ATR72
AR1	Avro RJ100
AR7	Avro RJ70
AR8	Avro RJ85
B11	BAe (BAC) 1-11
HS7	BAe (HS) 748
14F	BAe 146 (Freighter)
146	BAe 146 Passenger
141	BAe 146-100 Passenger
142	BAe 146-200 Passenger
143	BAe 146-300 Passenger
ATP	BAe ATP
JST	BAe Jetstream
J31	BAe Jetstream 31
J32	BAe Jetstream 32
J41	BAe Jetstream 41
BE1	Beechcraft 1900 Airliner
BES	Beechcraft 1900C Airliner
BEH	Beechcraft 1900D Airliner
BE9	Beechcraft C99 Airliner
D1Y	Boeing (douglas) DC10 - 30 /40 Freighter
D1F	Boeing (douglas) DC10 (Freighter)
D1C	Boeing (douglas) DC10-30 /40 (Pax)
D8F	Boeing (douglas) DC8 Freighter
DC8	Boeing (Douglas) DC-8 Passenger
D8Y	Boeing (douglas) DC-8-71 /72 /73 Frtr
D9F	Boeing (douglas) DC9 Freighter
DC9	Boeing (douglas) DC-9 Passenger
D93	Boeing (douglas) DC9-30 Passenger
D95	Boeing (douglas) DC9-50 Passenger
M1F	Boeing (douglas) MD-11 (Freighter)
M11	Boeing (douglas) MD-11 Passenger

OAG Specific Aircraft	OAG Specific Aircraft Name
75W	Boeing 757-200 (winglets) Passenger
752	Boeing 757-200 Passenger
75F	Boeing 757-200PF (Freighter)
753	Boeing 757-300 Passenger
76F	Boeing 767 Freighter
767	Boeing 767 Passenger
762	Boeing 767-200 Passenger
76Y	Boeing 767-300 Freighter
763	Boeing 767-300 Passenger
764	Boeing 767-400 Passenger
777	Boeing 777 Passenger
772	Boeing 777-200 Passenger
77L	Boeing 777-200LR
773	Boeing 777-300 Passenger
77W	Boeing 777-300ER Passenger
CRA	Canadair Crj Series 705
CRJ	Canadair Regional Jet
CR1	Canadair Regional Jet 100
CR2	Canadair Regional Jet 200
CR7	Canadair Regional Jet 700
CR9	Canadair Regional Jet 900
CS5	CASA /Nusantara CN-235
CS2	CASA C212 /Nusantara NC-212 Aviocar
DHT	De Havilland DHC-6 Twin Otter
DH8	De Havilland DHC-8 Dash 8
DH1	De Havilland DHC-8-100 Dash 8 /8q
DH2	De Havilland DHC-8-200 Dash 8 /8q
DH3	De Havilland DHC-8-300 Dash 8 /8q
EMB	Embraer 110 Bandeirante
EM2	Embraer 120 Brasilia
E70	Embraer 170
EMJ	Embraer 170 /195
E75	Embraer 175
E90	Embraer 190
E95	Embraer 195
ERJ	Embraer RJ 135 /140 /145
ER3	Embraer RJ135
ERD	Embraer RJ140
ER4	Embraer RJ145
D28	Fairchild Dornier 228
D38	Fairchild Dornier 328-100
FRJ	Fairchild Dornier 328jet
SWM	Fairchild Sa26 /Sa226 /Sa277 Merlin /Me
100	Fokker 100
F50	Fokker 50
F70	Fokker 70
F27	Fokker F27 Friendship /Fairchild F27
F28	Fokker F28 Fellowship
F24	Fokker F28-4000 Fellowship
IL8	Ilyushin IL-18
IL6	Ilyushin Il-62
IL7	Ilyushin IL-76
ILW	Ilyushin IL-86

OAG Specific Aircraft	OAG Specific Aircraft Name	OAG Specific Aircraft	OAG Specific Aircraft Name
M80	Boeing (douglas) MD-80	IL9	Ilyushin Il-96 Passenger
M81	Boeing (douglas) MD-81	WWP	Israel Aircraft Ind.1124 Westwind
M82	Boeing (douglas) MD-82	LRJ	Learjet
M83	Boeing (douglas) MD-83	L4T	Let 410
M87	Boeing (douglas) MD-87	L1F	Lockheed L1011 Tristar Freighter
M88	Boeing (douglas) MD-88	L10	Lockheed L1011 Tristar Passenger
M90	Boeing (douglas) MD-90	LOH	Lockheed L182 /1282 /1382 (L100) Hercul
707	Boeing 707 /720 Passenger	LOF	Lockheed L188 Electra (Freighter)
717	Boeing 717-200	D9S	McD-Douglas DC9 30 /40 /50
72F	Boeing 727 (Freighter)	PL2	Pilatus PC-12
72M	Boeing 727 (Mixed Configuration)	S20	Saab 2000
727	Boeing 727 (Passenger)	SF3	Saab 340
721	Boeing 727-100 (Passenger)	SH3	Shorts 330 (sd3-30)
72S	Boeing 727-200 (Pax)	SH6	Shorts 360 (sd3-60)
722	Boeing 727-200 Passenger	TU3	Tupolev TU134
73F	Boeing 737 (Freighter)	TU5	Tupolev TU154
73M	Boeing 737 (Mixed Configuration)	T20	Tupolev TU-204 /tu-214
73S	Boeing 737 Advanced all Series	YK4	Yakovlev Yak-40
737	Boeing 737 Passenger	YK2	Yakovlev Yak-42 /142
73X	Boeing 737-200 (Freighter)		
732	Boeing 737-200 Passenger		
73Y	Boeing 737-300 (Freighter)		
73C	Boeing 737-300 (winglets) Passenger		
733	Boeing 737-300 Passenger		
73P	Boeing 737-400 (Freighter)		
73Q	Boeing 737-400 (Mixed Configuration)		
734	Boeing 737-400 Passenger		
735	Boeing 737-500 Passenger		
736	Boeing 737-600 Passenger		
73W	Boeing 737-700 (winglets) Passenger		
73G	Boeing 737-700 Passenger		
73H	Boeing 737-800 (winglets) Passenger		
738	Boeing 737-800 Passenger		
739	Boeing 737-900 Passenger		

Die für Juni 2008 ebenfalls in der OAG-Datenbank enthaltenen, aber bei der Anwendung der Clusteranalyse nicht verwendeten Flugzeugtypen sind in Tabelle 5-2 dargestellt. Die Boeing 737-200 (Mixed Configuration) wird nicht berücksichtigt, weil sie auf keinem der in der OAG-Datenbank enthaltenen geplanten Flüge eine Transportkapazität anbietet. Für die wenigen geplanten Flüge, denen dieser Flugzeugtyp zugeordnet ist, ist in der OAG-Datenbank weder eine Anzahl verfügbarer Sitzplätze, noch eine Anzahl verfügbarer Frachttonnen angegeben. Dabei handelt es sich möglicherweise um einen Fehler in den Flugplandaten der OAG-Datenbank. Die anderen in Tabelle 5-2 dargestellten Flugzeugtypen werden nicht berücksichtigt, weil kein vergleichbares Flugzeug in der BADA-Datenbank identifiziert werden konnte.

Tabelle 5-2: Bei der Anwendung der Clusteranalyse nicht berücksichtigte OAG Specific Aircraft

OAG Specific Aircraft	Grund für Nichtberücksichtigung
73L Boeing 737-200 (Mixed Configuration)	Keine Transportkapazität
A40 Antonov AN-140	Kein Entsprechung in der BADA-Datenbank
CNC Cessna Light Aircraft(Single Turboprop)	Kein Entsprechung in der BADA-Datenbank
CV5 Convair 580 Passenger	Kein Entsprechung in der BADA-Datenbank
CVF Convair 440 /580 /600 /640 (Freighter)	Kein Entsprechung in der BADA-Datenbank
DH7 De Havilland DHC-7 Dash 7	Kein Entsprechung in der BADA-Datenbank
DHB De Havilland DHC2 Beaver/Turbo Beaver	Kein Entsprechung in der BADA-Datenbank
DHL De Havilland DHC-3 Turbo Otter	Kein Entsprechung in der BADA-Datenbank
DHS De Havilland DHC 3 Otter	Kein Entsprechung in der BADA-Datenbank
I14 Ilyushin IL114	Kein Entsprechung in der BADA-Datenbank
YN2 Harbin Yunshuji Y12	Kein Entsprechung in der BADA-Datenbank
YN7 Xian Yunshuji Y7 /Ma60	Kein Entsprechung in der BADA-Datenbank
YS1 NAMC YS-11	Kein Entsprechung in der BADA-Datenbank

Nur relativ wenige „Specific Aircraft“ aus der OAG-Datenbank sind in identischer Konfiguration in der BADA-Datenbank enthalten und konnten direkt zugeordnet werden. Dies liegt vor allem daran, dass in der OAG-Datenbank relativ fein zwischen verschiedenen Konfigurationen differenziert wird. Beispielsweise werden oft Fracht- und Passagiermaschinen sowie teilweise gemischte Konfigurationen unterschieden. Diese Unterscheidung wird in der BADA-Datenbank üblicherweise nicht getroffen. In diesen Fällen wird das ähnlichste verfügbare Flugzeug, in der Regel aus der gleichen Flugzeugfamilie, verwendet. Für einige Familien von Flugzeugtypen ist kein Typ in der BADA-Datenbank enthalten. In diesem Fall ist in der BADA-Datenbank teilweise ein äquivalentes Flugzeug einer anderen Bauart, das sogenannte BADA-Äquivalent, angegeben. In diesen Fällen wird das BADA-Äquivalent für die Zuordnung in dieser Arbeit verwendet.

Die Zuordnungen der Flugzeugtypen aus der BADA-Datenbank zu den „Specific Aircraft“ aus der OAG-Datenbank sowie Kommentare zum Zustandekommen der Zuordnungen finden sich in Anlage 1.

5.2 Ergebnisse der Clusteranalyse

In diesem Abschnitt werden die Ergebnisse der durchgeführten Clusteranalyse vorgestellt. In Abschnitt 5.2.1 werden mögliche geeignete Clusterzahlen identifiziert. Auf dieser Basis werden in den Abschnitten 5.2.2 und 5.2.3 zwei ausgewählte Lösungen vorgestellt.

