

*Bachelorarbeit*

# *Visualisierung und Vergleich der Clusterverfahren anhand von QEBS-Daten*



zur Erlangung des Grades  
**Bachelor of Science**

**von Sophia Hendriks**

(Matrikelnummer: 182984)

Studiengang Statistik

eingereicht bei **Prof. Dr. Wolfgang Härdle**

Juni 2007



Humboldt Universität zu Berlin  
Wirtschaftswissenschaftliche Fakultät  
Institut für Statistik und Ökonometrie

## **Eigenständigkeitserklärung:**

Hiermit versichere ich, die vorliegende Arbeit "Visualisierung und Vergleich der Clusterverfahren anhand von QEBS-Daten" eigenständig verfasst und alle verwendeten Hilfsmittel und Quellen angegeben zu haben.

# Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>1</b>
<b>2</b>	<b>Grundlagen der Clusteranalyse</b>	<b>3</b>
2.1	Proximitätsmaße . . . . .	3
2.1.1	Binäre Daten . . . . .	4
2.1.2	Metrische Daten . . . . .	7
2.2	Clusterverfahren . . . . .	8
2.2.1	Hierarchische Klassifikationsverfahren . . . . .	9
2.2.2	Partitionierende Verfahren . . . . .	12
<b>3</b>	<b>Methodik der Datenanalyse</b>	<b>13</b>
3.1	Idee . . . . .	13
3.2	Koeffizienten zur Beurteilung von Clusterstrukturen . . . . .	15
3.2.1	Kappa-Koeffizient . . . . .	15
3.2.2	Kophenetischer Korrelationskoeffizient . . . . .	16
<b>4</b>	<b>Vergleich der Clusterverfahren anhand von Kontingenzta- bellen</b>	<b>17</b>
4.1	Vergleich der Distanzmaße . . . . .	18
4.1.1	Tanimoto und Dice-Koeffizient . . . . .	18
4.1.2	Euklidische Distanz und City-Block-Metrik . . . . .	20
4.1.3	Ordinale Distanzmaße und metrische Distanzmaße . . . . .	20
4.1.4	Entwicklung der Kappa-Koeffizienten in Abhängigkeit von der Clu- sterzahl . . . . .	21
4.2	Vergleich der Algorithmen . . . . .	25
<b>5</b>	<b>Analyse der Clusterverfahren anhand der kophenetischen Korrelationskoeff- fizienten</b>	<b>29</b>
5.1	Vergleich der Distanzmaße . . . . .	29

## *Inhaltsverzeichnis*

5.1.1	Bemerkungen . . . . .	33
5.2	Vergleich der Verfahren . . . . .	34
<b>6</b>	<b>Interpretation der Clusterstrukturen</b>	<b>36</b>
6.1	Ergebnisse . . . . .	37
6.1.1	Single-Linkage . . . . .	37
6.1.2	Complete Linkage . . . . .	38
6.1.3	Average Linkage . . . . .	38
<b>7</b>	<b>Vergleich mit Ergebnissen einer Faktorenanalyse</b>	<b>41</b>
<b>8</b>	<b>Zusammenfassung</b>	<b>46</b>
<b>9</b>	<b>Literatur</b>	<b>48</b>
<b>A</b>	<b>Verzeichnis der Dateien</b>	<b>49</b>
A.1	Datensätze . . . . .	49
A.2	Clusterzuordnungen und Kontingenztabelle . . . . .	49
A.3	Faktorenanalyse und Vergleich mit Clusterlösungen . . . . .	50
A.4	Graphiken . . . . .	51
A.5	Sonstiges . . . . .	52

# Abbildungsverzeichnis

3.1	Formatierung von Kontingenztabelle . . . . .	14
4.1	Tanimoto- und Dice-Koeffizient in Abhängigkeit von der Anzahl der positiven Übereinstimmungen . . . . .	18
4.2	Dendrogramme für Average-Linkage-Verfahren unter Tanimoto und Dice . . . . .	19
4.3	Vergleich von Tanimoto und Dice mittels Kontingenztabelle bei Anwendung des Average-Linkage-Verfahrens . . . . .	19
4.4	Dendrogramme für das Single-Linkage-Verfahren unter Euklidischer Distanz und City-Block-Metrik . . . . .	21
4.5	Entwicklung des Kappa-Koeffizienten in Abhängigkeit von der Clusterzahl unter Anwendung von Average-Linkage . . . . .	22
4.6	Entwicklung des Kappa-Koeffizienten in Abhängigkeit von der Clusterzahl unter Anwendung von Complete-Linkage . . . . .	23
4.7	Entwicklung des Kappa-Koeffizienten in Abhängigkeit von der Clusterzahl unter Anwendung von Single-Linkage . . . . .	24
4.8	Paarweiser Vergleich der Clusteralgorithmen unter Verwendung von Tanimoto mittels Kontingenztabelle . . . . .	26
4.9	Dendrogramme für das Single-Linkage-Verfahren unter Verwendung von Tanimoto und Euklidischer Distanz . . . . .	27
4.10	Dendrogramme für das Complete-Linkage- und Average-Linkage-Verfahren unter Tanimoto-Koeffizienten . . . . .	27
4.11	Entwicklung des Kappa-Koeffizienten in Abhängigkeit von der Clusterzahl unter Anwendung von Tanimoto und Euklidischer Distanz . . . . .	28
5.1	Kophenetische Korrelationen . . . . .	30
5.2	Streudiagramme der kophenetischen Distanzen bei Anwendung des Average-Linkage-Verfahrens auf Basis des Tanimoto-Koeffizienten und des Dice-Koeffizienten . . . . .	31

## Abbildungsverzeichnis

5.3	Streudiagramme der kophenetischen Distanzen bei Anwendung des Average-Linkage-Verfahrens auf Basis der Euklidischen Distanz und der City-Block-Metrik . . . . .	32
5.4	Vergleich der Distanzmaße in einer Scatterplotmatrix . . . . .	33
5.5	Streudiagramme der kophenetischen Distanzen auf Basis von Tanimoto . . . . .	34
5.6	MDS . . . . .	35
7.1	Screeplot für den Datensatz "impute1" . . . . .	42
7.2	Faktorladungen der "gemeinsamen Gruppen" . . . . .	44

# Tabellenverzeichnis

5.1	Statistiken verschiedener Distanzmaße . . . . .	30
6.1	Clusterzuordnungen unter Single-Linkage-Verfahren . . . . .	37
6.2	Clusterzuordnungen unter Complete-Linkage-Verfahren . . . . .	39
6.3	Clusterzuordnungen unter Average-Linkage-Verfahren . . . . .	39
7.1	Einteilung der Variablen in Faktoren . . . . .	43
7.2	gemeinsame Gruppen . . . . .	44

# 1 Einleitung

Gegenstand dieser Arbeit ist die nähere Betrachtung und Analyse verschiedener (hierarchischer) Clusterverfahren sowie insbesondere der Vergleich unterschiedlicher Distanzmaße.

Der zugrunde liegende Originaldatensatz stützt sich dabei auf eine im Sommer 2006 durchgeführte Lehrerbefragung bezüglich der Schulprogrammarbeit und Evaluation an berufsbildenden Schulen im Rahmen des Berliner Modellprojektes "Qualitätsentwicklung in den Berufsschulen" (QEBS). Die Konzeption und Auswertung der zugrunde liegenden Fragebögen erfolgte durch das Institut für Erziehungswissenschaften der Humboldt Universität Berlin.

Die nachfolgenden Analysen basieren auf einem Teildatensatz bestehend aus 67 Variablen mit ordinalem Skalenniveau und 862 Beobachtungen. Die Befragten (Lehrer an Berufsschulen) sollten dabei auf einer Skala von 1 ("trifft gar nicht zu") bis 6 ("trifft völlig zu") Aussagen unter anderem zu innerschulischer Organisation, Evaluation und Arbeitsklima treffen sowie den Einfluss Vorgesetzter beurteilen.

Die Variablen sind in 4 Gruppen eingeteilt:

f1001-f1018: Konstatierungen bezüglich des eigenen Fachbereiches

f1201-f1220: Konstatierungen bezüglich des eigenen Fachbereiches

f1401-f1419: Konstatierungen bezüglich des eigenen Fachbereiches

f1601-f1610: Konstatierungen bezüglich der übergeordneten Ebene

Der vollständige Fragebogen ist dem Anhang A.1 entnehmbar.

## **Behandlung fehlender Werte**

Aufgrund der Vielzahl fehlender Werte im Datensatz wurde das Verfahren der MRI

## 1 Einleitung

(Multiple Random Imputation) angewendet. Dieses Verfahren basiert auf der Generierung von Datensätzen ohne fehlende Werte. Da die Ersetzung dieser zufällig erfolgt, werden mehrere Datensätze imputiert und zur endgültigen Auswertung herangezogen. Zu den Grundlagen der MRI sei an dieser Stelle auf *Schafer (1997)*<sup>1</sup> verwiesen. Für die Analysen der vorliegenden Arbeit wurden mir 5 imputierte Datensätze ("impute1" bis "impute5") zur Verfügung gestellt.

Die Tatsache, dass zu Interpretationszwecken sämtliche Analysen mit allen generierten Datensätzen durchgeführt werden müssen und die anschließende Auswertung stets einen "Kompromiss" zwischen den Ergebnissen der Einzelanalysen darstellt, hat zur Folge, dass die Verfahren nicht mehr unmittelbar miteinander vergleichbar sind. Aus diesem Grund werden die Untersuchungen, die sich ausschließlich auf den Vergleich von *Strukturunterschieden* (ohne Intention einer inhaltlichen Deutung) in den Ergebnissen der Algorithmen beziehen, nur anhand *eines* Datensatzes durchgeführt.

Dies betrifft die Analysen in Kapitel 4 und Kapitel 5. Alle dort getroffenen Aussagen beziehen sich auf den Datensatz "impute1". In Kapitel 4 werden die Algorithmen paarweise anhand von Kontingenztabelle verglichen, Kapitel 5 beinhaltet die Analysen der Algorithmen auf Basis des kophenetischen Korrelationskoeffizienten (s. Kapitel 3.2.2). Auf die genaue Methodik (und Problematik) dieser Untersuchungen wird in Kapitel 3 eingegangen.

Eine inhaltliche Interpretation der beobachteten Clusterstrukturen sowie der Vergleich mit Ergebnissen einer zuvor durchgeführten Faktorenanalyse finden sich in den Kapiteln 6 und 7. Die dort aufgeführten Ergebnisse beziehen sich -wenn nicht anders angemerkt- auf die "Synthese" der 5 imputierten Datensätze.

Kapitel 8 liefert schließlich eine Zusammenfassung der Hauptergebnisse.

Der Anhang dieser Arbeit befindet sich in einer beigefügten CD. Diese beinhaltet alle Datensätze, sämtliche Graphiken, SPSS-Outputs, verwendeten Matlab-Funktionen und aufgestellte Kontingenztabelle. Unter Anhang A ist die Auflistung aller Dateien aufgeführt.

---

<sup>1</sup>J.L.Schafer: *Analysis of incomplete Multivariate Data*, Chapman and Hall (1997)

## 2 Grundlagen der Clusteranalyse

Die Clusteranalyse gehört zu den strukturentdeckenden Verfahren. Sie dient der Aufteilung gegebener Objekte in verschiedene Gruppen mit dem Ziel, dass diese Gruppen in sich möglichst homogen (ähnlich) und untereinander möglichst heterogen (unähnlich) sind.

Eine Clusteranalyse gliedert sich in drei Ablaufschritte <sup>1</sup>:

- 1 Bestimmung der Distanz zwischen den einzelnen Variablen
- 2 Wahl eines geeigneten Fusionierungsalgorithmus
- 3 Bestimmung der optimalen Clusteranzahl

### 2.1 Proximitätsmaße

Zur Bestimmung der Ähnlichkeit bzw. Distanz zwischen zwei Objekten  $x_i$  und  $x_j$  werden sogenannte Proximitätsmaße verwendet. Diese unterscheiden sich je nach Vorliegen von Daten mit nominaler Struktur oder Daten mit metrischer Struktur. Während zwischen zwei nominalen Variablen meist die Ähnlichkeit gemessen wird, werden im Falle metrischer Daten im allgemeinen Distanzmaße genutzt. Dabei lassen sich jedoch Ähnlichkeitsmaße oft durch geeignete Transformation in Distanzmaße umformen. Da sich die Werte der Koeffizienten  $k_{i,j}$  zur Bestimmung der Ähnlichkeit in den meisten Fällen zwischen Null (keine Ähnlichkeit) und Eins (vollkommene Ähnlichkeit) befinden, kann durch die Transformation  $1 - k_{i,j}$  eine Umwandlung der Ähnlichkeitsmaße in Distanzmaße  $d_{ij}$  erreicht werden. Diese bilden die Grundlage der hierarchischen Clusterverfahren.

---

<sup>1</sup>vgl. Backhaus, Erichson, Plinke, Weiber: *Multivariate Analysemethoden* (2003), S.481 ff

### 2.1.1 Binäre Daten

Weisen die zugrunde liegenden Variablen eine binäre Struktur auf (0/1-Variablen), kodiert ein Wert von Null das Fehlen der definierten Eigenschaft und entsprechend ein Wert von Eins das Vorhandensein derselbigen. Die zugehörigen Proximitätsmaße sind meist Ähnlichkeitsmaße. Ihre Bestimmung basiert auf dem Vergleich der Anzahl der Übereinstimmungen (bzw. Nicht-Übereinstimmungen) bezüglich der betrachteten Variablen. Dabei sind bei einem paarweisen Vergleich folgende Kombinationen möglich:

$a_{11}$ : beide Variablen weisen die Eigenschaft auf (11 – Kodierung)

$a_{10}$ : nur die erste Variable weist die Eigenschaft auf (10 – Kodierung)

$a_{01}$ : nur die zweite Variable weist die Eigenschaft auf (01 – Kodierung)

$a_{00}$ : keine der beiden Variablen weist die Eigenschaft auf (00 – Kodierung)

Bei Vorliegen von mehrkategorialen (oBdA n-kategorialen) Variablen muss eine Transformation in Binärvariablen erfolgen. Dazu stehen mehrere Möglichkeiten zur Verfügung<sup>2</sup>:

Die erste Möglichkeit besteht darin, mehrere Kategorien zusammenzufassen, so dass letztendlich nur zwei Kategorien betrachtet werden. Zu beachten ist jedoch, dass die Zusammenfassung der Kategorien zum einen inhaltlich sinnvoll sein sollte und zum anderen selbst bei Interpretierbarkeit der neu entstandenen Kategorien ein hoher Informationsverlust entstehen kann.

Eine weitere Möglichkeit besteht darin, das Vorliegen der  $i$ -ten Kategorie durch die binäre Folge

$$[0, \dots, 0, \overbrace{1}^i, 0, \dots, 0]$$

zu kodieren. Bei dieser Vorgehensweise werden nur Übereinstimmungen bzgl. derselben Kategorie gezählt. Bei der Auswertung ordinalstrukturierter Daten mit hoher Kategorienganzahl muss daher geprüft werden, ob ein metrisches Distanzmaß zur Bestimmung der Proximität eventuell geeigneter wäre, da es auf der Bestimmung absoluter Distanzen basiert.

---

<sup>2</sup>vgl. Moosbrugger, Frank: *Clusteranalytische Methoden in der Persönlichkeitsforschung*, Verlag Hans Huber (2002)

Eine dritte Methode zur Codierung mehrkategorialer Variablen in Binärvariablen stellt die "Niveau-Regression" dar: Nimmt ein ordinalskaliertes Objekt den  $i$ -ten Rangplatz der geordneten Skala ein, so werden den ersten  $i$  Variablen ein Wert von 1 zugeordnet, die verbleibenden Variablen werden mit 0 kodiert. Der Sinn der Verwendung dieser Methode hängt jedoch stark von der zugrundeliegenden Fragestellung ab.

