Funktionales Clustern von Transaktionsverläufen

Birgit Oellinger
Funktionales Clustern von Transaktionsverläufen

Master Thesis

verfasst von

Birgit Oellinger

Gutachter:

Prof. Dr. Friedrich Leisch
AG Computationale Statistik

Institut für Statistik
Ludwig–Maximilians–Universität München

25. Januar 2010
Inhaltsverzeichnis

Abbildungsverzeichnis iv
Tabellenverzeichnis viii

1. Einführung 1

2. Funktionale Daten 2

3. Anpassung funktionaler Daten 4
   3.1. Splineexpansion 4
      3.1.1. Fourier-Basis-System 6
      3.1.2. Polynomsplines 7
      3.1.3. Smoothing Splines 9
   3.2. Karhunen-Loève-Expansion 10

4. Clusterverfahren 12
   4.1. Heuristische Ansätze 12
   4.2. Modellbasierte Verfahren 15

5. Methoden zum Clustern funktionaler Daten 18
   5.1. Modellbasierte Verfahren 19
   5.2. Heuristische Ansätze 20

6. Erweiterungen zum Clustern funktionaler Daten mit dem k-medoids-Algorithmus 24
   6.1. Poissonstanz 24
   6.2. Splineclustern 25

7. Funktionales Clustern von Transaktionsverläufen 29
   7.1. Daten und Vorgaben 30
Abbildungsverzeichnis

3.1. Fourier-Basis mit fünf Basisfunktionen auf dem Intervall $T = [0, 12]$ 6
3.2. B-Spline-Basis der Ordnung drei mit fünf inneren Knoten auf dem Intervall $T = [0, 12]$ 8

6.1. Beobachtungen eines Verlaufsmusters mit Centroidvektor und Centroidspline 26
6.2. Ausgewählte Beobachtung eines Verlaufsmusters mit Centroidvektor und Centroidspline I 27
6.3. Ausgewählte Beobachtung eines Verlaufsmusters mit Centroidvektor und Centroidspline II 27

7.1. Histogramm der Einkäufe des Produktes 61 nach Monat 31
7.2. Anzahl gekauftes Produkt 61 über die Zeit mit angepasstem B-Spline der Ordnung 3 und 2 inneren Knoten, KW-Ebene 33
7.3. Anzahl gekauftes Produkt 63 über die Zeit mit angepasstem B-Spline der Ordnung 3 und 2 inneren Knoten, KW-Ebene 34
7.4. Anzahl gekauftes Produkt 61 über die Zeit mit angepasstem B-Spline der Ordnung 3 und 2 inneren Knoten, Monatsebene 35
7.5. Scree-Plot für $k = 3$ bis $k = 10$ Cluster bei Clusterverfahren 4, Produkt 61, KW-Ebene 36
7.6. Verlauf der Clusterzentren, Clustern der Rohdaten mit k-means und euklidischer Distanz, Produkt 63, KW-Ebene 37
7.7. Anzahl gekauftes Produkt 63 nach KW, aufgeteilt nach Cluster, Stichprobe von 50 HH – Clustern der Rohdaten mit k-means (k=4) und euklidischer Distanz; Clusterzentrum in rot eingezeichnet 38
7.8. Verlauf der Clusterzentren – Clustern der Rohdaten mit k-means und Poisson-Distanz, Produkt 63, KW-Ebene 39
<table>
<thead>
<tr>
<th>Abbildungsverzeichnis</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.9. Verlauf der Clusterzentren – Clustern der Splinekoeffizienten (B-Splines Grad 3, 2 innere Knoten) mit k-means (k=4) und euklidischer Distanz, Produkt 63, KW-Ebene</td>
</tr>
<tr>
<td>7.10. Angepasste Splines nach Clusterzuordnung – Clustern der Splinekoeffizienten (B-Splines Grad 3, 2 innere Knoten) mit k-means (k=4) und euklidischer Distanz, Produkt 63 – Stichprobe von 100 HH; Clusterzentren in rot eingezeichnet</td>
</tr>
<tr>
<td>7.11. Verlauf der Clusterzentren – Clustern der vorhergesagten Werte der Splines (B-Splines Grad 3, 2 innere Knoten) an den Erhebungszeitpunkten mit k-means (k=4) und euklidischer Distanz, Produkt 63, KW-Ebene</td>
</tr>
<tr>
<td>7.12. Verlauf der Clusterzentren – Clustern der Rohdaten mit Splineclustern (B-Splines Grad 3, 2 innere Knoten, df=6) und euklidischer Distanz (k=4), Produkt 63, KW-Ebene</td>
</tr>
<tr>
<td>7.13. Verlauf der Clusterzentren – Clustern der Rohdaten mit Splineclustern (B-Splines Grad 3, 2 innere Knoten, df=6) und Poisson-Distanz (k=4), Produkt 63, KW-Ebene</td>
</tr>
<tr>
<td>7.14. Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 61, Monatsebene</td>
</tr>
<tr>
<td>7.15. Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 44, KW-Ebene</td>
</tr>
<tr>
<td>7.16. Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 19, Monatsebene</td>
</tr>
<tr>
<td>7.17. angepasste Splines nach Clusterzuordnung – Clustern der Splinekoeffizienten (B-Splines Grad 3, 2 innere Knoten) mit k-means (k=4) und euklidischer Distanz, poissonverteilt simulierte Daten, Produkt 63, Stichprobe von 100 HH</td>
</tr>
<tr>
<td>7.18. Verlauf der Clusterzentren der als Basis dienenden Cluster für die simulierten Daten, KW-Ebene</td>
</tr>
<tr>
<td>7.19. Verlauf des Randindex für die einzelnen Clusterlösungen nach steigender Anzahl simulierter Beobachtungen – poissonverteilt simulierte Daten, Produkt 61, KW-Ebene</td>
</tr>
<tr>
<td>7.20. Verlauf der Clusterzentren für die einzelnen Clusterlösungen, Produkt 61, poissonverteilt simulierte Daten, KW-Ebene</td>
</tr>
<tr>
<td>7.21. Verlauf der Clusterzentren für die einzelnen Clusterlösungen, Produkt 63, poissonverteilt simulierte Daten, KW-Ebene</td>
</tr>
<tr>
<td>7.22. Verlauf der Clusterzentren für die einzelnen Clusterlösungen, Produkt 61, normalverteilt simulierte Daten, KW-Ebene</td>
</tr>
</tbody>
</table>
Abbildungsverzeichnis

7.23. Verlauf der Clusterzentren – Clustern der Splinekoeffizienten (B-Splines Grad 3, 2 innere Knoten) mit k-means (k=4) und euklidischer Distanz, poissonverteilt simulierte Daten, Produkt 63, KW-Ebene mit eingezeichneten wahren Verläufen der Simulationsbasis .......... 63

7.24. Verlauf der Clusterzentren – Clustern der Rohdaten mit Splineclustern (B-Splines Grad 3, 2 innere Knoten, df=6) und Poisson-Distanz (k=4), poissonverteilt simulierte Daten, Produkt 63, KW-Ebene mit eingezeichneten wahren Clusterzentroiden der Simulationsbasis .......... 64

7.25. Verlauf der Clusterzentren – Clustern der Rohdaten mit Splineclustern (B-Splines Grad 3, 2 innere Knoten, df=6) und euklidischer Distanz (k=4), normalverteilt simulierte Daten, Produkt 63, KW-Ebene mit eingezeichneten wahren Clustercentroiden der Simulationsbasis .......... 65


7.27. Boxplots der Randindizes der Verfahren i und j zueinander, aufgeteilt nach Datensituation E und F .......... 68

7.28. Boxplots des Randindex .......... 70

7.29. Boxplots des Randindex für die unterschiedlichen Verfahren aufgeteilt nach Aggregationsniveaus .......... 72

7.30. multipler Test nach TukeyHSD .......... 73

A.1. Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 63, KW-Ebene .......... 81

A.2. Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 61, KW-Ebene .......... 82

A.3. Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 19, KW-Ebene .......... 83

A.4. Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 63, Monatsebene .......... 84

A.5. Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 44, Monatsebene .......... 85

C.1. Verlauf der Clusterzentren der als Basis dienenden Cluster für die simulierten Daten, Monatsebene .......... 90

D.1. Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 44, poissonsimulierte Daten, KW-Ebene .......... 92

D.2. Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 19, poissonsimulierte Daten, KW-Ebene .......... 93
Abbildungsverzeichnis

D.3. Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 63, poissonsimulierte Daten, Monatsebene ........................................ 94
D.4. Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 61, poissonsimulierte Daten, Monatsebene ........................................ 95
D.5. Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 44, poissonsimulierte Daten, Monatsebene ........................................ 96
D.6. Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 19, poissonsimulierte Daten, Monatsebene ........................................ 97
Tabellenverzeichnis

5.1. Übersicht der Clusterverfahren für funktionale Daten .................. 18

7.1. Vergleich der Clusterlösungen der unterschiedlichen Verfahren, Produkt 63, KW-Ebene, k=4 .................................................. 45

7.2. Randindex zwischen Clusterverfahren i und Clusterverfahren j aufgeteilt nach Produkt und Aggregationsniveau ................................... 50

7.3. Randindex zwischen wahrer Clusterzugehörigkeit und der entsprechenden Clusterlösung nach Anzahl simulierter Beobachtungen pro Cluster, poissonverteilte Daten, Produkt 61, KW-Ebene .......................... 55

7.4. Randindex zwischen den verschiedenen Clusterlösungen, n=500, poissonverteilte Daten .................................................. 58

7.5. Randindex zwischen Clusterverfahren i und Clusterverfahren j aufgeteilt nach Produkt und Aggregationsniveau, poissonverteilte Daten ........ 59

7.6. Randindex zwischen Clusterverfahren i und Clusterverfahren j aufgeteilt nach Produkt und Aggregationsniveau, normalverteilte Daten .... 62

7.7. Randindex zwischen wahrer Klassenzugehörigkeit und den einzelnen Verfahren, poissonverteilte Daten .................................... 64

7.8. Randindex zwischen wahrer Klassenzugehörigkeit und den einzelnen Verfahren, normalverteilte Daten ...................................... 66

7.9. Randindex zwischen verschiedenen Clusterverfahren i – j bei verschiedenen Produkten und Daten; A: Daten, KW-Ebene B: Daten, Monatsebene C: poissonverteilte Daten, KW-Ebene D: poissonverteilte simulierten Daten, Monatsebene E: normalverteilte simulierte Daten, KW-Ebene F: normalverteilte simulierte Daten, Monatsebene ........................................ 67

7.10. Kenngrößen der Randindizes der einzelnen Verfahren ....................... 71

7.11. Test linearer Kontraste .......................................................... 72
<table>
<thead>
<tr>
<th>B.1. Vergleich der Clusterlösungen, Produkt 61, KW-Ebene</th>
<th>86</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.2. Vergleich der Clusterlösungen, Produkt 44, KW-Ebene</td>
<td>86</td>
</tr>
<tr>
<td>B.3. Vergleich der Clusterlösungen, Produkt 19, KW-Ebene</td>
<td>87</td>
</tr>
<tr>
<td>B.4. Vergleich der Clusterlösungen, Produkt 63, Monatsebene</td>
<td>88</td>
</tr>
<tr>
<td>B.5. Vergleich der Clusterlösungen, Produkt 61, Monatsebene</td>
<td>88</td>
</tr>
<tr>
<td>B.6. Vergleich der Clusterlösungen, Produkt 44, Monatsebene</td>
<td>89</td>
</tr>
<tr>
<td>B.7. Vergleich der Clusterlösungen, Produkt 19, Monatsebene</td>
<td>89</td>
</tr>
<tr>
<td>E.1. Randindizes der einzelnen Verfahren zueinander aufgeteilt nach den zugrunde liegenden Daten, Produkt 63</td>
<td>98</td>
</tr>
<tr>
<td>E.2. Randindizes der einzelnen Verfahren zueinander aufgeteilt nach den zugrunde liegenden Daten, Produkt 61</td>
<td>99</td>
</tr>
<tr>
<td>E.3. Randindizes der einzelnen Verfahren zueinander aufgeteilt nach den zugrunde liegenden Daten, Produkt 44</td>
<td>99</td>
</tr>
<tr>
<td>E.4. Randindizes der einzelnen Verfahren zueinander aufgeteilt nach den zugrunde liegenden Daten, Produkt 19</td>
<td>100</td>
</tr>
</tbody>
</table>

Im Bereich der funktionalen Clusteranalyse existieren mehrere unterschiedliche Ansätze, die in ihrer Ursprungsform mehr oder minder gut geeignet erscheinen, Transaktionsverläufe zu clustern. So sollen zunächst verschiedene Möglichkeiten der funktionalen Clusteranalyse als Teilgebiet der funktionalen Datenanalyse vorgestellt werden und anschließend ein bestimmtes Verfahren so modifiziert werden, dass die Vorgehensweise besser mit den speziell betrachteten Verläufen abgestimmt ist. Abschließend erfolgt eine Analyse von realen Transaktionsverläufen unter Verwendung verschiedener heuristischer Clustermethoden. Dabei richtet sich der Fokus auf neue Möglichkeiten Transaktionsverläufe angemessen zu partitionieren.
Neben Wachstumskurven stellen auch unter anderem ökonometrische Daten, meteorologische Messungen, Genexpressionsdaten, Reaktionszeiten oder Preisentwicklungen funktionale Verläufe dar. In der Praxis liegen solche Daten univariat oder multivariat als Vektor von Beobachtungen in diskreter Zeit vor, entstehen aber eigentlich als Funktion über die Zeit t oder allgemeiner als Funktion einer stetigen Variablen. Eine Funktion x wird somit nicht für jeden Wert von t erfasst, sondern vielmehr werden funktionale Daten normalerweise als T Paare \((t_j, y_j), j = 1, \ldots, T\), beobachtet und aufgezeichnet, wobei mit \(y_j\) der möglicherweise mit einem Fehler \(\epsilon_j\) behaftete Messwert zum Zeitpunkt \(t_j\) bezeichnet wird.
Das signal-to-noise-Modell sieht folgendermaßen aus:
\[
y_j = x(t_j) + \epsilon_j
\] (2.1)

Die Grundlage der funktionalen Datenanalyse ist es, diese „beobachtete“ Funktion selbst als einzelne Beobachtung \(x\) zu betrachten. Besondere Betonung liegt dabei in der Glättungsanforderung an die latente Funktion \(x\). Ohne diese bestünde kein zusätzlicher Nutzen funktionaler gegenüber multivariater Betrachtung der Daten. Ist die Funktion \(x\) glatt in

dem Sinn, dass sie eine oder mehrere Ableitungen besitzt, bedeutet dies, dass zwei auf- 
einanderfolgende Werte $y_j$ und $y_{j+1}$ zwangsläufig in einem gewissen Umfang miteinander 
verbunden sind und sich bei geringem Abstand nicht stark unterscheiden. Die diskreten 
Daten $y_j$, $j = 1, \ldots, T$ werden dabei üblicherweise dazu verwendet, die Funktion $x$ und 
gleichzeitig eine bestimmte Anzahl ihrer Ableitungen zu schätzen.

Im Allgemeinen sieht man sich mit einer Stichprobe von funktionalen Daten konfron-
tiert und nicht nur mit einer einzelnen Beobachtung $x$. So besteht eine Beobachtung $x_i$, 
$i = 1, \ldots, n$ aus $T_i$ Wertepaaren $(t_{ij}, y_{ij})$ mit $j = 1, \ldots, T_i$. Die Argumente $t_{ij}$ können für 
jede Beobachtung gleich sein, aber genausogut von einer Beobachtung zur anderen varie-
ren. Dies gilt auch für das Intervall $T$, über dem die Daten gesammelt wurden, selbst.
Bei den meisten praktischen Problemen liegen – wie schon erwähnt – von einer Funktion nur endlich viele diskret gemessene Werte $y_1, \ldots, y_m$ vor und somit muss diese Funktion $x_i$ mit Werten $x_i(t)$ als erstes interpoliert oder approximiert werden, um für jedes Argument $t$ den zugehörigen Funktionswert berechnen zu können. Da die Messwerte häufig mit Fehlern belastet sind und die zugrundeliegende Funktion durch den Messprozess verrauscht und dadurch nicht mehr erkennbar sein kann, ist es oft sinnvoll, bei der Schätzung der Funktion Smoothing-Techniken mit einzubeziehen und nicht ausschließlich zu interpolieren.


### 3.1. Splineexpansion

Mit Hilfe einer Splineexpansion lässt sich nun die Schätzung einer Funktion auf ein endlich-dimensionales Problem reduzieren. Splines teilen sich mit der Anpassung von Polynomen den Vorteil, nur eine geringe Rechenintensität zum Speichern von Informationen von Kurven aufbringen zu müssen. Gleichzeitig sind aber Splines den Polynomen noch hinsichtlich der Flexibilität bei der Darstellungsmöglichkeit überlegen. Das sogenannte Basisfunktio-
nensystem einer Splineexpansion besteht aus einer Menge bekannter Funktionen \( \phi_l \), die unabhängig voneinander sind und die Eigenschaft besitzen, jede Funktion beliebig gut zu approximieren, indem eine ausreichend große Anzahl \( L \) dieser Funktionen als Linearkombination zusammengesetzt wird. Eine solche Funktionsanpassung lässt sich folgendermaßen ausdrücken:

\[
x(t) = \sum_{l=1}^{L} \eta_l \phi_l(t),
\]

mit \( \eta_l \) als zugehörige Basiskoeffizienten.

Bezeichnet man mit \( \eta \) den \( L \)-dimensionalen Koeffizientenvektor und mit \( \phi \) den Vektor bzw. die \( n \times L \) – Matrix der Splinefunktionen \( \phi_l(t_j) \) schreibt sich Gleichung 3.1 in Matrixnotation als:

\[
x = \eta' \phi = \phi' \eta.
\]

Bei der Wahl von \( L = n \) können die Koeffizienten \( \eta_l \) so gewählt werden, dass \( x(t_j) = y_j \) für alle \( j \) und somit eine exakte Darstellung der Kurve vorliegt. Im Gegensatz zu dieser Interpolation bestimmt sich bei einer Approximation der Grad, zu dem die Daten \( y_j \) geübert werden, nach der Anzahl \( L \) an Basisfunktionen. Es wird deutlich, dass \( L \) selbst einen Parameter darstellt und nach der Beschaffenheit der Daten gewählt werden muss.

Die Schätzung der Splinekoeffizienten \( \eta_l \) kann mit Hilfe der Kleinste-Quadrate-Schätzung erfolgen. Die zu minimierende Zielfunktion \( SSE \) (sum of squared errors) lautet demnach:

\[
SSE = \sum_{j=1}^{n} [y_j - \sum_{l=1}^{L} \eta_l \phi_l(t_j)]^2.
\]

Als Schätzer für den Vektor der Basiskoeffizienten ergibt sich:

\[
\hat{\eta} = (\Phi' \Phi)^{-1} \Phi y,
\]

mit \( y \) als Vektor der tatsächlich beobachteten Werte.

Diese Schätzmethode ist in Situationen geeignet, in denen das Standard-Error-Modell angenommen wird, also die Residuen \( \epsilon_j \) unabhängig und identisch verteilt sind, mit \( \mathbb{E}(\epsilon_j) = 0 \) und konstanter Varianz \( \text{Var}(\epsilon_j) = \sigma^2, \ j = 1, \ldots, T \).

3.1 Splineexpansion

3.1.1. Fourier-Basis-System

Bei der Fourier-Basis-Expansion für periodische Daten ergeben sich die Basisfunktionen als $\phi_0(t) = 1$, $\phi_{2r-1}(t) = \sin r\omega t$ und $\phi_{2r}(t) = \cos r\omega t$, wobei der Parameter $\omega$ die Periode $\frac{2\pi}{\omega}$ bestimmt. Die Darstellung als Basisfunktionenexpansion wie in (3.1) ist somit:

$$\hat{x}(t) = \eta_0 + \eta_1 \sin \omega t + \eta_2 \cos \omega t + \eta_3 \sin 2\omega t + \eta_4 \cos 2\omega t + \cdots$$  \hspace{1cm} (3.5)

Ein Charakteristikum der Fourierbasen ist, dass sie, wenn die Werte $t_j$ äquidistant auf dem Intervall $T$ verteilt sind und die Periode gleich der Länge dieses Intervalls ist, orthogonal sind und somit die Kreuzproduktmatrix $\Phi^T\Phi$ diagonal ist. Diese Eigenschaft spielt beim Clustern funktionaler Daten eine Rolle.