5.2.1 Anzahl der Cluster

Tabelle 5-3 zeigt die Bewertung der Qualität der Clusterstrukturen für die verschiedenen Anzahlen von Clustern, für die die Clusteranalyse durchgeführt wurde, anhand der in Abschnitt 4.4 vorgestellten Hilfsmittel.

Tabelle 5-3: Bewertung der Anzahl der Cluster

Cluster-anzahl	Abnahme der Summe der Unähnlichkeiten	Average Silhouette Width
5		0.389
6	7.1%	0.443
7	7.4%	0.378
8	4.2%	0.276
9	5.1%	0.146
10	3.5%	0.146
11	2.7%	0.130
12	3.7%	0.130
13	3.1%	0.130
14	3.2%	0.130
15	3.1%	0.141
16	2.8%	0.141
17	2.9%	0.141
18	2.0%	0.141
19	2.7%	0.141
20	2.6%	0.130
21	2.6%	0.141
22	3.1%	0.178
23	3.0%	0.222
24	2.5%	0.276
25	2.3%	0.276

Die Abnahme der Summe der Unähnlichkeiten bewertet die Qualität der verschiedenen Clusterstrukturen relativ zueinander. Es ist erkennbar, dass die Lösung mit 7 Clustern nach diesem Maß die beste Lösung ist. Der Grund ist, dass die Summe der Unähnlichkeiten zwischen der 6-Cluster-Lösung und der 7-Cluster-Lösung um 7,4% abfällt, während die folgende Verbesserung bei Erhöhung der Anzahl der Cluster auf 8 mit 4,2% deutlich geringer ist. Weitere relativ gute Clusterlösungen sind entsprechend der Abnahme der Summe der Unähnlichkeiten die Lösungen mit 9 und 17 Clustern.

Die Average Silhouette Width stellt eine absolute Bewertung der Qualität der verschiedenen Clusterlösungen dar. Die Ergebnisse zeigen, dass die Average Silhouette Width bei einer Anzahl der Cluster von 6 den maximalen Wert von 0,443 erreicht. Das Vorliegen einer Clusterstruktur in den Daten ist in diesem Bereich fraglich. Ab einer Anzahl der Cluster von 8 ist die Average Silhouette Width geringer oder nur minimal größer als 0,25, was darauf hindeutet, dass in den Daten eindeutig keine entsprechende Clusterstruktur vorliegt.

Zur Anzahl der Cluster lässt sich zusammenfassend sagen, dass die 7-Cluster-Lösung im Bereich der geringen Clusterzahlen am besten erscheint. Die Lösung mit 9 Clustern stellt dabei möglicherweise eine Alternative zu dieser Lösung dar. Im Bereich der höheren Clusterzahlen scheint die 17-Cluster-Lösung den angrenzenden Lösungen überlegen zu sein. Eine wahre Clusterstruktur liegt den Daten allem Anschein nach nicht zugrunde. Ange-

sichts der Natur der Daten ist das Vorliegen einer wahren Clusterstruktur auch nicht zu erwarten. Vielmehr sollten die Daten Zufallsverteilungen unterliegen.

Die Lösungen mit 7 und 17 Clustern werden in den folgenden Abschnitten näher betrachtet.

5.2.2 7-Cluster-Lösung

In diesem Abschnitt wird die Clusterstruktur mit 7 Clustern dargestellt und beschrieben. Dazu dienen vor allem Box Plots zur Darstellung der Eigenschaften der in den Clustern enthaltenen Flugzeuge.

Jeder Box Plot zeigt die Ausprägungen eines Clustering-Parameters für die verschiedenen Cluster. Das obere Ende des ausgefüllten Balkens markiert das 90. Perzentil, die Mittellinie im Balken den Median und die untere Begrenzung des Balkens das 10. Perzentil der Flugzeuge, die im entsprechenden Cluster enthalten sind, für den jeweils dargestellten Parameter. Das bedeutet, dass 80% der Flugzeuge eines Clusters innerhalb des ausgefüllten Balkens liegen. Die Fehlerbalken ober- und unterhalb des ausgefüllten Balkens reichen bis zu den Flugzeugen mit höchsten, beziehungsweise niedrigsten Werten für den dargestellten Parameter im entsprechenden Cluster. Die Lage des Repräsentanten eines Clusters ist als kleines Kreuz dargestellt.

Abbildung 5-1 zeigt die Anzahl der verfügbaren Sitzplätze und Abbildung 5-2 die verfügbaren Frachttonnen für die Clusterstruktur mit 7 Clustern. Die Cluster 4 und 6 bestehen ausschließlich aus Frachtflugzeugen, wobei beide Cluster hinsichtlich der Frachtkapazitäten der enthaltenen Flugzeuge klar voneinander abgegrenzt sind. Cluster 4 enthält viele kleinere Frachtflugzeuge wie die Frachtversion der Boeing 767, während Cluster 6 sehr große Frachtflugzeuge wie die Boeing 747 und die MD-11 enthält. Cluster 2 enthält alle und ausschließlich Turboprop-Flugzeuge, die relativ geringe Passagier- und Frachtkapazitäten haben. Während die bisher genannten Cluster hinsichtlich Passagier- und Frachtkapazität sehr gut voneinander abgegrenzt sind, unterscheiden sich die Cluster 1 und 3 und insbesondere die Cluster 5 und 7 weniger deutlich voneinander. Die Cluster 1 und 3 sind Cluster großer Passagierflugzeuge mit typischerweise mehr als 200 verfügbaren Sitzplätzen. Die Cluster 5 und 7 enthalten kleinere Passagier- und Frachtflugzeuge mit typischerweise weniger als 150 Sitzplätzen und einer geringen Frachtkapazität.

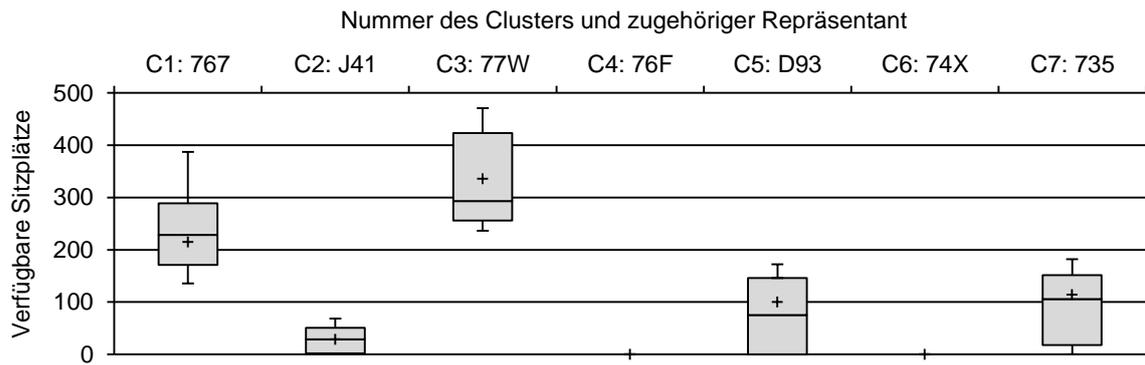


Abbildung 5-1: Box Plot der verfügbaren Anzahl der Sitzplätze für 7 Cluster

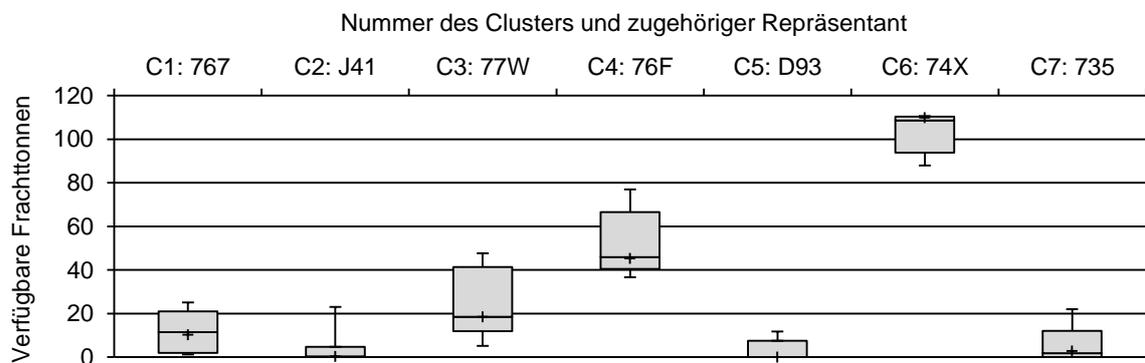


Abbildung 5-2: Box Plot der verfügbaren Frachtonnen für 7 Cluster

Abbildung 5-3 zeigt die verfügbare Nutzlast, die sich gemäß der Berechnung in Abschnitt 5.1.1.2 aus der verfügbaren Anzahl der Sitzplätze und den verfügbaren Frachtonnen ergibt.

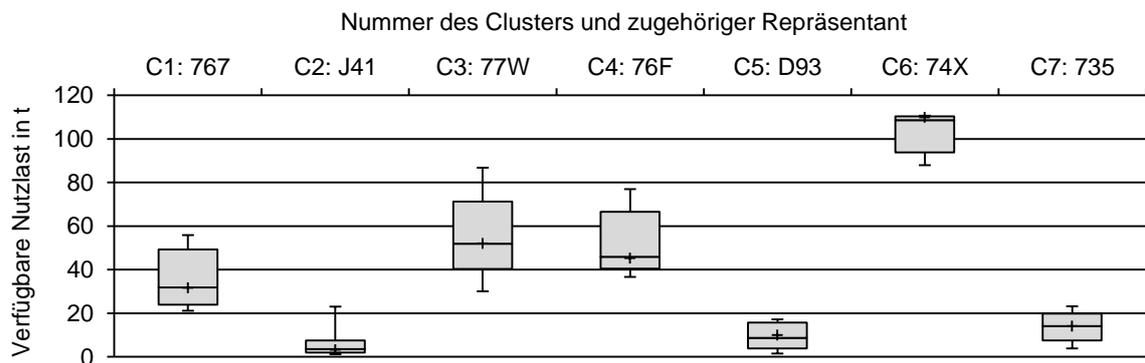


Abbildung 5-3: Box Plot der verfügbaren Nutzlast für 7 Cluster

Abbildung 5-4 ist der Box Plot der durchschnittlichen Flugstrecke. Während sich das Cluster 2 (Turboprop-Flugzeuge), sowie die beiden Frachtflugzeugcluster 4 und 6 hinsichtlich

der durchschnittlichen Flugstrecke gut voneinander unterscheiden lassen, sind sich die Cluster 1 und 3 sowie 5 und 7 auch in diesem Parameter relativ ähnlich.