Die bekanntesten Koeffizienten zur Ermittlung der Ähnlichkeit zwischen den Objekten  $x_i$  und  $x_j$  werden nachfolgend erläutert<sup>34</sup>.

### **Tanimoto**

Der Tanimoto-Koeffizient misst den Anteil der gemeinsam vorkommenden Eigenschaften (positive Übereinstimmungen) an der Anzahl aller Variablen, die die Eigenschaft aufweisen. Er ist definiert als

$$k_{i,j} = \frac{a_{11}}{a_{11} + a_{10} + a_{01}}$$

Da hier die Anzahl der negativen Übereinstimmungen nicht berücksichtigt wird, ist der Gebrauch des Tanimoto-Koeffizienten nicht sinnvoll, wenn es nur darum geht, (negative und positive) Übereinstimmungen zwischen den Objekten zu bewerten. Liegen dagegen ordinalskalierte Variablen vor, die der oben genannten zweiten Methode der Binärtransformation unterzogen wurden, sind hier negative Übereinstimmungen überproportional vorhanden und sollten daher nicht in die Distanzmessung miteinbezogen werden. In diesem Fall stellt der Tanimoto-Koeffizient ein sinnvolles Maß zur Ermittlung der Ähnlichkeit dar.

### **Russel & Rao (RR)**

Der RR-Koeffizient ist definiert als

$$k_{i,j} = \frac{a_{11}}{a_{11} + a_{10} + a_{01} + a_{00}}$$

Er misst den Gesamtanteil aller positiven Übereinstimmungen. Auch hier wird das gemeinsame Vorhandensein einer Eigenschaft höher bewertet als das gemeinsame Fehlen dieser Eigenschaft.

---

<sup>3</sup>vgl. Backhaus, Erichson, Plinke, Weiber: *Multivariate Analysemethoden*, Springer(2003), S.485-490

<sup>4</sup>vgl. Härdle, Simar: *Applied Multivariate Statistical Analysis*, Springer (2002), S.304

### Simple Matching (M)

Mit dem M-Koeffizienten wird der Gesamtanteil aller positiven und negativen Übereinstimmungen ermittelt:

$$k_{i,j} = \frac{a_{11} + a_{00}}{a_{11} + a_{10} + a_{01} + a_{00}}$$

Er kommt dann zur Geltung, wenn positive und negative Matchings dieselbe Wertigkeit besitzen und ist daher nicht sinnvoll anzuwenden, wenn beispielsweise ordinalskalierte Variablen einer Binärtransformation unterzogen wurden.

### Dice

Bei Anwendung des Dice-Koeffizienten werden positive Gemeinsamkeiten sehr stark gewichtet, während das gemeinsame Fehlen der definierten Eigenschaft vernachlässigt wird. Dieses Ähnlichkeitsmaß ist daher mit dem Tanimoto-Koeffizienten zu vergleichen. Es wird definiert durch

$$k_{i,j} = \frac{2a_{11}}{2a_{11} + a_{10} + a_{01}}$$

### Kulczynski

Der Kulczynski-Koeffizient misst den Anteil aller positiven Übereinstimmungen gemessen an der Anzahl aller Nicht-Übereinstimmungen:

$$k_{i,j} = \frac{a_{11}}{a_{10} + a_{01}}$$

Er ist daher stets größer als der Tanimoto- oder RR-Koeffizient.

## 2.1.2 Metrische Daten

Weisen die zu klassifizierenden Variablen metrisches Skalenniveau auf, wird ihre Ähnlichkeit im allgemeinen mittels eines Distanzmaßes bestimmt. Dieses nimmt bei großer

Ähnlichkeit Werte nahe Null an. Im Gegensatz zu den Ähnlichkeitskoeffizienten bei binären Variablen, deren Werte sich meist im Intervall  $[0, 1]$  befinden, basiert eine Vielzahl der metrischen Distanzmaße auf absoluten Abständen, die Werte im Bereich der positiven reellen Zahlen annehmen.

Gängige Distanzmaße stellen beispielsweise die  $L_r$  – Normen dar:

$$d_{i,j} = \left( \sum_{k=1}^p |x_{ik} - x_{jk}|^r \right)^{\frac{1}{r}}$$

Zwei dieser  $L_r$  – Normen werden im Folgenden vorgestellt:

### **City-Block-Metrik**

Die City-Block-Metrik (auch Taxifahrer- oder Manhattan-Metrik genannt) entspricht der  $L_1$  – Norm. Hier wird nicht die Luftlinie zwischen zwei Punkten als Distanzmaß verwendet, sondern die Summe der absoluten Abstände zwischen den Objekten herangezogen.

### **(Quadrierte) Euklidische Distanz**

Sie entspricht der (quadrierten)  $L_2$  – Norm und ist ein häufig verwendetes Distanzmaß. Bei Vorliegen einer Quadrierung werden große Distanzen stärker gewichtet als geringe Distanzen zwischen den Objekten.

Der Nachteil der  $L_r$  – Normen ist, dass sie nicht skaleninvariant sind. Die Objekte sollten daher in vergleichbarer Größendimension vorliegen oder einer Standardisierung unterworfen werden.

Als weitere Proximitätsmaße für Daten mit metrischem Skalenniveau können auch Korrelationskoeffizienten herangezogen werden.

## **2.2 Clusterverfahren**

In der Clusteranalyse existiert eine Vielzahl verschiedener Verfahren zur Klassifikation von Objekten. Zwei Algorithmengruppen sind dabei von besonderer Bedeutung: Die hierarchischen Verfahren und die partitionierenden Verfahren. Unter dem Gesichtspunkt des

Clusterbildungsprozesses lassen sich bei den partitionierenden Verfahren iterative und nicht-iterative Methoden unterscheiden, bei hierarchischen Verfahren gibt es die Einteilung in agglomerative und divisive Algorithmen.

Ein besonderes Augenmerk soll in diesem Abschnitt auf die hierarchisch-agglomerativen Verfahren gelegt werden.

## 2.2.1 Hierarchische Klassifikationsverfahren

### Agglomerative Verfahren

Agglomerative Verfahren starten bei der Clusterbildung mit der feinsten Partition. Das bedeutet, dass jedes der zu clusternden Objekte  $x_i$  einen Cluster darstellt. Im nächsten Schritt werden die beiden Objekte, die die geringste Distanz (die mittels eines Proximitätsmaßes zuvor berechnet wurde) zueinander aufweisen, zu einer Gruppe  $K$  zusammengefasst. Anschließend wird eine neue Distanzmatrix erstellt, die die Distanz zwischen dem so gebildeten Objekt  $K$  und den noch verbleibenden Variablen  $x_i$  enthält. Durch die Art und Weise, wie diese neue Distanzberechnung erfolgt, unterscheiden sich die einzelnen Verfahren.

Iterativ werden dann so lange neue Gruppierungen gebildet, bis nur noch ein Cluster, der alle Objekte umfasst, besteht. Der Verlauf der Clusterbildung ist beispielsweise anhand eines Dendrogrammes (Baumdiagrammes) ablesbar.

Der Algorithmus der hierarchischen Verfahren läuft also wie folgt ab<sup>5</sup>:

1. Bestimmung der Distanzmatrix (Ähnlichkeitsmaße  $k$  werden einer geeigneten Transformation unterworfen)
2. Fusionierung der Objekte (Gruppen), die die geringste Distanz zueinander aufweisen, die Anzahl der zu clusternden Gruppen verringert sich damit um 1
3. Berechnung der reduzierten Distanzmatrix, dann zurück zu 2.

Die Bestimmung der neuen Distanzen in Schritt 3 unterscheidet sich je nach verwendeten Algorithmus.

---

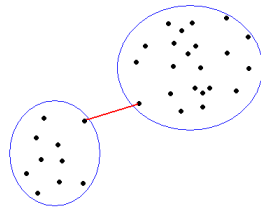
<sup>5</sup>vgl. Härdle, Simar: *Applied Multivariate Statistical Analysis*, Springer (2002), S.309

Drei dieser Algorithmen werden im Folgenden vorgestellt<sup>6</sup>:

### Single-Linkage-Verfahren

Beim Single-Linkage-Verfahren wird als Distanz zwischen zwei Clustern  $A$  und  $B$  der minimale Abstand zwischen zwei Elementen  $x_A$  und  $x_B$  aus  $A$  und  $B$  verwendet:

$$d(A, B) = \min_{x_A \in A, x_B \in B} (d(x_A, x_B))$$

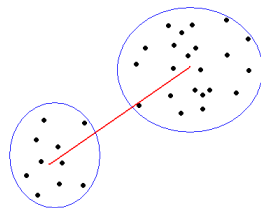


Das Single-Linkage-Verfahren hat den Nachteil, dass es bei unzureichend voneinander isolierten Clustern oder ungünstig liegenden Objekten zu Kettenbildung und Entstehung großer Cluster kommen kann.

### Average-Linkage-Verfahren

Als Distanz zwischen zwei Clustern  $A$ ,  $B$  wird der durchschnittliche Abstand aller Elementpaare aus beiden Clustern verwendet:

$$d(A, B) = \frac{1}{|A| \cdot |B|} \cdot \sum_{x_A \in A, x_B \in B} (d(x_A, x_B))$$



Die entstandenen Cluster weisen häufig kleine Varianzen auf.

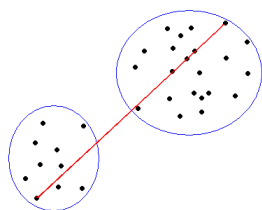
---

<sup>6</sup>vgl. <http://de.wikipedia.org/wiki/Clusteranalyse> (15.03.2007)

### Complete-Linkage-Verfahren

Beim Complete-Linkage-Verfahren wird als Distanz zwischen zwei Clustern  $A$  und  $B$  der maximale Abstand zwischen zwei Elementen  $x_A$  und  $x_B$  aus  $A$  und  $B$  verwendet:

$$d(A, B) = \max_{x_A \in A, x_B \in B} (d(x_A, x_B))$$



Unter dem Complete-Linkage-Verfahren besteht die Tendenz zur Bildung kleiner und kompakter Cluster, es ist jedoch anfällig für Ausreißer. Dieses Verfahren ist geeignet, wenn die Gruppen zwar in sich homogen, jedoch aufgrund ungünstiger Objekte nicht stark voneinander isoliert sind.

### Divisive Verfahren

Divisive Verfahren beginnen mit der größten Clusterunterteilung, das heißt, alle Objekte befinden sich zunächst in einem Cluster, und unterteilen sukzessive die vorhandenen Gruppen in mehrere Cluster. Man unterscheidet zwischen monothetischen und polythetischen Verfahren. Die meisten monothetischen Verfahren finden ihre Anwendung bei Vorliegen binärer Daten, die Clusterbildung stützt sich auf das Vorhandensein oder Nicht-Vorhandensein eines Divisionsmerkmals. Da die Aufteilung bei monothetischen Verfahren nur anhand dieses einen Merkmals verläuft, sind die gebildeten Gruppen zwar diesbezüglich homogen, jedoch besteht die Möglichkeit, dass sich die Objekte innerhalb eines Clusters bezüglich anderer Merkmale stark voneinander unterscheiden. Der Nachteil divisiv-polythetischer Verfahren, die alle Merkmale berücksichtigen, ist die Erfordernis eines (im Vergleich zu agglomerativen Verfahren) hohen Rechenaufwands.

### 2.2.2 Partitionierende Verfahren

Im Gegensatz zu den hierarchischen Verfahren bildet die Ausgangsbasis der partitionierenden Clusterverfahren eine vorgegebene Gruppierung der untersuchten Objekte. Die zugrunde liegenden Algorithmen sind dadurch gekennzeichnet, dass sie diese Cluster schrittweise so umschichten, bis eine optimale Gruppeneinteilung erreicht ist. Die vorgegebene Clusterzahl ändert sich dabei nicht. Zur Bestimmung der optimalen Gruppierung wird ein bestimmtes Gütekriterium herangezogen, das Verfahren bricht dann ab, wenn keine Verbesserung der Güte mehr eintritt.

Ein Vorteil der partitionierenden Verfahren liegt in der Flexibilität bzgl. des Clusterbildungsprozesses: Im Gegensatz zu den hierarchischen Verfahren ist eine Auflösung bereits bestehender Gruppen noch möglich, sofern dadurch eine Verbesserung des Gütekriteriums erreicht werden kann. Von Nachteil ist jedoch die Voraussetzung der Wahl einer bestimmten Clusterzahl.

# 3 Methodik der Datenanalyse

## 3.1 Idee

Ein Ziel der vorliegenden Datenanalyse ist der Vergleich der unterschiedlichen Clusterstrukturen, die mittels verschiedener Verfahren gewonnen werden, sowie die Untersuchung der einzelnen Clusterbildungsprozesse. Hierbei soll festgestellt werden, inwieweit die angewendeten Fusionierungsalgorithmen und Distanzmaße in Abhängigkeit von der gewählten Clusterzahl übereinstimmen.

Wie erfolgt aber die Messung solcher Übereinstimmungen? Eine mögliche Methodik ist der paarweise Vergleich zweier Verfahren oder Distanzmaße durch die Bildung von  $k \times k$ -Kontingenztabellen  $(a_{i,j})_{i,j=1,\dots,k}$ , wobei  $k$  die Anzahl der gebildeten Cluster ist. Ein Element  $a_{i,j}$  der Tabelle gibt an, wieviele Objekte unter Verfahren  $A$  dem  $i$ -ten Cluster zugeordnet wurden und gleichzeitig unter Verfahren  $B$  dem  $j$ -ten Cluster. Liegt bei beiden Verfahren dieselbe Gruppierung vor, sollte daher in jeder Zeile und Spalte nur je ein positiver Eintrag existieren. Durch geeignete Umdefinierung der Clusternummern wird erreicht, dass sich diese positiven Einträge gerade in der Hauptdiagonalen der Kontingenztafel befinden (s. Abb 3.1). Die Randhäufigkeiten  $a_{i,\cdot}$ ,  $a_{\cdot,j}$ ,  $i, j = 1, \dots, k$ , geben an, wieviele Objekte insgesamt unter Verfahren  $A$  bzw.  $B$  dem  $i$ -ten bzw.  $j$ -ten Cluster zugeordnet wurden.

Ein geeigneter Test, durch den die Übereinstimmung der betrachteten Verfahren bewiesen werden kann, ist auf Basis der Annahme einer auf Multinomialverteilung basierenden Kontingenztafel nicht möglich. Der Grund dafür ist, dass die zu testende Hypothese aus der Aussage, dass nur die Hauptdiagonale positive Einträge aufweist (das bedeutet, dass die Wahrscheinlichkeit eines positiven Eintrags in den übrigen Zellen Null wäre), bestünde. Dies hat zur Folge, dass die Hypothese bereits nicht mehr aufrechterhalten werden kann, wenn die Randhäufigkeiten beider Verfahren bezüglich des  $i$ -ten Clusters nicht denselben Wert aufweisen.

Ein Chi-Quadrat-Test auf Unabhängigkeit der betrachteten Verfahren wäre zwar unter

	1	2	3
1	27	27	27
2	2	16	18
3	19	1	2
	21	28	18
	67		

→

	1	2	3
1	27	27	27
2	2	16	2
3	1	2	19
	28	18	21
	67		

Abbildung 3.1: Formatierung der Kontingenztabelle bei Vorgabe von 3 Clustern. Positive Einträge der Kontingenztabelle sind rot gekennzeichnet, die Randhäufigkeiten gelb. So werden beispielsweise unter Verfahren *A* insgesamt 22 Objekte dem Cluster 1 zugeordnet, davon befinden sich 19 Objekte auch unter Verfahren *B* im selben Cluster. Die Gesamtzahl der Objekte beträgt 67.

gegebenen Voraussetzungen durchführbar, jedoch für die behandelte Fragestellung von geringer Bedeutung, da der Nicht-Beweis der Unabhängigkeit nicht impliziert, dass die Verfahren auch zu gleichen Ergebnissen führen.