Abbildung 3.1.: Fourier-Basis mit fünf Basisfunktionen auf dem Intervall $T = [0, 12]$
### 3.1 Splineexpansion

#### 3.1.2. Polynomsplines

Etwas anders ist nun die allgemeine Struktur einer Splinefunktion bei der Anpassung nicht-periodischer Daten. Das Intervall $T$, über dem die zu approximierende Funktion $x(t)$ definiert ist, wird hierbei in $D$ Subintervalle geteilt. Diese Intervalle werden von den sogenannten Knoten $\tau_d$, $d = 1, \ldots, D - 1$, getrennt. Schließt man die Endpunkte als Knoten mit ein, ergeben sich die Knoten $\tau_0, \ldots, \tau_D$. Über jedem Subintervall ist ein Spline ein Polynom einer festgelegten Ordnung $m$ bzw. vom Grad $m - 1$. Die Knoten können dabei äquidistant gewählt, aber auch individuell gesetzt werden. Letzteres ist vor allem dann sinnvoll, wenn es Bereiche in der Funktion mit mehr Schwankungen und andere relativ geradlinige gibt. Darüberhinaus darf es keine Intervalle ohne Daten geben, da so der Funktionaverlauf aufgrund mangelnder Information nicht bestimmt werden kann.

Damit die einzelnen Polynome glatt ineinander übergehen, müssen die Funktionswerte an den Knoten gleich sein. Zusätzlich sind aufgrund der globalen Glattheitsanforderung bis zu $m - 2$ Ableitungen an diesen Verbindungsstellen identisch. Zusammengefasst besteht eine Splinefunktion aus zwei Elementen, der Ordnung der Polynomstücke $m$ und der Knotensequenz $\tau$. Die Anzahl der Parameter, die $x(t)$ bestimmen, summiert sich aus Polynomordnung $m - 1$ und der Anzahl innerer Knoten $D$ zu $L = m + D - 1$.

Nach Fahrmeir et al. (2009) ist ein Polynomspline wie folgt definiert:

**Definition 1 (Polynom-Spline)** Eine Funktion $f : [a,b] \rightarrow \mathbb{R}$ heißt Polynom-Spline vom Grad $m - 1 \geq 0$ zu den Knoten $a = \tau_0 < \cdots < \tau_d = b$, falls sie die folgenden Bedingungen erfüllt:

1. $f(t)$ ist $(m - 2)$-mal stetig differenzierbar. Für $m = 2$ entspricht dies der Forderung, dass $f(t)$ stetig ist, für $m = 1$ werden keine Glattheitsanforderungen an $f(t)$ gestellt.

2. $f(t)$ ist auf den durch die Knoten gebildeten Intervallen $[\tau_j, \tau_{j+1})$ ein Polynom vom Grad $m - 1$.

Die äußerst bekannte B-Spline-Basis ist eine mögliche Darstellung solcher Polynomsplines. Die Funktion $x(t)$ lässt sich analog zu Gleichung (3.1) als B-Spline-Expansion darstellen:

$$
\hat{x}(t) = \sum_{l=1}^{m+D-1} \eta_l B_l(t) \quad (3.6)
$$

Im Gegensatz zu beispielsweise den Fourier-Basen sind B-Spline-Basen nur lokal definiert. Das bedeutet, dass eine B-Spline-Basisfunktion nur über $m+1$ benachbarten Intervallen existiert und an einer beliebigen Stelle des Intervalls $T$ nur $m$ Basisfunktionen positiv sind. Dadurch können auch Veränderungen in kleineren Bereichen deutlich gemacht werden. Insgesamt besteht die Basis aus $m$ Polynomstücken vom Grad $m-1$, die an den Knoten $m-2$-mal stetig differenzierbar zusammengesetzt sind. Bei äquidistanten Knoten haben die mittleren Basisfunktionen dieselbe Form, während sie sich bei nicht-äquidistanten unterscheiden. Die $m-1$ Basisfunktionen am rechten und linken Rand von $T$ distinguierten sich von den Übrigen. Am Rand von $T$ steigt die Anzahl an aufeinanderfolgenden Intervallen, über denen die Basisfunktionen positiv sind, von eins zu $m$ in Richtung des Zentrums an, besitzen aber immer noch den zweimal stetig differenzierbaren Übergang zu Null. Nach Außen hin verlieren die Basisfunktionen an Differenzierbarkeit, was die Tatsache unterstützt, dass der weitere Funktionsverlauf von $x(t)$ jenseits der Ränder nicht bekannt ist.

Abbildung 3.2.: B-Spline-Basis der Ordnung drei mit fünf inneren Knoten auf dem Intervall $T = [0, 12]$
3.1.3. Smoothing Splines

Sind die zugrundeliegenden Daten nicht oder nicht annähernd gleichabständig, ist es sinnvoller, an jeden $j$-ten Datenpunkt einen Knoten zu setzen. Bei den Smoothing-Splines wird diese maximale Knotenmenge verwendet und dadurch Regionen mit hoher Datendichte begünstigt. Zudem erfolgen keine Annahmen über die nichtparametrische Funktion $x(t)$, außer dass sie zweimal stetig differenzierbar ist. Anders als vorher gibt es keine Beschränkung auf den durch Basisfunktionen gebildeten Funktionenraum mehr. Das SSE-Kriterium aus (3.3) erweitert sich um einen Strafterm, der zu rauhe Funktionsschätzungen, also eine zu starke Anpassung, vermeiden soll. Als Maß für die Krümmung einer Funktion wird ihre quadrierte zweite Ableitung verwendet. Die zu minimierende Zielfunktion bei den Smoothing Splines ist das *penalisierte Kleinste-Quadrate-Kriterium*:

$$ PSSE = \sum_{j=1}^{n} [y_j - x(t_j)]^2 + \lambda \int [x''(t)]^2 dt $$ (3.7)

Aus der Gleichung wird deutlich, dass der erste Term die Nähe zu den Daten misst, während der zweite Term die Krümmung bestraft. Die Stärke der Bestrafung wird dabei durch den *Smoothingparameter* $\lambda \geq 0$ beeinflusst. Für $\lambda = 0$ kann $x(t)$ jede Funktion sein, die die Datenpunkte interpoliert, für $\lambda \to \infty$ ergibt sich ein Polynom vom Grad $m-1$ als Schätzung für $x(t)$. Gleichung (3.7) besitzt mit dem *natürlichen kubischen Spline* mit Knoten an den einzelnen Datenpunkten eine eindeutige, endlich-dimensionale Lösung. Nach [Fahrmeir et al. 2009] ist ein *natürlicher kubischer Spline* wie folgt definiert:

**Definition 2 (Natürlicher kubischer Spline)** Zu einer vorgegebenen Knotenmenge $a < \tau_0 < \cdots < \tau_d < b$ ist die Funktion $f(t)$ genau dann ein natürlicher kubischer Spline, wenn gilt

1. $f(t)$ ist ein kubischer Polynom-Spline zur obigen Knotenmenge und
2. $f(t)$ erfüllt die Randbedingungen $f''(\alpha) = f''(b) = 0$, d.h. $f(t)$ ist linear in den Intervallen $[a, \tau_0]$ und $[\tau_d, b]$.

Ein natürlicher kubischer Spline ist demnach ein kubischer Polynomspline mit einer speziellen, der *natürlichen*, Randbedingung.
3.2. Karhunen-Loève-Expansion


\[ x(t) = \mu(t) + \sum_{l=1}^{\infty} \xi_l(x) \rho_l(t), \]  
\[ (3.8) \]

wobei \( \mu(t) = E[x(t)] \) der marginale Mittelwert, \( p_l \) die Eigenfunktionen zur Kovarianz \( \Gamma \) mit \( \Gamma(s,t) = \text{cov}\{x(s), x(t)\} \), und \( \xi_l(x) \) zufällige, unkorrelierte Koeffizienten mit Mittelwert Null und Varianz \( \lambda_l \) sind.

Diese Repräsentation einer Funktion entspricht der einer funktionalen Hauptkomponentenanalyse. Analog zum multivariaten Fall, in der es Ziel der Hauptkomponentenanalyse ist, die Eigenwerte und Eigenvektoren der Kovarianzmatrix aufzufinden, ist dies auch Aufgabe der funktionalen Betrachtung. Nur entsprechen die Eigenvektoren nun Eigenfunktionen und dienen als Basisfunktionen für die Expansion. Dementsprechend sind diese Basisfunktionen orthonormal, d.h. \( \int \rho_j \rho_k = 0 \) \( \forall j < k \). Die Eigenfunktionen \( \rho_j \) zur Kovarianz \( \Gamma \) mit den zugehörigen Eigenwerten \( \lambda_j \) werden durch das Eigenwertproblem festgelegt, so dass jede Eigenwertfunktion, bzw. im Hauptkomponentenkontext, jede Principal-Component-Gewichtungs-Funktion \( \rho_l(t) \) für einen bestimmten Eigenwert \( \lambda \) folgende Gleichung erfüllt:

\[ \int \Gamma(s,t) \rho_l(t) dt = \lambda_l \rho_l(t) \]  
\[ (3.9) \]

Daraus entsteht eine sogenannte Integraltransformation \( \Gamma \) der Eigenfunktion \( \rho \), definiert durch

\[ \Gamma \rho = \int \Gamma(\cdot,t) \rho(t) dt, \]  
\[ (3.10) \]

ist, lässt sich das Eigenwertproblem wie im multivariaten Fall darstellen als

\[ \Gamma \rho = \lambda \rho, \]  
\[ (3.11) \]

wobei \( \rho \) nun eine Funktion und keinen Vektor beschreibt.

---


Die gesuchten Eigenwert-Eigenfunktions-Paare ergeben sich z.B. mit Hilfe diskreter Approximation durch Lösung der Gleichung
\[ \int_0^T \hat{\Gamma}(s,t) \hat{\rho}_l(s) \, ds = \hat{\lambda}_l \hat{\rho}_l(t) \]
unter der Bedingung
\[ \int_0^T \hat{\rho}_j(t) \hat{\rho}_l(t) \, dt = 0 \quad \text{für} \quad j = t \quad \text{und} \quad 1 \quad \text{für} \quad j \neq t \quad \text{(vgl.} \text{Chiou and Li (2007))}. \]
Andere Möglichkeiten finden sich bei Ramsay and Silverman 2006, Kapitel 8.4.

Die Koeffizienten \( \xi_l(x) \) der Expansion sind unkorrelierte Zufallsvariable mit Mittelwert Null und Varianz \( \lambda_l \), so dass
\[
\xi_l(x) = \langle x - \mu, \rho_l \rangle
\]
und damit eine Projektion der zentrierten Funktion \( x - \mu \) in Richtung der \( l \)-ten Eigenfunktion \( \rho_l \) ist. Eine mögliche Approximation der funktionalen Principal-Component-Scores findet sich in Chiou and Li 2007, S.695, Anhang A.3.

4.1. Heuristische Ansätze

Beim Clustern funktionaler Daten kommen i.d.R. unter den heuristischen Ansätzen nicht-hierarchische partitionierende Verfahren wie der bekannte $k$-means bzw. $k$-medoids Algorithmus zum Einsatz. Bei den partitionierenden Clusterverfahren wird – ausgehend von einer festgesetzten Clusteranzahl – eine vorgegebene Startlösung, also eine anfängliche Klasseneinteilung der Daten, durch schrittweises Verschieben von Objekten zwischen den Clustern so verändert, dass ein Heterogenitätskriterium $H(C)$ für eine Partition $C$ optimiert wird. Die beste Partition unter zwingender Vorgabe der Clusteranzahl wird bestimmt durch $H(C_{opt}) = \min_{\tilde{C}} H(\tilde{C})$. Vorteil nicht-hierarchischer Verfahren ist die Austauschbarkeit der Objekte zwischen den Clustern, die bei hierarchischen Verfahren nicht gegeben ist. Kritisch bei den partitionierenden Clustermethoden ist dagegen die Beeinflussung der Ergebnisse durch die subjektiv festgelegte Startposition und die Zielfunktion. So können immer nur lokale Optima erreicht werden, keine globalen.

$k$-means Die gängigste Methode unter den partitionierenden Clusterverfahren ist der $k$-means-Algorithmus. Dieser Algorithmus hat die Aufgabe, $n$ Objekte mit je $T$ Messungen in
4.1 Heuristische Ansätze

Eine fest vorgegebene Anzahl $K$ Gruppen $(C_1, C_2, \ldots, C_K)$ zuzuordnen. Als $C_k$ bezeichnet sich die Menge der $n_k$ Objekte im Cluster $k$. Die Partition wird dadurch gebildet, dass die Distanz zwischen dem Datenvektor $x_i$ und dem Centroidvektor des jeweiligen Clusters $c(x_i)$ minimiert wird:

$$d(X_i, C_k) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, c(x_i)) \rightarrow \min_{C_k}$$  (4.1)

Dabei berechnet sich der Centroid in Cluster $C_k$ als Durchschnitt der jeweiligen Variablenwerte $T$ über alle Objekte, die sich im Cluster befinden. So ergibt sich der Centroidwert für die $j$-te Variable im $k$-ten Cluster als

$$\bar{x}_j^{(k)} = \frac{1}{n_k} \sum_{i \in C_k} x_{ij}$$  (4.2)

Der vollständige Centroidvektor $\bar{x}^{(k)} = (\bar{x}_1^{(k)}, \bar{x}_2^{(k)}, \ldots, \bar{x}_T^{(k)})$ für das Cluster $C_k$ setzt sich aus den einzelnen Durchschnittswerten der Variablen zusammen. Das iterative Vorgehen beim $k$-means-Algorithmus läuft wie folgt ab:

**Algorithmus 1 (k-means-Algorithmus)**

1. **Starte mit** $K$ zufälligen $T$-dimensionalen Centroidvektoren $c^{(k)} = (c_1^{(k)}, c_2^{(k)}, \ldots, c_T^{(k)})$.

2. **Berechne die Distanz** $d(i, k)$ zwischen dem $i$-ten Objekt und dem $k$-ten Cluster mit der quadrierten euklidischen Distanz:

$$d(i, k) = \sum_{j=1}^{T} (x_{ij} - c_j^{(k)})^2$$  (4.3)

Vergleiche die Objekte mit jedem (neuen) Centroid bzgl. ihrer Distanz $d(i, k)$ zueinander und ordne sie ihrem nächstgelegenen Cluster zu.

3. **Ermittle nach dieser (anfänglichen) Zuordnung der Objekte die neuen Centroide für jedes Cluster gemäß (4.2).**

4. **Wiederhole die Schritte 2 bis 3 solange, bis kein Objekt mehr zwischen den Clustern wechselt.**

Der $k$-means Algorithmus startet also mit $K$ Clustermittelwerten, die i.d.R. zufällig aus den Datenpunkten gezogen werden. Danach werden alle einzelnen Datenpunkte demjenigen Cluster zugeordnet, das bzgl. der quadrierten euklidischen Distanz am nächsten liegt. Die Clustermittelwerte werden nach der vollständigen Zuordnung aller Datenpunkte als Durchschnittswert aller sich im jeweiligen Cluster befindenden Objekte neu berechnet und

---

*In Anlehnung an Steinley 2006*
der Algorithmus solange iteriert, bis keine Objekte mehr zwischen den Clustern wechseln. Insgesamt wird bei der Optimierung der Partition versucht, die Fehlerquadratsumme (SSE) zu minimieren:

\[ SSE = \sum_{j=1}^{T} \sum_{k=1}^{K} \sum_{i \in C_k} (x_{ij} - \bar{x}_j^{(k)})^2 \]  

(4.4)

Innerhalb des \( k \)-means-Algorithmus ist es auch möglich, andere Distanzen als die euklidische Distanz zu verwenden. Die allgemeine Vorgehensweise ohne konkrete Festlegung des Distanzmaßes kann als \textit{verallgemeinerter} \( k \)-\textit{means-Algorithmus} für ein \( k \)-Centroid-Clusterproblem bezeichnet werden.

\textbf{\( k \)-medians} Eine Alternative zum klassischen \( k \)-means Algorithmus ist der \( k \)-medians Algorithmus (vgl. [Steinley 2006] und [Leisch (2006)]). Dieser ist weniger empfindlich gegenüber Ausreißern. Der Unterschied dieses speziellen \( k \)-Centroid-Algorithmus zum \( k \)-means besteht also im Distanzmaß. Während \( k \)-means die quadrierte euklidische Distanz verwendet, bedient sich der \( k \)-medians Algorithmus der Manhattan-Distanz:

\[ d(i,k) = \sum_{j=1}^{T} \left| x_{ij} - c_j^{(k)} \right|. \]  

(4.5)

Als Verlustfunktion ergibt sich die Fehlerquadratsumme als

\[ SSE = \sum_{j=1}^{T} \sum_{k=1}^{K} \sum_{i \in C_k} \left| x_{ij} - med_j^{(k)} \right|. \]  

(4.6)

Die Centroide entsprechen demnach dem Median der Beobachtungen innerhalb des jeweiligen Clusters \( k \).

\footnote{\textit{siehe} Leisch 2006}
4.2. Modellbasierte Verfahren

Die zweite große Klasse beim Clustern sind die modellbasierten Verfahren. Hierbei wird angenommen, dass die Beobachtungen \( x_1, \ldots, x_n \) Realisationen des Zufallsvektors \( \mathbf{x} \) sind und \( \mathbf{x} \) in jeder Klasse eine andere Verteilung besitzt. Die Daten stammen also aus einer sogenannten *Mischverteilung* mit \( K \) Komponenten, wobei die Klassenzugehörigkeiten \( z_i \) nicht bekannt sind. Grundsätzlich gibt es zwei mögliche Vorgehensweisen die Clusterzugehörigkeiten \( z_i \) zu behandeln. Dies mündet in zwei unterschiedliche Ansätze, der *classification likelihood* und der *mixture likelihood*.

**classification likelihood** Bei der *classification likelihood* werden die \( z_i \) als Parameter gesehen und das Modell durch die Maximierung der folgenden Likelihood angepasst:

\[
L_C(\theta_1, \ldots, \theta_K; \mathbf{z}_1, \ldots, \mathbf{z}_n | \mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} f_{z_i}(x_i | \theta_{z_i}) \tag{4.7}
\]

mit \( f_{z_i}(x_i | \theta_{z_i}) \) als Dichte von Cluster \( z_i \) mit den zugehörigen Parametern im Cluster \( \theta_{z_i} \).

**mixture likelihood** Betrachtet man die Clusterzugehörigkeit \( z_i \) dagegen als fehlenden Wert und nimmt an, dass \( z_i \) multinomial verteilt ist mit Parametern \((\pi_1, \ldots, \pi_K)\) mit \( \pi_k \) als Wahrscheinlichkeit, dass eine Beobachtung zum \( k \)-ten Cluster gehört, dann lautet die zu maximierende sogenannte *mixture likelihood*:

\[
L_M(\theta_1, \ldots, \theta_K; \pi_1, \ldots, \pi_K | \mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(x_i | \theta_k). \tag{4.8}
\]

Dabei stellt \( f_k(x_i | \theta_k) \) die Dichte und \( \theta_k \) den unbekannten Parametervektor der \( k \)-ten Mischungskomponenten, sowie \( \pi_k \) mit \( \pi_k \geq 0 \) und \( \sum_{k=1}^{K} \pi_k = 1 \pi_k \) den unbekannten Mischungsanteil dar.


\[\text{\footnotesize \cite{Banfield and Raftery 1993, Fraley and Raftery 2002, James and Sugar 2003}}\]
EM-Algorithmus für Mischverteilungen

Notation:

- \( y_i \) beobachtete (unvollständige) Daten
- \( z_i \) unbeobachtete Daten bzw. latente Variablen
- \( x_i = (y_i, z_i) \) vollständige Daten


**Algorithmus 2 (EM-Algorithmus)**

*Wähle Startwerte der Parameter \( \theta^{(0)} \).*

**E-Step:**

Berechne den bedingten Erwartungswert der Log-Likelihood der vollständigen Daten, gegeben die beobachteten Daten und die aktuellen Parameterschätzer:

\[
Q(\theta) = Q(\theta|\theta^{(0)}) = \mathbb{E}[l(\theta; y, z, \theta^{(0)})]
\]

**M-Step:**

Schätze die Parameter \( \theta^{(1)} \), die diese erwartete Log-Likelihood des E-Step maximieren:

\[
Q(\theta^{(1)}) = \arg \max_\theta Q(\theta)
\]

*Iteriere E-/M-Steps bis zur Konvergenz.*

Bei Mischverteilungen sind – wie bereits erwähnt – die unbekannten, latenten Variablen die Klassenzugehörigkeiten \( z_i = (z_{i1}, \ldots, z_{iK}) \) mit

\[
z_{ik} = \begin{cases} 
1, & \text{falls } x_i \text{ zur Klasse } k \text{ gehört} \\
0, & \text{sonst.}
\end{cases}
\]

Die Log-Likelihood der beobachteten Daten \( y \) ist

\[
l(\theta; y) = \sum_{i=1}^{n} \log \left[ \sum_{k=1}^{K} \pi_k f_k(y_i|\theta_k) \right] \quad (4.9)
\]

mit \( \pi_k \) – wie zuvor definiert – als unbekannte Mischungsanteile.
Die Log-Likelihood der vollständigen Daten $x = (y, z)$ enthält die unbekannte Indikatorvariable $z_i$ und ergibt sich als

$$l(\theta; y, z) = \sum_{i=1}^{n} \log f(x_i, z_i | \theta)$$  \hspace{1cm} (4.10)

$$= \sum_{i=1}^{n} \log [f(x_i | z_i; \theta) \cdot f(z_i)]$$  \hspace{1cm} (4.11)

$$= \sum_{i=1}^{n} [\log f_{z,i}(x_i | \theta) + \log \pi_{z,i}]$$  \hspace{1cm} (4.12)

Der Algorithmus läuft wie folgt ab:

**Algorithmus 3 (EM-Algorithmus für Mischverteilungen)**

Wähle geeignete Startwerte $\theta_1, \ldots, \theta_K$ und a-priori Wahrscheinlichkeiten $\pi_1, \ldots, \pi_K$.