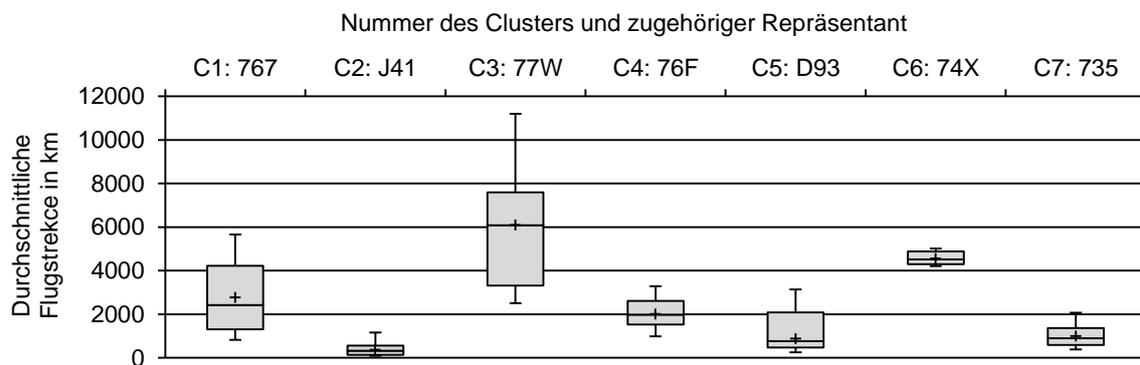


Abbildung 5-4: Box Plot mit der durchschnittlichen Flugstrecke für 7 Cluster

Abbildung 5-5 zeigt die Reisefluggeschwindigkeit aus der OAG-Datenbank. Die Darstellung wird von den deutlich unterschiedlichen Reisefluggeschwindigkeiten von Turboprop-Flugzeugen und Flugzeugen mit Jet-Antrieb dominiert. Die Cluster der Flugzeuge mit Jet-Antrieb sind einander sehr ähnlich, wobei die Cluster 5 und 7 jene Flugzeugtypen mit Jet-Antrieb beinhalten, die eine Reisefluggeschwindigkeit von weniger als 500 Meilen pro Stunde haben. Die Cluster 1 und 3 sowie 5 und 7 sind sich auch hinsichtlich der Reisefluggeschwindigkeit jeweils sehr ähnlich.

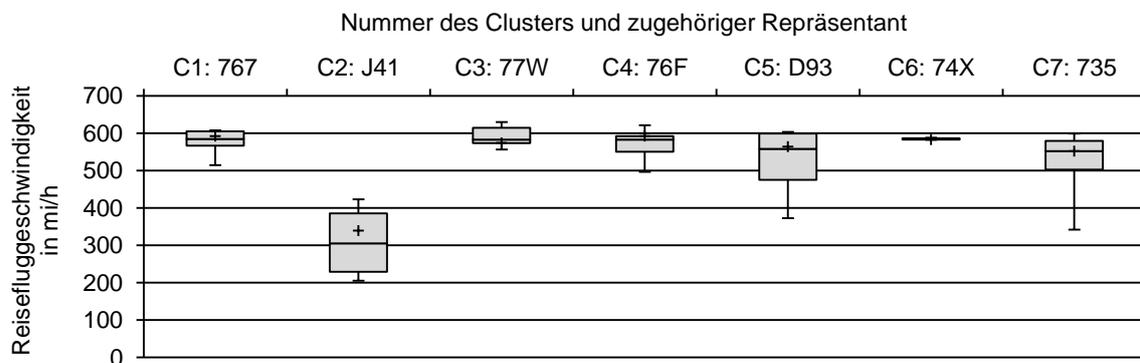


Abbildung 5-5: Box Plot mit der Reisefluggeschwindigkeit aus der OAG-Datenbank für 7 Cluster

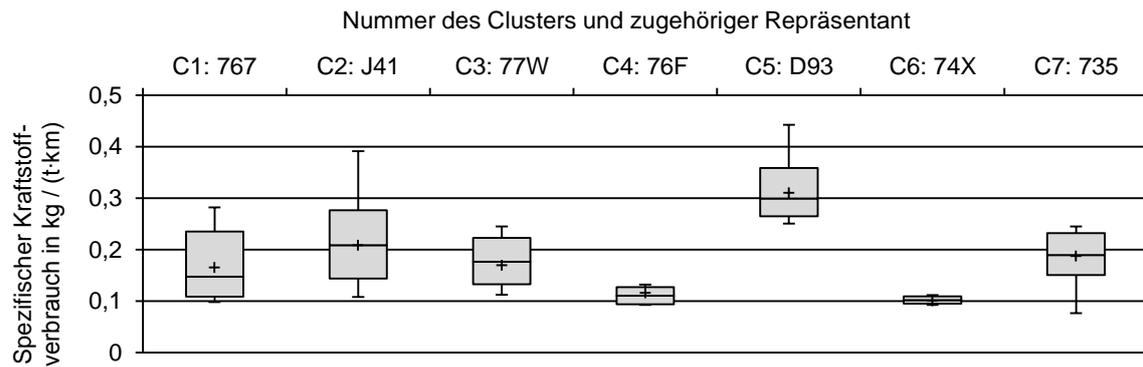


Abbildung 5-6 ist der Box Plot des spezifischen Kraftstoffverbrauchs für die 7-Cluster-Lösung. Lediglich Cluster 5 enthält ausschließlich Flugzeugtypen mit einem relativ hohen spezifischen Kraftstoffverbrauch und unterscheidet sich in diesem Parameter deutlich vom sonst sehr ähnlichen Cluster 7. Die anderen Cluster sind weniger klar voneinander abgegrenzt. Die spezifischen Kraftstoffverbräuche erstrecken sich für diese Cluster über relativ weite Bereiche. Auffällig ist der geringe Kraftstoffverbrauch aller Frachtflugzeuge.

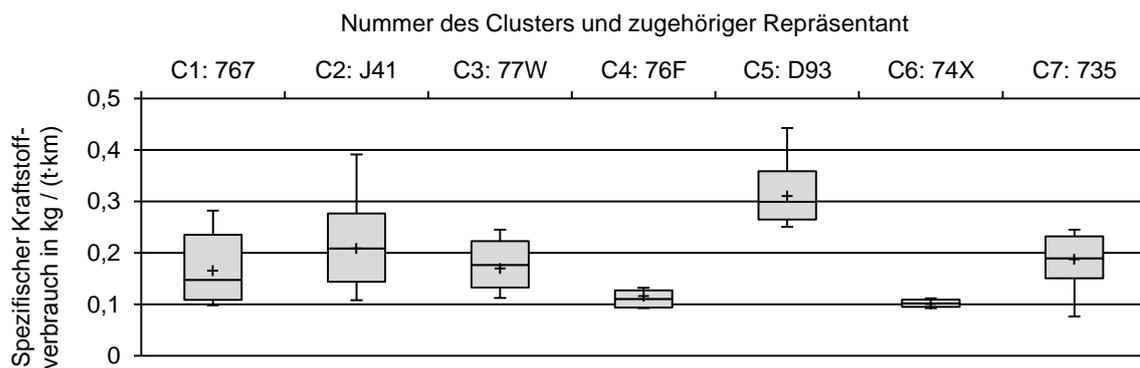


Abbildung 5-6: Box Plot des spezifischen Kraftstoffverbrauchs für 7 Cluster

5.2.3 17-Cluster-Lösung

In diesem Abschnitt wird die Clusterstruktur mit 17 Clustern dargestellt und beschrieben. Wie im vorangegangenen Abschnitt zur 7-Cluster-Lösung erfolgt die Darstellung vor allem mithilfe von Box Plots der in den Clustern enthaltenen Flugzeuge für die verwendeten Clustering-Parameter.

Abbildung 5-7 ist der Box Plot für die Anzahl der verfügbaren Sitzplätze. Es ist erkennbar, dass die Cluster 1, 4 und 6 ausschließlich aus Frachtflugzeugen bestehen. Die Cluster überschneiden sich hinsichtlich der verfügbaren Sitzplätze der enthaltenen Flugzeuge deutlich.

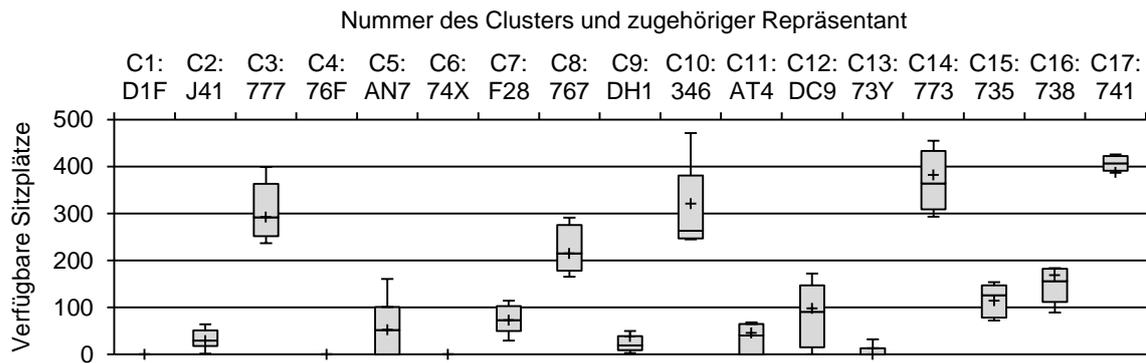


Abbildung 5-7: Box Plot der verfügbaren Anzahl der Sitzplätze für 17 Cluster

Abbildung 5-8 zeigt den Box Plot für die verfügbaren Frachttonnen. Die Cluster der Frachtflugzeuge sowie die Cluster 3 und 14 fallen durch die hohen Frachtkapazitäten der enthaltenen Flugzeugtypen auf. Alle anderen Cluster enthalten vor allem Flugzeuge mit weniger als 20 verfügbaren Frachttonnen. Diese Cluster überschneiden sich hinsichtlich der verfügbaren Frachttonnen zumeist deutlich.