Die Messung des Grades der Übereinstimmungen kann hier nur durch ein Bestimmtheitsmaß erfolgen. In den weiteren Analysen findet der symmetrische Kappa-Koeffizient (s. Kapitel 3.2.1) Verwendung.

Allgemein besteht das Problem bei der Untersuchung der Ähnlichkeit von iterativen Clusterverfahren auf Basis von Kontingenztabelle bezüglich einzelner Iterationsschritte darin, dass lediglich eine "Momentaufnahme" des Clusterbildungsprozesses vorliegt. Um fundierte Aussagen treffen zu können, müsste daher jede Stufe des Prozesses analysiert werden. Dies hat die Nachteile, dass zum einen ein erheblicher Rechenaufwand erforderlich ist und zum anderen die Interpretation der Clusterstrukturen vorab klar definiert werden muss. Schließlich werden sämtliche Fusionierungsalgorithmen sowohl unter Betrachtung der feinsten Partition (im vorliegenden Fall sind dies 67 Cluster) als auch unter Betrachtung der größten Partition (ein Cluster) dieselben Ergebnisse hervorbringen (der Kappa-Koeffizient wird hier also stets einen Wert von Eins aufweisen).

Wann werden also bestimmte Verfahren als ähnlich angesehen? Der Kappa-Koeffizient allein kann auf diese Frage keine Antwort liefern. Allgemein erweist es sich vorab als sinnvoll, anhand der graphischen Repräsentation der Fusionierungsprozesse (zum Beispiel einzelne Dendrogramme) gemeinsame Strukturen aufzudecken.

Eine weitere Möglichkeit zur Beurteilung von Clusterverfahren bietet der kophenetische Korrelationskoeffizient (s. Kapitel 3.2.2).

### Vergleich mit den Ergebnissen einer Faktorenanalyse

Da auch die Faktorenanalyse zu den strukturentdeckenden Verfahren gehört, werden in Kapitel 7 die durch eine Faktorenanalyse ermittelten Faktoren mit den Clusterstrukturen eines Clusterverfahrens verglichen. Allgemein erweist sich ein solcher Vergleich zum Teil als schwierig, da die Distanzen zwischen je zwei Objekten meist so definiert sind, dass stark negativ korrelierenden Variablen ein hoher Distanzwert zugeordnet wird. Diese Variablen würden nach Durchführung einer Faktorenanalyse bezüglich eines Faktors sehr hohe Faktorladungen aufweisen, während sie nach Anwendung eines hierarchischen Clusterverfahrens unterschiedlichen Clustern zugeordnet würden. Bei vorliegender Datenstruktur fällt dieser "Fehler" deutlich ins Gewicht, daher sollte er bei der Interpretation und dem Vergleich der Gruppierungen nicht unbeachtet bleiben.

Eine weitere Schwierigkeit der Analyse besteht zum einen darin, dass die Faktorstruktur von der gewählten Rotation der Faktorladungen abhängt, zum anderen handelt es sich bei der Faktorenanalyse um eine *Regression* der Variablen auf die einzelnen Faktoren. Das bedeutet, dass es nicht genügt, die Variablen dem Faktor zuzuordnen, auf den sie am höchsten laden, denn allein dies ist nicht das Ergebnis einer Faktorenanalyse. Für einen sinnvollen Vergleich der entstandenen Gruppen müsste die Gesamtheit der Faktorladungen betrachtet werden.

## 3.2 Koeffizienten zur Beurteilung von Clusterstrukturen

### 3.2.1 Kappa-Koeffizient

Der Kappa-Koeffizient ist ein Bestimmtheitsmaß für nominale Daten. Er misst den Grad der Übereinstimmungen zweier Objekte  $A$  und  $B$  bezüglich der Kategorien einer oder mehrerer Variablen. Die Häufigkeiten  $a_{i,j}$ , mit denen Objekt  $A$  der Kategorie  $i$  und Objekt  $B$  der Kategorie  $j$  zugeordnet wird, sind dabei in einer quadratischen Kontingenztafel aufgeführt. Formal ist Cohens Kappa-Koeffizient folgendermaßen definiert:

$$\kappa = \frac{P_a - P_e}{1 - P_e}$$

wobei

$P_a$  = relative beobachtete Häufigkeit an Übereinstimmungen

$P_e$  = relative erwartete Häufigkeit an Übereinstimmungen

Der Anteil der beobachteten Übereinstimmungen wird also um den zufällig zu erwartenden Anteil bereinigt. Die zu erwartenden relativen Häufigkeiten  $P_e$  lassen sich dabei anhand der Randverteilungen berechnen.

Cohens Kappa-Koeffizient nimmt Werte im Bereich -1 (völlige Nicht-Übereinstimmung) und 1 (völlige Übereinstimmung) an. Hat  $\kappa$  einen Wert nahe Null, wird die Übereinstimmung als zufällig angesehen<sup>1</sup>.

### 3.2.2 Kophenetischer Korrelationskoeffizient

Der kophenetische Korrelationskoeffizient ist ein Maß für die Güte von Clusterlösungen. Er beschreibt den Zusammenhang zwischen den Einträgen  $d_{i,j}$  der Distanzmatrix  $D$  und den Werten der kophenetischen Matrix  $D^*$ .

Die kophenetische Matrix führt dabei die Distanzen  $d_{i,j}^*$  auf, bei denen unter dem angewendeten hierarchisch-agglomerativen Clusterverfahren erstmals je zwei Objekte in einem Cluster fusioniert werden. Diese Distanzen sind auch anhand des Dendrogramms ablesbar.

Je höher die Korrelation zwischen den  $d_{i,j}$  und  $d_{i,j}^*$  ist, desto besser werden die ursprünglich gebildeten Distanzen zwischen den einzelnen Objekten in der endgültigen Clusterstruktur abgebildet. Demnach sollte das Verfahren angewendet werden, bei dem der kophenetische Korrelationskoeffizient die höchsten Werte aufweist<sup>2</sup>.

---

<sup>1</sup>vgl. B.Rönz: *Skript zu "Computergestützte Statistik II"* (2000), S.77/78

<sup>2</sup>vgl. A.Handl: *Multivariate Analysemethoden*, Springer (2002), S.380

## 4 Vergleich der Clusterverfahren anhand von Kontingenztabellen

Aufgrund der Vielzahl von Fusionierungsalgorithmen und Distanzmaßen wird in den weiteren Untersuchungen das Augenmerk nur auf eine Auswahl der Verfahren gelegt. Als Distanzmaße für binäre Daten werden der Tanimoto-Koeffizient und der Dice-Koeffizient, für metrische Daten die Euklidische Distanz und die City-Block-Metrik verwendet. Der Tanimoto- und Dice-Koeffizient werden gewählt, da die Daten ein ordinales Skalenniveau aufweisen und der in Kapitel 2.1.1 erläuterten Binärtransformation unterworfen wurden, was zur Folge hat, dass die Anzahl der negativen Übereinstimmungen ohne Bedeutung ist und daher Distanzmaße ohne Berücksichtigung dieser in Betracht gezogen werden sollten.

Es stellt sich jedoch die Frage, ob ein Distanzmaß für binäre (bzw. ordinale) Daten überhaupt sinnvoll ist. Im vorliegenden Fall scheint es aufgrund der Anzahl der Beurteilungsstufen (1 bis 6) ebenso plausibel, ein Distanzmaß für metrische Daten zu verwenden.

Sei zum Beispiel Variable  $A$  immer mit "6" bewertet worden, Variable  $B$  immer mit "5" und Variable  $C$  einmal mit "6" und sonst immer mit "1". Unter einem Proximitätsmaß, das lediglich die Übereinstimmungen zählt, würden die Variablen  $A$  und  $B$  eine höhere Distanz aufweisen als die Variablen  $A$  und  $C$ . Unter der Euklidischen Distanz wäre es umgekehrt, was jedoch in diesem Fall angebracht erscheint.

Da die untersuchten Fusionierungsalgorithmen für metrische *und* binäre Distanzmaße interpretierbar sein sollen, wird sowohl von partitionierenden Verfahren als auch von hierarchisch-agglomerativen Verfahren wie Ward-Algorithmus oder Zentroid-Methode abgesehen. Verwendet werden der Average-Linkage-Algorithmus, der Complete-Linkage-Algorithmus und der Single-Linkage-Algorithmus.

Da ein Vergleich der Algorithmen nur unter Verwendung *eines* Datensatzes möglich ist, beziehen sich die folgenden Analysen nur auf einen ("impute1") der fünf imputierten Datensätze. Die Auswertung der Ergebnisse aus allen Datensätzen ist dagegen erst für

## 4 Vergleich der Clusterverfahren anhand von Kontingenztabellen

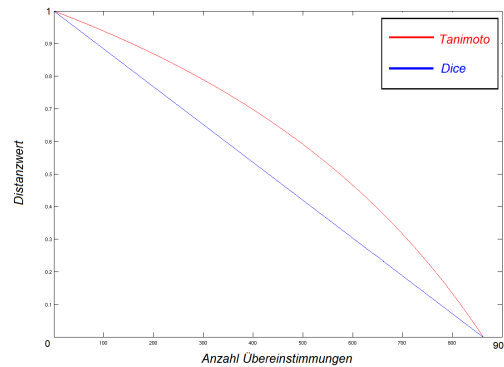


Abbildung 4.1: Tanimoto- und Dice-Koeffizient in Abhängigkeit von der Anzahl der positiven Übereinstimmungen

die Interpretation der Clusterstrukturen bezogen auf den originalen Datensatz notwendig.

Die nachfolgenden Analysen basieren auf der Interpretation von Kontingenztabellen, die die Übereinstimmungen je zweier Verfahren bezüglich der Variablen-Gruppierung bei vorgegebener Clusterzahl aufführen. Die Clusterzahl variiert dabei zwischen 1 und 6.

## 4.1 Vergleich der Distanzmaße

### 4.1.1 Tanimoto und Dice-Koeffizient

Abb.4.1 stellt den Wert des Tanimoto- und Dice-Koeffizienten (als Distanzmaß) in Abhängigkeit von der Anzahl der positiven Übereinstimmungen dar.

Es zeigt sich, dass der Tanimoto-Koeffizient niemals niedrigere Werte annimmt als der Dice-Koeffizient. Bei Anwendung des Single-Linkage-Verfahrens ("nächster Nachbar") und des Complete-Linkage-Verfahrens ("entferntester Nachbar") entstehen aufgrund dieser Monotonie daher unter beiden Proximitätsmaßen die gleichen Clusterstrukturen.

Unter dem Average-Linkage-Verfahren dagegen können hier geringe Unterschiede auftreten, die bei den vorliegenden imputierten Datensätzen jedoch keine bedeutende Rolle spielen. Abbildung 4.2 zeigt die Dendrogramme unter Anwendung des Average-Linkage-

#### 4 Vergleich der Clusterverfahren anhand von Kontingenztabelle

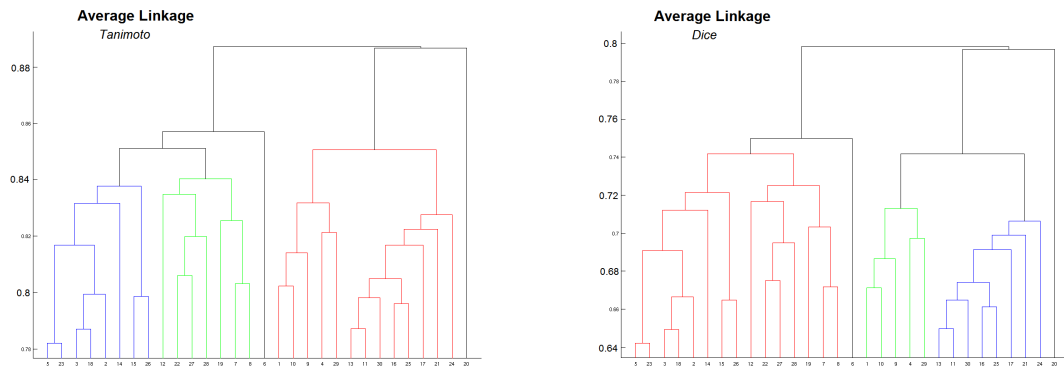


Abbildung 4.2: Dendrogramme für Average-Linkage-Verfahren unter Anwendung von Tanimoto (links) und Dice (rechts)

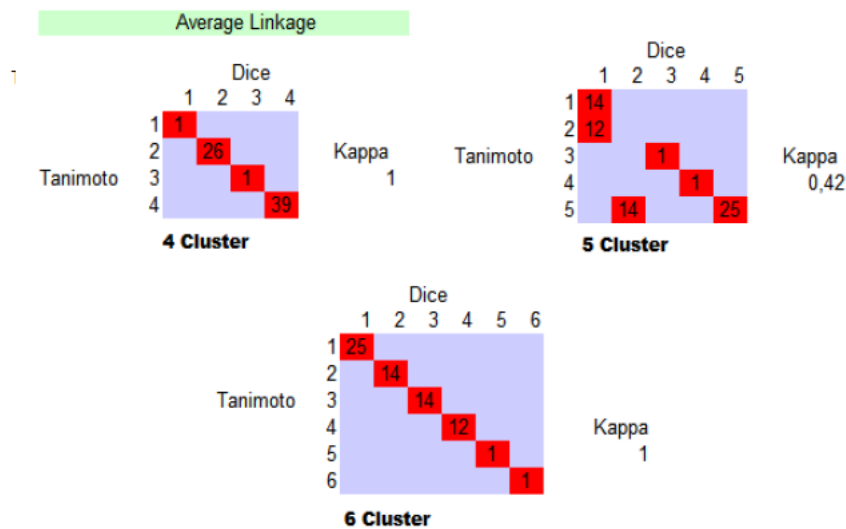


Abbildung 4.3: Vergleich von Tanimoto und Dice mittels Kontingenztabelle bei Anwendung des Average-Linkage-Verfahrens. Hier wird deutlich, dass der Kappa-Koeffizient nur eine Momentaufnahme des Fusionierungsprozesses widerspiegelt: Während beide Distanzmaße bei Betrachtung von 6 Clustern gleiche Ergebnisse hervorbringen, werden in der nachfolgenden Iteration unter den Koeffizienten verschiedene Gruppen fusioniert. Unter Tanimoto sind dies die Cluster mit 25 und 14 Objekte, unter Dice die Cluster mit 14 und 12 Objekten. Anschließend (Bildung von 4 Clustern) ist die Struktur bei beiden Distanzmaßen wieder gleich.

Verfahrens jeweils mit dem Tanimoto- und Dice-Koeffizienten als zugrunde liegendes Distanzmaß. An dieser Graphik ist zum einen ersichtlich, dass sich die Ergebnisse bei beiden Distanzmaßen nur minimal voneinander unterscheiden, zum anderen werden die bereits erwähnten Probleme der Interpretation des Kappa-Koeffizienten deutlich: Wie in den Graphiken zu sehen, bewirkt ein minimaler Unterschied in der kophenetischen Distanz, dass bei Vorgabe von 5 Clustern unter den beiden Distanzmaßen eine unterschiedliche Clusterung entsteht. Der Kappa-Koeffizient weist also an dieser Stelle einen Wert unter 1 auf, obwohl sich die Ergebnisse insgesamt bei Anwendung der beiden Proximitätsmaße kaum voneinander unterscheiden.

### 4.1.2 Euklidische Distanz und City-Block-Metrik

Zwischen den verwendeten Distanzmaßen für metrische Daten bestehen bei Betrachtung von 2 bis 6 Clustern im Wesentlichen keine größeren Unterschiede. Die einzelnen Dendrogramme (s. *Anhang A.4.1*) weisen unter Anwendung des Average-Linkage-Algorithmus und des Complete-Linkage-Algorithmus ähnliche Strukturen auf.