**E-Step:**

Berechne für $i = 1, \ldots, n$ die a-posteriori Wahrscheinlichkeiten

$$\hat{p}_{ik} = \mathbb{E}[I(z_i = k)|y_i, \theta]$$  \hspace{1cm} (4.13)

$$= \frac{\hat{\pi}_k f_k(y_i | \hat{\theta}_k)}{\sum_{m=1}^{K} \hat{\pi}_m f_m(y_i | \hat{\theta}_m)}$$  \hspace{1cm} (4.14)

für $k = 1, \ldots, K$ und daraus die a-priori Klassenwahrscheinlichkeit

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_{ik}.$$  \hspace{1cm} (4.15)

**M-Step:**

Berechne $\hat{\theta}_k|z$ für $k = 1, \ldots, K$.

Somit wird im **E-Step** des EM-Algorithmus die Indikatorvariable durch ihren Erwartungswert $\mathbb{E}(z_i)$ ersetzt. Der **M-Step** dient zur Maximierung der Log-Likelihood der vollständigen Daten bzgl. $\pi_{z,i}$ und $\theta_{z,i}$, wobei $z_{ik}$ fest auf den im **E-Step** berechneten Wert gesetzt wird. Analog werden wiederum die **E-** und **M-Schritte** solange iteriert, bis der Algorithmus konvergiert.
Methoden zum Clustern funktionaler Daten


<table>
<thead>
<tr>
<th>heuristische Verfahren</th>
<th>modellbasierte Verfahren</th>
</tr>
</thead>
<tbody>
<tr>
<td>(k-means)</td>
<td></td>
</tr>
<tr>
<td>Orthonormale Basis</td>
<td>Serban and Wasserman (2005)</td>
</tr>
<tr>
<td></td>
<td>Shimizu and Mizuta (2008)</td>
</tr>
<tr>
<td></td>
<td>Tarpey and Kinateder (2003)</td>
</tr>
<tr>
<td></td>
<td>Tarpey et al. (2003)</td>
</tr>
<tr>
<td></td>
<td>Tarpey (2007a)</td>
</tr>
<tr>
<td>B-Splines</td>
<td>Abraham et al. (2003)</td>
</tr>
<tr>
<td></td>
<td>Garcia-Escudero and Gordaliza (2005)</td>
</tr>
<tr>
<td></td>
<td>Jank and Shmueli (2006) (penalisiert, k-medians)</td>
</tr>
<tr>
<td></td>
<td>Tarpey (2007a)</td>
</tr>
<tr>
<td></td>
<td>Hitchcock et al. (2007) (k-medoids mit PAM)</td>
</tr>
<tr>
<td>natürliche Splines</td>
<td>James and Sugar (2003)</td>
</tr>
<tr>
<td>Smoothing Splines</td>
<td>Ma et al. (2006)</td>
</tr>
<tr>
<td>Karhunen-Loève-Expansion</td>
<td>Chiou and Li (2007)</td>
</tr>
<tr>
<td></td>
<td>Chiou and Li (2008)</td>
</tr>
</tbody>
</table>

Tabelle 5.1.: Übersicht der Clusterverfahren für funktionale Daten
5.1 Modellbasierte Verfahren


Ein übliches Mixed-Effects-Model für eine Beobachtung zum Zeitpunkt $t_{ij}$, gegeben die beobachtete Kurve $i$ ist in Cluster $k$, lässt sich schreiben als:

$$y_i = X_i \beta + Z_i \gamma_i + \epsilon_i$$ (5.1)

wobei $\beta$ den Vektor der fixen Effekte und $\gamma_i$ den Vektor der zufälligen Effekte darstellt.

Das bekannte lineare Modell wird damit um einen weiteren Term $Z_i \gamma_i$ ergänzt, der dazu dient, die erwähnte Abweichung jeder Kurve von ihrem populationsbezogenen mittleren Kurvenverlauf zu berücksichtigen. Normalerweise wird dabei angenommen, dass die zufälligen Effekte i.i.d. $\gamma_i \sim N(0, \Gamma)$ und die Fehler i.i.d. und unabhängig von $\gamma_i$, $\epsilon_i \sim N(0, \sigma^2 I)$ verteilt sind.

Die einzelnen Verfahren unterscheiden sich in der Modellierung des festen Effekts, aber vor allem in der des zufälligen Effekts. So wird beim Functional Clustering Model (FCM) von James and Sugar (2003) sowohl für die Modellierung der populationsspezifischen Kurve eines Clusters als auch für die Zufallsabweichung eine Splineexpansion $y_i(t) = S_i \eta_i + \epsilon_i$ mit $\eta_i = \mu_{zi} + \gamma_i$ gewählt:

$$y_i(t) = S_i(t) \mu_{zi} + S_i(t) \gamma_i + \epsilon_i$$ (5.2)

mit $S_i(t)$ als natürliche Splinebasis, $\gamma_i \sim N(0, \Gamma)$ und $\epsilon_i \sim N(0, \sigma^2 I)$.

Bei Luan and Li (2003) wird die typische Kurve im $k$-ten Cluster sowie die zufällige Abweichung mit Hilfe von B-Spline-Basen dargestellt:

$$y_i(t_{ij}) = \sum_{l=1}^{p} \beta^{(k)}_l \tilde{B}_l(t_{ij}) + \sum_{l=1}^{q} \gamma_{il} B_l(t_{ij}) + \epsilon_{ij}$$ (5.3)

mit denselben Verteilungsannahmen für $\gamma_i$ und $\epsilon_i$.

Für eine Beschreibung von Mixed-Effects-Models siehe Laird and Ware (1982).
5.2 Heuristische Ansätze


5.2. Heuristische Ansätze


Orthonormale Basisexpansion und k-means Clustern

5.2 Heuristische Ansätze


$$x^{(k)}(t) = \mu^{(k)}(t) + \sum_{l=1}^{L_k} \xi_l^{(k)}(x) \hat{\rho}_l^{(k)}(t)$$

wird. Falls nun Kurve $x$ tatsächlich zu Cluster $k$ gehört, entspricht $x^{(k)}(t)$ der Karhunen-Loève-Expansion von $x(t)$, ist dies nicht der Fall unterscheiden sich die beiden Funktionen. Das Clustern der Kurven erfolgt ähnlich zum k-means Algorithmus. Im ersten Schritt werden die marginalen FPC-Scores (functional principal component scores) aller Kurven $\hat{\xi} = (\hat{\xi}_i^1, \ldots, \hat{\xi}_i^L)$ mit dem k-means-Algorithmus geclustert und man erhält für eine vorgegebene Clusteranzahl $K$ die Klassenzugehörigkeiten der einzelnen Kurven. Im zweiten Schritt werden nun die Objekte iterativ reklassifiziert. Als optimale Clusterzuordnung für eine Kurve dient als Kriterium die $L^2$-Distanz zwischen der Kurve und der Clusterzentrumskurve $\tilde{x}^{(k)}$, dargestellt als Karhunen-Loève-Expansion wie in Gleichung (5.4):

$$k^*(x) = \arg \min_{k \in \{1, \ldots, K\}} \| x - \tilde{x}^{(k)} \|$$

Die neuen Clusterzentren berechnen sich anschließend als Expansion aller Kurven im jeweiligen Cluster.

Ein weiteres Verfahren von Chiou and Li (vgl. Chiou and Li (2008)) fokussiert sich auf die Kurvenform selbst. So sollen Funktionen, die sich nur durch eine vertikale Verschiebung und durch einen multiplikativen Effekt bzgl. der Skalierung unterscheiden, demselben Cluster zugeordnet werden. Dafür entwickelten sie ein functional multiplicative random-effects model in Verbindung mit einer Karhunen-Loève Expansion. Die Idee ist ähnlich dem vorhergehenden Ansatz. Falls $x_i$ tatsächlich zum Cluster $k$ gehört, ist die Approximation im Cluster $k$ gleich der trunkierten stochastischen Expansion mit $M_k$ Komponenten, verbunden mit einem multiplikativen Skaleneffekt. Die Ähnlichkeit wird in diesem Fall durch die funktionale Korrelation zwischen $x_i$ und $x_{i(c)}^{M_k}$ gemessen.
5.2 Heuristische Ansätze


Nicht selten gibt es in vorliegenden Daten Ausreißer, die – gerade bei k-means, das für seine Ausreißerempfindlichkeit bekannt ist – sehr starken Einfluss auf die Clusterergebnisse nehmen können, so auch bei funktionalen Daten. Deshalb kann es nützlich sein, solche auffälligen Daten aufzudecken und aus dem Clusterprozess auszuschließen. García-Escudero and Gordaliza (2005) stellen dazu das *Robust Curve Clustering* vor, das mit dem sogenannten *trimmed k-means* arbeitet. Beim *trimmed k-means* werden die Clusterzentren nur aus \( \lfloor n(1 - \alpha) \rfloor \) Datenpunkten berechnet, wobei \( \alpha \in [0, 1] \) der sogenannte Trimmungsparameter ist, der angibt, wieviel Prozent der Daten weggeschnitten werden. Die Trimmung selbst ergibt sich aus den jeweiligen Daten und deren Kontaminationslevel. Der typische *trimmed k-means Algorithmus* läuft für ein vorgegebenes \( \alpha \) wie folgt ab (in Anlehnung an García-Escudero et al. 2003):

**Algorithmus 4 (trimmed k-means-Algorithmus)**

1. Starte mit \( K \) zufälligen \( T \)-dimensionalen Centroidvektoren \( c^{(k)} = (c_1^{(k)}, c_2^{(k)}, \ldots, c_T^{(k)}) \).
2. Berechne die Distanz jeder Beobachtung zu ihrem nächstgelegenen Zentrum gemäß:

   \[
   d(i, k) = \sum_{j=1}^{T} (x_{ij} - c_j^{(k)})^2
   \]

   und behalte die Menge \( H \) mit insgesamt \( \lfloor n(1 - \alpha) \rfloor \) Beobachtungen mit den niedrigsten Distanzen \( d(i, k) \).

   Teile \( H \) in \( H = \{H_1, \ldots, H_K\} \), wobei die Punkte in \( H_k \) diejenigen sind, die näher an \( c_k \) liegen als zu einem anderen Zentrum.

3. Wiederhole Schritt 2 einige Male.

Nach dieser Iteration, berechne die endgültige Zielfunktion

\[
\frac{1}{n(1-\alpha)} \sum_{j=1}^{T} \sum_{i \in H_k} \sum_{l=1}^{K} (x_{ij} - c_{j}^{(k)})^2
\]

(5.7)

Ebenfalls ein abweichendes Distanzmaß bei ansonsten gleicher Vorgehensweise verwen-

den Hitchcock et al. (2007). Die Distanz basiert hier auf paarweisen Unähnlichkeiten und
lässt sich bei \( n \) Messungen über einen Zeitraum \( T = [t_1, \ldots, t_T] \) schreiben als:

\[
d(i, k) = \frac{t_T - t_1}{T - 1} \left( \frac{\hat{x}_i(t_1) - \hat{x}_k(t_1)}{2} + \sum_{j=1}^{T-1} [\hat{x}_i(t_j) - \hat{x}_k(t_j)] + \frac{\hat{x}_i(t_T) - \hat{x}_k(t_T)}{2} \right), \tag{5.8}
\]

wobei \( \hat{x} \) die geglättete Kurve darstellt. Jank and Shmueli (2006) weichen hingegen sowohl
im Distanzmaß als auch in der Splineanpassung vom Standardverfahren ab. So verwen-
den sie anstelle des \( k \)-means Algorithmus den in Abschnitt 4.1 ebenfalls beschriebenen
\( k \)-medians Algorithmus, zur Kurvenanpassung verwenden sie \( P \)-Splines mit selbst fest-
gelegten, nicht-äquidistanten Knoten und Smoothingparameter \( \lambda \). Die Kurvenanpassung
erfolgt hier im Grunde vergleichbar mit der bei den Smoothing Splines aus Abschnitt 3.1.3.
Ziel ist es, das penalisierte KQ-Kriterium ähnlich Gleichung (3.7):

\[
PSSE = \sum_{j=1}^{T} [y_j - \sum_{l=1}^{L} \eta_l B_l(t_j)]^2 + \lambda \sum_{l=d+1}^{L} (\Delta^d \eta_l)^2
\]

(5.9)

zu minimieren, wobei mit \( \Delta^d \) Differenzen \( d \)-ter Ordnung bezeichnet werden.\footnote{Für eine genauere Beschreibung von \( P \)-Splines siehe Fahrmeir et al. 2009}
Erweiterungen zum Clustern funktionaler Daten mit dem k-medoids-Algorithmus

6.1. Poisson-Distanz


Da die Ähnlichkeit der Beobachtungen mit zunehmender Likelihood steigt, sollte als Distanzmaß, als Maß der Unähnlichkeit, die negative Log-Likelihood der beobachteten Daten passend sein. Nimmt man die Beobachtungen über die Zeitpunkte als unabhängig an, ergibt die negative Log-Likelihood \( l \) aufsummiert über alle Zeitpunkte ein entsprechendes Distanzmaß. Somit berechnet sich die Distanz der beobachteten Daten \( x_{ij} \) zum jeweiligen Centroidvektor \( c_j^{(k)} \) im Cluster \( k \) durch:

\[
d(i, k) = \sum_{j=1}^{T} \left[ -l(x_{ij}, c_j^{(k)}) \right] = \sum_{j=1}^{T} -\log \left[ f(x_{ij}|c_j^{(k)}) \right]
\]
und unter der Annahme von poissonverteilten Daten:

\[
d(i, k) = \sum_{j=1}^{T} \left[ -\log \left( \frac{c_j^{(k)x_i}}{x_{ij}!} e^{-c_j^{(k)}} \right) \right]
\]

(6.3)

Die neuen Centroide berechnen sich bei poissonverteilten Daten analog zu \[4.2\]. Der \textit{k-means Algorithmus} läuft ansonsten gleich ab\[10\].

### 6.2. Splineclustern


**Algorithmus 5 (Splineclustern)**

1. Stärte mit \( K \) zufälligen \( T \)-dimensionalen Centroidvektoren \( c^{(k)} = (c_1^{(k)}, c_2^{(k)}, \ldots, c_T^{(k)}) \).

2. Lege einen Spline durch den Centroidvektor:

\[
c^{(k)} = \eta^{(k)} S
\]

(6.4)

Schätze die Splinekoeffizienten mittels KQ-Schätzung:

\[
\hat{\eta}^{(k)} = (S' S)^{-1} S' c^{(k)}
\]

(6.5)

und berechne den Splinewert an den einzelnen Messzeitpunkten:

\[
c_j^{(k),\text{neu}} = \sum_{l=1}^{L} \hat{\eta}_l^{(k)} S_l(t)
\]

(6.6)

Berechne die Distanz \( d(i, k) \) zwischen dem \( i \)-ten Objekt und dem \( k \)-ten Clustercen-
troid. Vergleiche die Objekte mit jedem (neuen) Centroid bzgl. ihrer Distanz \( d(i, k) \)
zueinander und ordne jede Beobachtung dem nächstgelegenen Cluster zu.

3. Ermittle nach dieser (anfänglichen) Zuordnung der Objekte die neuen Centroidvek-
toren für jedes Cluster gemäß \[4.2\]:

\[
\bar{x}_j^{(k)} = \frac{1}{n_k} \sum_{i \in C_k} x_{ij}.
\]

4. Wiederhole die Schritte 2 bis 3 solange, bis kein Objekt mehr zwischen den Clustern wechselt.

\[10\] Die Poisson-Distanz ist in einer Erweiterung des \textit{flexclust} package als Funktion \textit{distPoisson} hinterlegt, ©Friedrich Leisch.
Als Distanzen $d(i,k)$ sind beispielsweise die euklidische Distanz, aber auch die in Abschnitt 6.1 beschriebene Poissondistanz denkbar.


Abbildung 6.1.: Beobachtungen eines Verlaufsmusters mit Centroidvektor und Centroid-spline

Jetzt wird exemplarisch ein Beobachtungsvektor herausgegriffen. Dieser ist in Abbildung 6.2 als schwarze durchgehende Linie geplotet.

Die gängige Methode funktionale Daten zu clustern, beinhaltet als ersten Schritt die Splineexpansion der Beobachtungsvektoren. Diese ist für den speziellen Datenvektor als gestrichelte schwarze Linie dargestellt. Weiterhin beruht der partitionierende $k$–means-
Abbildung 6.2.: Ausgewählte Beobachtung eines Verlaufsmusters mit Centroidvektor und Centroidspline I

Abbildung 6.3.: Ausgewählte Beobachtung eines Verlaufsmusters mit Centroidvektor und Centroidspline II

In diesem Kapitel sollen nun verschiedene, partitionierende Clustermethoden, basierend auf dem $k$-means Algorithmus aus Abschnitt 4.1 anhand realer und simulierter Daten durchgeführt werden. Ziel ist es, herauszufinden, ob und wie sich die betrachteten Clusterverfahren für die vorliegenden Daten, nämlich Transaktionsverläufe, unterscheiden. Wie verhält sich das Standardverfahren der funktionalen Clusteranalyse – $k$-means Clustern von Basiskoeffizienten einer Splineexpansion – bei dieser Art von Daten? Bringt die neue Poissonändistanz aus Abschnitt 6.1 einen Vorteil gegenüber der herkömmlich verwendeten euklidischen Distanz und wie wichtig ist es, die Funktionalität der Daten überhaupt zu berücksichtigen?

7.1 Daten und Vorgehen

Als Ausgangspunkt für die Analyse lagen Informationen von 4424 Haushalten über deren Einkäufe von 65 verschiedenen Produkten bzw. Produktgruppen innerhalb eines Kalenderjahres vor. Von jedem Haushalt (im folgenden auch mit HH abgekürzt) wurden Kalendertag und Anzahl der jeweils bei diesem Einkauf erworbenen Produkte aufgezeichnet. Aus den 65 Produktkategorien erfolgte für die weitere Untersuchung eine Auswahl von vier Typen:

- Produkt 61: Mineralwasser
- Produkt 44: Alkoholfreie Getränke ohne Kohlensäure (Fruchthaltige)
- Produkt 63: Eiscreme (Haushaltspackungen)
- Produkt 19: Zahnpasta

Bei Mineralwasser, alkoholfreien Getränken und Eiscreme sollten, nach eigener Überlegung, mehr Schwankungen bei der Anzahl der Einkäufe über die Sommer- und Wintermonate auftreten, als bei Zahnpasta, bei deren Transaktionsverlauf eher ein konstanteres Verhalten über die Zeit vermutet wurde.

Um äquidistante Messungen der Einkaufsmenge zu erhalten, wurden in einem ersten Schritt alle Einkäufe der jeweiligen Produkte getrennt für jeden Haushalt sowohl auf Kalenderwochen (im folgenden auch mit KW abgekürzt), als auch auf Monatsebene aggregiert.