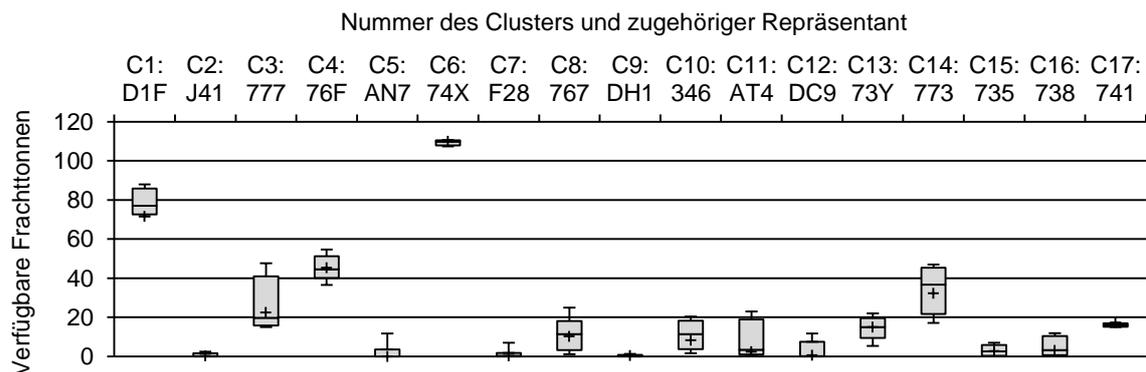


Abbildung 5-8: Box Plot der verfügbaren Frachttonnen für 17 Cluster

Abbildung 5-9 ist der Box Plot der verfügbaren Nutzlast, die sich aus den verfügbaren Sitzplätzen und den verfügbaren Frachttonnen ergibt.

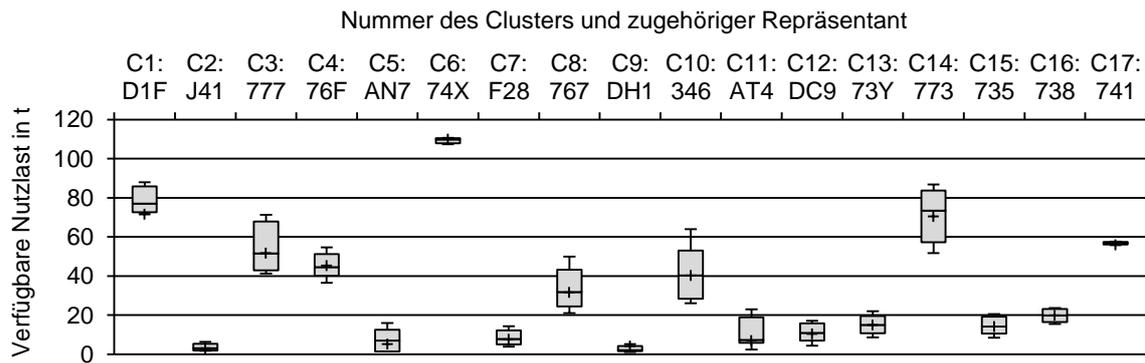


Abbildung 5-9: Box Plot der verfügbaren Nutzlast für 17 Cluster

Abbildung 5-10 zeigt die durchschnittlichen Flugstrecken der Flugzeugtypen in den 17 Clustern. Viele Cluster enthalten nur Flugzeuge mit einer durchschnittlichen Flugstrecke von weniger als 2.000 Kilometern, daher ist die Überschneidung der Cluster in diesem Bereich sehr deutlich.

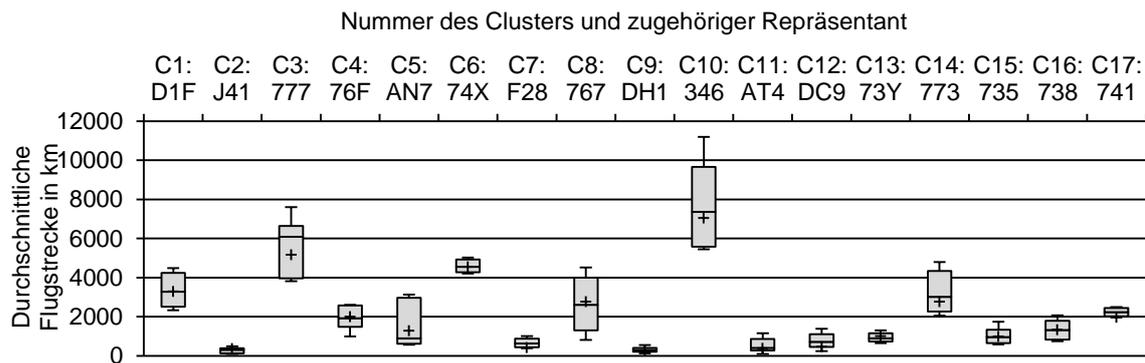


Abbildung 5-10: Box Plot mit der durchschnittlichen Flugstrecke für 17 Cluster

Abbildung 5-11 zeigt den Box Plot für die Reisefluggeschwindigkeit der Flugzeugtypen. Auffällig sind die niedrigen Reisefluggeschwindigkeiten der Turboprop-Flugzeuge. Die Cluster 2, 9 und 10 bestehen aus allen in der Clusteranalyse enthaltenen Turboprop-Flugzeugen.

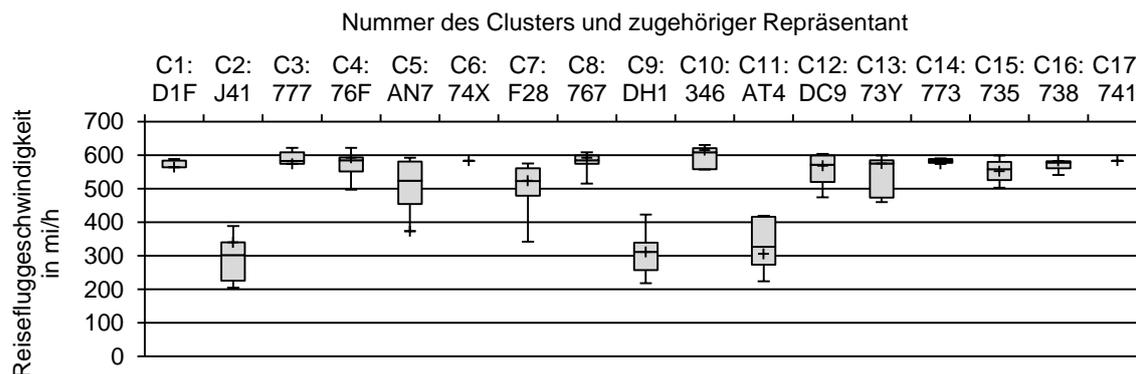


Abbildung 5-11: Box Plot mit der Reisefluggeschwindigkeit aus der OAG-Datenbank für 17 Cluster

Abbildung 5-12 ist der Box Plot des spezifischen Kraftstoffverbrauchs der Flugzeugtypen. Die spezifischen Kraftstoffverbräuche der Flugzeugtypen innerhalb eines Clusters bewegen sich meist in einem relativ engen Bereich, sodass die Cluster relativ gut voneinander abgegrenzt sind. Angesichts der großen Anzahl der Cluster liegen jedoch meist trotzdem mehrere Cluster in ähnlichen Bereichen.

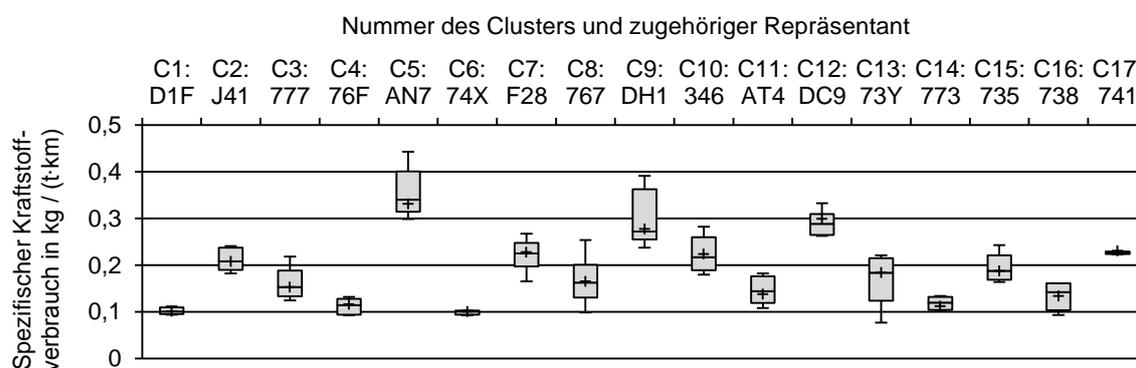


Abbildung 5-12: Box Plot des spezifischen Kraftstoffverbrauchs für 17 Cluster

5.3 Kritische Betrachtung des Clusteranalyse-Tools und der Ergebnisse

In diesem Abschnitt wird das entwickelte Clusteranalyse-Tool hinsichtlich der Erfüllung der gestellten Anforderungen bewertet (Abschnitt 5.3.1) und die Eignung des Vorgehens für die Reduktion der Komplexität bei Betrachtungen zukünftiger Flugzeugflotten bewertet (Abschnitt 5.3.2). Diese Bewertung der Eignung zur Lösung der Problemstellung dieser Arbeit basiert auf den bei der Anwendung des Clusteranalyse-Tools gewonnenen Erkenntnissen.

5.3.1 Erfüllung der an das Clusteranalyseverfahren gestellten Anforderungen

In Abschnitt 3.2 werden die Anforderungen an das in dieser Arbeit zu verwendende Clusteranalyseverfahren definiert. In diesem Abschnitt wird überprüft, inwieweit das auf Basis des ausgewählten Clusteranalyseverfahrens entwickelte Tool zur Clusteranalyse diese Anforderungen erfüllt.