Eine Auffälligkeit ist nur unter dem Single-Linkage-Verfahren zu beobachten: Anhand der Dendrogramme (s. *Abb.4.4*) sind bei beiden Distanzmaßen grob gesehen zwei größere Cluster zu erkennen. Der Unterschied besteht aus dem Zeitpunkt des Fusionierungsprozesses, zu dem diese beiden Gruppen zu einem großen Cluster zusammengeführt werden. Während dieser Zusammenschluss auf Basis der quadrierten euklidischen Distanz bereits bei Betrachtung von 6 Clustern erfolgt ist, findet er unter Anwendung der City-Block-Metrik erst statt, wenn der Clusterbildungsprozess so weit fortgeschritten ist, dass nur noch drei Cluster bestehen. Identifiziert man die Cluster, die nur aus einer Variablen bestehen, als mögliche Ausreißer, bedeutet dies, dass unter der quadrierten euklidischen Distanz 5 Variablen als Ausreißer erkannt werden, unter der City-Block-Metrik jedoch nur 2 Variablen.

### 4.1.3 Ordinale Distanzmaße und metrische Distanzmaße

Beim Vergleich der Proximitätsmaße für ordinale Daten mit den Distanzmaßen für metrische Daten fällt bei Betrachtung der Dendrogramme auf, dass zwar Unterschiede bezüglich der Clusterstrukturen vorhanden sind, jedoch folgen diese Differenzen im All-

## 4 Vergleich der Clusterverfahren anhand von Kontingenztabellen

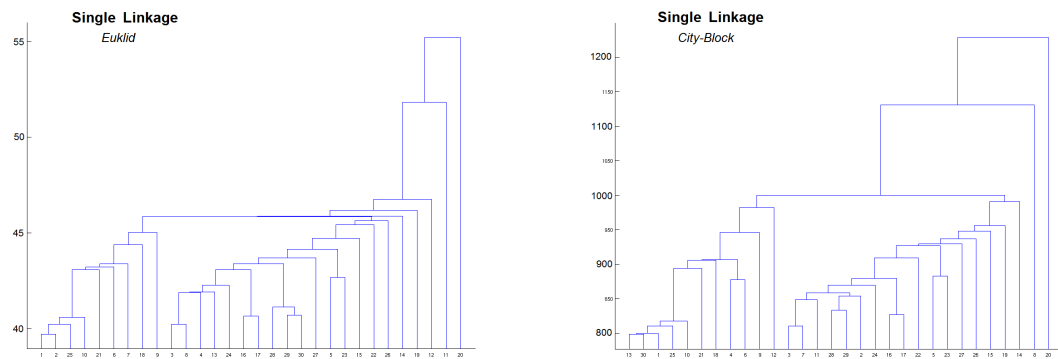


Abbildung 4.4: Dendrogramme für das Single-Linkage-Verfahren unter Verwendung der Euklidischen Distanz (links) und der City-Block-Metrik (rechts)

gemeinen keinem bestimmten Muster. Das bedeutet, dass bei vorliegender Datenstruktur generell keine Aussagen beispielsweise zur Gleichmäßigkeit oder Strukturklarheit der Cluster gemacht werden können. Welches Distanzmaß verwendet werden sollte, hängt zum einen davon ab, mit welchem Gewicht Unterschiede in den Beurteilungen der Variablen versehen werden sollten und zum anderen, ob bei gegebener Datenstruktur diese Gewichtung sinnvoll ist.

Eine Deutung der Gruppierungen soll an dieser Stelle nicht erfolgen, da sich die obigen Vergleiche nicht auf den originalen Datensatz beziehen, sondern nur auf einen der 5 imputierten Datensätze.

### 4.1.4 Entwicklung der Kappa-Koeffizienten in Abhängigkeit von der Clusterzahl

Da der Kappa-Koeffizient zur Beurteilung der Übereinstimmung der Verfahren immer nur unter Vorgabe einer bestimmten Clusterzahl  $k$  berechnet werden kann, spiegelt er nicht die Übereinstimmungen der Algorithmen im gesamten Iterationsprozess wider. Die Abbildungen 4.5-4.7 zeigen die Entwicklungen der Kappakoeffizienten für alle 5 imputierten Datensätze in Abhängigkeit von der vorausgesetzten Clusterzahl, welche im vorliegenden Fall Werte zwischen 2 und 6 annimmt. Dabei sind die Graphiken wie folgt zu verstehen: Sei  $G = (g_{i,j})_{i,j=1,\dots,k}$  die gesamte Matrix. Dann stellt die Graphik  $g_{i,j}$  den

#### 4 Vergleich der Clusterverfahren anhand von Kontingenztabellen

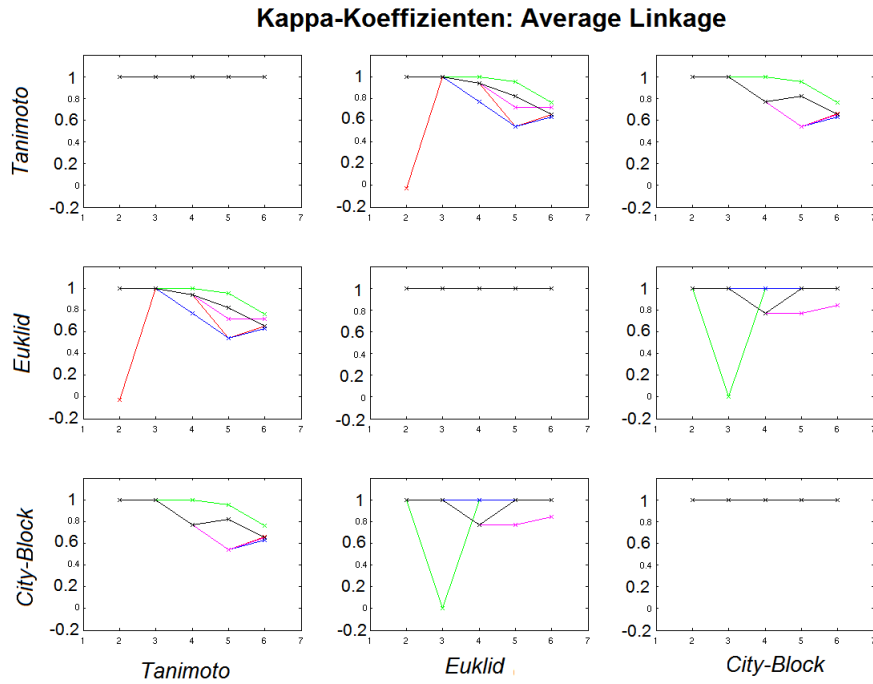


Abbildung 4.5: Entwicklung des Kappa-Koeffizienten in Abhängigkeit von der Clusterzahl unter Anwendung von Average-Linkage für alle Datensätze (rot: impute1, grün: impute2, blau: impute3, magenta: impute4, schwarz: impute5)

Verlauf der Kappa-Koeffizienten für alle imputierten Datensätze bezüglich des  $i$ -ten und  $j$ -ten Verfahrens (bzw. Distanzmaßes) dar. Auf der Hauptdiagonalen nimmt  $\kappa$  immer einen Wert von 1 an, da hier die Übereinstimmungen eines Verfahrens mit sich selbst bewertet werden.

Abb.4.5 stellt die Kappakoeffizienten für den Vergleich der Distanzmaße unter Anwendung des Average-Linkage-Verfahrens dar. Da der Tanimoto-Koeffizient und der Dice-Koeffizient zu annähernd gleichen Ergebnissen führen, findet an dieser Stelle der Dice-Koeffizient keine Berücksichtigung.

Zunächst ist zu beobachten, dass die Kappa-Koeffizienten im allgemeinen sehr hohe Werte annehmen (zwischen 0,6 und 1). Anhand dieser Darstellung könnte man zu der Vermutung gelangen, dass die metrischen Distanzmaße untereinander zu ähnlicheren Ergebnissen führen ( $\kappa > 0.77$ ) als ein metrisches Distanzmaß verglichen mit dem Tanimoto-Koeffizienten. Zu berücksichtigen ist jedoch, dass eine genauere Analyse die Entwicklung der Kappa-Koeffizienten im gesamten Fusionierungsverlauf erfordert.

#### 4 Vergleich der Clusterverfahren anhand von Kontingenztabellen

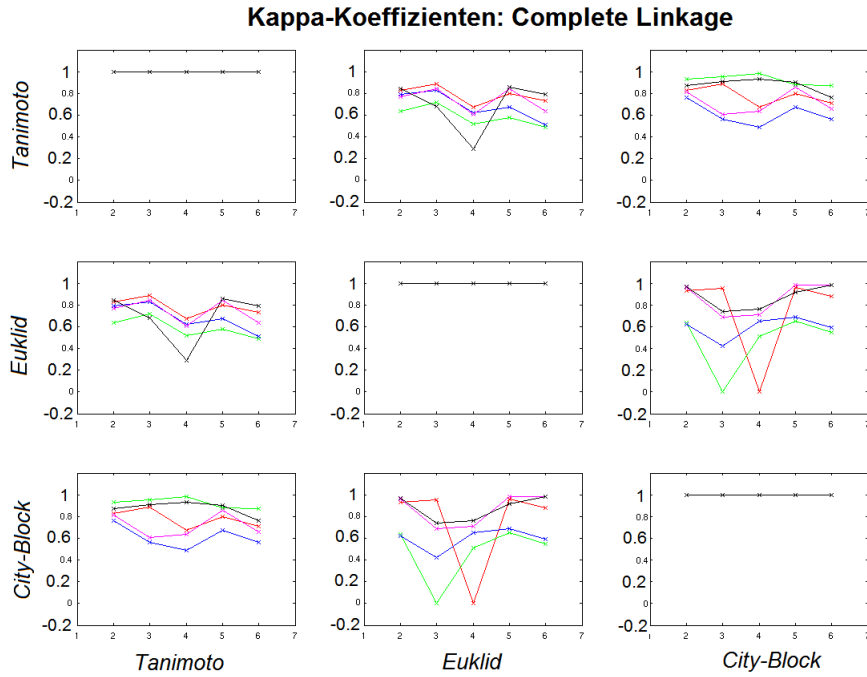


Abbildung 4.6: Entwicklung des Kappa-Koeffizienten in Abhängigkeit von der Clusterzahl unter Anwendung von Complete-Linkage für alle Datensätze (rot: impute1)

In Abb. 4.6 wird deutlich, dass die obige Vermutung nicht allgemein haltbar ist. Betrachtet man beispielsweise die grün gekennzeichnete Trajektorie (Datensatz "impute3"), so ist die Übereinstimmung der Gruppierungen beim Complete-Linkage-Verfahren unter dem Tanimoto-Koeffizienten und der City-Block-Metrik nahezu perfekt ( $\kappa > 0.87$ , s. *Anhang A.2.1*), während die Übereinstimmung zwischen Euklidischer Distanz und City-Block-Metrik nur mittelmäßig ( $\kappa \in [0.51, 0.73]$ , s. *Anhang A.2.1*) ist.

Die Abbildungen 4.5-4.7 sollen mehr aussagen: Zum einen wird verdeutlicht, dass der Kappa-Koeffizient ein sprunghaftes Verhalten zeigt, so dass es nicht ausreichend ist, die Verfahren oder Distanzmaße unter Vorgabe *einer* bestimmten Clusterzahl zu vergleichen. Das Sprungverhalten deutet vielmehr darauf hin, dass zwar in einem bestimmten Iterationsschritt bei den betrachteten Verfahren ein unterschiedlicher Gruppenzusammenschluss erfolgt, dieser Unterschied jedoch im nächsten Iterationsschritt wieder ausgeglichen wird.

#### 4 Vergleich der Clusterverfahren anhand von Kontingenztabellen

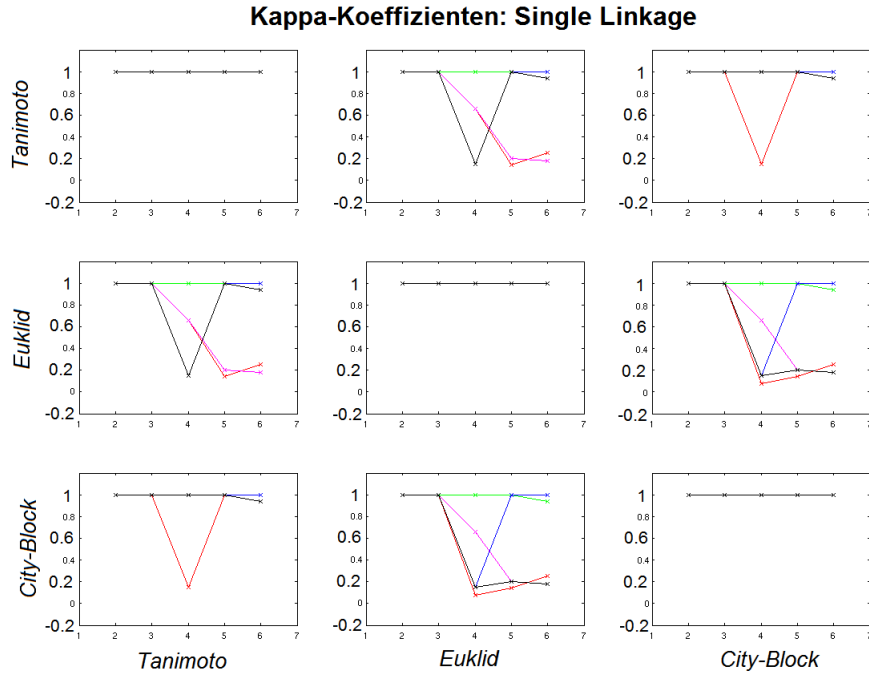


Abbildung 4.7: Entwicklung des Kappa-Koeffizienten in Abhängigkeit von der Clusterzahl unter Anwendung von Single-Linkage für alle Datensätze (rot: impute1)

Beispiel:

Situation: 4 Cluster  $C_1, C_2, C_3, C_4$

	Ausgangslage	i-te Iteration	(i+1)-te Iteration
Verfahren A	$\{C_1\}, \{C_2\}, \{C_3\}, \{C_4\}$	$\{C_1 + C_2\}, \{C_3\}, \{C_4\}$	$\{C_1 + C_2\}, \{C_3 + C_4\}$
Verfahren B	$\{C_1\}, \{C_2\}, \{C_3\}, \{C_4\}$	$\{C_1\}, \{C_2\}, \{C_3 + C_4\}$	$\{C_1 + C_2\}, \{C_3 + C_4\}$
$\kappa$	1	$< 1$	1

Weiterhin kann hier aufgezeigt werden, dass der Clusterbildungsprozess bei den imputierten Datensätzen unterschiedlich verlaufen kann, so dass bei Zusammenführung der Ergebnisse ein hoher Informationsverlust entstehen kann. Dieses Problem wird besonders bei Anwendung des Single-Linkage-Verfahrens deutlich. Abb. 4.7 zeigt, wie stark die entstandenen Cluster bei den einzelnen imputierten Datensätzen differieren können.

## 4.2 Vergleich der Algorithmen

Bei dem Vergleich der Clusteralgorithmen mittels Kontingenztabellen treten auch die zuvor erwähnten Probleme auf. Abb. 4.8 zeigt die Kontingenztabellen für den paarweisen Vergleich der Verfahren bei Vorgabe von 4 Clustern und unter Verwendung des Tanimoto-Koeffizienten. Deutlich zu erkennen sind hier die Neigungen der Algorithmen zu unterschiedlichen Clustergrößen. Während beim Complete-Linkage-Verfahren in dem betrachteten Iterationsschritt 3 große Cluster gebildet werden, beträgt diese Anzahl beim Average-Linkage nur 2, unter dem Single-Linkage kommt es nur zur Bildung einer großen Gruppe, 3 Variablen werden hier als Ausreißer identifiziert.