Im vorliegenden Fall stellen die Einkäufe eines Haushalts zu einem Zeitpunkt oder aggregiert über einen bestimmten Zeitraum Zähldaten dar. Die Verteilung der Einkaufsmenge aller Haushalte sieht über die einzelnen Monate betrachtet jeweils für ein bestimmtes Produkt sehr ähnlich aus. Klar erkennbar ist eine Dominanz sehr niedriger Einkaufsmengen und die damit einhergehende rechtsschiefe Verteilung der Transaktionszahl. Oft werden die Produkte gar nicht oder nur einmal pro Woche bzw. Monat besorgt. Die durchschnittliche Anzahl an gekauften Produkten pro Kalenderwoche liegt bei Produkt 19 – Zahnpasta ungefähr bei 0.2, bei den Getränken (Produkt 44 – Alkoholfreie Getränke (Fruchthaltige) und Produkt 61 – Mineralwasser) bei jeweils 0.4 und bei der Eiscreme (Produkt 63) bei 0.1. Auf Monatsebene führen ebenfalls die Getränke mit einem durchschnittlichen Einkaufsvolumen von 1.6 bzw. 1.7, die Zahnpasta wird ca. einmal pro Monat und die Eiscreme im Schnitt 0.5 mal eingekauft. Das aus Abbildung 7.1 deutlich werdende Einkaufsverhalten und die damit verbundene Schiefe der Verteilung ist bei allen vier Produkten zu erkennen.[11] Somit

ist es fragwürdig gängige \( k \)-means-Algorithmen, die auf symmetrischen Distanzmaßen beruhen, zum Clustern dieser Daten zu verwenden. Ein Fokus der weiteren Betrachtung ist demzufolge das in Abschnitt 6.1 vorgestellte Distanzmaß, das die theoretische Verteilung der Daten berücksichtigen kann.

Neben der mangelnden Symmetrie existiert eine weitere Besonderheit von Transaktionsdaten. So lässt sich bei der Betrachtung von Einkaufsverhalten ein glatter Verlauf der getätigten Transaktionen eines Haushaltes, also ein funktionaler Verlauf, unterstellen. Grundgedanke dabei ist, dass die Anzahl eines gekauften Produkts von vorherliegenden und nachfolgend geplanten Käufen beeinflusst ist und einem gewissen, glatt darstellbaren Einkaufsmuster folgt. Die unterschiedliche Behandlung der Daten im Hinblick auf die Einbeziehung der funktionalen Form ist demnach beim Clustern von Transaktionsdaten ebenfalls von großem Interesse. Entweder bleibt die Funktionalität dabei gänzlich unberücksichtigt oder sie wird in Form von Splines zur Darstellung der Kurven (vgl. Abschnitt 3.1) innerhalb oder im Vorfeld des Clusterns bedacht.

Mit den in Abschnitt 6 besprochenen Erweiterungen entstehen sechs weiter verfolgte Möglichkeiten die zugrundeliegenden Daten unter Verwendung des \( k \)-meines-Algorithmus, kombiniert mit den verschiedenen Distanzmaßen (euklidisches vs. Poisson) sowie Vorgehensweisen (klassisches \( k \)-means-Verfahren vs. Splineclustern), zu partitionieren:
7.1 Daten und Vorgehen

Möglichkeiten zum Clustern:

1. Clustern der Rohdaten mit $k$-means und euklidischer Distanz
2. Clustern der Rohdaten mit $k$-means und Poisson-Distanz
3. Clustern der Koeffizienten der Splineexpansion mit $k$-means und euklidischer Distanz
4. Clustern der mittels Splineexpansion vorhergesagten Werte an den Messzeitpunkten mit $k$-means und euklidischer Distanz
5. Clustern der Rohdaten mit Splineclustern und euklidischer Distanz
6. Clustern der Rohdaten mit Splineclustern und Poisson-Distanz


Sowohl für die Anpassung der Kurven mit Splines, als auch für das Splineclustern wurden kubische B-Splines (vgl. Abschnitt 3.1.2 zu B-Splines) gewählt. Nach inhaltlichen Überlegungen über abzubildende Einkaufsmuster, erfolgte eine Festlegung der Anzahl an Frei-

12Die Berechnungen erfolgten mit einer bisher unveröffentlichten Erweiterung des Pakets flexclust, ©Freidrich Leisch, in R unter Verwendung der Funktion stepFlexclust.
heitsgraden auf $df = 6$, wobei sich bei kubischen B-Spline-Basen zwei innere Knoten auf dem betrachteten Intervall ergeben. Mit zwei inneren Knoten und Polynomstücken vom Grad 3 lassen sich ausreichend Extrema darstellen, um saisonale Effekte im Einkaufsverhalten abbilden zu können.

Abbildung 7.2.: Anzahl gekauftes Produkt 61 über die Zeit mit angepasstem B-Spline der Ordnung 3 und 2 inneren Knoten, KW-Ebene

Bei der Splineexpansion im Vorfeld zur Clusteranalyse wurde, als erster Schritt, durch jeden Beobachtungsvektor eines Haushaltes ein Spline gelegt. Abbildung 7.2 zeigt die Ausgangskurve der Einkäufe von Produkt 61 – Mineralwasser – vier einzelner Haushalte mit zugehöriger Splineexpansion in rot eingezeichnet. Mit der Wahl von nur zwei inneren Knoten erfolgt eine eher starke Glättung, die das Erkennen von groben Strukturen zulässt. So zeigt beispielsweise Graphik 7.2, dass die Haushalte 112709, 102128 und 100964 ähnliche Muster in ihrem Kaufverhalten für Mineralwasser (Produkt 61) aufweisen. Es gibt sowohl im Frühjahr einen mehr oder weniger stark ausgeprägten Anstieg, der um die Kalenderwochen 15 – 22, also April bis Mai, zurückgeht und im Sommer bis zum Herbst hin wieder stärker ansteigt. Dagegen wird bei Haushalt 100756 das ganze Frühjahr sowie den Sommer...
über stärker eingekauft und in den Wintermonaten schwächer.

Abbildung 7.3.: Anzahl gekauftes Produkt 63 über die Zeit mit angepasstem B-Spline der Ordnung 3 und 2 inneren Knoten, KW-Ebene


Aus den beiden Abbildungen 7.2 und 7.3 soll deutlich werden, dass die Anpassung eines geeigneten Splines mit ausreichend inneren Knoten von der Fragestellung abhängt. In unserem Fall, in dem die einzelnen Käufe eine untergeordnete Rolle spielen und nur ein allgemeiner Trend im Vordergrund steht, ist eine so geringe Knotenanzahl ausreichend. Mit der Beschränkung auf die hier gewählte Knotenanzahl von zwei können beispielsweise etwaige Gruppen saisonaler Einkaufsmuster aufgefunden werden. Differenziertere Aussagen

34
7.1 Daten und Vorgehen

auf beispielsweise wöchentlichen Änderungen sind nicht zu treffen.

Abbildung 7.4.: Anzahl gekauftes Produkt 61 über die Zeit mit angepasstem B-Spline der Ordnung 3 und 2 inneren Knoten, Monatsebene


auf Basis eines Scree-Plots, der als graphisches Hilfsmittel zur Festlegung einer geeignetem Gruppenanzahl bei partitionierenden Clusteralgorithmen dient.

Abbildung 7.5.: Scree-Plot für $k = 3$ bis $k = 10$ Cluster bei Clusterverfahren 4, Produkt 61, KW-Ebene


$^{13}$Die Scree-Plots für die einzelnen Clusteralgorithmen finden sich in Teilen in Anhang F.
7.2. Reale Transaktionsdaten

7.2.1. Verläufe der Clusterzentren

7.2.1.1. Produkt 63 – Eiscreme (Haushaltspackungen)


Abbildung 7.6.: Verlauf der Clusterzentren, Clustern der Rohdaten mit \( k \)-means und euklidischer Distanz, Produkt 63, KW-Ebene

Das vierte Zentrum in Abbildung 7.6 liegt um die 0.2, das zweite schwankt um 0.5 und das dritte um einen Einkauf pro Woche. Alle diese drei Zentren weisen eine leicht ansteigende Konsummenge bis ins späte Frühjahr auf, mit dem Maximum im Hochsommer. Danach fällt die Einkaufsmenge in den Herbst- und Wintermonaten auf ein Basisniveau ab. Daneben existiert noch ein Cluster von Nichtkäufern, die über das Jahr hinweg keinerlei Eiscreme kaufen.

Um zu erkennen, welche unterschiedlichen Einkaufsverläufe den jeweiligen Gruppen bei Clusterverfahren 1 zugeordnet wurden, sind in Abbildung 7.7 50 zufällig ausgewählte Verläufe mit ihrem zugehörigen Zentrum (in rot eingezeichnet) dargestellt. Hieraus wird auch die Problematik deutlich, aus großen Datensätzen mit vielen Verläufen einzelne Struktu-
Abbildung 7.7.: Anzahl gekauftes Produkt 63 nach KW, aufgeteilt nach Cluster, Stichprobe von 50 HH – Clustern der Rohdaten mit \( k \)-means (\( k=4 \)) und euklidischer Distanz; Clusterzentrum in rot eingezeichnet.


**Clusterverfahren 2**  Clusterverfahren 2 mit der Poisson-Distanz anstelle der euklidischen bringt die in Graphik 7.8 dargestellten typischen Transaktionsverläufe für *Eiscreme* auf Kalenderwochenebene hervor.

Rein optisch unterscheiden sich die einzelnen Verläufe zueinander und auch im Hinblick auf Clusterverfahren 1 nur in einer Verschiebung im Niveau und nicht im Verlauf selbst. Ebenfalls erkennbar ist die Nichtkäufer-Gruppe, danach folgen zwei Gruppen mit relativ konstanter Ausgangseinkäufe von 0.1 und 0.2 gekauften Packungen pro Woche und eine etwas höher angesiedelte mit 0.5 Packungen mit jeweils starkem Hoch, mit bis zu 100% Wachstum, im späten Frühjahr bis in den Sommer, analog zu Verfahren 1.
7.2 Reale Transaktionsdaten

Die Betrachtung auf Monatsebene bringt auch hier in etwa gleiche Verläufe bei insgesamt höherem Einkaufsvolumen hervor (vgl. Anhang A).

**Clusterverfahren 3** Das Standardverfahren beim Clustern von Kurven – Clustern der Splinekoeffizienten mit $k$-means – liefert im Gegensatz zu Verfahren 1 und 2 teilweise unterschiedliche Transaktionsmuster (vgl. Abbildung 7.9). Hier entsteht sowohl auf Kalenderwochen- als auch auf Monatsebene wiederum ein konstantes Einkaufsverhaltensmuster bei den Nullkäufern. Zusätzlich gibt es nun eine Gruppe von Frühjahrskäufern, die um Mitte Mai herum ihr Maximum von ca. 0.7 Packungen pro Woche bzw. 3 pro Monat erreicht. Darüberhinaus werden zwei Klassen von Sommerkäufern aufgedeckt: Eine relativ stark ausgeprägte Sommergruppe, die um den August herum 1 Packung Eiscreme pro Woche bzw. 5 Packungen pro Monat kauft und eine etwas unterhalb verlaufende Klasse mit wöchentlich maximal 0.3 bzw. monatlich ca. 2 Packungen.

Bei Clusterverfahren 3 lassen sich zudem die an die Transaktionen der Haushalte angepassten Splines, nach ihrer Clusterzuordnung getrennt, abbilden (siehe Abbildung 7.10). Im Gegensatz zu beispielsweise Abbildung 7.7 auf Seite 38 sind hier die verschiedenen Kurvenverläufe wesentlich deutlicher erkennbar. Durch die starke Glättung kristallisieren sich die grundlegenden Unterschiede klarer heraus, als dies beim Plotten der Rohdaten der Fall ist.

Die Splines der Haushalte zeigen beispielsweise in Cluster 3 das markante Frühjahrschöck und in Klasse 1 und 4 den mehr oder weniger deutlich ausgeprägten Anstieg im Hochsommer. In Cluster 2 dagegen sammeln sich die konstanten Kurvenverläufe der Wenig- und Nichtkäufer.

Abbildung 7.8.: Verlauf der Clusterzentren – Clustern der Rohdaten mit $k$-means und Poisson-Distanz, Produkt 63, KW-Ebene
Abbildung 7.9.: Verlauf der Clusterzentren – Clustern der Splinekoeffizienten (B-Splines Grad 3, 2 innere Knoten) mit $k$-means ($k=4$) und euklidischer Distanz, Produkt 63, KW-Ebene

Abbildung 7.10.: Angepasste Splines nach Clusterzuordnung – Clustern der Splinekoeffizienten (B-Splines Grad 3, 2 innere Knoten) mit $k$-means ($k=4$) und euklidischer Distanz, Produkt 63 – Stichprobe von 100 HH; Clusterzentren in rot eingezeichnet.
**Clusterverfahren 4**  
Das Partitionieren der mit Hilfe der B-Spline-Expansion angepassten Funktionswerte an den Erhebungszeitpunkten ergibt auf Kalenderwochenebene für Produkt 63 die in Abbildung 7.11 dargestellte Lösung.

Abbildung 7.11.: Verlauf der Clusterzentren – Clustern der vorhergesagten Werte der Splines (B-Splines Grad 3, 2 innere Knoten) an den Erhebungszeitpunkten mit k-means (k=4) und euklidischer Distanz, Produkt 63, KW-Ebene

Die Zentren verlaufen sowohl auf Wochen- als auch auf Monatsebene (vgl. Anhang A) ähnlich zu denen aus Clusterverfahren 1 – Clustern der Rohdaten mit dem k-means Algorithmus unter Verwendung der euklidischen Distanz – mit dem Unterschied, dass sie geglättet sind. Dies entspricht genau der dahinterliegenden Idee von Verfahren 4 im Gegensatz zu Verfahren 1, der Glättung der Rohdaten aufgrund eines unterstellten funktionalen Verlaufs. So ergibt sich sowohl auf Wochen- als auch auf Monatsebene das klassische Cluster der Nichtkäufer und drei weitere, gleichförmig verlaufende Centroide auf unterschiedlichem Niveau. Diese besitzen eine Basiseinkaufsmenge über den Winter von wöchentlich 0.7 (Cluster 4), 0.2 (Cluster 1) und 0.1 (Cluster 3) Packungen Eiscreme, die ab dem Frühjahr bis zum Sommer, Mitte Juli, auf eine Menge von 1.4, 0.7 und 0.2 Packungen ansteigt. Auf Monatsebene liegt das Niveau dieser drei Cluster etwas höher bei ca. 0.5, 1 und 3.5 mit einem Anstieg auf 1, 3 und 6 Packungen pro Monat.

**Clusterverfahren 5**  
Das neue Verfahren des Splineclusterns unter Verwendung der euklidischen Distanz ergibt auf den ersten Blick auf beiden Aggregationsstufen (Kalenderwochen- und Monatsebene) sehr ähnliche Centroide wie Verfahren 4 und auch Verfahren 1. Bei den drei obersten Clusterverläufen in Abbildung 7.12 für die Kalenderwochenebene wird wiederum das Hoch um den Monat Juli herum deutlich. Ansonsten verlaufen diese Zentren fast
Abbildung 7.12.: Verlauf der Clusterzentren – Clustern der Rohdaten mit Splineclustern (B-Splines Grad 3, 2 innere Knoten, df=6) und euklidischer Distanz (k=4), Produkt 63, KW-Ebene

identisch über das Jahr hinweg auf jeweils unterschiedlichem Basisniveau. Die einzelnen Ausgangs- und Maximaleinkaufsmengen entsprechen den Einkäufen bei Clusterverfahren 4. Bis auf das Nichtkäufercluster steigen die Einkäufe von Eiscreme ab Anfang März an, bis sie Mitte Juli mit einer über 100%-igen Steigerung im Vergleich zum Winterniveau ihr Maximum erreichen und danach langsam wieder bis Ende September abfallen (Die Abbildung der Clusterzentren auf Monatsebene befinden sich in Anhang A).

**Clusterverfahren 6** Eine Kombination des Splineclusterns mit der Poisson-distanz, das sowohl den funktionalen Verlauf als auch die Verteilung der Daten berücksichtigt, ergibt bei wöchentlicher Betrachtung die Centrpoide in Graphik 7.13. Hier zeigt sich im Gegensatz zu den Clusterverfahren mit euklidischer Distanz eine etwas „gestauchte“ Clusterlösung. Die stärkeren Käufe in den Sommermonaten bei den oberen zwei Clusterzentren sind in diesem Fall weniger stark ausgeprägt. Vom Niveau entsprechen diese, jeweils die Aggregationsebenen für sich betrachtet, dem Clusterverfahren 2, dem Clustern der Rohdaten mit der neu entwickelten Poisson-Distanz. Das Basisniveau bei den drei neben den Nichtkäufern existierenden Clustern liegt bei ca. 0.1, 0.2 bzw. 0.5 wöchentlichen bzw. 0, 0.5 und 2 monatlichen Käufen (vgl. Anhang A für die Monatebene) mit ebenfalls einem i.d.R. um bis zu 100% stärkerem Konsum um den Monat Juli herum.
7.2 Reale Transaktionsdaten

Abbildung 7.13.: Verlauf der Clusterzentren – Clustern der Rohdaten mit Splinecluster (B-Splines Grad 3, 2 innere Knoten, df=6) und Poisson-Distanz (k=4), Produkt 63, KW-Ebene

**Vergleich** Anhand der Graphiken der Clusterzentren wird bereits deutlich, dass sich die einzelnen Clusterverfahren mehr oder minder stark unterscheiden. So stellt sich die Frage nach einer Maßzahl, die die Ähnlichkeit der Ergebnisse der betrachteten Möglichkeiten miteinander oder mit dem Standardverfahren geeignet vergleichen kann.

**Randindex** Die bekannteste Methode zwei gegebene Partitionen miteinander zu vergleichen ist der sogenannte Randindex (vgl. [Rand (1971)]). Der Randindex beruht auf dem Vergleich, wie Objektpaare, also jeweils zwei einzelne Beobachtungen, geclustert werden. Dabei wird erfasst, ob die beiden betrachteten Objekte bei zwei Partitionen $\mathcal{C}(X)$ und $\mathcal{C}(Y)$ in jeweils die gleiche Gruppe gelangen oder beispielsweise in Partition $\mathcal{C}(X)$ schon und in $\mathcal{C}(Y)$ nicht. Allgemein differenziert man zwischen Übereinstimmungen und Nichtübereinstimmungen. Als Übereinstimmung zählen Objektpaare, bei denen beide Objekte sowohl in $\mathcal{C}(X)$ als auch in $\mathcal{C}(Y)$ im gleichen oder nicht im gleichen Cluster sind. Dagegen ergeben sich Nichtübereinstimmungen bei Objekten, die in $\mathcal{C}(X)$ im selben und in $\mathcal{C}(Y)$ nicht im selben Cluster verzeichnet werden, oder umgekehrt. Zwei nach dem Randindex ähnliche Partitionen haben eine relativ hohe Anzahl an Übereinstimmungen, während eher unterschiedliche Partitionen eine hohe Anzahl an Nichtübereinstimmungen aufweisen.
7.2 Reale Transaktionsdaten

Berechnung des Randindex:

Ausgangspunkt für die Berechnung des Randindex sind die in der Kontingenztafel über die Partitionen \( \mathcal{C}(X) = \{ \mathcal{C}_1(X), \mathcal{C}_2(X), \ldots, \mathcal{C}_R(X) \} \) und \( \mathcal{C}(Y) = \{ \mathcal{C}_1(Y), \mathcal{C}_2(Y), \ldots, \mathcal{C}_S(Y) \} \) zusammengesetzten Überschneidungen:

<table>
<thead>
<tr>
<th>( \mathcal{C}_1(X) )</th>
<th>( \mathcal{C}_2(X) )</th>
<th>( \ldots )</th>
<th>( \mathcal{C}_S(X) )</th>
<th>( \mathcal{C}_1(Y) )</th>
<th>( \mathcal{C}_2(Y) )</th>
<th>( \ldots )</th>
<th>( \mathcal{C}_S(Y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_{11} )</td>
<td>( n_{12} )</td>
<td>( \ldots )</td>
<td>( n_{1S} )</td>
<td>( n_{21} )</td>
<td>( n_{22} )</td>
<td>( \ldots )</td>
<td>( n_{2S} )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \ddots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \ddots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( \mathcal{C}_R(X) )</td>
<td>( n_{R1} )</td>
<td>( n_{R2} )</td>
<td>( \ldots )</td>
<td>( n_{RS} )</td>
<td>( n_{R+} )</td>
<td>( n_{++} = n )</td>
<td></td>
</tr>
</tbody>
</table>

mit \( n_{ij} = |\mathcal{C}_i(X) \cap \mathcal{C}_j(Y)| \) als Anzahl gemeinsamer Objekte in Gruppe \( \mathcal{C}_i(X) \) und \( \mathcal{C}_j(Y) \).