Die für die vorgesehene Anwendung des Clusteranalyse-Tools zwingend zu erfüllenden Muss-Anforderungen aus Abschnitt 3.2.1 werden vollumfänglich erfüllt. Mit dem entwickelten Tool wird eine objektorientierte Clusteranalyse für Flugzeuge durchgeführt. Im Ergebnis wird jedes Flugzeug genau einem Cluster zugeordnet. Die für die Clusteranalyse verwendeten Clustering-Parameter können frei gewählt und definiert werden. Dabei können Variablen aller Skalenniveaus, jeweils mit beliebiger Anzahl verschiedener Ausprägungen, verwendet werden. Die Anzahl der verwendeten Clustering-Parameter ist nicht begrenzt. Ebenso wie die Clustering-Parameter können die verwendeten Objekte flexibel ausgewählt und variiert werden. Zum Einlesen der benötigten Daten greift das Tool auf Datenbankdateien sowie exportierte Dateien aus Datenbanken zu. Aus diesen wird im Rahmen des in VBA implementierten Algorithmus die Unähnlichkeitsmatrix für die analysierten Objekte berechnet.

In Abschnitt 3.2.2 werden die Soll-Anforderungen an das Clusteranalyseverfahren und dessen Umsetzung im Clusteranalyse-Tool definiert. Demnach soll das Verfahren eine sinnvolle Anzahl der Cluster ermitteln. Für die Bewerksstellung dieser Anforderung wurde die Darstellung zweier Maße für die Bewertung der Qualität der ermittelten Clusterstrukturen implementiert: Die Average Silhouette Width und die Abnahme der Summe der Unähnlichkeiten. Die Entscheidung über die favorisierte Anzahl der Cluster wird jedoch dem Nutzer überlassen. In die letztendliche Entscheidung sollte neben der Bewertung der angezeigten Maßzahlen auch eine subjektive Bewertung der Sinnhaftigkeit der Einteilung der Cluster durch den Nutzer eingehen. Möglicherweise ist eine Erweiterung des Tools um zusätzliche Maße zur Bewertung der Qualität der ermittelten Clusterstrukturen sinnvoll, um zu einem noch differenzierteren Bild von der Qualität der Ergebnisse zu gelangen.

Die Ergebnisse der Clusteranalyse werden im Clusteranalyse-Tool sowohl tabellarisch als auch grafisch in Box Plots für die verschiedenen Parameter dargestellt. Die Charakteristika der einzelnen Cluster sowie die Unterschiede zwischen verschiedenen Clustern sind so relativ schnell erkennbar.

Um die Unterschiede zwischen ausgewählten Clustern noch schneller erfassen zu können, könnte eine vergleichende Darstellung weniger ausgewählter Cluster hinsichtlich mehrerer Parameter hilfreich sein. Mithilfe einer multidimensionalen Skalierung könnten zudem weitere Erkenntnisse über die Relationen zwischen den Clustern und den enthaltenen Flugzeugtypen gewonnen werden.

Die Durchführung einer Clusteranalyse ist mit dem entwickelten Tool mit sehr geringem Arbeitsaufwand und ohne detaillierte Kenntnisse der Clusteranalyse möglich. Lediglich für die Auswahl einer Clusterstruktur (Anzahl der Cluster) ist ein gewisses Hintergrundwissen von Vorteil.

Bei identischen Einstellungen und Daten wird das Tool immer identische und damit reproduzierbare Ergebnisse liefern. Über die Abhängigkeit der Qualität der Ergebnisse von den verwendeten Daten lässt sich ohne entsprechende detaillierte Analysen keine Aussage treffen.

5.3.2 Eignung der Clusteranalyse zur Lösung der Problemstellung dieser Arbeit

In diesem Abschnitt wird die Eignung der Clusteranalyse für die Vereinfachung von Flugplänen und für die Reduktion der Komplexität von Flottenbetrachtungen im Allgemeinen bewertet. Die Bewertung basiert vor allem auf den aus der ersten Anwendung des entwickelten Clusteranalyse-Tools zur Vereinfachung von Flugplänen gewonnenen Erkenntnissen. Zunächst werden dafür die Auswirkungen der Einschränkungen bezüglich der Datenverfügbarkeit sowie die Auswirkungen der getroffenen Annahmen in Bezug auf die durchgeführte Anwendung bewertet. Anschließend wird ein allgemeines Fazit zur Eignung der Clusteranalyse für die Komplexitätsreduktion bei Betrachtungen zukünftiger Flottenentwicklungen getroffen. Außerdem wird Potential für mögliche Verbesserungen und Validierungen aufgezeigt.

Aufgrund der mangelnden Verfügbarkeit einiger Flugzeugtypen in der BADA-Datenbank konnten bei der Anwendung der Clusteranalyse nicht alle in der verwendeten OAG-Datenbank enthaltenen Flugzeugtypen verwendet werden. 13 Flugzeugtypen mussten aus diesem Grund ausgeschlossen werden. Die verbleibenden Flugzeugtypen haben jedoch entsprechend den Daten aus der OAG-Datenbank im entsprechenden Zeitraum 99,99% der Transportkapazität für Passagiere und für Fracht angeboten. Die Eignung der Clusteranalyse für die Lösung der Problemstellung wird daher durch die mangelnde Verfügbarkeit von Daten für einige Flugzeugtypen nicht beeinflusst. Der überwiegende Anteil der Transportkapazität wird von nur wenigen Flugzeugen erbracht. Selbst wenn nur die 100 bezüglich der Transportkapazität bedeutendsten Flugzeugtypen berücksichtigt würden, würden noch mehr als 99,9% der Transportkapazitäten für Passagiere und Fracht durch die berücksichtigten Flugzeugtypen abgedeckt werden.

Nachteilig für die Qualität der Ergebnisse ist jedoch die Tatsache, dass auch für die verwendeten Flugzeuge nicht in jedem Fall eigene Daten in der BADA-Datenbank verfügbar waren. Häufig mussten aufgrund der begrenzten Anzahl der in der BADA-Datenbank enthaltenen Flugzeugtypen die gleichen BADA-Flugzeugtypen mehreren OAG-Flugzeugtypen zugeordnet werden. Dadurch werden die Unähnlichkeiten zwischen den

Flugzeugtypen verzerrt. Insbesondere sind die ermittelten Unähnlichkeiten zwischen den betroffenen Flugzeugen geringer als die tatsächlichen Unähnlichkeiten.

Für die Berechnung der verfügbaren Nutzlast wurde angenommen, dass ein verfügbarer Sitzplatz hinsichtlich Nutzlast äquivalent zu 100 kg angebotener Frachtkapazität ist. Diese Annahme beeinflusst direkt, beziehungsweise indirekt die verwendete verfügbare Nutzlast und den verwendeten spezifischen Kraftstoffverbrauch und hat daher großen Einfluss auf die Ergebnisse des Clustering.

Basierend auf der Average Silhouette Width der gefundenen Clusterstrukturen lässt sich feststellen, dass in den verwendeten Daten offenbar keine Clusterstruktur vorhanden ist, was angesichts der Natur der Daten auch nicht zu erwarten war. Allerdings lassen sich bei Betrachtung der Abnahme der Summe der Unähnlichkeiten einige Clusterstrukturen erkennen, die scheinbar eine deutlich bessere Qualität haben als andere Clusterstrukturen.

Für die Sinnhaftigkeit der identifizierten Clusterstrukturen spricht die gute Separation von Flugzeugtypen mit unterschiedlichen Antriebsystemen (Turboprop-Antrieb, Jet-Antrieb) sowie von Frachtflugzeugen und Passagierflugzeugen. Sowohl bei der Clusterlösung mit 7 Clustern als auch bei jener mit 17 Clustern finden sich separate Cluster für Flugzeugtypen mit Turboprop-Antrieb und für Frachtflugzeuge, in denen ausschließlich entsprechende Flugzeuge enthalten sind. Nur wenige Frachtflugzeugtypen wurden Clustern zugeordnet, in denen sonst vorwiegend Passagierflugzeuge enthalten sind.

Hinsichtlich der anderen verwendeten Clustering-Parameter sind die charakteristischen Unterschiede zwischen den Clustern in einigen Fällen nicht intuitiv erkennbar. Dies trifft jedoch vor allem auf die Lösung mit 17 Clustern zu und ist in diesem Fall angesichts der großen Anzahl der Cluster nicht überraschend.

Trotz der offenbar nicht vorhandenen Clusterstruktur in den Daten kann die Clusteranalyse für die in dieser Arbeit verfolgte Reduktion der Komplexität durch eine Reduktion der Anzahl der betrachteten Flugzeugtypen weiterhin sinnvoll sein. Die Bedingung für die Eignung der Clusteranalyse ist, dass die vorgenommene Vereinfachung mit keinem unzulässig großen Fehler in der Abbildung der Realität verbunden ist. Angesichts der bei subjektiver Beurteilung sinnvoll wirkenden Einteilung der Flugzeuge ist ein geringer resultierender Fehler durchaus vorstellbar.

Um die Validität der vorgesehenen Vereinfachung des Flugplans tatsächlich zu beurteilen, scheint die Ermittlung und Analyse der mit der Vereinfachung einhergehenden Fehler hinsichtlich Transportkapazität und Kraftstoffverbrauch sinnvoll.

Um eine hohe Qualität der Ergebnisse sicherzustellen scheint eine Analyse der Plausibilität der verwendeten Daten, insbesondere aus der OAG-Datenbank angebracht. Da die Boeing 737-200 (Mixed Configuration) laut OAG-Datenbank auf ihren Flügen keine Transportleistung anbietet, wurde sie von der Clusteranalyse in dieser Arbeit ausgeschlossen. Mög-

licherweise finden sich neben diesem eindeutigen Fall weitere Anomalien in der OAG-Datenbank, die einer Korrektur bedürfen. Beispielsweise wäre zu untersuchen, inwieweit mit Flugzeugtypen, die in der Datenbank als Frachtflugzeuge bezeichnet werden, auch tatsächlich Frachttransport angeboten wird.