Aufgrund dieser unterschiedlichen Clusterstrukturen nimmt der Kappa-Koeffizient zwischen Complete- und Average-Linkage in diesem Fall den höchsten Wert (0.59) an, zwischen Complete- und Single-Linkage-Algorithmus ist ein Wert nahe Null zu beobachten.

Unter Verwendung des Tanimoto-Koeffizienten ist zu beobachten, dass dieses Verhalten des Kappa-Koeffizienten auch bei Untersuchung von 2 bis 6 Clustern zu beobachten ist. Abb.4.11 zeigt die Entwicklung der Kappa-Koeffizienten unter Tanimoto. Die exakten Werte sind dem Anhang A.2.1 zu entnehmen. Bei dem Vergleich der Entwicklung von  $\kappa$  unter Tanimoto mit der Entwicklung von  $\kappa$  unter der Euklidischen Distanz (Abb.4.11) fällt jedoch auf, dass die obigen Beobachtungen keine allgemeine Gültigkeit besitzen (also unabhängig vom gewählten Distanzmaß sind). Unter der Euklidischen Distanz verläuft der Clusterbildungsprozess etwas anders als unter Tanimoto, was zu teilweise erheblichen Differenzen in den einzelnen Kappa-Koeffizienten führen kann. Im vorliegenden Fall liegt der konkrete Grund in der sich unterscheidenden Anzahl identifizierter "Ausreißer" beim Single-Linkage-Verfahren unter Tanimoto (3 alleinstehende Variablen) bzw. Euklidischer Distanz (5 alleinstehende Variablen). Abb.4.9 stellt die zugehörigen Dendrogramme dar.

Diese Unterschiede in der Anzahl an möglichen Ausreißern (alleinstehende Variablen) sind auch allgemein als Ursache für niedrige Kappa-Koeffizienten anzusehen. Abb.4.10 zeigt die Dendrogramme von Complete-Linkage- und Average-Linkage Algorithmus auf Basis des Tanimoto-Koeffizienten. Bei Anwendung des Complete-Linkage-Verfahrens sind hier grob gesehen 3-4 Cluster zu erkennen, unter dem Average-Linkage-Verfahren kommt es zu größerer Clusterbildung, die Struktur des Dendrogramms weist auf 2-3 Cluster hin, außerdem werden zwei Variablen als mögliche Ausreißer identifiziert.

Um die Clusterstrukturen der Verfahren ohne Berücksichtigung von einzelnen isolierten Variablen unterscheidbar zu machen, müssen diese Variablen eliminiert werden. Erst dann kann ein sinnvoller paarweiser Vergleich der Algorithmen auf einer vorgegebe-

#### 4 Vergleich der Clusterverfahren anhand von Kontingenztabelle

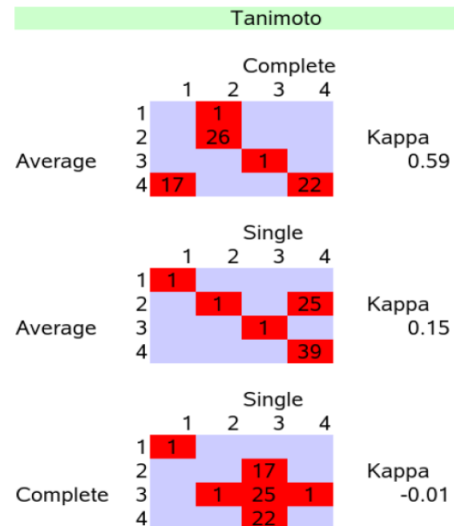


Abbildung 4.8: Paarweiser Vergleich der Clusteralgorithmen unter Verwendung von Tanimoto mittels Kontingenztabelle

nen Stufe der Fusionierungsprozesse (zum Beispiel auf Basis der zugehörigen Kappa-Koeffizienten) stattfinden. Eine Analyse der Kappa-Koeffizienten nach Eliminierung von Ausreißern, die durch das Single-Linkage-Verfahren identifiziert wurden, lässt eine deutliche Erhöhung Kappas erkennen (s. *Anhang A.2.1*).

Da der in diesem Kapitel analysierte Datensatz "impute1" jedoch nur einen Teil der Gesamtanalyse darstellt und allein kein Repräsentant des Originaldatensatzes ist, ist die Eliminierung von Ausreißern hier nicht sinnvoll. Diese Ausreißer müssten zur Gesamtanalyse bei allen 5 imputierten Datensätzen gestrichen werden, was zu Verzerrungen innerhalb der einzelnen Datensätze und damit in der gesamten Interpretation führen könnte.

#### 4 Vergleich der Clusterverfahren anhand von Kontingenztabellen

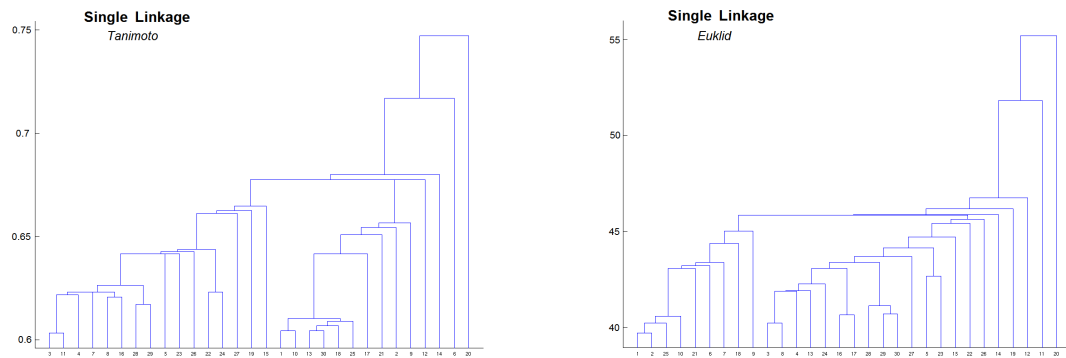


Abbildung 4.9: Dendrogramme für das Single-Linkage-Verfahren unter Verwendung von Tanimoto (links) und Euklidischer Distanz (rechts)

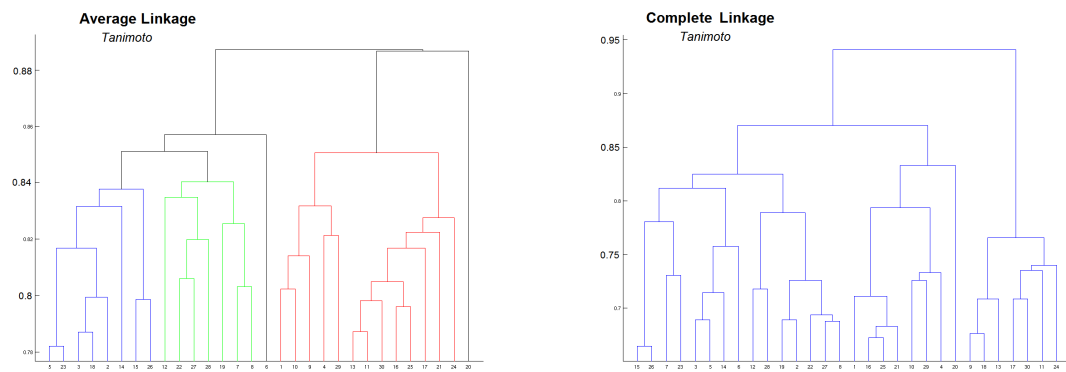


Abbildung 4.10: Dendrogramme für das Complete-Linkage-Verfahren (links) und das Average-Linkage-Verfahren (rechts) unter Verwendung des Tanimoto-Koeffizienten

#### 4 Vergleich der Clusterverfahren anhand von Kontingenztabellen

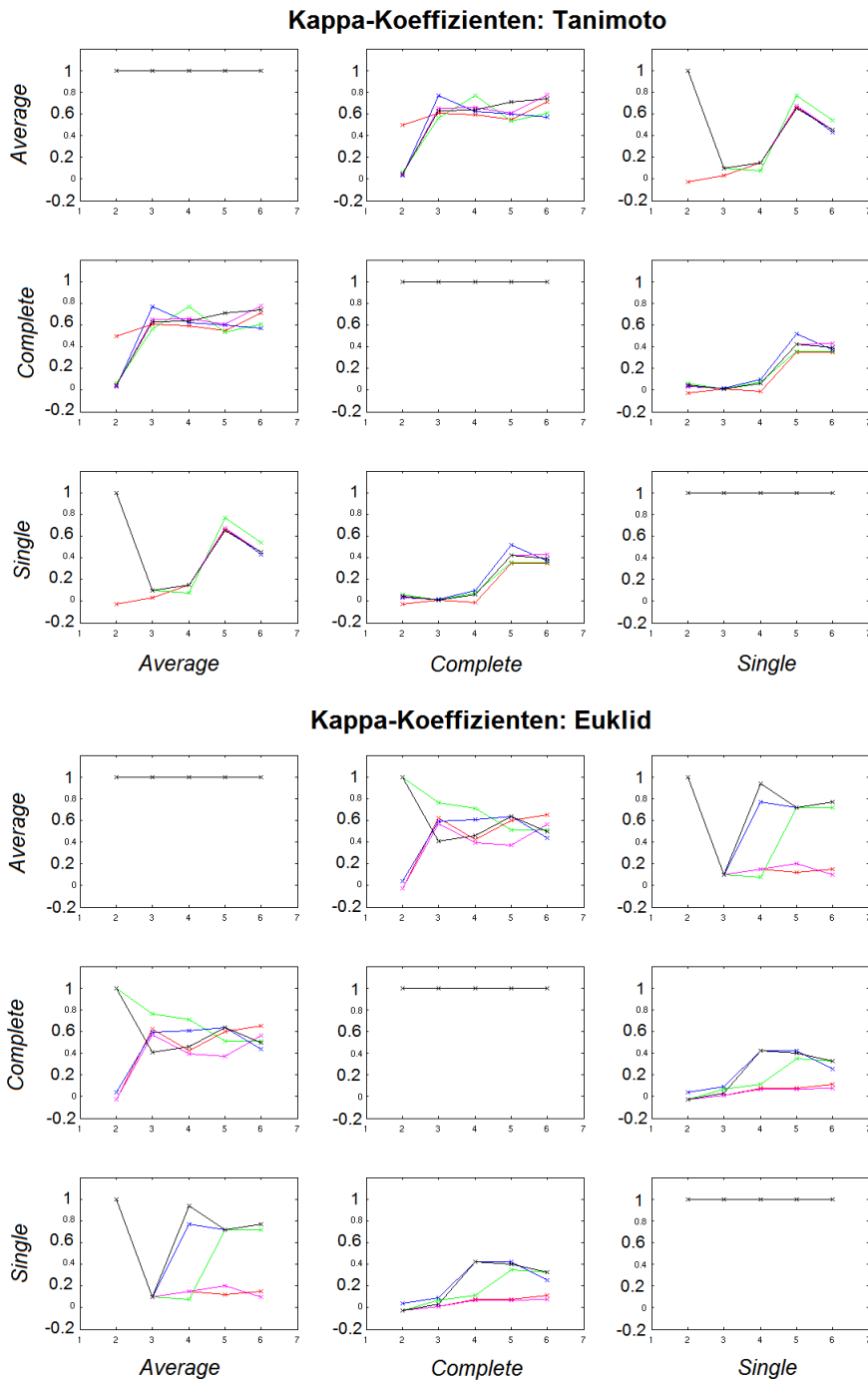


Abbildung 4.11: Entwicklung des Kappa-Koeffizienten in Abhängigkeit von der Clusterzahl unter Anwendung von Tanimoto (oben) und Euklidischer Distanz (unten) für alle Datensätze (rot: impute1)

# 5 Analyse der Clusterverfahren anhand der kophenetischen Korrelationskoeffizienten

Der kophenetische Korrelationskoeffizient gibt im Gegensatz zum Kappa-Koeffizienten Aufschluss über den gesamten Clusterbildungsprozess und wird daher als Kriterium für die Güte eines Fusionierungsalgorithmus verwendet. Abb.5.1 führt die kophenetischen Korrelationskoeffizienten für jedes Verfahren und Distanzmaß bezogen auf alle 5 imputierten Datensätze sowie den Mittelwert über diese auf.

Anhand der Tabelle können zum einen die Clusteralgorithmen und Distanzmaße miteinander verglichen werden, zum anderen werden Unterschiede zwischen den einzelnen generierten Datensätzen deutlich.

## 5.1 Vergleich der Distanzmaße

Bei Betrachtung von Abb.5.1 fallen unter dem Aspekt des Vergleiches der Distanzmaße folgende Fakten auf:

- (i) Die kophenetischen Korrelationskoeffizienten für die Clusterverfahren auf Basis von metrischen Distanzmaßen unterscheiden sich deutlich von den Korrelationen auf Basis binärer Distanzmaße.
- (ii) Unter dem Pearsonschen Korrelationskoeffizienten als Distanzmaß (bzw. Ähnlichkeitsmaß) spiegeln die Clusterlösungen aller Verfahren am besten die ursprünglich gegebenen Distanzen zwischen den Variablen wider.

Kophenetische Korrelationskoeffizienten						
		Tanimoto	Dice	Euklid	City-Block	Korrelation
Average Linkage	impute1	0.8173	0.8004	0.7663	0.7584	0.8857
	impute2	0.8235	0.8078	0.7695	0.7599	0.8887
	impute3	0.8142	0.7975	0.7666	0.7576	0.8852
	impute4	0.8178	0.8011	0.7537	0.7576	0.8870
	impute5	0.8178	0.8010	0.7721	0.7619	0.8860
	Mittelwert	0.8181	0.8016	0.7656	0.7591	0.8865
Complete Linkage	impute1	0.6772	0.6514	0.5926	0.5586	0.8582
	impute2	0.6903	0.6659	0.6952	0.5591	0.8565
	impute3	0.7436	0.7266	0.6686	0.5789	0.8533
	impute4	0.7341	0.7132	0.5588	0.5411	0.8408
	impute5	0.7075	0.6846	0.5944	0.6054	0.8634
	Mittelwert	0.7105	0.6883	0.6219	0.5686	0.8544
Single Linkage	impute1	0.6960	0.6763	0.6335	0.6520	0.7103
	impute2	0.7013	0.6825	0.6576	0.6523	0.7476
	impute3	0.6800	0.6902	0.6330	0.6517	0.7146
	impute4	0.6998	0.6799	0.6111	0.6326	0.7023
	impute5	0.7060	0.6900	0.6769	0.6679	0.7792
	Mittelwert	0.6966	0.6838	0.6424	0.6513	0.7308
Ward	impute1	0.6895	0.7117	0.6421	0.6611	0.8318
	impute2	0.7099	0.7087	0.6507	0.6656	0.8373
	impute3	0.7127	0.6938	0.6474	0.6634	0.8346
	impute4	0.6884	0.7098	0.6619	0.6650	0.8377
	impute5	0.7014	0.7216	0.6524	0.6693	0.8371
	Mittelwert	0.7004	0.7091	0.6509	0.6649	0.8357

Abbildung 5.1: Kophenetische Korrelationen für 5 imputierte Datensätze sowie Mittelwerte. Die Extremwerte (Maximum und Minimum) sind besonders gekennzeichnet (blau: niedrigster Wert, orange: höchster Wert).

	Mittelwert	Standardabw.	Minimum	Maximum	Spannweite
Tanimoto	0.8574	0.0472	0.5938	0.9695	0.3757
Dice	0.7534	0.0704	0.4223	0.9408	0.5186
Euklid	55.2911	8.7136	30.7409	86.683	55.9425
City-Block	1246.7	258.1	510.0	2246.0	1736.0
Korrelation	0.9596	0.1792	0.3240	1.4513	1.1273

Tabelle 5.1: Statistiken verschiedener Distanzmaße

- (iii) Obwohl die Clusterstruktur unter Anwendung von Tanimoto-Koeffizienten und Dice-Koeffizienten nahezu gleich ist (bzw. sehr starke Ähnlichkeit aufweist), wird die Distanzstruktur unter Anwendung des Tanimoto-Koeffizienten in der endgültigen Clusterlösung im Allgemeinen besser repräsentiert als unter Anwendung des Dice-Koeffizienten.