Mit dem Randindex wird die Wahrscheinlichkeit der Übereinstimmung bezeichnet. Diese berechnet sich als Anzahl der Übereinstimmungen relativ zur Anzahl der Übereinstimmungen und Nichtübereinstimmungen.

Die Anzahl der Objekte, die in beiden Partitionen im selben Cluster sind, berechnet sich gemäß

\[
\frac{1}{2} \sum_{i=1}^{R} \sum_{j=1}^{S} n_{ij}(n_{ij} - 1) \quad (7.1)
\]

und die der Objekte, die in beiden Partitionen in unterschiedlichen Clustern sind, nach

\[
\frac{1}{2} \left[ \sum_{i=1}^{R} \sum_{j=1}^{S} n_{ij}^2 - \left( \sum_{i=1}^{R} n_{i+}^2 + \sum_{j=1}^{S} n_{+j}^2 \right) \right] \quad (7.2)
\]

Die Anzahl der Übereinstimmungen (entspricht der Summe aus (7.1) und (7.2)) und Nichtübereinstimmungen berechnet sich als Binomialkoeffizient \( \binom{n}{2} \).

Insgesamt ergibt sich für den Randindex (RI):

\[
RI = \frac{\binom{n}{2} + \sum_{i=1}^{R} \sum_{j=1}^{S} n_{ij}^2 - \frac{1}{2} \left( \sum_{i=1}^{R} n_{i+}^2 + \sum_{j=1}^{S} n_{+j}^2 \right) \binom{n}{2}}{\binom{n}{2}} \quad (7.3)
\]

Der Randindex besitzt einen Wertebereich von \( RI \in [0,1] \). Er nimmt den Wert 0 an, falls zwei Partitionen keinerlei Ähnlichkeit besitzen und den Wert 1, falls sie vollkommen identisch sind.
Eine Erweiterung des Randindex von [Rand (1971)](Rand) ist der sogenannte *adjusted Randindex* (vgl. [Arabie (1985)](Arabie)). Dieser korrigiert den Randindex bezüglich Zufall. Es wird angenommen, dass die vorliegende Kontingenztafel zweier Clusterlösungen einer hypergeometrischen Verteilung entspringt. Dies bedeutet, dass die beiden Partitionen \( \mathcal{C}(X) \) und \( \mathcal{C}(Y) \) zufällig, unter der Bedingung der echten Anzahl an Klassen und Objekten in jeder Partition, gezogen wurden. Der erwartete Index, unter dieser Annahme, wird nun in die Berechnung miteinbezogen, und der adjusted Randindex ergibt sich als:

### Berechnung des adjusted Randindex (aRI):

(gegl. Arabie [1985])

Die allgemeine Formel eines Index mit Zufallskorrektur lautet:

\[
\text{aRI} = \frac{\text{Index} - \text{ExpectedIndex}}{\text{MaximumIndex} - \text{ExpectedIndex}}
\]

Beim Einsetzen der konkreten Formeln für die Berechnung des Randindex und des erwarteten Index, sowie des maximalen Index ergibt sich:

\[
aRI = \frac{\frac{\sum_{i=1}^{R} \sum_{j=1}^{S} \binom{n_{ij}^2}{2} - \sum_{i=1}^{R} \binom{n_{i+}}{2} \sum_{j=1}^{S} \binom{n_{+j}}{2}}{2}}{\frac{\sum_{i=1}^{R} \binom{n_{i+}}{2} + \sum_{j=1}^{S} \binom{n_{+j}}{2}}{2} - \sum_{i=1}^{R} \binom{n_{i+}}{2} \sum_{j=1}^{S} \binom{n_{+j}}{2}}}
\]

Auch der *adjusted Randindex* besitzt den Wertebereich \( \text{aRI} \in [0, 1] \). Beim Wert 0 entspricht der aRI dem erwarteten Index.

In Tabelle 7.1 werden die (adjusted) Randindizes der verschiedenen Verfahren zueinander als Kreuztabelle dargestellt.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rohdaten, k-means, eukl</td>
<td>0.00</td>
<td>0.41</td>
<td>0.48</td>
<td>0.88</td>
<td>0.90</td>
<td>0.40</td>
</tr>
<tr>
<td>Rohdaten, k-means, poi</td>
<td>0.41</td>
<td>1.00</td>
<td>0.30</td>
<td>0.42</td>
<td>0.41</td>
<td>0.89</td>
</tr>
<tr>
<td>Splinekoeffizienten, k-means, eukl</td>
<td>0.48</td>
<td>0.30</td>
<td>1.00</td>
<td>0.48</td>
<td>0.48</td>
<td>0.28</td>
</tr>
<tr>
<td>Splinewerte, k-means, eukl</td>
<td>0.88</td>
<td>0.42</td>
<td>0.48</td>
<td>1.00</td>
<td>0.96</td>
<td>0.40</td>
</tr>
<tr>
<td>Rohdaten, Splinecluster, eukl</td>
<td>0.90</td>
<td>0.41</td>
<td>0.48</td>
<td>0.96</td>
<td>1.00</td>
<td>0.40</td>
</tr>
<tr>
<td>Rohdaten, Splinecluster, poi</td>
<td>0.40</td>
<td>0.89</td>
<td>0.28</td>
<td>0.40</td>
<td>0.40</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Tabelle 7.1.: Vergleich der Clusterlösungen der unterschiedlichen Verfahren, Produkt 63, KW-Ebene, k=4

---

14: Der adjusted Randindex ist in der Funktion `Rand` im Package `flexclust` in R implementiert.  
15: Im folgenden ist immer mit `Randindex der adjusted Randindex` gemeint.
Auffällig sind die hohen Randindizes zwischen Verfahren 1 – Clustern der Rohdaten mit euklidischer Distanz – mit Verfahren 4 – Clustern der Splinewerte mit euklidischer Distanz – (0.88) und Verfahren 5 – Splineclustern mit euklidischer Distanz – (0.90). Somit werden die bereits graphisch deutlichen Ähnlichkeiten beim Centroidverlauf bestätigt. Auch bei den die Poisson-Distanz verwendenden Clustermöglichkeiten 2 – Clustern der Rohdaten – und 6 – Splineclustern – zeigt sich ein hoher Index von 0.89. Ebenso wurden in diesem Fall bereits optische Gemeinsamkeiten der Clusterverläufe festgestellt. Beim Standardverfahren, dem Clustern der Splinekoeffizienten, das bei den Centroiden von allen anderen Verfahren stark abweichende Strukturen aufdeckte, zeigen sich bestätigende Werte des Randindex. So liegt dieser bei einem Vergleich mit den übrigen Clustermethoden nur zwischen 0.28 und 0.48. Dies spricht für eine größere Unähnlichkeit bei der Clusterzuordnung. Ebenfalls bestehen solche Unterschiede beim Vergleich zwischen auf Poisson-Distanz und auf euklidischer Distanz basierenden Methoden. Das Clustern der Rohdaten mit Poisson-Distanz ergibt mit einem Randindex von 0.30 bis 0.42 zu den Verfahren mit klassischem Distanzmaß, gewichtigere Abweichungen in der Clusterzuordnung. Das gleiche Bild zeigt die Gegenüberstellung mit dem Splineclustern unter der Poisson-Distanz (Verfahren 6), hier liegt der Index zwischen 0.28 und 0.40. Diese Auffälligkeiten werden in Abschnitt 7.2.2 näher untersucht.

7.2.1.2. Produkt 61 – Mineralwasser

7.2.1.3. Produkt 44 – Alkoholfreie Getränke ohne Kohlensäure (Fruchthaltige)

Bei den alkoholfreien Getränken ohne Kohlensäure kommt es auf Wochenebene zu den in Graphik 7.15 dargestellten Centroiden.

Die vertraute Gruppe der Nichtkäufer ist auch hier auf Kalenderwochen- und Monatsebene über alle Verfahren hinweg vertreten (für die Ergebnisse auf Monatsebene vgl. Anhang A). Augenfällig sind wiederum die Zusammenhänge der Lösungen zwischen den Verfahren 1, 4 und 5 mit der euklidischen Distanz sowie zwischen 2 und 6 mit der Poisson-Distanz. Bei diesen Lösungen existieren wiederum zwei mittlere, relativ parallel und konstant über die Zeit verlaufende Einkaufsmuster, die bei den ersten Verfahren wiederum etwas höher liegen als bei den Verfahren mit der Poisson-Distanz. So liegen beispielsweise die Zentren auf Monatsebene im ersten Fall bei ca. 3 und 6 Einkäufen pro Monat, während sie im zweiten Fall bei 1 und 3.5 liegen. Das obere Cluster, das meist mehr Struktur aufweist, zeigt auch bei diesem Produkt den markantesten Verlauf. Bei den alkoholfreien Geträn-
Abbildung 7.15.: Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 44, KW-Ebene

*ken ohne Kohlensäure* verzeichnet sich ein ausgeprägtes Frühjahrs- hoch und ein weiterer kleiner Anstieg in den Transaktionen im Frühsommer zwischen Mitte Mai und Mitte Juli. Diese Hochpunkte nivellieren sich etwas bei den Clusterverfahren 2 und 6, sowie bei einer groberen Betrachtungsweise auf Monatsebene.

### 7.2.1.4. Produkt 19 – Zahnpasta

Der Verlauf der Zentren auf Monatsebene bei der als relativ konstant bezüglich des Kaufs eingeschätzten *Zahnpasta* (Produkt 19) ist Abbildung 7.16 zu entnehmen. Markant ist hier, dass die einzelnen typischen Transaktionsverläufe sehr flach verlaufen und auf Monatsebene nur zwischen keinem und maximal 4 Einkäufen pro Monat differenzieren. Das Cluster mit den *Nullkäufern* ist wiederum in allen Lösungen über beide Aggregationsstufen hinweg vorhanden. Daneben sind zwei mittlere Gruppen, die beide um einen Einkauf pro Monat schwanken und eine letzte Gruppe, die sich bei Verfahren 1, 4, und 5 bei 4 Zahnpastakäu-
fen pro Monat bzw. bei den restlichen Verfahren bei ca. 3 bewegt. Auch in diesem Fall ist das gegenläufige Verhaltensmuster der beiden mittleren Centroiden beim Clustern der Splinekoeffizienten in Verfahren 3 hervorstechend. Auf Kalenderwochenebene lassen sich kaum noch sinnvolle Unterschiede zwischen dem Einkaufsverhalten erkennen, da die einzelnen Gruppen alle nur keines oder ein Produkt in der Woche kaufen (für die Ergebnisse auf Kalenderwochenebene vgl. Anhang A).

Abbildung 7.16.: Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 19, Monatsebene

7.2.2. Vergleich der Clusterlösungen der Transaktionsdaten

Bereits aus Tabelle 7.1 auf Seite 45 wurde klar, dass die einzelnen betrachteten Clusterverfahren mehr oder weniger starke Abweichungen bei der Klasseneinteilung der Haushalte besitzen. Interessant ist hier, ob die Ähnlichkeiten der Verfahren zueinander über alle betrachteten Produkte und auch über die beiden Aggregationsniveaus, also sowohl auf
7.2 Reale Transaktionsdaten


<table>
<thead>
<tr>
<th>Clusterverfahren i – j</th>
<th>Produkt</th>
<th>63</th>
<th>61</th>
<th>44</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Produkt</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Aggregationsniveau</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 – 2</td>
<td>KW</td>
<td>0.41</td>
<td>0.33</td>
<td>0.45</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>Monat</td>
<td>0.40</td>
<td>0.37</td>
<td>0.47</td>
<td>0.21</td>
</tr>
<tr>
<td>1 – 5</td>
<td>KW</td>
<td>0.90</td>
<td>0.97</td>
<td>0.97</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>Monat</td>
<td>0.93</td>
<td>0.98</td>
<td>0.98</td>
<td>0.59</td>
</tr>
<tr>
<td>4 – 5</td>
<td>KW</td>
<td>0.96</td>
<td>0.99</td>
<td>0.98</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>Monat</td>
<td>0.96</td>
<td>0.98</td>
<td>0.99</td>
<td>0.97</td>
</tr>
<tr>
<td>2 – 6</td>
<td>KW</td>
<td>0.89</td>
<td>0.82</td>
<td>0.98</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>Monat</td>
<td>0.93</td>
<td>0.96</td>
<td>0.99</td>
<td>0.49</td>
</tr>
<tr>
<td>5 – 6</td>
<td>KW</td>
<td>0.40</td>
<td>0.37</td>
<td>0.45</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td>Monat</td>
<td>0.40</td>
<td>0.36</td>
<td>0.46</td>
<td>0.41</td>
</tr>
</tbody>
</table>

Tabelle 7.2.: Randindex zwischen Clusterverfahren i und Clusterverfahren j aufgeteilt nach Produkt und Aggregationsniveau

Verfahren 4 – 5: Unterschied Clustern der Splinewerte und Splineclustern bei euklidischer Distanz Tatsächlich lassen sich, wie vermutet, die größten Randindizes zwischen Clusterverfahren 4 und 5, also Clustern der Splinewerte und Splineclustern mit euklidischer Distanz feststellen (vgl. Tabelle 7.2). Einzig auffällig ist hier die Betrachtung des Produktes 19 auf Kalenderwochenebene. Dort liegt der Index beim Vergleich dieser beiden Verfahren auf wöchentlicher Basis zwar immer noch am höchsten, ist aber generell mit 0.58
7.2 Reale Transaktionsdaten


**Verfahren 1 – 5 und 2 – 6: Unterschied Splineclustern und Clustern der Rohdaten**

Tabelle 7.2 zeigt außerdem, dass auch das Clustern der Rohdaten mit der euklidischen Distanz (Verfahren 1) starke Überschneidungen hinsichtlich der Klasseneinteilung mit dem Splineclustern (Verfahren 5) hat. Hier schwanken, bis auf die strukturarme *Zahnpasta*, die Randindizes zwischen 0.90 und 0.98, also zwischen extrem ähnlicher bis nahezu identischer Gruppenbildung. Insgesamt ist der Zusammenhang in diesem Fall nicht ganz so hoch wie zwischen Verfahren 4 und 5.

Der Unterschied zwischen der funktionalen Herangehensweise beim Splineclustern auf der einen Seite und einem Ignorieren dieser funktionalen Form beim Clustern der Rohdaten auf der anderen Seite, lässt sich bei einem zusätzlichen Vergleich von Verfahren 2 mit Verfahren 6 beurteilen. So zeigt sich bei den strukturstärkeren Produkten 63, 61 und 44 große Ähnlichkeit zwischen dem Clustern der Rohdaten mit der neuen Poisson-Distanz und dem Splineclustern mit dieser. Der Randindex variiert hier von 0.82 bis sogar 0.99, also fast vollkommen identischer Zuordnung. Die Beachtung des glatten Kurvenverlaufs der Transaktionen durch das Splineclustern scheint somit bei den realen Daten, sowohl bei Verwendung der euklidischen als auch der Poisson-Distanz, keine größeren Abweichungen bei der Klasseneinteilung hervorzubringen.

**Verfahren 1 – 2 und 5 – 6: Unterschied im Distanzmaß**

Bei Betrachtung der Randindizes, die den Zusammenhang der Clusterlösungen 1 mit 2 bzw. 5 mit 6, also die Clusterverfahren ohne Berücksichtigung der funktionalen Struktur der Daten und, als Gegenstück, dem Splineclustern mit der Berücksichtigung, zwischen den beiden Distanzmaßen widergeben, lässt sich ein deutlicher Unterschied erkennen. Anders als beim Vergleich der Verfahren innerhalb gleicher Distanzen, führt ein unterschiedliches Distanzmaß zu weit größeren Differenzen in der Klassenbildung. Die Indizes liegen hierbei nie über 0.47.

**Verfahren 3: Unterschied zum Standardverfahren**

Abschließend soll kurz auf das Standardverfahren, dem Clustern von Splinekoefizienten mit dem $k$-means Algorithmus, eingegangen werden. Bereits anhand der Verläufe der Clusterzentren ließen sich größere Abweichungen bei oft zwei von vier völlig anderes verlaufenden Centroiden feststellen. Auf Wochenebene erreicht der Randindex des Verfahrens 3 zu je einer aller anderen Cluster-

51
7.3 Simulierte Daten

Um nun nicht nur die Ähnlichkeit der Verfahren zueinander, sondern auch deren Leistung in Hinsicht auf das Auffinden der wahren Klassenzugehörigkeit beurteilen zu können, müssen die einzelnen Gruppen im Vorfeld des Clusterns bekannt sein. Aus diesem Grund wurden für diese Fragestellung Daten simuliert. Wie bereits in Abschnitt 3.1.2 erwähnt, ist der Splinegrad und auch die Anzahl und die Orte der gesetzten Knoten sehr entscheidend für die Funktionsanpassung. Beim Testen der einzelnen Clusterlösungen sollten diejenigen Möglichkeiten, die eine splineinterpolation beinhalten, den Vorteil erhalten, denselben Splinegrad und dieselbe Knotenmenge zu verwenden, die in den Daten steckten. Also sollten nur Kurven geclustert werden, von denen auch die Extrema durch diese Vorgaben ausreichend dargestellt werden können. Um dies zu erreichen, wurden für die Simulation die Clusterergebnisse für das Clustern der Splinekoeffizienten mit dem k-means-Algorithmus (Verfahren 3) aus der vorhergehenden Analyse verwendet. Die Splineexpansionen in den entsprechenden Verfahren sind damit mit den simulierten Daten hinsichtlich der Form, die sie darstellen können, abgestimmt. Bei einer solchen Herangehensweise wird die größtmögliche Aufdeckung der wahren Cluster erwartet. Entsprechen Splinegrad und Knoten dagegen nicht den Vorgaben, ist es für das Verfahren aufgrund mangelder oder auch zu großer Flexibilität in der Splineanpassung sicher schwerer, die richtigen Klassen zu finden.

Abbildung \[7.17\] verdeutlicht dies am Beispiel des Clusterns von Splinekoeffizienten bei poissonverteilt simulierten Daten auf Basis von Produkt 63 – Eiscreme. Dadurch, dass die simulierten Daten auf denselben Vorgaben beruhen, wie das Clustern selbst, entsprechen
sich die Kurvenverläufe der Splineexpansionen der simulierten Daten und der Kurvenverlauf der Clusterzentren viel mehr, als dies z.B. bei den realen Daten in Abbildung 7.10 der Fall ist.

Abbildung 7.17.: angepasste Splines nach Clusterzuordnung – Clustern der Splinekoeffizienten (B-Splines Grad 3, 2 innere Knoten) mit k-means (k=4) und euklidischer Distanz, poissonverteilt simulierten Daten, Produkt 63, Stichprobe von 100 HH

Bei den simulierten Daten ist es nun möglich eine eher den Daten angemessene Splineexpansion zu verwenden, da die Muster der Simulationsbasis entsprechend kurvig verlaufen.