Eine Steigerung der Ergebnisqualität kann möglicherweise durch eine zunächst wenig intuitiv wirkende Reduktion der Anzahl der berücksichtigten Flugzeugtypen erzielt werden. Durch die Streichung von Flugzeugen, die hinsichtlich ihrer Transportkapazität eine unbedeutende Rolle spielen, kann die Qualität der Clusterergebnisse bei gleichbleibender Anzahl der Cluster gesteigert werden, weil die Cluster nicht durch unbedeutenden Flugzeugtypen beeinflusst werden. Der gleiche Effekt kann möglicherweise auch ohne Streichung unbedeutender Flugzeuge erzielt werden, indem die verschiedenen Flugzeugtypen bei der Clusteranalyse nach ihrer Bedeutung (zum Beispiel gemessen an der Transportkapazität) gewichtet werden.

6 Zusammenfassung und Ausblick

In dieser Arbeit wurde ein Tool zur Clusteranalyse für zivile Verkehrsflugzeuge entwickelt und angewendet. Ziel der Clusteranalyse ist die Reduktion der Komplexität von Analysen zur zukünftigen Entwicklung von Flugzeugflotten und ziviler Luftfahrt, die sich durch die große Anzahl verschiedener Flugzeugtypen ergibt. Dazu werden mit dem als Clusteranalyseverfahren ausgewählten Medoid-Verfahren Gruppen von Flugzeugtypen, die sich nach bestimmten Parametern ähnlich sind, identifiziert. Außerdem wird für jedes identifizierte Cluster ein Flugzeugtyp ermittelt, der dieses Cluster am besten repräsentiert (Repräsentant).

Zunächst wurden zur Lösung der Problemstellung der Arbeit verschiedene bestehende Klassifikationen von Flugzeugen identifiziert und untersucht. Anwendungen einer Clusteranalyse konnten dabei nicht gefunden werden. Bisherige Einteilungen von Flugzeugtypen sind relativ einfach und stark auf die zugrundeliegende Zielsetzung ausgerichtet. Aufgrund der beschränkten Anzahl von Klassifikationen von Flugzeugtypen wurden zudem einige Klassifikationen aus dem Bereich der Automobile untersucht, um weitere Erkenntnisse für die Umsetzung eines Clusteranalyse-Tools zu gewinnen.

Nach der Definition der Anforderungen an das in dieser Arbeit verwendete Clusteranalyseverfahren wurden die verschiedenen Gruppen von Clusteranalyseverfahren vorgestellt und geeignete Verfahren identifiziert. Besonders geeignet sind das K-Means-Verfahren und das Medoid-Verfahren. Auf Basis einer Aufstellung der Vor- und Nachteile beider Verfahren wurde das Medoid-Verfahren für die Verwendung in dieser Arbeit ausgewählt. Ausschlaggebend waren dabei vor allem die Bestimmung eines real existierenden Repräsentanten für jedes Cluster und die Reproduzierbarkeit der Ergebnisse.

Das auf Basis des Medoid-Verfahrens entwickelte Tool zur Clusteranalyse liest die benötigten Daten aus verschiedenen relevanten Datenbanken ein und ermittelt die Clusterstrukturen für beliebige Anzahlen von Clustern. Außerdem wird eine Hilfestellung zur Auswahl der Clusteranzahl gegeben und die Ergebnisse werden tabellarisch und grafisch dargestellt.

Das Tool wurde im Rahmen der Arbeit für die Ermittlung einer Clusterstruktur für die Vereinfachung eines OAG-Flugplans angewendet. Im OAG-Flugplan sollen dazu alle Flugzeuge durch die Repräsentanten ihrer Cluster ersetzt werden. Dabei sollen die Transportkapazität sowie der Kraftstoffverbrauch möglichst unverändert bleiben. Für diese Problemstellung wurden die folgenden Clustering-Parameter ausgewählt: Die Anzahl der verfügbaren Sitzplätze, die verfügbaren Frachtkontainer, die verfügbare Nutzlast, die Reisefluggeschwindigkeit, die durchschnittliche Flugstrecke, der spezifische Kraftstoffverbrauch und der Antriebstyp. Die Clusteranalyse wurde mit 185 Flugzeugtypen aus der OAG-Datenbank für Juni 2008 durchgeführt. 13 Flugzeugtypen konnten, primär aufgrund der mangelnden Verfügbarkeit entsprechender Flugzeugdaten in der BADA-Datenbank, nicht

verwendet werden. Die verwendeten Flugzeuge haben entsprechend der OAG-Datenbank 99,99% der Transportleistung für Passagiere sowie für Fracht erbracht.

Die Ergebnisse der Clusteranalyse zeigen, dass den Flugzeugtypen keine Clusterstruktur zugrunde zu liegen scheint. Allerdings lassen sich auf Basis der relativen Abnahme der Summe der Unähnlichkeiten Cluster-Lösungen ermitteln, die eine deutlich höhere Qualität aufweisen als andere Lösungen. Diese Cluster-Lösungen erscheinen bei individueller Betrachtung sinnvoll. Die Flugzeugtypen sind hinsichtlich Antriebstyp und der Art der Transportleistung sehr trennscharf in separate Cluster eingeteilt: Turboprop-Flugzeuge und Frachtflugzeuge sind jeweils in eigenen, teilweise auch mehreren Clustern enthalten.

Die an das implementierte Clusteranalyseverfahren gestellten Anforderungen konnten sehr gut erfüllt werden. Es verbleiben jedoch mehrere Herausforderung im Rahmen einer Clusteranalyse, die durch zukünftige Erweiterungen möglicherweise abgemildert werden können. Die größten Herausforderungen sind die Bestimmung der optimalen Anzahl der Cluster und die Zuordnung der Flugzeugtypen aus den verschiedenen verwendeten Datenbanken.

Um eine geeignete Anzahl von Clustern auszuwählen, wird die Qualität der gefundenen Clusterstrukturen mit zwei Hilfsmitteln absolut sowie relativ zu anderen Clusterstrukturen bewertet. Diese Hilfsmittel können jedoch keine klare Aussage über eine optimale Clusterstruktur treffen. Die Bewertung der Lösungen und Entscheidung für eine Cluster-Lösung muss durch den Nutzer vorgenommen werden. Die Integration weitere Maße zur Bewertung der verschiedenen Cluster-Lösungen könnte dem Nutzer möglicherweise eine noch bessere Hilfestellung bei dieser Entscheidung bieten.

Bei der Zuordnung der Flugzeuge aus verschiedenen Datenbanken besteht die Herausforderung im unterschiedlichen Umfang der enthaltenen Flugzeuge und dem abweichenden Detaillierungsgrad im Hinblick auf spezifische Flugzeugkonfigurationen. Einige Flugzeugtypen können daher nicht berücksichtigt werden. Für andere Flugzeuge müssen teilweise Daten von vergleichbaren Flugzeugtypen verwendet werden.

Die Interpretation der Ergebnisse der Clusteranalyse könnte durch die Implementierung weiterer Darstellungsformen der Ergebnisse weiter vereinfacht werden. Beispielsweise könnten eine vergleichende Darstellung weniger ausgewählter Cluster hinsichtlich mehrerer Parameter oder eine Darstellung mittels multidimensionaler Skalierung hilfreich sein.

Unabhängig davon, ob den verwendeten Daten tatsächlich eine Clusterstruktur zugrunde liegt, sollte die Validität der Verwendung der Clusteranalyse für konkrete Problemstellungen anhand der mit der Vereinfachung einhergehenden Fehler bewertet werden. Im Fall der Anwendung in dieser Arbeit sollten daher die mit der Flottenvereinfachung einhergehenden Fehler hinsichtlich Transportkapazität und Kraftstoffverbrauch untersucht und bewertet werden. Zudem sollte die getroffene Annahme der Nutzlastäquivalenz von einem

Passagier und 100 kg Fracht kritisch bewertet, beziehungsweise validiert werden, da diese Annahme einen deutlichen Einfluss auf die Ergebnisse der Clusteranalyse haben kann.

Die verwendeten Daten sollten angesichts festgestellter Anomalien für spätere Anwendungen intensiv auf ihre Plausibilität geprüft und bei Bedarf gefiltert oder korrigiert werden.

Viele der verwendeten Flugzeugtypen haben nur einen sehr geringen Anteil an der gesamten angebotenen Transportleistung. Aufgrund der Datenverfügbarkeit sind die verwendeten Daten für viele dieser Flugzeugtypen zudem weniger genau als für die bedeutendsten Typen. Um den Einfluss dieser Flugzeugtypen und der geringeren Datenqualität auf die Ergebnisse des Clustering zu begrenzen, sind ein Ausschluss dieser Flugzeugtypen aus der Analyse oder die Entwicklung eines Systems zur Gewichtung der Flugzeugtypen, beispielsweise nach ihrer Transportleistung, im Clusteranalyseverfahren denkbar.