Abb.5.2 zeigt das Streudiagramm der Distanzen  $d_{i,j}$  der Distanzmatrix und den kophenetischen Distanzen  $d_{i,j}^*$  bei Anwendung des Average-Linkage-Verfahrens auf Basis des Tanimoto-Koeffizienten und des Dice-Koeffizienten. Die Strukturähnlichkeit ist hier deutlich zu erkennen. Da jedoch aufgrund der insgesamt niedrigen Anzahl an Überein-

## 5 Analyse der Clusterverfahren anhand der kophenetischen Korrelationskoeffizienten

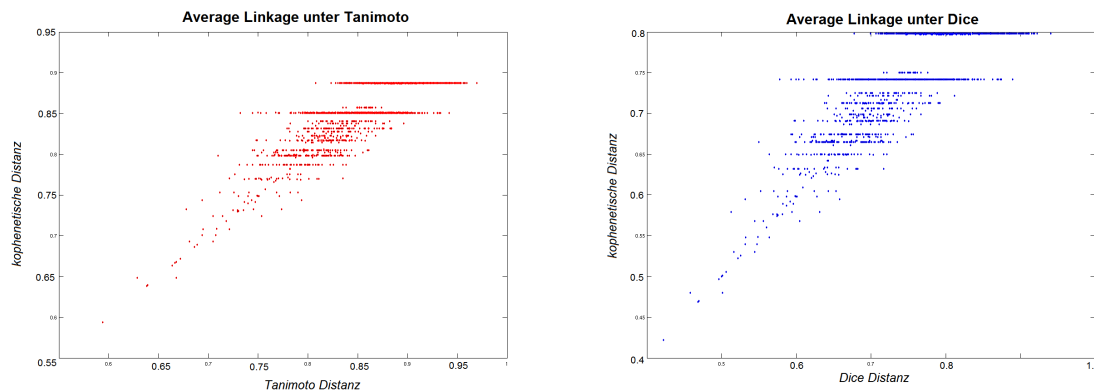


Abbildung 5.2: Streudiagramm der Distanzen  $d_{i,j}$  der Distanzmatrix und den kophenetischen Distanzen  $d_{i,j}^*$  bei Anwendung des Average-Linkage-Verfahrens auf Basis des Tanimoto-Koeffizienten (links) und des Dice-Koeffizienten (rechts)

stimmungen in den Variablen die Spannweite des Dice-Koeffizienten größer ist als die des Tanimoto-Koeffizienten (s. *Tabelle 5.1, Abb.5.2*), weist der kophenetische Korrelationskoeffizient bei Verwendung des Dice-Koeffizienten im Allgemeinen etwas kleinere Werte auf als bei Verwendung des Tanimoto-Koeffizienten.

Auch die Streudiagramme bezüglich der Euklidischen Distanz und der City-Block-Metrik weisen eine ähnliche Struktur auf. Zu beobachten ist jedoch, dass die Werte (relativ) stärker streuen als bei Verwendung von binären Distanzmaßen, was dazu führt, dass der kophenetische Korrelationskoeffizient niedrigere Werte aufweist.

Im allgemeinen sind die Unterschiede zwischen den bisher betrachteten Distanzmaßen bezüglich der erhaltenen Clusterlösungen für den vorliegenden Datensatz nicht sehr groß. *Abb.5.4* zeigt diese Tatsache sehr deutlich: Die Graphik stellt eine Scatterplotmatrix dar, in der die paarweisen Distanzen bezüglich der einzelnen Proximitätsmaße gegeneinander geplottet werden. Zugrunde liegender Datensatz ist wieder der Datensatz "impute1". Deutlich zu erkennen ist im vorliegenden Fall eine lineare Abhängigkeit zwischen der Euklidischen Distanz und der City-Block-Metrik sowie eine (fast) lineare Abhängigkeit zwischen Tanimoto- und Dice-Koeffizienten. Letztere steht nicht im Widerspruch zur *Abb.4.1*, da durch die Binärtransformation, der die ordinalskalierten Daten unterworfen wurden, der Anteil an positiven Übereinstimmungen sehr gering ist und aufgrund dessen nur ein sehr kleiner Ausschnitt aus der in *Abb.4.1* dargestellten Kurve zur Geltung kommt. Daher kann die nichtlineare Abhängigkeit zwischen Tanimoto- und Dice-Koeffizienten nicht unmittelbar beobachtet werden.

## 5 Analyse der Clusterverfahren anhand der kophenetischen Korrelationskoeffizienten

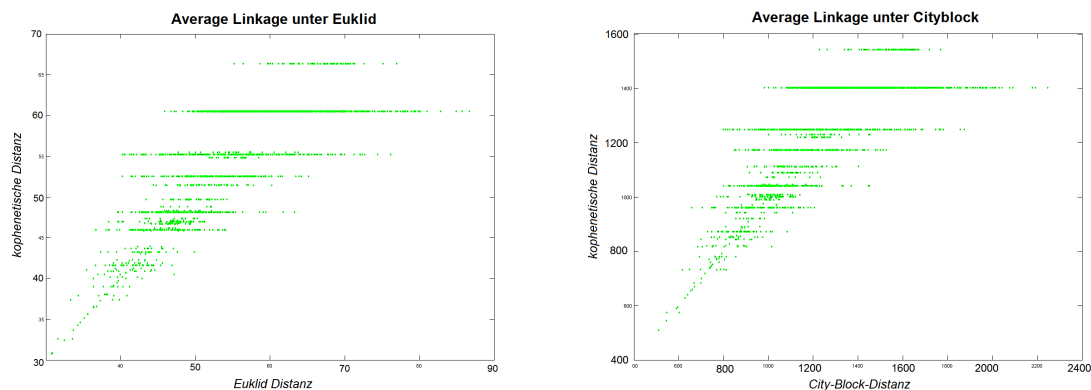


Abbildung 5.3: Streudiagramm der Distanzen  $d_{i,j}$  der Distanzmatrix und den kophenetischen Distanzen  $d_{i,j}^*$  bei Anwendung des Average-Linkage-Verfahrens auf Basis der Euklidischen Distanz (links) und der City-Block-Metrik (rechts)

Interessant ist auch die Abhängigkeitsstruktur zwischen den binären und metrischen Distanzmaßen: Sie scheint schwach quadratischer Natur zu sein. Dies ist nicht selbstverständlich, denn die Anzahl positiver Übereinstimmungen in den Beurteilungen der Variablen sagt nichts über die Größe der Differenz zwischen diesen Werten aus.

Allgemein lässt sich also feststellen, dass die Wahl der in Kapitel 4 analysierten Distanzmaße aufgrund der gegebenen Abhängigkeitsstruktur keinen wesentlichen Einfluss auf die endgültigen Clusterlösungen hat.

Anders verhält es sich bei Verwendung des Korrelationskoeffizienten von Pearson. So ist in Abb.5.4 zu erkennen, dass zwischen diesem Proximitätsmaß und den bisher untersuchten binären und metrischen Proximitätsmaßen kein deutlicher (linearer) Zusammenhang besteht. Die Ursache dafür ist die Verteilung der Korrelationen  $r_{i,j}$  zwischen den Variablen: So treten nur selten Korrelationen nahe bei Null auf, was zur Folge hat, dass bei Bildung der Distanz zwischen den Variablen ( $d_{i,j} = 1 - r_{i,j}$ ) Werte um Eins seltener zu beobachten sind als Werte um 0.8 ( $\cong r_{i,j} = 0.2$ ) oder 1.2 ( $\cong r_{i,j} = -0.2$ ). Die Hauptdiagonale der Scatterplotmatrix zeigt die Häufigkeitsverteilungen der jeweiligen Distanzen. Aufgrund dieser Struktur erfolgt eine klarere Trennung der Variablen in 2 Gruppen. Die Variablen innerhalb einer Gruppe weisen untereinander positive Korrelation auf, die Variablen zwischen den Gruppen negative Korrelation. Die deutlichere Trennung der Objekte kann eine mögliche Ursache für die hohen Werte der zugehörigen kophenetischen Korrelationskoeffizienten sein.

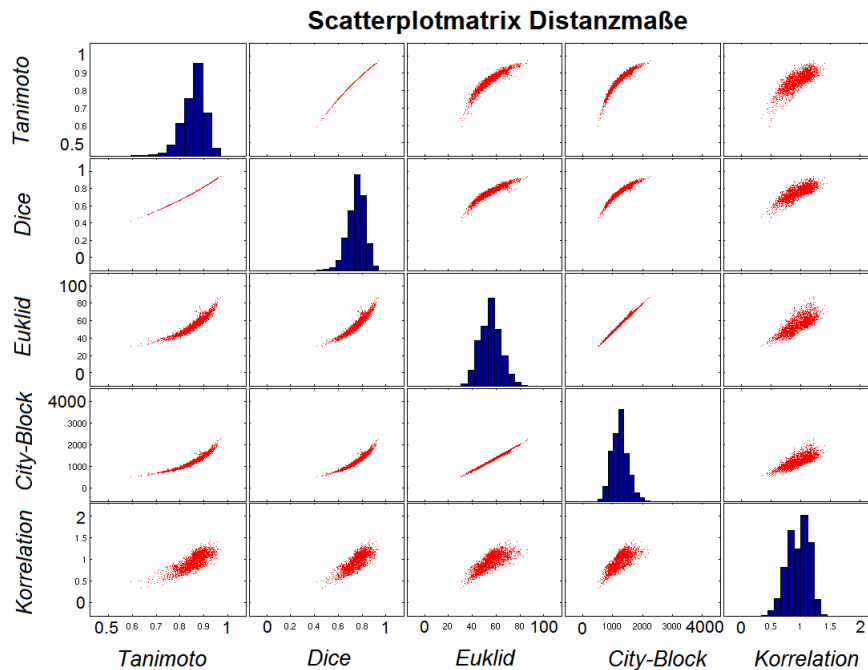


Abbildung 5.4: Vergleich der Distanzmaße in einer Scatterplotmatrix: Plotten der paarweisen Distanzen bzgl. der Proximitätsmaße

### 5.1.1 Bemerkungen

Das Problem der Wahl eines geeigneten Distanzmaßes muss vor der Datenanalyse aufgrund interpretatorischer Überlegungen gelöst werden: Wie soll die Definition einer Distanz zwischen zwei Variablen erfolgen? Ist es überhaupt sinnvoll, ein ordinales Skalenniveau anzunehmen, wenn die Anzahl der Kategorien "groß genug" ist? In den hier untersuchten Datensätzen ("impute1" bis "impute5") stellt sich die Frage, einen Korrelationskoeffizienten als Proximitätsmaß zu verwenden. Auf dessen Basis findet eine deutlichere Trennung zwischen negativ korrelierten Variablen statt, was dazu führt, dass die Objekte stärker voneinander isoliert werden und daher der kophenetische Korrelationskoeffizient insgesamt höhere Werte aufweist als unter Anwendung der bisherigen Distanzmaße.

Für den Vergleich mit den Ergebnissen einer Faktorenanalyse, die auf Bildung des Pearsonschen Korrelationskoeffizienten basiert, wird daher dieser auch als Proximitätsmaß für die Clusteranalyse verwendet.

Die endgültige Interpretation der Clusterlösungen (*s. Kapitel 6*) aus Kapitel 4 erfolgt jedoch unter Voraussetzung ordinalen Skalenniveaus auf Basis des Tanimoto-Koeffizienten.

## 5 Analyse der Clusterverfahren anhand der kophenetischen Korrelationskoeffizienten

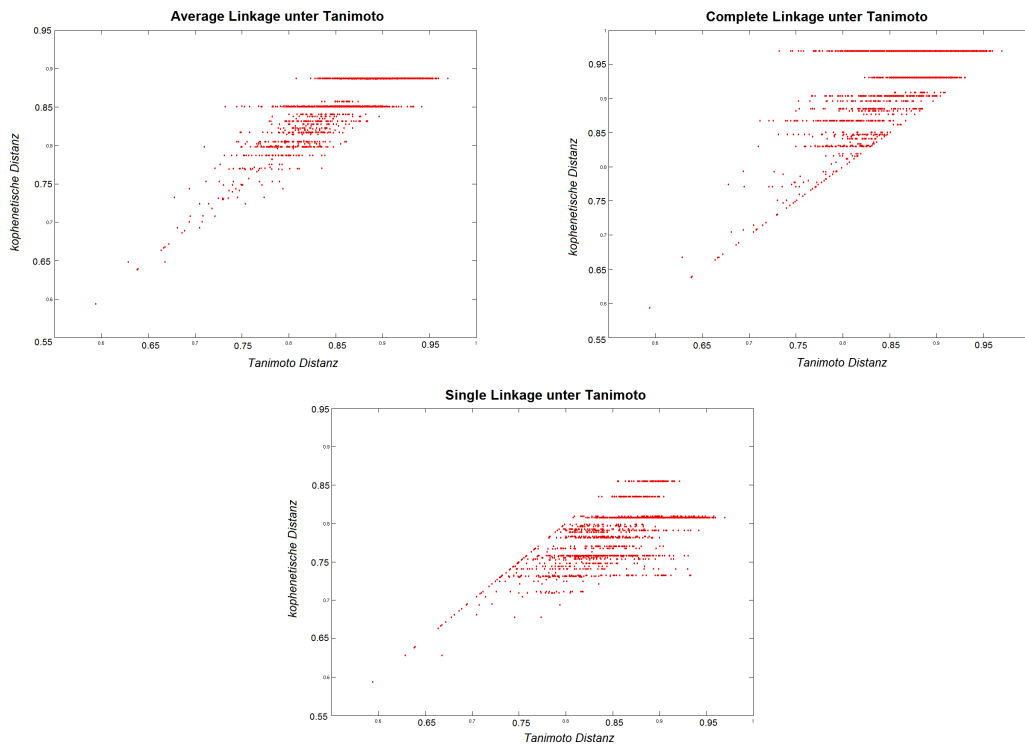


Abbildung 5.5: Streudiagramm der Distanzen  $d_{i,j}$  der Distanzmatrix und den kophenetischen Distanzen  $d_{i,j}^*$  bei Anwendung des Average-Linkage- (links), des Single-Linkage- (mitte) und des Complete-Linkage-Verfahrens (rechts) auf Basis des Tanimoto-Koeffizienten

### 5.2 Vergleich der Verfahren

Ausgehend von den kophenetischen Korrelationskoeffizienten liefert das Average-Linkage-Verfahren die beste Anpassung an die ursprüngliche Distanzmatrix (s. Abb.5.1). Hier ist zu erkennen, dass sich die kophenetischen Korrelationen für das Single-Linkage-Verfahren und das Complete-Linkage-Verfahren im allgemeinen (mit Ausnahme der Verwendung der City-Block-Metrik) nicht stark voneinander unterscheiden. Abb.5.5 stellt die Streudiagramme der Tanimoto-Distanzen gegen die kophenetischen Distanzen jeweils unter Anwendung des Average-Linkage-, des Complete-Linkage- und des Single-Linkage-Algorithmus dar.