In Abbildung 7.18 sind die Verläufe der Clusterzentroide, die nun die Grundlage für die Simulation stellen, für alle vier Produktgruppen auf der Kalenderwochenebene aufgezeichnet. Die Verläufe auf Monatsebene finden sich in Anhang C. Die Clusterzentren für Produkt 61 – Mineralwasser und Produkt 44 – Alkoholfreie Getränke ohne Kohlensäure (Fruchthaltige) ähneln sich dabei im Verlauf. Es gibt jeweils ein Zentrum, das relativ nahe bei 0 verläuft und ein Zentrum mit einem deutlich höheren Verlauf. Bei Letzterem bewegt sich die Einkaufsmenge auf Kalenderwochenebene bei beiden Produkten um 2 Stück und auf Monatsebene bei ca. 10 bzw. ca. 7 Käufen. Zwei weitere Centroiden verlaufen auf etwa gleichem Niveau, aber wechseln sich dabei in der Schwankung ab. Während der Transaktionsverlauf bei einem Zentrum am Messzeitpunkt hoch ist, ist dieser beim anderen niedrig.
Diese Schwankungen liegen in einem Fenster von ungefähr 0.5 Einkäufen auf Kalenderwochenbene und ca. einem bei monatlicher Betrachtung. Produkt 63 weist ebenfalls einen Clusterverlauf nahe bei 0 auf, zwei Zentren, die abwechselnd schwanken und eines, dass sich auf etwas höherem Niveau ungefähr gleichläufig zu einem der beiden mittleren Cluster bewegt. Insgesamt finden die Bewegungen im Einkaufsverhalten in einem kleinen Bereich von 0 bis 1 Einkauf auf Kalenderwochenbene, sowie zwischen 0 und 5 Einkäufen auf Monatsebene, statt. Produkt 19 weist keine großen Unterschiede in der Anzahl der Einkäufe auf. Die Zentren bewegen sich alle nur in einem Spektrum von 0 bis 0.5 Einkäufe pro Woche bzw. 0 bis ca. 2.5 pro Monat. Es existieren dennoch zwei relativ konstant laufende Centroiden, einer am unteren Rand und einer am oberen Rand. Die beiden mittleren Cluster schwanken wiederum abwechselnd. Trotz des unterschiedlichen Verlaufs, lässt sich aufgrund der geringen Schwankungsbreite, diese Simulationsbasis als strukturarm bezeichnen und dient dazu, das Verhalten der Clusterlösungen bei solchen Daten zu untersuchen.

<table>
<thead>
<tr>
<th></th>
<th>50</th>
<th>100</th>
<th>150</th>
<th>200</th>
<th>250</th>
<th>300</th>
<th>350</th>
<th>400</th>
<th>450</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.59</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.61</td>
<td>0.62</td>
<td>0.66</td>
<td>0.68</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.67</td>
<td>0.73</td>
<td>0.68</td>
<td>0.71</td>
<td>0.73</td>
<td>0.74</td>
<td>0.73</td>
<td>0.70</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.35</td>
<td>0.36</td>
<td>0.30</td>
<td>0.34</td>
<td>0.33</td>
<td>0.32</td>
<td>0.35</td>
<td>0.33</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.71</td>
<td>0.76</td>
<td>0.73</td>
<td>0.74</td>
<td>0.78</td>
<td>0.75</td>
<td>0.77</td>
<td>0.74</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.71</td>
<td>0.76</td>
<td>0.75</td>
<td>0.74</td>
<td>0.61</td>
<td>0.75</td>
<td>0.77</td>
<td>0.74</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.73</td>
<td>0.77</td>
<td>0.73</td>
<td>0.76</td>
<td>0.80</td>
<td>0.75</td>
<td>0.76</td>
<td>0.76</td>
<td>0.77</td>
<td></td>
</tr>
</tbody>
</table>

Tabelle 7.3.: Randindex zwischen wahrer Clusterzugehörigkeit und der entsprechenden Clusterlösung nach Anzahl simulierter Beobachtungen pro Cluster, poissonverteilte simulierte Daten, Produkt 61, KW-Ebene


16 Die Funktion zur Simulation der poissonverteilten Daten mit entsprechender Klassenzuordnung ist in einer bisher unveröffentlichten Erweiterung des *flexclust* Paketes als Funktion *expPoisson* implementiert, © Friedrich Leisch
17 Die Tabellen und Graphiken für alle Produkte und Aggregationsebenen finden sich in Anhang F.
7.3 Simulierte Daten

Abbildung 7.19.: Verlauf des Randindex für die einzelnen Clusterlösungen nach steigender Anzahl simulierter Beobachtungen – poissonverteilt simulierte Daten, Produkt 61, KW-Ebene

normalverteilte Daten Zusätzlich zu poissonverteilt simulierten Daten wurden auch normalverteilte Daten mit Hinblick auf das Verhalten des neu entwickelten Splineclusterns aus Abschnitt 6.2 untersucht. Bei der Simulation wurden, analog zur poissonverteilt simulierten Vorgehensweise, die Beobachtungen zu den Zeitpunkten als unabhängig gesehen und aus einer Normalverteilung mit $\mu = c^{(k)}(t_j)$ gleich dem $k$ – ten Centroid zu dem entsprechenden Zeitpunkt $t_j$ und einer festen Varianz $\sigma^2 = 1$ gezogen.

7.3.1. Verläufe der Clusterzentren mit Vergleich

7.3.1.1. poissonverteilt simulierte Daten

Auch bei den poissonverteilt simulierten Kurven wurden alle sechs vorgestellten Möglichkeiten des Clusterns (siehe Box auf Seite 32) durchgeführt. Wie schon bei den realen Transaktionsdaten ähneln sich einige Clusterzentrenverläufe von einzelnen Verfahren mehr als andere. Abbildung 7.20 zeigt exemplarisch die Clusterlösungen für die poissonverteilt simulierten Daten nach der Vorlage von Produkt 61 – Mineralwasser auf Wochenebene.\footnote{Für die Clusterlösungen aller Produkte bei poissonverteilt simulierten Daten siehe Anhang D}

\footnotetext[18]{Für die Clusterlösungen aller Produkte bei poissonverteilt simulierten Daten siehe Anhang D}
Deutlich erkennbar sind Zusammenhänge zwischen Verfahren 1, 2, 4 und 6. Anders als vorher ergibt Clusterlösung 5 ein differentes Bild zu dem meist ähnlich verlaufenden Muster beim Clustern der Rohdaten mit euklidischer Distanz (Verfahren 1). Dagegen bleibt der bekannte Unterschied des Standardverfahrens (Lösung 3) mit allen anderen Verfahren auch bei den poissonverteilten simulierten Daten bestehen.

Eine Beurteilung der Zusammenhänge anhand des Randindex ist in Tabelle 7.4 aufgeführt. Klar ersichtlich sind die höheren Indizes zwischen den Verfahren 1 und 2, 4 und 6, sowie die zwischen 2 und 4 bzw. 2 und 6. Das Splineclustern mit der euklidischen Distanz (Verfahren 5) hat bei diesen Daten nur noch Indizes bis zu 0.63 im Vergleich zu anderen Verfahren. Die Ähnlichkeiten zwischen Verfahren mit gleichen Distanzmaßen, wie sie bei den realen Transaktionsdaten in Abschnitt 7.2.1.1 auf Seite 43 festgestellt wurden, kann bei den hier simulierten Daten nicht bestätigt werden. Die Frage, die sich stellt, ist, ob bei den poissonverteilten simulierten Daten insgesamt eventuell andere oder auch gleiche Ähnlichkeiten, wie bei den realen Daten, entdeckt werden können. So sollte noch eine weitere Lösung
7.3 Simulierte Daten

betrachtet werden, beispielsweise die für das Produkt 63 in Abbildung 7.21.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rohdaten, k-means, eukl</td>
<td>1.00</td>
<td>0.79</td>
<td>0.34</td>
<td>0.76</td>
<td>0.63</td>
<td>0.75</td>
</tr>
<tr>
<td>Rohdaten, k-means, poi</td>
<td>0.79</td>
<td>1.00</td>
<td>0.34</td>
<td>0.81</td>
<td>0.62</td>
<td>0.82</td>
</tr>
<tr>
<td>Splinekoeffizienten, k-means, eukl</td>
<td>0.34</td>
<td>0.34</td>
<td>1.00</td>
<td>0.39</td>
<td>0.31</td>
<td>0.40</td>
</tr>
<tr>
<td>Splinewerte, k-means, eukl</td>
<td>0.76</td>
<td>0.81</td>
<td>0.39</td>
<td>1.00</td>
<td>0.63</td>
<td>0.92</td>
</tr>
<tr>
<td>Rohdaten, Splinecluster, eukl</td>
<td>0.63</td>
<td>0.62</td>
<td>0.31</td>
<td>0.63</td>
<td>1.00</td>
<td>0.62</td>
</tr>
<tr>
<td>Rohdaten, Splinecluster, poi</td>
<td>0.75</td>
<td>0.82</td>
<td>0.40</td>
<td>0.92</td>
<td>0.62</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Tabelle 7.4.: Randindex zwischen den verschiedenen Clusterlösungen, n=500, poissonverteilt simulierte Daten, Produkt 61, KW-Ebene

Abbildung 7.21.: Verlauf der Clusterzentren für die einzelnen Clusterlösungen, Produkt 63, poissonverteilt simulierte Daten, KW-Ebene
Nach rein optischen Gesichtspunkten erscheinen bei diesen simulierten Daten alle Verfahren bis auf das Standardverfahren (Verfahren 3) sehr ähnlich zu sein (vgl. Abbildung 7.21). Aber wie wird der Zusammenhang der einzelnen Methoden durch den Randindex erklärt? Um für alle betrachteten Produktgruppenbasen die Ähnlichkeit zueinander beurteilen zu können, wurden wiederum die Randindizes einzelner Verfahren zueinander, analog zu Tabelle 7.2 für die echten Transaktionsdaten, zusammengefasst (siehe Tabelle 7.5).

<table>
<thead>
<tr>
<th>Clusterverfahren i – j</th>
<th>Produkt</th>
<th>Aggregationsniveau</th>
<th>63</th>
<th>61</th>
<th>44</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KW</td>
<td>0.58</td>
<td>0.79</td>
<td>0.93</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Monat</td>
<td>0.73</td>
<td>0.90</td>
<td>0.62</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>1 – 2</td>
<td>KW</td>
<td>0.82</td>
<td>0.63</td>
<td>0.93</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Monat</td>
<td>0.97</td>
<td>0.94</td>
<td>0.97</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>1 – 5</td>
<td>KW</td>
<td>0.97</td>
<td>0.63</td>
<td>0.99</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Monat</td>
<td>0.98</td>
<td>0.99</td>
<td>1.00</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>4 – 5</td>
<td>KW</td>
<td>0.94</td>
<td>0.82</td>
<td>0.92</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Monat</td>
<td>0.96</td>
<td>0.97</td>
<td>0.63</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>2 – 6</td>
<td>KW</td>
<td>0.68</td>
<td>0.62</td>
<td>0.95</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Monat</td>
<td>0.77</td>
<td>0.90</td>
<td>0.95</td>
<td>0.74</td>
<td></td>
</tr>
</tbody>
</table>

Tabelle 7.5.: Randindex zwischen Clusterverfahren i und Clusterverfahren j aufgeteilt nach Produkt und Aggregationsniveau, poissonverteilt simulierte Daten

Verfahren 3: Unterschied zum Standardverfahren und Auffälligkeiten im Aggregationsniveau Analog zu den realen Transaktionsdaten erreicht auch hier das Standardverfahren 3 im Vergleich zu allen anderen Clustermöglichkeiten keine hohen Randindizes, sie liegen zwischen 0.34 und 0.62 (vgl. Anhang E). Dagegen zeigen sich insgesamt bei den poissonverteilt simulierten Daten der strukturstärkeren Produkte (Produkte 63, 61 und 44), anders als zuvor, zwischen den einzelnen Aggregationsebenen selbst größere Abweichungen. Der Vergleich zweier Verfahren auf Wochen- oder Monatsebene ist nicht immer gleich. Um die wahren Auswirkungen einer größeren Betrachtungsweise beurteilen zu können, wurden hier die verschiedenen Ebenen getrennt voneinander auf den Kalenderwochen- und Monatsbasen simuliert und die monatliche Stufe stellt nicht, wie zuvor, nur eine Aggregation der Wochendaten dar.
7.3 Simulierte Daten


Verfahren 1 – 2 und 5 – 6: Unterschied im Distanzmaß Ein Vergleich der unterschiedlichen Distanzen (euklidische oder Poisson) unter Berücksichtigung der funktionalen Form beim Splinecluster (5 – 6) oder ohne Berücksichtigung durch Clustern der Rohdaten (1 – 2) zeigt einen angleichenden Effekt bei den simulierten Daten. Die Abweichungen in der Klassenzuordnung, die in einem unterschiedlichen Distanzmaß begründet sind, erweisen sich als nicht mehr so drastisch, wie bei den realen Daten. Lag vorher der Randindex zwischen Verfahren 1 und 2 bzw. 5 und 6 nie höher als 0.47 (vgl. Tabelle 7.2 auf Seite 50), erreicht er hier sogar Werte bis 0.95.

7.3.1.2. normalverteilt simulierte Daten


In diesem speziellen Fall erkennen alle vier Verfahren ein oberes relativ konstant verlaufendes Cluster bei ca. 2 Einheiten pro Woche. Beim Cluster der Splinekoeffizienten (Verfahren 3) ergibt sich, anders als bei den drei übrigen Verfahren, kein Cluster nahe bei 0, sondern nur eines, dass parallel auf etwas höherem Niveau verläuft. Auffällig sind die fast identisch verlaufenden Lösungen des Splineclusters (Verfahren 5) und des Clusters der angepassten Splinewerte (Verfahren 4). Inwieweit das Clustern der Rohdaten den Klassenzuordnungen der anderen Clustermöglichkeiten entspricht, ist vor allem bei Betrachtung des Verlaufs der beiden mittleren Cluster, nicht ganz deutlich. Das Standardverfahren (Ver-
Abbildung 7.22.: Verlauf der Clusterzentren für die einzelnen Clusterlösungen, Produkt 61, normalverteilte simulierter Daten, KW-Ebene

fahren 3) zeigt im Gegensatz zu Verfahren 4 und 5 etwas ausgeprägtere Schwankungen bei diesen Centroiden.

**Verfahren 3: Unterschied zum Standardverfahren** Ob und wie stark sich Verfahren 3, wie schon bei den poissonverteilten und den realen Daten, bei allen Produkten über beide Aggregationsebenen unterscheidet, soll wiederum anhand der Randindizes zueinander entschieden werden. Diese liegen auf Wochenebene zwischen 0.13 und 0.45, also insgesamt noch etwas niedriger als bei den poissonverteilten simulierten Daten. Dagegen zeigt die monatliche Betrachtung ein durchwachsenes Bild. Hier liegen die Indizes tendenziell hoch. Sie reichen von 0.48 bis 0.92 (vgl. Tabellen im Anhang E). Tabelle 7.6 beinhaltet die Indizes aller Produkte für einen Vergleich zwischen dem Splineclustern und dem Clustern der Rohdaten, sowie zu dem Clustern der angepassten Splinewerte.
7.3 Simulierte Daten

Verfahren 1 – 5: Unterschied Clustern der Rohdaten und Splineclustern  
Bei den vorliegenden Daten sind Verfahren 1 und 5 auf Monatsebene gemessen am Randindex fast identisch, während sich auf Wochenebene niedrigere Ähnlichkeiten abzeichnen. Vor allem bei den strukturarmen Verläufen der Basen von Produkt 63 und 19, die sich auf Wochenebene meist nur zwischen 0 und einer Einheit bewegen, sind die am Randindex bewerteten Unterschiede bei den beiden Verfahren hoch. Die Ähnlichkeit ist bei differenzierteren Transaktionsverläufen dagegen stärker gegeben.

Verfahren 4 – 5: Unterschied Clustern der Splinewerte und Splineclustern  
Ob es einen Unterschied zwischen Verfahren 4 und 5 gibt, lässt sich ebenfalls nicht so eindeutig beantworten. Bei den normalverteilt simulierten Daten ähneln sich die Clusterzuordnungen manchmal mehr und manchmal weniger. Das durchschnittliche Bild hängt von den jeweiligen Ausgangsdaten ab. Oft sind beide Verfahren aber in der Clusterzuordnung so gut wie identisch, wie beispielsweise auf Monatsebene bei den Produkten 63, 61 und 19.

<table>
<thead>
<tr>
<th>Clusterverfahren i – j</th>
<th>Aggregationsniveau</th>
<th>Produkt</th>
<th>63</th>
<th>61</th>
<th>44</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 – 5</td>
<td>KW</td>
<td>0.72</td>
<td>0.97</td>
<td>1.00</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Monat</td>
<td>0.99</td>
<td>1.00</td>
<td>0.63</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>1 – 5</td>
<td>KW</td>
<td>0.45</td>
<td>0.71</td>
<td>0.90</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Monat</td>
<td>0.98</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
<td></td>
</tr>
</tbody>
</table>

Tabelle 7.6.: Randindex zwischen Clusterverfahren i und Clusterverfahren j aufgeteilt nach Produkt und Aggregationsniveau, normalverteilt simulierte Daten

7.3.2. Vergleich der einzelnen Verfahren hinsichtlich der richtigen Clusterzuordnung

Die Verläufe der Centroide der einzelnen Verfahren lassen sich aufgrund der bekannten Clustercentroide der Simulationsbasis gut beurteilen.

7.3.2.1. poissonverteilt simulierte Daten

Abbildung 7.21 enthält, getrennt für die sechs betrachteten Clustermöglichkeiten, die Verläufe aller Zentren für Produkt 63 – Eiscreme (Haushaltspackungen) auf Wochenebene. Ein Vergleich der Zentren mit der Simulationsbasis in Abbildung 7.18 gibt Auskunft über
die Leistungsfähigkeit der entsprechenden Clusterverfahren. Deshalb werden in den Abbildungen 7.23 und 7.24 die wahren Centroidverläufe der Basis und beispielsweise die des Clusterergebnisses des Standardverfahrens bzw. des Splineclusterings mit der Poissondistanz übereinandergelegt.


Abbildung 7.24.: Verlauf der Clusterzentren – Clustern der Rohdaten mit Splineclustern (B-Splines Grad 3, 2 innere Knoten, df=6) und Poisson-Distanz (k=4), poissonverteilte simulierte Daten, Produkt 63, KW-Ebene mit eingezeichneten wahren Verläufen der Simulationsbasis.

Tabelle 7.7.: Randindex zwischen wahrer Klassenzugehörigkeit und den einzelnen Verfahren, poissonverteilte simulierte Daten.
Splinekoeffizienten bei den poissonverteilt simulierten Daten wird nun auch deutlich, dass es dem Verfahren am schlechtesten gelingt die echten Klassen aufzudecken. Der Index erreicht maximal den Wert 0.59. Das neue Verfahren des Splineclusterns in Verbindung mit der Poissonstrecke zeigt bei den vorliegenden Daten ein gutes Ergebnis. Der Randindex weist bei allen Produkten über beide Aggregationsstufen den höchsten oder zumindest mit den höchsten Wert auf und liegt insgesamt zwischen 0.60 und 0.85.

### 7.3.2.2. normalverteilt simulierte Daten

Bei den normalverteilt simulierten Kurven schneidet Verfahren 3 auch in den meisten Fällen am schlechtesten ab. So liegt der Randindex auf der Wochen-ebene bei maximal 0.42. Auf Monatsebene erreicht er dagegen bei dieser Art von Daten weit höhere Werte als bei den poissonverteilt simulierten Daten. Das Analogon zu Verfahren 6 bei poissonverteilt simulierten Beobachtungen ist hier das Splineclustern mit euklidischer Distanz.

Abbildung 7.25.: Verlauf der Clusterzentren – Clustern der Rohdaten mit Splineclustern (B-Splines Grad 3, 2 innere Knoten, df=6) und euklidischer Distanz (k=4), normalverteilt simulierte Daten, Produkt 63, KW-Ebene mit eingezeichneten wahren Clustercentroiden der Simulationsbasis

Auch hier lässt sich feststellen, dass das Splineclustern die größten oder zumindest gleich große Indizes produziert. Insgesamt ist die Aufdeckung der wahren Cluster auf Kalenderwochen-ebene eher schlecht. Auf Monatsebene scheinen die Verfahren 1, 4 und 5 in vielen Fällen für diese Daten gleich gut geeignet zu sein.
7.4 Vergleich der Clusterlösungen

7.4.1. Vergleich der Ähnlichkeiten

In Abschnitt 7.2.2 und 7.3.1 wurden bereits jeweils die Ähnlichkeiten der einzelnen Verfahren zueinander für die realen und die simulierten Daten besprochen. Um zu sehen, ob gewisse Gemeinsamkeiten zwischen den Clusterlösungen auch über beide Datenlagen (real vs. simuliert) hinweg bestehen, enthält Tabelle 7.2 die Randindizes der Vergleiche zwischen den unterschiedlichen Distanzmaßen (1 – 2 und 5 – 6), sowie den unterschiedlichen funktionalen Betrachtungsweisen (1 – 5, 2 – 6 und 4 – 5).

Die Abbildungen 7.26 und 7.27 enthalten die Boxplots der Randindizes zwischen diesen Verfahren, aufgeteilt nach den zugrundeliegenden Daten. Mit A und B werden die realen Transaktionsdaten jeweils auf Kalenderwochen- und Monatsebene und mit C und D bzw. E und F die entsprechend poissonverteilte bzw. normalverteilte simulierten Daten bezeichnet.