Literaturverzeichnis

ACRISS - Association of Car Rental Industry Systems Standards 2011: Ihr Quick Reference Guide um Mietwagen mit Vertrauen zu buchen, http://www.acriss.org/pdfs/Vehicle_Guide_German_2011_Sep.pdf, abgerufen am 31.05.2012

Air France-KLM S.A. 2012: The Air France KLM fleet, <http://www.airfranceklm.com/en/fleet/long-haul/>, abgerufen am 22.03.2012

ANOTEC Consulting, S.L. 2003: Study on Current and Future Aircraft Noise Exposure at and around Community Airports. Final Report

Bacher, Johann/Pöge, Andreas/Wenzig, Knut 2010: Clusteranalyse. Anwendungsorientierte Einführung in Klassifikationsverfahren, 3. Aufl., Oldenbourg Wissenschaftsverlag, München 2010

Backhaus, Klaus/Erichson, Bernd/Plinke, Wulff/Weiber, Rolf 2011: Multivariate Analysemethoden. Eine anwendungsorientierte Einführung, 13. Aufl., Springer Verlag, Berlin 2011

Deutsche Lufthansa AG 2012: Flotte, <http://konzern.lufthansa.com/flotte.html>, abgerufen am 22.03.2012

Eyers, C. J. et al. 2004: AERO2k Global Aviation Emissions Inventories for 2002 and 2025, http://aero-net.info/fileadmin/aeronet_files/links/documents/AERO2K_Global_Aviation_Emissions_Inventories_for_2002_and_2025.pdf

Federal Aviation Administration 2011a: Air Traffic Control. Order JO 7110.65U, <http://www.faa.gov/documentLibrary/media/Order/ATC.pdf>, abgerufen am 26.03.2012

Federal Aviation Administration 2011b: Interim Procedures for Airbus A388 Flights. Notice N JO 7110.567, <http://www.faa.gov/documentLibrary/media/Notice/N7110.541.pdf>, abgerufen am 26.03.2012

IHS Global Insight 2009: European Light Vehicle Industry Forecast Report, hausinterne Studie

International Civil Aviation Organization 2006: Wake Vortex Aspects of A380 aircraft

International Civil Aviation Organization 2007: Air Traffic Management. Doc 4444

Kaufman, Leonard/Rousseeuw, Peter J. 1990: Finding Groups in Data. An Introduction to Cluster Analysis, Verlag John Wiley & Sons, New York 1990

Tan, Pang-Ning/Steinbach, Michael/Kumar, Vipin 2005: Introduction to data mining, 1. Aufl., Verlag Addison Wesley, Boston 2005

United Continental Holdings, Inc. 2012: United Continental Holdings, Inc. Fleet Plan, <http://phx.corporate-ir.net/External.File?item=UGFyZW50SUQ9MTI0MTU2fENoaWxkSUQ9LTF8VHlwZT0z&t=1>, abgerufen am 22.03.2012

Anlagen

Anlage 1 (Zuordnung der Flugzeuge aus OAG- und BADA-Datenbank)

OAG Specific Aircraft	OAG Specific Aircraft Name	BADA File Name	Kommentar zur BADA Zuordnung
ND2	Aerospatale (nord) 262	JS41__	Approximiert durch BADA-Äquivalent Jetstream 41
ABF	Airbus A300 (Freighter)	A30B__	Approximiert durch A300B4-200
AB3	Airbus A300 Passenger	A30B__	Approximiert durch A300B4-200
ABY	Airbus A300-600 Freighter	A306__	Approximiert durch A300-600
AB6	Airbus A300-600 Passenger	A306__	In BADA enthalten
AB4	Airbus A300B2 /B4 Passenger	A30B__	Approximiert durch A300B4-200
ABX	Airbus A300B4 /A300C4 /A300F4	A30B__	Approximiert durch A300B4-200
31F	Airbus A310 Freighter	A310__	Approximiert durch A310
310	Airbus A310 Passenger	A310__	In BADA enthalten
312	Airbus A310-200 Passenger	A310__	Approximiert durch A310
31Y	Airbus A310-300 Freighter	A310__	Approximiert durch A310
313	Airbus A310-300 Passenger	A310__	Approximiert durch A310
318	Airbus A318	A318__	In BADA enthalten
32S	Airbus A318 /319 /320 /321	A320__	Approximiert durch A320
319	Airbus A319	A319__	In BADA enthalten
320	Airbus A320	A320__	In BADA enthalten
321	Airbus A321	A321__	In BADA enthalten
330	Airbus A330	A332__	Approximiert durch A330-200
332	Airbus A330-200	A332__	In BADA enthalten
333	Airbus A330-300	A333__	In BADA enthalten
340	Airbus A340	A343__	Approximiert durch A340-300
342	Airbus A340-200	A342__	In BADA enthalten
343	Airbus A340-300	A343__	In BADA enthalten
345	Airbus A340-500	A345__	In BADA enthalten
346	Airbus A340-600	A346__	In BADA enthalten
380	Airbus A380-800 Passenger	A388__	In BADA enthalten
ANF	Antonov AN-12	C130__	Approximiert durch BADA-Äquivalent Lockheed Hercules
A40	Antonov AN-140		Keine Entsprechung in der BADA-Datenbank
AN4	Antonov AN-24	F27__	Approximiert durch BADA-Äquivalent Fokker Friendship
AN6	Antonov AN-26 /30 /32	E120__	Approximiert durch BADA-Äquivalent EMB-120 Brasilia

OAG Specific Aircraft	OAG Specific Aircraft Name	BADA File Name	Kommentar zur BADA Zuordnung
A28	Antonov An28 /Pzl Mielec M-28 Skytruck	D228__	In BADA enthalten
AN7	Antonov AN-72 /74	F28__	Approximiert durch BADA-Äquivalent Fokker Followship
AT4	ATR 42-300 /320	AT43__	In BADA enthalten
AT5	ATR 42-500	AT45__	In BADA enthalten
AT7	ATR 72	AT73__	Approximiert durch ATR 72-210
ATR	ATR42 /ATR72	AT45__	Approximiert durch ATR 42-500
AR1	Avro RJ100	RJ1H__	In BADA enthalten
AR7	Avro RJ70	RJ85__	Approximiert durch BADA-Äquivalent Avro RJ85
AR8	Avro RJ85	RJ85__	In BADA enthalten
B11	BAe (BAC) 1-11	BA11__	In BADA enthalten
HS7	BAe (HS) 748	ATP__	Approximiert durch Avro 748
14F	BAe 146 (Freighter)	B462__	Approximiert durch BAe 146-200
146	BAe 146 Passenger	B462__	Approximiert durch BAe 146-200
141	BAe 146-100 Passenger	B462__	Approximiert durch BADA-Äquivalent BAe 146-200
142	BAe 146-200 Passenger	B462__	In BADA enthalten
143	BAe 146-300 Passenger	B463__	In BADA enthalten
ATP	BAe ATP	ATP__	In BADA enthalten
JST	BAe Jetstream	JS32__	Approximiert durch Jetstream Super 31
J31	BAe Jetstream 31	JS32__	Approximiert durch BADA-Äquivalent Jetstream Super 31
J32	BAe Jetstream 32	JS32__	In BADA enthalten
J41	BAe Jetstream 41	JS41__	In BADA enthalten
BE1	Beechcraft 1900 Airliner	JS32__	Approximiert durch BADA-Äquivalent Jetstream Super 31
BES	Beechcraft 1900C Airliner	JS32__	Approximiert durch BADA-Äquivalent Jetstream Super 31
BEH	Beechcraft 1900D Airliner	JS32__	Approximiert durch BADA-Äquivalent Jetstream Super 31
BE9	Beechcraft C99 Airliner	BE99__	In BADA enthalten
D1Y	Boeing (douglas) DC10 - 30 /40 Freichte	DC10__	Approximiert durch DC10
D1F	Boeing (douglas) DC10 (Freighter)	DC10__	Approximiert durch DC10
D1C	Boeing (douglas) DC10-30 /40 (Pax)	DC10__	Approximiert durch DC10
D8F	Boeing (douglas) DC8 Freighter	DC87__	Approximiert durch DC8-70
DC8	Boeing (Douglas) DC-8 Passenger	DC87__	Approximiert durch DC8-70
D8Y	Boeing (douglas) DC-8-71 /72 /73 Frtr	DC87__	Approximiert durch DC8-70
D9F	Boeing (douglas) DC9 Freighter	DC94__	Approximiert durch DC9-40

OAG Specific Aircraft	OAG Specific Aircraft Name	BADA File Name	Kommentar zur BADA Zuordnung
DC9	Boeing (douglas) DC-9 Passenger	DC94__	Approximiert durch DC9-40
D93	Boeing (douglas) DC9-30 Passenger	DC94__	Approximiert durch DC9-40
D95	Boeing (douglas) DC9-50 Passenger	DC94__	Approximiert durch DC9-40
M1F	Boeing (douglas) MD-11 (Freighter)	MD11__	Approximiert durch MD-11
M11	Boeing (douglas) MD-11 Passenger	MD11__	In BADA enthalten
M80	Boeing (douglas) MD-80	MD82__	Approximiert durch MD-82
M81	Boeing (douglas) MD-81	MD82__	Approximiert durch BADA-Äquivalent MD-82
M82	Boeing (douglas) MD-82	MD82__	In BADA enthalten
M83	Boeing (douglas) MD-83	MD83__	In BADA enthalten
M87	Boeing (douglas) MD-87	MD82__	Approximiert durch BADA-Äquivalent MD-82
M88	Boeing (douglas) MD-88	MD82__	Approximiert durch BADA-Äquivalent MD-82
M90	Boeing (douglas) MD-90	MD83__	Approximiert durch BADA-Äquivalent MD-83
707	Boeing 707 /720 Passenger	B752__	Approximiert durch BADA-Äquivalent B757-200 für die B707-100
717	Boeing 717-200	B712__	In BADA enthalten
72F	Boeing 727 (Freighter)	B722__	Approximiert durch B727-200
72M	Boeing 727 (Mixed Configuration)	B722__	Approximiert durch B727-200
727	Boeing 727 (Passenger)	B722__	Approximiert durch B727-200
721	Boeing 727-100 (Passenger)	B752__	Approximiert durch BADA-Äquivalent B757-200
72S	Boeing 727-200 (Pax)	B722__	In BADA enthalten
722	Boeing 727-200 Passenger	B722__	In BADA enthalten
73F	Boeing 737 (Freighter)	B733__	Approximiert durch B737-300
73M	Boeing 737 (Mixed Configuration)	B737__	Approximiert durch B737-700
73S	Boeing 737 Advanced all Series	B732__	Approximiert durch B737-200
737	Boeing 737 Passenger	B738__	Approximiert durch B737-800
73X	Boeing 737-200 (Freighter)	B732__	Approximiert durch B737-200
73L	Boeing 737-200 (Mixed Configuration)	B732__	Approximiert durch B737-200
732	Boeing 737-200 Passenger	B732__	In BADA enthalten
73Y	Boeing 737-300 (Freighter)	B733__	Approximiert durch B737-300