Hier sind die Charakteristika der einzelnen Strukturbildungsprozesse ("nächster Nach-

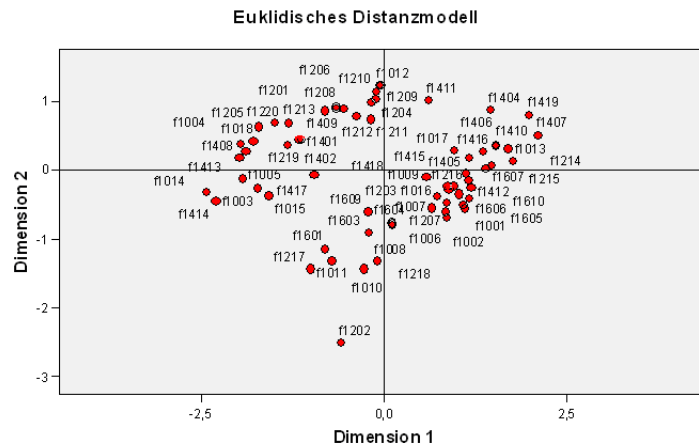


Abbildung 5.6: Relative Lage der Variablen im euklidischen Raum auf Basis der Euklidischen Distanzen

bar”, ”entferntester Nachbar”,...) deutlich erkennbar. Die Tatsache, dass der kophenetische Korrelationskoeffizient sowohl unter Anwendung des Single-Linkage-Verfahrens als auch unter Anwendung des Complete-Linkage-Verfahrens eher mittlere Werte annimmt, lässt die Vermutung zu, dass die Variablengruppen weder voneinander isoliert sind (so dass Kettenbildung erfolgen kann) noch im Innern eine kohärente Struktur aufweisen (also visuell keine klaren Cluster erkennbar sind).

Abb.5.6 stellt ein mittels Multidimensionaler Skalierung gewonnenes Distanzmodell zur Veranschaulichung der auf Basis der Euklidischen Distanzen bestehenden relativen Lagen der einzelnen Variablen zueinander dar. Die Graphik wurde mit dem Programm Alscal in SPSS 14.0 erzeugt. Hier ist zum einen eine Trennung zweier Gruppen zu erkennen, wobei jedoch zum anderen deutlich wird, dass diese Trennung nicht ”klar” verläuft, das heißt, die Existenz von ungünstigen Objekten zwischen den Gruppen bewirkt eine Kettenbildung.

## 6 Interpretation der Clusterstrukturen

Während in den vorhergehenden Kapiteln der Schwerpunkt der Untersuchungen auf mögliche Differenzen innerhalb der Clusterbildungsprozesse der hierarchischen Fusionsalgorithmen gelegt und anhand dessen der Ansatz einer Beurteilung der Verfahren unternommen wurde, sollen in diesem Abschnitt die Unterschiede der Verfahren bezüglich der Interpretation der beobachteten Clusterstrukturen herausgearbeitet werden.

Aufgrund der bestehenden Zusammenhänge zwischen den einzelnen Distanzmaßen (s. *Abb.5.1*) wird an dieser Stelle nur der Tanimoto-Koeffizient als Proximitätsmaß für die vorhandene ordinale Skalierung der Daten herangezogen.

Eine Interpretation der entstandenen Gruppierungen ist nur sinnvoll, wenn sie sich auf den originalen Datensatz bezieht. Die Analyse eines imputierten Datensatzes reicht daher nicht aus, um allgemeinere Aussagen treffen zu können. Vielmehr werden die Verfahren auf jeden imputierten Datensatz angewendet und anschließend entschieden, welche Variable insgesamt bei vorgegebener Clusterzahl welchem Cluster zugeordnet werden soll. Die Entscheidung einer solchen Zuordnung wird dann getroffen, wenn sie das Ergebnis in mehr als der Hälfte der generierten Datensätze (im vorliegenden Fall also mindestens 3) ist.

Eine Aussage über die Anzahl der entstandenen Cluster für jedes Verfahren wird hier mittels der visuellen Veranschaulichung durch die jeweiligen Dendrogramme (aller imputierten Daten) gewonnen. So lassen sich unter Anwendung des Single-Linkage-Verfahrens 2 Cluster erkennen, unter dem Average-Linkage-Algorithmus sind es 2 bis 3 Gruppen und unter dem Complete-Linkage-Verfahren 3 bis 4 Cluster (s. *Anhang A.4.1*).

Cluster 1	Cluster 2	ohne Zuordnung
f1001, f1002, f1006-f1011 f1013, f1016, f1017,	f1003-f1005, f1012, f1015, f1018,	f1014
f1203, f1207, f1214-f1218,	f1201, f1204-f1206, f1208-f1213, f1219,	f1202 f1220
f1404-f1407, f1410-f1412, f1415, f1416 f1418-f1419,	f1401-f1403, f1408, f1409, f1413, f1414, f1417	
f1601-f1610		

Tabelle 6.1: Clusterzuordnungen unter Single-Linkage-Verfahren

## 6.1 Ergebnisse

### 6.1.1 Single-Linkage

Unter dem Single-Linkage-Verfahren ist bei allen 5 Datensätzen eine starke Kettenbildung zu beobachten, da die Variablen nicht stark genug voneinander isoliert sind. Aus diesem Grund lassen sich nur zwei Cluster herauskristallisieren. Tabelle 6.1 führt die Variablen in diesen beiden Clustern sowie alleinstehende Objekte (können als Ausreißer angesehen werden) auf. Die hinter den Variablennamen stehenden Aussagen sind dem Anhang A.1 zu entnehmen.

Von Interesse ist nun die Interpretation dieser Gruppen und die Fragestellung, ob die betrachteten Aussagen innerhalb eines Clusters eine inhaltliche Struktur besitzen.

Die beiden Gruppen lassen sich folgendermaßen beschreiben:

Cluster 1: enthält Variablen mit einer positiven Grundkonnotation

Cluster 2: enthält Variablen mit einer negativen bzw. eher negativ zu bewertenden Grundkonnotation

Hier wird das bereits in Kapitel 3.1 angedeutete Problem der Distanzdefinition deutlich: Da unter den verwendeten Distanzmaßen nur Variablen als ähnlich angesehen werden, deren Beurteilungswerte nahe beinander liegen bzw. hohe Übereinstimmungen zeigen,

weisen Variablen mit vergleichbarem Inhalt und aber entgegengesetzter Konnotation hohe Distanzwerte auf.

So werden beispielsweise die Variablen

f1004 "Es wird oft über meinen Kopf hinweg entschieden" (negativ zu bewertende Konnotation) und

f1207 "Ich habe hinreichende Mitbestimmungsmöglichkeiten" (positive Konnotation)

aufgrund ihrer hohen Distanzwerte verschiedenen Clustern zugeordnet, obwohl sie inhaltliche Zusammenhänge zeigen.

### 6.1.2 Complete Linkage

Unter dem Complete-Linkage-Algorithmus sind visuell 3 bis 4 Cluster zu erkennen. Tabelle 6.2 zeigt die jeweiligen Zuordnungen der Variablen.

Cluster 1: enthält Variablen mit einer positiven Grundkonnotation hinsichtlich der innerschulischen Qualitätskontrolle und Reflexion der pädagogischen Arbeit

Cluster 2: Variablen mit positiver Grundkonnotation hinsichtlich der Bewertung des Arbeitsklimas, der individuellen Lehrerkompetenz und der innerschulischen Hierarchie

Cluster 3: Variablen mit negativer Konnotation, die tendenziell ein statisches "Verharren" in eingefahrenen Denk- und Handlungsmustern begünstigen

### 6.1.3 Average Linkage

Ein Vergleich der Tabellen 6.1-6.3 zeigt, dass alle Verfahren im Wesentlichen dieselben Gruppierungen erzeugen. So sind die Cluster 1 und 2(a und b) des Average-Linkage-Verfahrens konform mit denen des Single-Linkage-Verfahrens. Die Cluster 2a und 2b

6 Interpretation der Clusterstrukturen

Cluster 1	Cluster 2	Cluster 3a	Cluster 3b	ohne Zuordnung
f1001, f1002 f1006-f1008 f1010, f1011,	f1009, f1013 f1016, f1017	f1003-f1005, f1014, f1015, f1018	f1012,	
f1217, f1218,	f1203, f1207, f1214-f1216,	f1205, f1208, f1220,	f1201, f1204, f1206, f1219, f1209-f1213,	f1202
	f1404-f1407 f1410-f1412, f1415, f1416, f1418, f1419,	f1401, f1402, f1408, f1413, f1414, f1417	f1403, f1409	
f1601, f1603, f1604, f1609	f1602, f610 f1605-f1608,			

Tabelle 6.2: Clusterzuordnungen unter Complete-Linkage-Verfahren

Cluster 1	Cluster 2a	Cluster 2b	ohne Zuordnung
f1001, f1002 f1006-f1011 f1013, f1016, f1017,	f1003-f1005, f1014, f1015, f1018,	f1012,	
f1203, f1207, f1214-f1218,	f1205, f1208, f1213, f1219,	f1201, f1204, f1206, f1209-f1212,	f1202 f1220
f1404-f1407, f1410-f1412, f1415, f1416, f1418, f1419,	f1408, f1413	f1401-f1403, f1409	
f1601-f1610			

Tabelle 6.3: Clusterzuordnungen unter Average-Linkage-Verfahren

## 6 Interpretation der Clusterstrukturen

entsprechen größtenteils den Clustern 3a und 3b des Complete-Linkage-Algorithmus. Lediglich die Reihenfolge, in der die Unterteilung erfolgt und die Menge der "Ausreißer" unterscheidet sich hier.

Letztere Differenz ist der Grund dafür, dass der Kappa-Koeffizient zur Beurteilung der Übereinstimmungen der Algorithmen bei Analyse der Kontingenztabellen unter Vorgabe einer bestimmten Clusterzahl oft niedrige Werte annimmt. Hier zeigt sich wiederum, dass eine solche Vergleichsmethodik bei iterativen Prozessen nicht sinnvoll und nur schwer interpretierbar ist.

# 7 Vergleich mit Ergebnissen einer Faktorenanalyse

Die Faktorenanalyse gehört wie die Clusteranalyse zu den strukturentdeckenden Verfahren. Ziel einer Faktorenanalyse ist jedoch nicht die Aufteilung von Objekten in möglichst in sich homogene und untereinander heterogene Gruppen, sondern die Entdeckung von hinter den Variablen stehenden Faktoren. Im Unterschied zu einer Clusteranalyse weisen die Variablen zu jedem Faktor eine bestimmte "Korrelation" (Faktorladung) auf. Das heißt, eine Trennung der Faktoren im Sinne einer Trennung von Variablengruppen findet hier nicht statt.

Dadurch, dass die betrachteten Objekte nicht ausschließlich einem Faktor zugeordnet werden können, ist ein direkter Vergleich der beiden Methodiken nicht möglich. Ein weiteres Problem besteht in der Distanzdefinition bei den Clusterverfahren: Während hier negative Korrelationen zwischen Variablen zu einem hohen Distanzwert und damit zur Einteilung in unterschiedliche Cluster führen, tritt dieses Problem bei einer Faktorenanalyse nicht auf, da die Variablen auch hohe *negative* Faktorladungen besitzen können.

Von Interesse bei dem Vergleich der beiden Methodiken ist die Entdeckung gleicher Strukturelemente und die Untersuchung des Zusammenhanges zwischen den Faktorladungen der Variablen und ihrer Gruppenzuordnung bei den Clusteralgorithmen.

## **Vorgehen bei der Faktorenanalyse:**

Die hier mit SPSS 14.0 durchgeführte Faktorenanalyse basiert auf der Voraussetzung metrischer Daten. Als zugrunde liegendes Korrelationsmaß dient der Pearsonsche Korrelationskoeffizient, die Anzahl der Faktoren wird anhand des Screeplots bestimmt. Der Screeplot weist in allen 5 imputierten Datensätzen einen "Knick" an der Stelle 8 auf (s. *Anhang A.3.1*), dies sei die gewählte Anzahl der Faktoren. Abb.7.1 zeigt den Screeplot für den Datensatz "impute1".

Die Faktorladungsmatrix wird mittels *Varimax* rotiert. Die so erhaltenen Faktorladungen sowie SPSS-Outputs sind dem *Anhang A.3.1* entnehmbar.

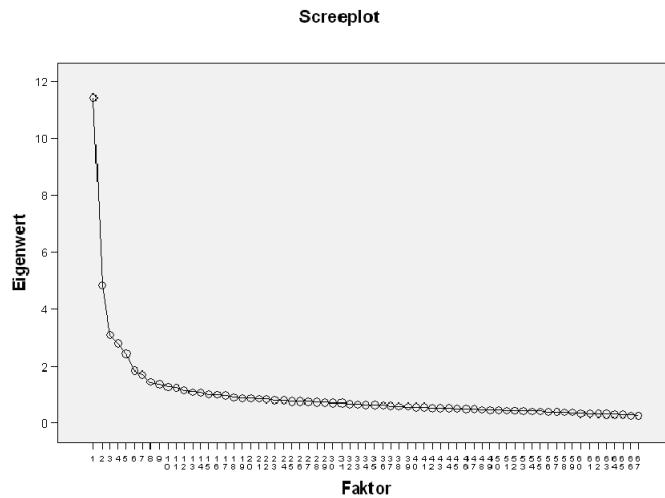


Abbildung 7.1: Screeplot für den Datensatz "impute1"

Um einen Vergleich der Methodiken möglich zu machen, werden die einzelnen Variablen den Faktoren zugeordnet, zu denen sie die höchsten Faktorladungen aufweisen. Tabelle 7.1 zeigt die Zuordnung der Variablen zu den einzelnen Faktoren. Variablen mit insgesamt niedrigen Faktorladungen (absolute Werte  $< 0.4$ ) bzw. daraus resultierender nicht eindeutiger Faktorzugehörigkeit sind in der Tabelle kursiv gedruckt.

Damit die Ergebnisse der Faktorenanalyse mit denen der Clusteranalyse vergleichbar sind, wird bei der Beschreibung der Faktoren die Konnotation der Variablen angegeben (diese entspricht im Allgemeinen dem Vorzeichen der jeweiligen Faktorladungen).

- F1: Bewertung der innerschulischen Entscheidungsprozesse und allgemeinen Evaluationspraxis (positive Konnotation)
- F2: Bewertung der hierarchischen Grenzen und persönlichen Handlungsspielräume (negative Grundkonnotation)
- F3: Bewertung äußerer Vorschriften (negative Konnotation) und Bewertung des innerschulischen Austausches und Zusammenhaltes (positive Konnotation)
- F4: Bewertung der Notwendigkeit innerschulischer Modernisierung und der dafür notwendigen persönlichen Leistungsbereitschaft
- F5: Bewertung des Einflusses der übergeordneten Ebene

F1	F2	F3	F4	F5	F6	F7	F8	o.Z.
f1001	f1003	f1005	f1012	f1601	f1018,	f1219,	f1404	<i>f1220</i>
f1002	f1004	f1013	<i>f1014</i>	bis	f1205	f1401	f1405	<i>f1410</i>
f1006	f1015	<i>f1204</i>	f1017	f1610	<i>f1209</i>	bis	f1411	
bis	f1016,	f1214	f1201,		bis	f1403	f1419	
f1011,	<i>f1206</i>	bis	f1406		f1212	f1408		
<i>f1202</i>	bis	f1216,	f1412			f1409		
<i>f1203</i>	f1208	f1407	f1415			f1413		
f1217	f1213					f1414		
f1218,						f1417		
<i>f1416</i>								
f1418								

Tabelle 7.1: Einteilung der Variablen in Faktoren, zu denen sie die höchsten Faktorladungen aufweisen

F6: Bewertung der im Lehrerkollegium vorhandenen Veränderungsbereitschaft (negative Konnotation)

F7: Bewertung der individuell empfundenen Belastung durch den gegenwärtigen Zustand (tendenziell negative Konnotation)

F8: Beurteilung des eigenverantwortlichen Handelns

### Ergebnisse der Clusteranalyse

Als zugrunde liegendes Proximitätsmaß für die Clusteranalyse wird aufgrund der Vergleichbarkeit der Methodiken der Pearsonsche Korrelationskoeffizient verwendet. In der im Anhang A.3.2 gelisteten Datei werden die Clusterzuordnungen unter dem Average-Linkage-Verfahren und dem Complete-Linkage-Verfahren aufgeführt. Da Unterschiede bezüglich einzelner Variablen nur von geringer Bedeutung sind, werden in Tabelle 7.2 Variablengruppen aufgeführt, die sich sowohl unter Anwendung der Clusterverfahren als auch nach Durchführung einer Faktorenanalyse innerhalb einer Gruppe befinden.