Der Median der Randindizes zwischen dem Clustern der Splinewerte (Verfahren 4) und dem Splineclustern mit euklidischer Distanz (Verfahren 5) ist bei allen Daten am höchsten. Dort besteht demnach offensichtlich der größte Zusammenhang. Auffällig ist zudem, dass die geringe Übereinstimmung in der Klassenzuteilung, gemessen durch den niedrigen Index beim Vergleich von Verfahren mit unterschiedlichen Distanzmaßen (1 – 2 und 5 – 6) bei den realen Daten, bei den poissonsimulierten stark abgeschwächt ist. Zwar ist der Median dieser Indizes weiterhin überall am niedrigsten, aber erreicht dennoch Werte von 0.66 bis 0.84. Beim Vergleich des Clusterns der Rohdaten vs. Splineclustern (1 – 5 und 2 – 6) zeigt sich, dass die Gemeinsamkeiten bei den unterschiedlichen Daten immer recht hoch sind,

<table>
<thead>
<tr>
<th>Aggregations-</th>
<th>Verfahren</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>niveau</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Produkt 63</td>
<td>KW</td>
<td>0.25</td>
<td>-</td>
<td>0.09</td>
<td>0.29</td>
<td>0.30</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Monat</td>
<td>0.81</td>
<td>-</td>
<td>0.45</td>
<td>0.82</td>
<td>0.82</td>
<td>-</td>
</tr>
<tr>
<td>Produkt 61</td>
<td>KW</td>
<td>0.63</td>
<td>-</td>
<td>0.36</td>
<td>0.66</td>
<td>0.66</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Monat</td>
<td>0.87</td>
<td>-</td>
<td>0.75</td>
<td>0.87</td>
<td>0.87</td>
<td>-</td>
</tr>
<tr>
<td>Produkt 44</td>
<td>KW</td>
<td>0.72</td>
<td>-</td>
<td>0.42</td>
<td>0.73</td>
<td>0.73</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Monat</td>
<td>0.96</td>
<td>-</td>
<td>0.89</td>
<td>0.63</td>
<td>0.96</td>
<td>-</td>
</tr>
<tr>
<td>Produkt 19</td>
<td>KW</td>
<td>0.15</td>
<td>-</td>
<td>0.09</td>
<td>0.22</td>
<td>0.22</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Monat</td>
<td>0.82</td>
<td>-</td>
<td>0.51</td>
<td>0.82</td>
<td>0.82</td>
<td>-</td>
</tr>
</tbody>
</table>

Tabelle 7.8.: Randindex zwischen wahrer Klassenzugehörigkeit und den einzelnen Verfahren, normalverteilte simulierte Daten
### Tabelle 7.9.: Randindex zwischen verschiedenen Clusterverfahren i – j bei verschiedenen Produkten und Daten;

<table>
<thead>
<tr>
<th>Produkt</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Produkt 63</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 – 2</td>
<td>0.41</td>
<td>0.40</td>
<td>0.58</td>
<td>0.73</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1 – 5</td>
<td>0.90</td>
<td>0.93</td>
<td>0.82</td>
<td>0.97</td>
<td>0.45</td>
<td>0.98</td>
</tr>
<tr>
<td>4 – 5</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
<td>0.98</td>
<td>0.72</td>
<td>0.99</td>
</tr>
<tr>
<td>2 – 6</td>
<td>0.89</td>
<td>0.93</td>
<td>0.94</td>
<td>0.96</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5 – 6</td>
<td>0.40</td>
<td>0.40</td>
<td>0.68</td>
<td>0.77</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Produkt 61</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 – 2</td>
<td>0.33</td>
<td>0.37</td>
<td>0.79</td>
<td>0.90</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1 – 5</td>
<td>0.97</td>
<td>0.98</td>
<td>0.63</td>
<td>0.94</td>
<td>0.71</td>
<td>0.99</td>
</tr>
<tr>
<td>4 – 5</td>
<td>0.99</td>
<td>0.98</td>
<td>0.63</td>
<td>0.99</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>2 – 6</td>
<td>0.82</td>
<td>0.96</td>
<td>0.82</td>
<td>0.97</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5 – 6</td>
<td>0.37</td>
<td>0.36</td>
<td>0.62</td>
<td>0.90</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Produkt 44</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 – 2</td>
<td>0.45</td>
<td>0.47</td>
<td>0.93</td>
<td>0.62</td>
<td>0.90</td>
<td>1.00</td>
</tr>
<tr>
<td>1 – 5</td>
<td>0.97</td>
<td>0.98</td>
<td>0.93</td>
<td>0.97</td>
<td>0.90</td>
<td>1.00</td>
</tr>
<tr>
<td>4 – 5</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>0.63</td>
</tr>
<tr>
<td>2 – 6</td>
<td>0.98</td>
<td>0.99</td>
<td>0.92</td>
<td>0.63</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5 – 6</td>
<td>0.45</td>
<td>0.46</td>
<td>0.95</td>
<td>0.95</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Produkt 19</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 – 2</td>
<td>0.26</td>
<td>0.21</td>
<td>0.62</td>
<td>0.72</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1 – 5</td>
<td>0.54</td>
<td>0.59</td>
<td>0.84</td>
<td>0.97</td>
<td>0.28</td>
<td>0.99</td>
</tr>
<tr>
<td>4 – 5</td>
<td>0.58</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
<td>0.84</td>
<td>1.00</td>
</tr>
<tr>
<td>2 – 6</td>
<td>0.37</td>
<td>0.49</td>
<td>0.86</td>
<td>0.94</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5 – 6</td>
<td>0.47</td>
<td>0.41</td>
<td>0.64</td>
<td>0.74</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
7.4 Vergleich der Clusterlösungen

Abbildung 7.26.: Boxplots der Randindizes der Verfahren i und j zueinander, aufgeteilt nach Datensituation A, B, C und D

Abbildung 7.27.: Boxplots der Randindizes der Verfahren i und j zueinander, aufgeteilt nach Datensituation E und F
wobei eine Betrachtung auf Monatsebene (Daten B, D und E) zu weit ähnlicheren Klassenzuweisungen kommt, als dies bei einer Betrachtung auf wöchentlicher Basis der Fall ist. Der Vollständigkeit wegen, sei an dieser Stelle noch aufgeführt, dass die Gemeinsamkeiten der Verfahren zum Standardverfahren, dem Clustern der Splinekoeffizienten (Verfahren 3), sowohl bei den realen, als auch den simulierten, i.d.R. am geringsten sind.

### 7.4.2. Untersuchung der Leistungsfähigkeit der Verfahren bei poissonverteilt simulierten Daten

Welches Verfahren scheint nun am geeignetsten zum Clustern von Transaktionsverläufen, als spezieller Fall poissonverteilter Zähldaten? In Abschnitt 7.3.2 wurde bereits kurz auf die Ähnlichkeiten der einzelnen Clusterlösungen mit der wahren Klassenzugehörigkeit bei den simulierten Daten eingegangen. Nun soll abschließend geklärt werden, ob bei den poissonverteilten Daten ein Verfahren zum Clustern funktionaler Daten (signifikant) besser – gemessen am Randindex – ist, als ein anderes. Zudem soll auch kurz festgehalten werden, inwieweit ein unterschiedliches Aggregationsniveau eine Rolle beim Abschneiden der Verfahren spielt. Besonderer Fokus wird darüberhinaus auf das Standardverfahren und das neu entwickelte Splineclustern mit Poissondistanz, als speziell für diese Art von Daten entwickelte Methode, gelegt. Schneidet Verfahren 3, wie in Abschnitt 7.3.2 angedeutet, in jedem Fall schlechter ab, als alle anderen betrachteten Clustermethoden und stellt sich das Splineclustern mit der Poissondistanz als signifikant am besten für die Transaktionsdaten heraus?

In einem ersten Schritt lassen sich Unterschiede in Verfahren, Aggregationsniveau und produktspezifischen Verläufen hinsichtlich des Randindex zur wahren Clusterlösung betrachten (vgl. dazu Tabelle 7.7 auf Seite 64). Dazu sind in der Abbildung 7.28 die Boxplots der Randverteilungen dargestellt.


Die Unterscheidung hinsichtlich der Aggregationsebene bringt hervor, dass der Randindex bei einer größeren Betrachtung höher liegt und weniger streut. Dies lässt sich aus Abbildung 7.28 zur Verteilung des Randindex nach KW bzw. Monat im mittleren Bild entnehmen. Der Median des Randindex auf Monatsebene ist hier bei 0.75, während er auf Wochenebene nur bei 0.61 liegt. Eine richtige Klassenzuteilung scheint also durch eine
7.4 Vergleich der Clusterlösungen

Abbildung 7.28.: Boxplots des Randindex
größere Betrachtungsweise begünstigt zu sein.

Aufgeteilt nach den einzelnen Verfahren im untersten Bild wird bestätigt, dass Verfahren 3 wesentlich niedrigere Randindizes hervorbringt, als alle anderen Verfahren. Aus Tabelle 7.10 lässt sich entnehmen, dass der Median des Index hier bei 0.37 liegt, während dieser bei den anderen Verfahren Werte zwischen 0.68 bis hin zu 0.77 beim Splineclustern mit Poissondistanz erreicht. Die Verwendung der Poissondistanz zum Clustern der Rohdaten besitzt den zweithöchsten Medianwert. Diese drei bereits erwähnten Methoden (3, 2 und 6) zeigen einen geringeren Interquartilsabstand und damit eine geringere Streuung im Randindex, als die Verfahren 1, 4 und 5, den Verfahren mit klassischer euklidischer Distanz. Die Methoden mit Poissondistanz liegen demnach bei diesen Daten nicht nur insgesamt höher im Index, sondern erzeugen, über verschiedene Aggregationsniveaus und Transaktionsverläufen betrachtet, gleichmäßigere Ergebnisse.

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>0.466</td>
<td>0.613</td>
<td>0.700</td>
<td>0.683</td>
<td>0.784</td>
<td>0.837</td>
</tr>
<tr>
<td>V2</td>
<td>0.584</td>
<td>0.708</td>
<td>0.745</td>
<td>0.735</td>
<td>0.792</td>
<td>0.832</td>
</tr>
<tr>
<td>V3</td>
<td>0.296</td>
<td>0.334</td>
<td>0.366</td>
<td>0.389</td>
<td>0.406</td>
<td>0.592</td>
</tr>
<tr>
<td>V4</td>
<td>0.502</td>
<td>0.625</td>
<td>0.736</td>
<td>0.706</td>
<td>0.780</td>
<td>0.850</td>
</tr>
<tr>
<td>V5</td>
<td>0.509</td>
<td>0.611</td>
<td>0.681</td>
<td>0.692</td>
<td>0.782</td>
<td>0.854</td>
</tr>
<tr>
<td>V6</td>
<td>0.600</td>
<td>0.755</td>
<td>0.772</td>
<td>0.772</td>
<td>0.833</td>
<td>0.854</td>
</tr>
</tbody>
</table>

Tabelle 7.10.: Kenngrößen der Randindizes der einzelnen Verfahren

Abbildung 7.29 zeigt eine Aufteilung der Boxplots der Verfahren getrennt nach den beiden zeitlichen Betrachtungsebenen. Auch hier bleiben die grundsätzlichen, das Standardverfahren und die beiden poissondistanzbasierten Verfahren betreffenden Erkenntnisse bestehen: Das Clustern der Splinekoeffizienten bringt die niedrigsten Indizes hervor und die beiden anderen Methoden mit der neuen Distanz die höchsten Medianwerte.

Hilfe linearer Kontraste sollen nun die bestehenden Hypothesen zu den einzelnen Clustermethoden überprüft werden. Unterscheiden sich die Verfahren mit der Poisson-Distanz (2 und 6) signifikant von denen mit der euklidischen Distanz (1, 4 und 5)? Gibt es einen Unterschied zwischen der funktionalen Betrachtungsweise mittels Splineclustern (5 und 6) und dem Clustern der Rohdaten (1 und 2)? Bringt das Splineclustern gegenüber dem Clustern der Rohdaten unter Verwendung der Poisson-Distanz einen signifikanten Vorteil? Zeigt Verfahren 6 gegenüber 1, 2, 4 und 5 eine höhere Leistung in der Klassenzuordnung und ist das Standardverfahren bei dieser Datenlage signifikant schlechter als alle anderen Verfahren?

<table>
<thead>
<tr>
<th>Kontrast</th>
<th>Estimate</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verfahren V2, V6 vs V1, V4, V5</td>
<td>0.36</td>
<td>0.10</td>
</tr>
<tr>
<td>Verfahren V5, V6 vs V1, V2</td>
<td>0.09</td>
<td>0.56</td>
</tr>
<tr>
<td>Verfahren V2 vs. V6</td>
<td>-0.09</td>
<td>0.11</td>
</tr>
<tr>
<td>Verfahren V1, V2, V4, V5 vs. V6</td>
<td>-0.27</td>
<td>0.12</td>
</tr>
<tr>
<td>Verfahren V1, V2, V4, V5, V6 vs. V3</td>
<td>1.64</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Tabelle 7.11.: Test linearer Kontraste

Bei linearen Kontrasten ist die Nullhypothese, dass der durchschnittliche Mittelwert von bestimmten Gruppen dem durchschnittlichen Mittelwert bestimmter anderer Gruppen ents-
spricht. In Tabelle 7.11 sind die Ergebnisse der Tests mit dem jeweiligen p-value aufgeführt. Es lässt sich entnehmen, dass die poissondistanzbasierten Verfahren einen höheren mittleren Index besitzen, aber dieser Effekt nicht signifikant ist. Ein Vergleich der funktionalen Betrachtungsweise mit dem Clustern der Rohdaten über beide Distanzmaße bzw. nur mit dem Poisson-Distanzmaß ergibt einen positiven Effekt für die funktionale Herangehensweise, dieser ist aber ebenfalls nicht signifikant. Für das neue Verfahren 6 kann auch kein signifikanter Unterschied zu den anderen Methoden festgestellt werden, wenn auch insgesamt ein höherer mittlerer Index verzeichnet wird. Die einzig gesicherte Erkenntnis ist, dass für diese simulierten Daten das Standardverfahren, das Clustern der Splinekoeffizienten mit euklidischer Distanz, signifikant schlechter abschneidet als die übrigen Methoden.

Eine weitere Möglichkeit einen post-hoc-Test zur Analyse der Mittelwertsdifferenzen, der das multiple Signifikanzniveau hält, durchzuführen, ist der multiple Test nach Tukey (Tukey’s honestly significant difference (HSD))\(^{19}\) Dabei werden alle paarweisen Differenzen miteinander verglichen. Das Ergebnis ist in Abbildung 7.30 dargestellt.

95% family–wise confidence level

Abbildung 7.30.: multipler Test nach TukeyHSD

\[^{19}\text{vgl. Toutenburg (1994) und Everitt and Hothorn (2010).}\]
Ist in dieser Abbildung die Null in dem 95%-Intervall \((0 \in KI)\) eingeschlossen, so liegt kein signifikanter Unterschied zwischen den Mittelwerten der zwei miteinander verglichenen Verfahren vor. Lediglich Verfahren 3 zeigt mit den anderen Verfahren einen signifikanten Unterschied im mittleren Randindex. Bei Verfahren 6 ist der Randindex zwar im Mittel immer höher als bei den anderen Verfahren, dabei aber nicht signifikant. Darüberhinaus zeigt auch das Clustern der Rohdaten mit der Poisson-Distanz (Verfahren 2), bis auf einen Vergleich mit dem Splineclustern bei gleichem Distanzmaß, eine mittlere positive Auswirkung, dabei aber auch nicht signifikant.
Fazit und Ausblick


Auf Basis echter Transaktionsverlaufsdaten über Einkäufe von Haushalten, sowie bei auf deren Grundlage simulierten Daten, wurden in Abschnitt 7 die Ähnlichkeiten bestimmter Verfahren und deren Leistungsfähigkeit hinsichtlich der richtigen Klassenzuteilung untersucht. Die Überprüfung der Ähnlichkeiten erbrachte über die realen und simulierten Daten nur beschränkt einheitliche Ergebnisse. Offensichtlich ist, dass sich das Standardverfahren sowohl stark vom Clustern der Rohdaten, als auch vom Splinecluster unterscheidet. Die größten Gemeinsamkeiten, gemessen am Median des Randindex zwischen den Verfahren,
bestehen zwischen dem Clustern der Splinewerte an den Erhebungszeitpunkten (Verfahren 4) und dem Splinecluster (Verfahren 5), jeweils mit euklidischer Distanz. Darüberhinaus sind die Ähnlichkeiten zwischen Verfahren mit gleichem Distanzmaß aber unterschiedlicher funktionaler Betrachtungsweise (1 – 5, 2 – 6) i.d.R. höher, als bei einem unterschiedlichen Distanzmaß und gleicher Vorgehensweise (1 – 2, 5 – 6).

Für den speziellen Fall von poissonverteilten simulierten Transaktionsdaten ließ sich feststellen, dass das Standardverfahren wesentlich schlechter bzgl. der richtigen Klassenzuordnung abschnitt, als die anderen Verfahren. Außerdem schien in diesem Fall ein angemessen gewähltes Distanzmaß mehr zur Aufdeckung der wahren Cluster beizutragen, als eine funktionale Methodik wie das Splinecluster. Offen für weitere Untersuchungen bleibt demnach, ob das Splinecluster bei anderer Datenlage mehr Gewinn bringt. Dabei kann auch darüber diskutiert werden, inwieweit eine funktionale Herangehensweise überhaupt Vorteile gegenüber einem Clustern der Rohdaten bringt. Zumindest für die vorliegenden poissonverteilten simulierten Daten war der Randindex mit Poissondistanz, also die Kombination des angemessenen Distanzmaßes und einer funktionalen Clustermethode, wertmäßig immer am höchsten. Ob dies auch generell für andere Zähldaten gilt, wäre ebenfalls noch zu überprüfen.

Abschließend sollte auch noch erwähnt werden, dass die Idee des Splineclusters, anders als beispielsweise das klassische Verfahren, auf gleiche Messzeitpunkte beschränkt ist. Hier ließe sich noch überlegen, inwieweit dieses Verfahren hinsichtlich ungleicher Anzahl und nicht identischer Messzeitpunkte ausgeweitet werden kann.

Das Splinecluster und die Poisson-Distanz zeigen, dass eine den Daten angemessene Clustersystematik sehr bedeutend für die Qualität der Ergebnisse ist. Wie beim verallgemeinerten k-means-Algorithmus sollte auch beim funktionalen Clustern immer die Datenlage in die Entscheidung über ein geeignetes Distanzmaß miteinfließen. Die Berücksichtigung einer Funktionalität in den Daten lässt sich auf unterschiedlichste Weise erreichen. Ob das Splinecluster, als eine dieser Möglichkeiten, im Einzelfall bessere Ergebnisse liefert, ließ sich an dieser Stelle noch nicht eindeutig klären.