OAG Specific Aircraft	OAG Specific Aircraft Name	BADA File Name	Kommentar zur BADA Zuordnung
73C	Boeing 737-300 (wing-lets) Passenger	B733__	Approximiert durch B737-300
733	Boeing 737-300 Passen-ger	B733__	In BADA enthalten
73P	Boeing 737-400 (Freigh-ter)	B734__	Approximiert durch B737-400
73Q	Boeing 737-400 (Mixed Configuration)	B734__	Approximiert durch B737-400
734	Boeing 737-400 Passen-ger	B734__	In BADA enthalten
735	Boeing 737-500 Passen-ger	B735__	In BADA enthalten
736	Boeing 737-600 Passen-ger	B736__	In BADA enthalten
73W	Boeing 737-700 (wing-lets) Passenger	B737__	Approximiert durch B737-700
73G	Boeing 737-700 Passen-ger	B737__	In BADA enthalten
73H	Boeing 737-800 (wing-lets) Passenger	B738__	Approximiert durch B737-800
738	Boeing 737-800 Passen-ger	B738__	In BADA enthalten
739	Boeing 737-900 Passen-ger	B739__	In BADA enthalten
74F	Boeing 747 (Freighter)	B742__	Approximiert durch B747-200
74M	Boeing 747 (Mixed Con-figuration)	B744__	Approximiert durch B747-400
747	Boeing 747 (Passenger)	B744__	Approximiert durch B747-400
741	Boeing 747-100 (Passen-ger)	B743__	Approximiert durch BADA-Äquivalent B747-300
74X	Boeing 747-200 (Freigh-ter)	B742__	Approximiert durch B747-200
742	Boeing 747-200 (Passen-ger)	B742__	In BADA enthalten
743	Boeing 747-300 /747-100 /200 Sud (Pax)	B743__	In BADA enthalten
74D	Boeing 747-300 /747-200 Sud (Mxd Config	B743__	Approximiert durch B747-300
74E	Boeing 747-400 (Mixed Configuration)	B744__	Approximiert durch B747-400
744	Boeing 747-400 (Passen-ger)	B744__	In BADA enthalten
74Y	Boeing 747-400F (Freigh-ter)	B744__	Approximiert durch B747-400
74L	Boeing 747SP Passenger	B742__	Approximiert durch BADA-Äquivalent B747-200
757	Boeing 757 (Passenger)	B753__	Approximiert durch B757-300

OAG Specific Aircraft	OAG Specific Aircraft Name	BADA File Name	Kommentar zur BADA Zuordnung
75W	Boeing 757-200 (wing-lets) Passenger	B752__	Approximiert durch B757-200
752	Boeing 757-200 Passenger	B752__	In BADA enthalten
75F	Boeing 757-200PF (Freighter)	B752__	Approximiert durch B757-200
753	Boeing 757-300 Passenger	B753__	In BADA enthalten
76F	Boeing 767 Freighter	B762__	Approximiert durch B767-200
767	Boeing 767 Passenger	B763__	Approximiert durch B767-300
762	Boeing 767-200 Passenger	B762__	In BADA enthalten
76Y	Boeing 767-300 Freighter	B763__	Approximiert durch B767-300
763	Boeing 767-300 Passenger	B763__	In BADA enthalten
764	Boeing 767-400 Passenger	B764__	In BADA enthalten
777	Boeing 777 Passenger	B773__	Approximiert durch B777-300
772	Boeing 777-200 Passenger	B772__	Approximiert durch B777-200 ER
77L	Boeing 777-200LR	B77L__	Approximiert durch B777-200 ER
773	Boeing 777-300 Passenger	B773__	In BADA enthalten
77W	Boeing 777-300ER Passenger	B77W__	In BADA enthalten
CRA	Canadair Crj Series 705	CRJ9__	Approximiert durch CRJ-900
CRJ	Canadair Regional Jet	CRJ2__	Approximiert durch CRJ-200
CR1	Canadair Regional Jet 100	CRJ1__	In BADA enthalten
CR2	Canadair Regional Jet 200	CRJ2__	In BADA enthalten
CR7	Canadair Regional Jet 700	CRJ9__	Approximiert durch BADA-Äquivalent CRJ-900
CR9	Canadair Regional Jet 900	CRJ9__	In BADA enthalten
CS5	CASA /Nusantara CN-235	AT43__	Approximiert durch BADA-Äquivalent ATR 42-300
CS2	CASA C212 /Nusantara NC-212 Aviocar	D228__	Approximiert durch BADA-Äquivalent Dornier 228
CNC	Cessna Light Aircraft(Single Turboprop)		Keine Entsprechung in der BADA-Datenbank
CVF	Convair 440 /580 /600 /640 (Freighter)		Keine Entsprechung in der BADA-Datenbank
CV5	Convair 580 Passenger		Keine Entsprechung in der BADA-Datenbank
DHS	De Havilland DHC 3 Otter		Keine Entsprechung in der BADA-Datenbank
DHB	De Havilland DHC2 Beaver/Turbo Beaver		Keine Entsprechung in der BADA-Datenbank

OAG Specific Aircraft	OAG Specific Aircraft Name	BADA File Name	Kommentar zur BADA Zuordnung
DHL	De Havilland DHC-3 Turbo Otter		Keine Entsprechung in der BADA-Datenbank
DHT	De Havilland DHC-6 Twin Otter	D228__	Approximiert durch BADA-Äquivalent Dornier 228
DH7	De Havilland DHC-7 Dash 7		Keine Entsprechung in der BADA-Datenbank
DH8	De Havilland DHC-8 Dash 8	DH8A__	Approximiert durch DHC 8-100
DH1	De Havilland DHC-8-100 Dash 8 /8q	DH8A__	In BADA enthalten
DH2	De Havilland DHC-8-200 Dash 8 /8q	DH8C__	Approximiert durch DHC 8-300
DH3	De Havilland DHC-8-300 Dash 8 /8q	DH8C__	In BADA enthalten
EMB	Embraer 110 Bandeirante	SW4__	Approximiert durch BADA-Äquivalent Swearingen Merlin IV
EM2	Embraer 120 Brasilia	E120__	In BADA enthalten
E70	Embraer 170	E170__	Approximiert durch Embraer 175
EMJ	Embraer 170 /195	E170__	Approximiert durch Embraer 170
E75	Embraer 175	E170__	In BADA enthalten
E90	Embraer 190	E190__	In BADA enthalten
E95	Embraer 195	E190__	Approximiert durch Embraer 190
ERJ	Embraer RJ 135 /140 /145	E145__	Approximiert durch ERJ 145
ER3	Embraer RJ135	E135__	In BADA enthalten
ERD	Embraer RJ140	E145__	Approximiert durch ERJ 145
ER4	Embraer RJ145	E145__	In BADA enthalten
D28	Fairchild Dornier 228	D228__	In BADA enthalten
D38	Fairchild Dornier 328-100	D328__	In BADA enthalten
FRJ	Fairchild Dornier 328jet	D328__	In BADA enthalten
SWM	Fairchild Sa26 /Sa226 /Sa277 Merlin /Me	SW4__	In BADA enthalten
100	Fokker 100	F100__	In BADA enthalten
F50	Fokker 50	F50__	In BADA enthalten
F70	Fokker 70	F70__	In BADA enthalten
F27	Fokker F27 Friendship /Fairchild F27	F27__	In BADA enthalten
F28	Fokker F28 Fellowship	F28__	In BADA enthalten
F24	Fokker F28-4000 Fellowship	F28__	Approximiert durch Fokker Fellowship
YN2	Harbin Yunshuji Y12		Keine Entsprechung in der BADA-Datenbank
I14	Ilyushin IL114		Keine Entsprechung in der BADA-Datenbank
IL8	Ilyushin IL-18	C130__	Approximiert durch BADA-Äquivalent Lockheed Hercules
IL6	Ilyushin Il-62	A30B__	Approximiert durch BADA-Äquivalent A300B4-200

OAG Specific Aircraft	OAG Specific Aircraft Name	BADA File Name	Kommentar zur BADA Zuordnung
IL7	Ilyushin IL-76	B764__	Approximiert durch BADA-Äquivalent B767-400
ILW	Ilyushin IL-86	B763__	Approximiert durch BADA-Äquivalent B767-300
IL9	Ilyushin Il-96 Passenger	A342__	Approximiert durch BADA-Äquivalent A340-200
WWP	Israel Aircraft Ind.1124 Westwind	FA10__	Approximiert durch BADA-Äquivalent Dassault Falcon 10
LRJ	Learjet	LJ45__	In BADA enthalten
L4T	Let 410	D228__	Approximiert durch BADA-Äquivalent Dornier 228
L1F	Lockheed L1011 Tristar Freighter	L101__	Approximiert durch Lockheed L1011
L10	Lockheed L1011 Tristar Passenger	L101__	In BADA enthalten
LOH	Lockheed L182 /1282 /1382 (L100) Hercul	C130__	In BADA enthalten
LOF	Lockheed L188 Electra (Freighter)	C160__	Approximiert durch BADA-Äquivalent Transall C160
D9S	McD-Douglas DC9 30 /40 /50	DC94__	Approximiert durch DC9-40
YS1	NAMC YS-11		Keine Entsprechung in der BADA-Datenbank
PL2	Pilatus PC-12	BE9L__	Approximiert durch BADA-Äquivalent Beech King Air 90
S20	Saab 2000	SB20__	In BADA enthalten
SF3	Saab 340	SF34__	In BADA enthalten
SH3	Shorts 330 (sd3-30)	SH36__	Approximiert durch BADA-Äquivalent Shorts 360
SH6	Shorts 360 (sd3-60)	SH36__	In BADA enthalten
TU3	Tupolev TU134	T134__	In BADA enthalten
TU5	Tupolev TU154	T154__	In BADA enthalten
T20	Tupolev TU-204 /tu-214	T154__	Approximiert durch BADA-Äquivalent TU 154M für die TU 204
YN7	Xian Yunshuji Y7 /Ma60		Keine Entsprechung in der BADA-Datenbank
YK4	Yakovlev Yak-40	E120__	Approximiert durch BADA-Äquivalent EMB-120 Brasilia
YK2	Yakovlev Yak-42 /142	DC94__	Approximiert durch BADA-Äquivalent DC9-40 für YAK-42

Anlage 2 (Eidesstattliche Versicherung)

Ich versichere, dass ich diese Diplomarbeit selbstständig und nur unter Verwendung der angegebenen Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen.

Ort und Datum

Unterschrift