Der Fokus der Untersuchung soll nun auf diese gemeinsamen Gruppen gelegt werden. Es fällt auf, dass die Variablen innerhalb der Gruppen eine gleichgerichtete Konnotation, die sich in den jeweiligen Faktorladungen widerspiegelt, aufweisen.

Abb.7.2 führt die Faktorladungen der "gemeinsamen" Variablen bezüglich der Faktoren

7 Vergleich mit Ergebnissen einer Faktorenanalyse

F1	F5	F7	F6	F 3
f1001	f1601	f1219,	f1018,	f1013,
f1002	bis	f1401	f1205	f1214
f1006	f1610	bis	f1210	bis
bis		f1403	bis	f1216,
f1008		f1408	f1212	f1407
f1010		f1409		
f1011,		f1413		
f1217		f1414		
f1218		f1417		

Tabelle 7.2: Variablen, die sowohl bei Anwendung des Average-Linkage- und Complete-Linkage-Verfahrens als auch nach Durchführung einer Faktorenanalyse gemeinsamen Gruppen zugeordnet werden

**Faktorladungen gemeinsamer Gruppen**

F1	imp1	imp2	imp3	imp4	imp5
f1001	0.604	0.639	0.627	0.651	0.604
f1002	0.519	0.544	0.514	0.546	0.519
f1006	0.732	0.704	0.727	0.714	0.732
f1007	0.647	0.689	0.674	0.693	0.647
f1008	0.7	0.679	0.659	0.658	0.7
f1217	0.484	0.457	0.51	0.517	0.484
f1218	0.488	0.455	0.485	0.515	0.488
F3	imp1	imp2	imp3	imp4	imp5
f1013	0.542	0.596	0.155	0.22	0.542
f1214	0.575	0.636	0.186	0.37	0.575
f1215	0.615	0.628	0.172	0.367	0.615
f1216	0.46	0.517	0.08	0.162	0.46
f1407	0.484	0.454	0.349	0.411	0.484

Abbildung 7.2: Faktorladungen der "gemeinsamen Gruppen" bezüglich der Faktoren F1 und F3

F1 und F3 für jeden imputierten Datensatz auf (Die Ladungen für die weiteren Variablengruppen F5 bis F7 sind dem Anhang A.3.2 zu entnehmen). Diese Faktorladungen verzeichnen im Allgemeinen (bzw. in der Mehrheit der imputierten Datensätze) relativ hohe Werte. Negative Faktorladungen treten nicht auf. Der Grund dafür liegt in der bereits erläuterten Definition der Distanzen, auf denen die Clusterung der Variablen basiert.

Ein erstes Ergebnis des Vergleiches der Strukturentdeckung mittels Clusteranalyse und Faktorenanalyse lässt sich daher formulieren: Das Vorliegen hoher Faktorladungen impliziert eine deutlichere inhaltliche Trennung der einzelnen Faktoren voneinander. In einem solchen Fall führt die Anwendung einer Clusteranalyse auf Basis eines geeigneten Distanzmaßes zu ähnlichen Ergebnissen.

Im vorliegenden Fall liefern die Clusterverfahren aufgrund der ungeeigneten Distanzdefinition ein "verzerrtes" Bild der Gruppenstrukturen. Dies hat zum einen eine Begünstigung möglicher Fehlinterpretationen zur Folge, zum anderen sind die Ergebnisse der Clusterverfahren nicht mehr direkt vergleichbar mit denen einer Faktorenanalyse.

## 8 Zusammenfassung

Ergebnis der vorhergehenden Analysen ist weniger die Aufdeckung und Interpretation der mittels verschiedener Clusterverfahren erhaltenen Gruppenstrukturen als vielmehr die Untersuchung der Methodiken bezüglich des Vergleiches iterativer (Cluster-) Verfahren.

So lässt sich über den in Kapitel 4 durchgeführten Vergleich der Algorithmen mittels Kontingenztabellen konstatieren, dass diese Vergleichsmethodik (wie bereits in Kapitel 3.1 angedeutet) bei iterativen Prozessen nicht sinnvoll ist, wenn nicht der gesamte Fusionierungsprozess betrachtet wird.

Vor allem die Existenz von Ausreißern bzw. isolierten Variablen kann bei dem paarweisen Vergleich zweier Verfahren zu einer deutlichen Verminderung des Kappa-Koeffizienten führen, ohne dass jedoch Rückschlüsse auf die wesentlich zu erkennenden Strukturen und Gemeinsamkeiten der miteinander verglichenen Verfahren gemacht werden können.

Das Problem bei der durchgeführten Datenanalyse liegt darin, dass nicht nur *ein* Datensatz vorhanden ist, bei dem mögliche Ausreißer eliminiert werden können. Eine solche Eliminierung bei jedem Datensatz würde wiederum zu Verzerrungen in den einzelnen Fusionierungsprozessen führen.

Die mittels kophenetischen Korrelationskoeffizienten durchgeführten Analysen in Kapitel 5 liefern eine Vorstellung davon, wie gut die Clusterlösungen der einzelnen Verfahren die zugrunde liegenden Distanzen abbilden. Auf der Basis eines bestimmten Distanzmaßes stellt daher ein Vergleich der Güte der verschiedenen Algorithmen anhand dieses Koeffizienten eine durchaus sinnvolle Methode dar. Bei dem paarweisen Vergleich zweier Distanzmaße zeigt sich eine klare Abhängigkeitsstruktur zwischen den in Kapitel 4 behandelten Distanzmaßen. Daher ist hier die Wahl des Distanzmaßes nicht so entscheidend wie die Wahl des Clusterverfahrens. Eine solche Abhängigkeit ist jedoch nicht bei dem Vergleich des Pearsonschen Korrelationskoeffizienten mit den übrigen Distanzmaßen zu beobachten. Die Wahl des Distanzmaßes sollte daher ausschließlich aufgrund interpretatorischer Überlegungen getroffen werden.

Im vorliegenden Fall stellte sich heraus, dass die Struktur der Variablen (das Vorliegen von negativ konnotierten *und* positiv konnotierten Aussagen) die Wahl eines Distanzmaßes im herkömmlichen Sinne nur bedingt zulässt, da die Definition der Distanzen abhängig von dem Ziel und der Fragestellung der Analyse ist. So muss vorab geklärt werden, *wann* zwei Variablen inhaltlich eine hohe Ähnlichkeit oder eine hohe Distanz zueinander aufweisen. Eine Möglichkeit der Messung der Ähnlichkeit stellt zum Beispiel auch die Bildung von absoluten Korrelationen dar. In den hier behandelten Datensätzen streuen die Werte der absoluten Korrelationen nicht sehr stark, so dass die Frage gestellt werden muss, ob hier eine Clusteranalyse ein geeignetes Verfahren zur Aufdeckung bestimmter Strukturen darstellt.

Als wesentliches Ergebnis bezüglich einer solchen Strukturanalyse ist im vorliegenden Fall eine 2-Cluster-Lösung anzusehen. Diese begründet sich dadurch, dass unter den zugrunde liegenden Distanzmaßen negativ korrelierenden Variablen relativ hohe Distanzwerte zugeordnet werden. Dies hat zur Folge, dass sich auf der einen Seite Variablen mit negativer Konnotation innerhalb einer Gruppe befinden und auf der anderen Seite Variablen mit positiver Konnotation.

Zu Kapitel 7 (Vergleich mit Faktorenanalyse) ist zu sagen, dass die Intention der Methodiken Clusteranalyse und Faktorenanalyse sich deutlich voneinander unterscheiden. Ein direkter Vergleich ist daher nicht möglich. Jedoch zeigt sich im vorliegenden Fall, dass eine Faktorenanalyse bei gegebener Datenstruktur sinnvoller ist als die Anwendung einer Clusteranalyse, da letztere das Ziel einer klaren Trennung der Variablen verfolgt. Bei der inhaltlichen Interpretation der entstandenen Cluster wird allerdings deutlich, dass eine solche Trennung nicht direkt möglich ist.

## 9 Literatur

- [1] K.Backhaus, B.Erichson, W.Plinke, R.Weiber: *Multivariate Analysemethoden*, Springer (2003)
- [2] A.Bühl, P.Zöfel: *SPSS 12 Einführung in die moderne Datenanalyse unter Windows* (2004)
- [3] H.Büning, G.Trenkler: *Nichparametrische Statistische Methoden*, Verlag de Gruyter (1999)
- [4] A.Handl: *Multivariate Analysemethoden*, Springer (2002)
- [5] W.Härdle, L.Simar: *Applied Multivariate Statistical Analysis*, Springer (2003)
- [6] L.Fahrmeir, A.Hamerle, G.Tutz: *Multivariate statistische Verfahren*, de Gruyter (1996)
- [7] H.Moosbrugger, D.Frank: *Clusteranalytische Methoden in der Persönlichkeitsforschung*, Hans Huber Verlag (1992)
- [8] E.Pari Schatz: *Untersuchung der Ergebnisse der Faktoranalyse bei Anwendung auf ordinale Daten*, Masterarbeit (2005)
- [9] B.Rönz: *Skript zu "Computergestützte Statistik II"*
- [10] J.Schafer: *Analysis of Incomplete Multivariate Data*, Chapman and Hall (1997)
- [11] <http://de.wikipedia.org/wiki/Clusteranalyse> (15.03.2007)

# A Verzeichnis der Dateien

## A.1 Datensätze

Der Ordner enthält alle verwendeten imputierten Datensätze als Excel-Dateien sowie den originalen Gesamtdatensatz und zugehörigen Fragebogen.

## A.2 Clusterzuordnungen und Kontingenztabellen

enthält die Dateien "impute{i}\_analyse.ods",  $i=1,\dots,5$ , sowie die Datei "Gesamtanalyse.ods".

### 2.1: "impute{i}\_analyse.ods"

Die Dateien "impute{i}\_analyse.ods" beinhalten alle Kontingenztabellen für 2 bis 6 Cluster (Tabellenblätter: "2Cluster" bis "6Cluster") sowie die einzelnen Clusterzuordnungen für jede Variable unter jedem Verfahren.

Die Tabellenblätter "2Cluster\_red" bis "5Cluster\_red" führen die Kontingenztabellen nach Entfernung von in der Gesamtanalyse unter dem Single-Linkage-Verfahren identifizierten Ausreißern auf.

In dem jeweils letzten Tabellenblatt "Entwicklung Kappa" werden die jeweiligen Entwicklungen des Kappa-Koeffizienten in Abhängigkeit von der Clusterzahl aufgelistet.

### 2.2: "Gesamtanalyse.ods"

Die Datei "Gesamtanalyse.ods" enthält die durch die Synthese der 5 imputierten Datensätze erhaltenen Clusterzuordnungen jeder Variable bei Vorgabe von 2 bis 5 Clustern sowie die jeweilige Anzahl der Datensätze, bei denen diese Zuordnung vorlag (Tabellenblätter: "Ergebnisse 2Cluster" bis "Ergebnisse 5Cluster").

Desweiteren werden auch hier die insgesamt erhaltenen Kontingenztabellen aufgeführt (Tabellenblätter: "Kreuztab 2Cluster" bis "Kreuztab 5Cluster").

## A.3 Faktorenanalyse und Vergleich mit Clusterlösungen

Der Ordner enthält die SPSS-Viewer-Dateien "imp{i}\_faktor8",  $i=1, \dots, 5$ , sowie die Excel-Datei "ladungen.xls".

### 3.1 "imp{i}\_faktor8"

Die Dateien "imp{i}\_faktor8" beinhalten die SPSS-Outputs für den Datensatz impute{i} nach Durchführung einer Faktorenanalyse unter Vorgabe von 8 Faktoren (rotierte Komponentenmatrix, Screeplot,...).

### 3.2 "ladungen.xls"

Die Datei "ladungen.xls" beinhaltet die Auflistung aller (mit Varimax) rotierten Faktorladungen für jede Variable und jeden imputierten Datensatz (Tabellenblätter: "Ladungen1" bis "Ladungen5"). Dabei sind hohe Faktorladungen farblich gekennzeichnet (rot: über 0.7, blau: zwischen 0.5 und 0.7, gelb: zwischen 0.4 und 0.5).

Das Tabellenblatt "Vergleich mit Clusterlösung" führt auf, wie oft eine Variable einem Faktor zugeordnet wird. Dabei findet eine Zuordnung nur statt, wenn die maximalen Faktorladungen höher als 0.4 sind. Außerdem werden die einzelnen Faktoren sowie mittels Complete-Linkage- und Average-Linkage-Verfahren erhaltenen Cluster mit den in ihnen enthaltenen Variablen aufgelistet. Dabei findet eine Zuordnung aller Variablen statt, Variablen mit niedrigen Faktorladungen werden jedoch besonders gekennzeichnet (blauer Hintergrund).

Das letzte Tabellenblatt "Faktorladungen gem. Gruppen" enthält zu jedem Datensatz die Faktorladungen der Variablen, die sowohl nach Durchführung einer Faktorenanalyse als auch in den betrachteten Clusteranalysen denselben Gruppen zugeordnet wurden.

## A.4 Graphiken

”Graphiken” enthält 4 Unterordner: ”Dendrogramme”, ”Entwicklung Kappa”, ”Scatterplotmatrizen Distanz” und ”Scatterplot kophenetisch”.

### 4.1 Dendrogramme

Hier sind alle verwendeten Dendrogramme für jeden Datensatz enthalten. Die Graphiken sind so aufgebaut, dass sie jeweils 4 Dendrogramme zeigen: Je Verfahren auf Basis der 4 verwendeten Distanzmaße (Tanimoto, Dice, Euklidische Distanz, City-Block-Metrik). Je nach Datensatz und verwendeten Verfahren heißen die Dateien daher ”imp1\_average.png” bzw. ”imp2\_single.png” usw.

Im Ordner des ”impute1” sind zusätzlich einige (nicht alle!) Einzelgraphiken aufgeführt. Das jeweilige verwendete Verfahren und Distanzmaß geht aus dem Namen der Datei hervor (z.B. zeigt die Datei ”imp1dendaveragedice.png” das mit Average-Linkage-Verfahren und Dice erhaltene Dendrogramm).

### 4.2: Entwicklung Kappa

Der Ordner enthält die 6 Graphiken zur Entwicklung der Kappa-Koeffizienten.

### 4.3: Scatterplotmatrizen Distanz

Der Ordner enthält die Scatterplotmatrizen der gegeneinander geplotteten Distanzen für alle 5 imputierten Datensätze.

### 4.4: Scatterplot kophenetisch

Dieser Ordner beinhaltet alle Streudiagramme bezüglich der kophenetischen Distanzen. Dabei zeigt eine Graphik je 4 solcher Scatterplots: bei Anwendung des Average-Linkage-, Complete-Linkage-, Single-Linkage- und Ward-Verfahrens. Die zugrunde liegenden Distanzmaße gehen aus dem Dateinamen hervor (*tan* = *Tanimoto*, *dice* = *Dice*, *cor* = *Korrelation*, *euc* = *EuklidischeDistanz*, *cb* = *City – Block*).

Zusätzlich befindet sich im Ordner ”impute1” eine vollständige Aufführung aller Einzeldiagramme. Das jeweilige verwendete Verfahren und Distanzmaß geht aus dem Namen der Datei hervor (z.B. zeigt die Datei ”cophenetcompcbimp1.png” das mit Complete-Linkage-Verfahren und CityBlock-Metrik erhaltene Streudiagramm).

## **A.5 Sonstiges**

Die Datei "Distanz\_statistik.ods" enthält die Auflistung der in Kapitel 5.1 behandelten Statistiken über die verwendeten Distanzmaße.

"Matlab\_Funktionen.pdf" beschreibt die wichtigsten verwendeten Matlab-Befehle.

"bachelor.pdf" ist die pdf-Datei dieser Bachelorarbeit.