A.1. Kalenderwochenebene

Abbildung A.1.: Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 63, KW-Ebene
Abbildung A.2.: Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 61, KW-Ebene
Abbildung A.3.: Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 19, KW-Ebene
A.2. Monatsebene

Abbildung A.4.: Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 63, Monatsebene
Abbildung A.5.: Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 44, Monatsebene
### B.1. Kalenderwochenebene

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rohdaten, k-means, eukl - 1</td>
<td>1.00</td>
<td>0.33</td>
<td>0.48</td>
<td>0.96</td>
<td>0.97</td>
<td>0.37</td>
</tr>
<tr>
<td>Rohdaten, k-means, poi - 2</td>
<td>0.33</td>
<td>1.00</td>
<td>0.32</td>
<td>0.33</td>
<td>0.33</td>
<td>0.82</td>
</tr>
<tr>
<td>Splinekoeffizienten, k-means, eukl - 3</td>
<td>0.48</td>
<td>0.32</td>
<td>1.00</td>
<td>0.49</td>
<td>0.48</td>
<td>0.32</td>
</tr>
<tr>
<td>Splinewerte, k-means, eukl - 4</td>
<td>0.96</td>
<td>0.33</td>
<td>0.49</td>
<td>1.00</td>
<td>0.99</td>
<td>0.37</td>
</tr>
<tr>
<td>Rohdaten, Splineclustern, eukl - 5</td>
<td>0.97</td>
<td>0.33</td>
<td>0.48</td>
<td>0.99</td>
<td>1.00</td>
<td>0.37</td>
</tr>
<tr>
<td>Rohdaten, Splineclustern, poi - 6</td>
<td>0.37</td>
<td>0.82</td>
<td>0.32</td>
<td>0.37</td>
<td>0.37</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Tabelle B.1.: Vergleich der Clusterlösungen, Produkt 61, KW-Ebene

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rohdaten, k-means, eukl - 1</td>
<td>1.00</td>
<td>0.45</td>
<td>0.56</td>
<td>0.97</td>
<td>0.97</td>
<td>0.46</td>
</tr>
<tr>
<td>Rohdaten, k-means, poi - 2</td>
<td>0.45</td>
<td>1.00</td>
<td>0.30</td>
<td>0.45</td>
<td>0.45</td>
<td>0.98</td>
</tr>
<tr>
<td>Splinekoeffizienten, k-means, eukl - 3</td>
<td>0.56</td>
<td>0.30</td>
<td>1.00</td>
<td>0.56</td>
<td>0.56</td>
<td>0.30</td>
</tr>
<tr>
<td>Splinewerte, k-means, eukl - 4</td>
<td>0.97</td>
<td>0.45</td>
<td>0.56</td>
<td>1.00</td>
<td>0.98</td>
<td>0.46</td>
</tr>
<tr>
<td>Rohdaten, Splineclustern, eukl - 5</td>
<td>0.97</td>
<td>0.45</td>
<td>0.56</td>
<td>0.98</td>
<td>1.00</td>
<td>0.45</td>
</tr>
<tr>
<td>Rohdaten, Splineclustern, poi - 6</td>
<td>0.46</td>
<td>0.98</td>
<td>0.30</td>
<td>0.46</td>
<td>0.45</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Tabelle B.2.: Vergleich der Clusterlösungen, Produkt 44, KW-Ebene
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rohdaten, k-means, eukl - 1</td>
<td>1.00</td>
<td>0.26</td>
<td>0.18</td>
<td>0.65</td>
<td>0.54</td>
<td>0.41</td>
</tr>
<tr>
<td>Rohdaten, k-means, poi - 2</td>
<td>0.26</td>
<td>1.00</td>
<td>0.12</td>
<td>0.26</td>
<td>0.39</td>
<td>0.37</td>
</tr>
<tr>
<td>Splinekoeffizienten, k-means, eukl - 3</td>
<td>0.18</td>
<td>0.12</td>
<td>1.00</td>
<td>0.24</td>
<td>0.17</td>
<td>0.09</td>
</tr>
<tr>
<td>Splinewerte, k-means, eukl - 4</td>
<td>0.65</td>
<td>0.26</td>
<td>0.24</td>
<td>1.00</td>
<td>0.58</td>
<td>0.42</td>
</tr>
<tr>
<td>Rohdaten, Splinecluster, eukl - 5</td>
<td>0.54</td>
<td>0.39</td>
<td>0.17</td>
<td>0.58</td>
<td>1.00</td>
<td>0.47</td>
</tr>
<tr>
<td>Rohdaten, Splinecluster, poi - 6</td>
<td>0.41</td>
<td>0.37</td>
<td>0.09</td>
<td>0.42</td>
<td>0.47</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Tabelle B.3.: Vergleich der Clusterlösungen, Produkt 19, KW-Ebene
B.2. Monatsebene

Tabelle B.4.: Vergleich der Clusterlösungen, Produkt 63, Monatsebene

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rohdaten, k-means, eukl - 1</td>
<td>1.00</td>
<td>0.40</td>
<td>0.50</td>
<td>0.93</td>
<td>0.93</td>
<td>0.40</td>
</tr>
<tr>
<td>Rohdaten, k-means, poi - 2</td>
<td>0.40</td>
<td>1.00</td>
<td>0.27</td>
<td>0.40</td>
<td>0.40</td>
<td>0.93</td>
</tr>
<tr>
<td>Splinekoeffizienten, k-means, eukl - 3</td>
<td>0.50</td>
<td>0.27</td>
<td>1.00</td>
<td>0.49</td>
<td>0.49</td>
<td>0.26</td>
</tr>
<tr>
<td>Splinewerte, k-means, eukl - 4</td>
<td>0.93</td>
<td>0.40</td>
<td>0.49</td>
<td>1.00</td>
<td>0.96</td>
<td>0.41</td>
</tr>
<tr>
<td>Rohdaten, Splineclustern, eukl - 5</td>
<td>0.93</td>
<td>0.40</td>
<td>0.49</td>
<td>0.96</td>
<td>1.00</td>
<td>0.40</td>
</tr>
<tr>
<td>Rohdaten, Splineclustern, poi - 6</td>
<td>0.40</td>
<td>0.93</td>
<td>0.26</td>
<td>0.41</td>
<td>0.40</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Tabelle B.5.: Vergleich der Clusterlösungen, Produkt 61, Monatsebene

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rohdaten, k-means, eukl - 1</td>
<td>1.00</td>
<td>0.37</td>
<td>0.51</td>
<td>0.98</td>
<td>0.98</td>
<td>0.37</td>
</tr>
<tr>
<td>Rohdaten, k-means, poi - 2</td>
<td>0.37</td>
<td>1.00</td>
<td>0.33</td>
<td>0.37</td>
<td>0.36</td>
<td>0.96</td>
</tr>
<tr>
<td>Splinekoeffizienten, k-means, eukl - 3</td>
<td>0.51</td>
<td>0.33</td>
<td>1.00</td>
<td>0.50</td>
<td>0.50</td>
<td>0.33</td>
</tr>
<tr>
<td>Splinewerte, k-means, eukl - 4</td>
<td>0.98</td>
<td>0.37</td>
<td>0.50</td>
<td>1.00</td>
<td>0.98</td>
<td>0.37</td>
</tr>
<tr>
<td>Rohdaten, Splineclustern, eukl - 5</td>
<td>0.98</td>
<td>0.36</td>
<td>0.50</td>
<td>0.98</td>
<td>1.00</td>
<td>0.36</td>
</tr>
<tr>
<td>Rohdaten, Splineclustern, poi - 6</td>
<td>0.37</td>
<td>0.96</td>
<td>0.33</td>
<td>0.37</td>
<td>0.36</td>
<td>1.00</td>
</tr>
</tbody>
</table>
## Tabelle B.6.: Vergleich der Clusterlösungen, Produkt 44, Monatsebene

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rohdaten, k-means, eukl - 1</td>
<td>1.00</td>
<td>0.47</td>
<td>0.58</td>
<td>0.98</td>
<td>0.98</td>
<td>0.47</td>
</tr>
<tr>
<td>Rohdaten, k-means, poi - 2</td>
<td>0.47</td>
<td>1.00</td>
<td>0.31</td>
<td>0.46</td>
<td>0.46</td>
<td>0.99</td>
</tr>
<tr>
<td>Splinekoeffizienten, k-means, eukl - 3</td>
<td>0.58</td>
<td>0.31</td>
<td>1.00</td>
<td>0.58</td>
<td>0.58</td>
<td>0.31</td>
</tr>
<tr>
<td>Splinewerte, k-means, eukl - 4</td>
<td>0.98</td>
<td>0.46</td>
<td>0.58</td>
<td>1.00</td>
<td>0.99</td>
<td>0.46</td>
</tr>
<tr>
<td>Rohdaten, Splineclustern, eukl - 5</td>
<td>0.98</td>
<td>0.46</td>
<td>0.58</td>
<td>0.99</td>
<td>1.00</td>
<td>0.46</td>
</tr>
<tr>
<td>Rohdaten, Splineclustern, poi - 6</td>
<td>0.47</td>
<td>0.99</td>
<td>0.31</td>
<td>0.46</td>
<td>0.46</td>
<td>1.00</td>
</tr>
</tbody>
</table>

## Tabelle B.7.: Vergleich der Clusterlösungen, Produkt 19, Monatsebene

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rohdaten, k-means, eukl - 1</td>
<td>1.00</td>
<td>0.21</td>
<td>0.27</td>
<td>0.60</td>
<td>0.59</td>
<td>0.26</td>
</tr>
<tr>
<td>Rohdaten, k-means, poi - 2</td>
<td>0.21</td>
<td>1.00</td>
<td>0.15</td>
<td>0.22</td>
<td>0.22</td>
<td>0.49</td>
</tr>
<tr>
<td>Splinekoeffizienten, k-means, eukl - 3</td>
<td>0.27</td>
<td>0.15</td>
<td>1.00</td>
<td>0.28</td>
<td>0.27</td>
<td>0.17</td>
</tr>
<tr>
<td>Splinewerte, k-means, eukl - 4</td>
<td>0.60</td>
<td>0.22</td>
<td>0.28</td>
<td>1.00</td>
<td>0.97</td>
<td>0.40</td>
</tr>
<tr>
<td>Rohdaten, Splineclustern, eukl - 5</td>
<td>0.59</td>
<td>0.22</td>
<td>0.27</td>
<td>0.97</td>
<td>1.00</td>
<td>0.41</td>
</tr>
<tr>
<td>Rohdaten, Splineclustern, poi - 6</td>
<td>0.26</td>
<td>0.49</td>
<td>0.17</td>
<td>0.40</td>
<td>0.41</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Verlauf der Clusterzentren der Basen der simulierten Daten

Abbildung C.1.: Verlauf der Clusterzentren der als Basis dienenden Cluster für die simulierten Daten, Monatsebene
Verlauf der Clusterzentren, poissonverteilt simuliert Daten
D.1. Kalenderwochenebene

Abbildung D.1.: Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 44, poissan-simulierte Daten, KW-Ebene
Abbildung D.2.: Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 19, poissonsimulierte Daten, KW-Ebene
D.2. Monatsebene

Abbildung D.3.: Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 63, poissonsimulierte Daten, Monatsebene
Abbildung D.4.: Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 61, poissonsimulierte Daten, Monatsebene
Abbildung D.5.: Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 44, poissonsimulierte Daten, Monatsebene
Abbildung D.6.: Verlauf der Clusterzentren der einzelnen Clusterlösungen, Produkt 19, poissonsimulierte Daten, Monatsebene
Randindex der Clusterlösungen zueinander

<table>
<thead>
<tr>
<th></th>
<th>Daten, KW-Ebene</th>
<th>Daten, Monatsebene</th>
<th>poisson sim. Daten, KW-Ebene</th>
<th>poisson sim. Daten, Monatsebene</th>
<th>normalsim. Daten, KW-Ebene</th>
<th>normalsim. Daten, Monatsebene</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>0.41</td>
<td>0.40</td>
<td>0.58</td>
<td>0.73</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1-3</td>
<td>0.48</td>
<td>0.50</td>
<td>0.45</td>
<td>0.46</td>
<td>0.13</td>
<td>0.48</td>
</tr>
<tr>
<td>1-4</td>
<td>0.88</td>
<td>0.93</td>
<td>0.82</td>
<td>0.96</td>
<td>0.46</td>
<td>0.99</td>
</tr>
<tr>
<td>1-5</td>
<td>0.90</td>
<td>0.93</td>
<td>0.82</td>
<td>0.97</td>
<td>0.45</td>
<td>0.98</td>
</tr>
<tr>
<td>1-6</td>
<td>0.40</td>
<td>0.40</td>
<td>0.59</td>
<td>0.75</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2-3</td>
<td>0.30</td>
<td>0.27</td>
<td>0.38</td>
<td>0.41</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2-4</td>
<td>0.42</td>
<td>0.40</td>
<td>0.66</td>
<td>0.74</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2-5</td>
<td>0.41</td>
<td>0.40</td>
<td>0.67</td>
<td>0.75</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2-6</td>
<td>0.89</td>
<td>0.93</td>
<td>0.94</td>
<td>0.96</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3-4</td>
<td>0.48</td>
<td>0.49</td>
<td>0.44</td>
<td>0.46</td>
<td>0.17</td>
<td>0.48</td>
</tr>
<tr>
<td>3-5</td>
<td>0.48</td>
<td>0.49</td>
<td>0.44</td>
<td>0.46</td>
<td>0.16</td>
<td>0.48</td>
</tr>
<tr>
<td>3-6</td>
<td>0.28</td>
<td>0.26</td>
<td>0.38</td>
<td>0.42</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4-5</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
<td>0.98</td>
<td>0.72</td>
<td>0.99</td>
</tr>
<tr>
<td>4-6</td>
<td>0.40</td>
<td>0.41</td>
<td>0.68</td>
<td>0.75</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5-6</td>
<td>0.40</td>
<td>0.40</td>
<td>0.68</td>
<td>0.77</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Tabelle E.1.: Randindizes der einzelnen Verfahren zueinander aufgeteilt nach den zugrundeliegenden Daten, Produkt 63
<table>
<thead>
<tr>
<th></th>
<th>Daten, KW-Ebene</th>
<th>Daten, Monatsebene</th>
<th>poissonsim. Daten, KW-Ebene</th>
<th>poissonsim. Daten, Monatsebene</th>
<th>normalsim. Daten, KW-Ebene</th>
<th>normalsim. Daten, Monatsebene</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>0.33</td>
<td>0.37</td>
<td>0.79</td>
<td>0.90</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1-3</td>
<td>0.48</td>
<td>0.51</td>
<td>0.34</td>
<td>0.38</td>
<td>0.37</td>
<td>0.80</td>
</tr>
<tr>
<td>1-4</td>
<td>0.96</td>
<td>0.98</td>
<td>0.76</td>
<td>0.93</td>
<td>0.71</td>
<td>0.99</td>
</tr>
<tr>
<td>1-5</td>
<td>0.97</td>
<td>0.98</td>
<td>0.63</td>
<td>0.94</td>
<td>0.71</td>
<td>0.99</td>
</tr>
<tr>
<td>1-6</td>
<td>0.37</td>
<td>0.37</td>
<td>0.75</td>
<td>0.89</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2-3</td>
<td>0.32</td>
<td>0.33</td>
<td>0.34</td>
<td>0.39</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2-4</td>
<td>0.33</td>
<td>0.37</td>
<td>0.81</td>
<td>0.89</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2-5</td>
<td>0.33</td>
<td>0.36</td>
<td>0.62</td>
<td>0.89</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2-6</td>
<td>0.82</td>
<td>0.96</td>
<td>0.82</td>
<td>0.97</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3-4</td>
<td>0.49</td>
<td>0.50</td>
<td>0.39</td>
<td>0.39</td>
<td>0.43</td>
<td>0.80</td>
</tr>
<tr>
<td>3-5</td>
<td>0.48</td>
<td>0.50</td>
<td>0.31</td>
<td>0.38</td>
<td>0.42</td>
<td>0.80</td>
</tr>
<tr>
<td>3-6</td>
<td>0.32</td>
<td>0.33</td>
<td>0.40</td>
<td>0.39</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4-5</td>
<td>0.99</td>
<td>0.98</td>
<td>0.63</td>
<td>0.99</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>4-6</td>
<td>0.37</td>
<td>0.37</td>
<td>0.92</td>
<td>0.90</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5-6</td>
<td>0.37</td>
<td>0.36</td>
<td>0.62</td>
<td>0.90</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Tabelle E.2.: Randindizes der einzelnen Verfahren zueinander aufgeteilt nach den zugrundeliegenden Daten, Produkt 61

<table>
<thead>
<tr>
<th></th>
<th>Daten, KW-Ebene</th>
<th>Daten, Monatsebene</th>
<th>poissonsim. Daten, KW-Ebene</th>
<th>poissonsim. Daten, Monatsebene</th>
<th>normalsim. Daten, KW-Ebene</th>
<th>normalsim. Daten, Monatsebene</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>0.45</td>
<td>0.47</td>
<td>0.93</td>
<td>0.62</td>
<td>0.44</td>
<td>0.92</td>
</tr>
<tr>
<td>1-3</td>
<td>0.56</td>
<td>0.58</td>
<td>0.49</td>
<td>0.62</td>
<td>0.90</td>
<td>0.63</td>
</tr>
<tr>
<td>1-4</td>
<td>0.97</td>
<td>0.98</td>
<td>0.94</td>
<td>0.97</td>
<td>0.90</td>
<td>1.00</td>
</tr>
<tr>
<td>1-5</td>
<td>0.97</td>
<td>0.98</td>
<td>0.93</td>
<td>0.97</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1-6</td>
<td>0.46</td>
<td>0.47</td>
<td>0.91</td>
<td>0.93</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2-3</td>
<td>0.30</td>
<td>0.31</td>
<td>0.48</td>
<td>0.35</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2-4</td>
<td>0.45</td>
<td>0.46</td>
<td>0.90</td>
<td>0.62</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2-5</td>
<td>0.45</td>
<td>0.46</td>
<td>0.90</td>
<td>0.62</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2-6</td>
<td>0.98</td>
<td>0.99</td>
<td>0.92</td>
<td>0.63</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3-4</td>
<td>0.56</td>
<td>0.58</td>
<td>0.52</td>
<td>0.62</td>
<td>0.45</td>
<td>0.59</td>
</tr>
<tr>
<td>3-5</td>
<td>0.56</td>
<td>0.58</td>
<td>0.52</td>
<td>0.62</td>
<td>0.45</td>
<td>0.92</td>
</tr>
<tr>
<td>3-6</td>
<td>0.30</td>
<td>0.31</td>
<td>0.50</td>
<td>0.62</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4-5</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>0.63</td>
</tr>
<tr>
<td>4-6</td>
<td>0.46</td>
<td>0.46</td>
<td>0.95</td>
<td>0.95</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5-6</td>
<td>0.45</td>
<td>0.46</td>
<td>0.95</td>
<td>0.95</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Tabelle E.3.: Randindizes der einzelnen Verfahren zueinander aufgeteilt nach den zugrundeliegenden Daten, Produkt 44
<table>
<thead>
<tr>
<th></th>
<th>Daten, KW-Ebene</th>
<th>Daten, Monatsebene</th>
<th>poissonsim. Daten, KW-Ebene</th>
<th>poissonsim. Daten, Monatsebene</th>
<th>normalsim. Daten, KW-Ebene</th>
<th>normalsim. Daten, Monatsebene</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>0.26</td>
<td>0.21</td>
<td>0.62</td>
<td>0.72</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1-3</td>
<td>0.18</td>
<td>0.27</td>
<td>0.40</td>
<td>0.51</td>
<td>0.13</td>
<td>0.53</td>
</tr>
<tr>
<td>1-4</td>
<td>0.65</td>
<td>0.60</td>
<td>0.84</td>
<td>0.96</td>
<td>0.29</td>
<td>0.99</td>
</tr>
<tr>
<td>1-5</td>
<td>0.54</td>
<td>0.59</td>
<td>0.84</td>
<td>0.97</td>
<td>0.28</td>
<td>0.99</td>
</tr>
<tr>
<td>1-6</td>
<td>0.41</td>
<td>0.26</td>
<td>0.59</td>
<td>0.75</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2-3</td>
<td>0.12</td>
<td>0.15</td>
<td>0.37</td>
<td>0.44</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2-4</td>
<td>0.26</td>
<td>0.22</td>
<td>0.65</td>
<td>0.69</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2-5</td>
<td>0.39</td>
<td>0.22</td>
<td>0.67</td>
<td>0.71</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2-6</td>
<td>0.37</td>
<td>0.49</td>
<td>0.86</td>
<td>0.94</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3-4</td>
<td>0.24</td>
<td>0.28</td>
<td>0.42</td>
<td>0.51</td>
<td>0.19</td>
<td>0.54</td>
</tr>
<tr>
<td>3-5</td>
<td>0.17</td>
<td>0.27</td>
<td>0.43</td>
<td>0.51</td>
<td>0.19</td>
<td>0.54</td>
</tr>
<tr>
<td>3-6</td>
<td>0.09</td>
<td>0.17</td>
<td>0.36</td>
<td>0.45</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4-5</td>
<td>0.58</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
<td>0.84</td>
<td>1.00</td>
</tr>
<tr>
<td>4-6</td>
<td>0.42</td>
<td>0.40</td>
<td>0.63</td>
<td>0.72</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5-6</td>
<td>0.47</td>
<td>0.41</td>
<td>0.64</td>
<td>0.74</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Tabelle E.4.: Randindizes der einzelnen Verfahren zueinander aufgeteilt nach den zugrundeliegenden Daten, Produkt 19
ANHANG F

Elektronischer Anhang

elektronischer Anhang auf beigelegter CD-Rom
Hiermit versichere ich, dass ich die vorliegende Master Thesis selbständig und ohne Benutzung anderer als die angegebenen Hilfsmittel angefertigt habe.

München, den 25. Januar 2010

Birgit Oellinger