Clusteranalyse mit SAS für unterschiedlich lange Verlaufskurven mit verschiedenen Startpunkten

V. Schultze-Pawlitschko, M. Kersting
Forschungsinstitut für Kinderernährung
Heinstück 11
44225 Dortmund
pawl@fke-do.de, kersting@fke-do.de

Zusammenfassung

Die DONALD-Studie (Dortmunder Nutritional and Anthropometric Longitudinally Designed Study) ist eine Langzeitstudie bei Säuglingen, Kindern und Jugendlichen, in der verschiedene Ernährungs- und Wachstumsparameter in den ersten zwanzig Lebensjahren erhoben werden. Ein Ziel dieser Studie ist die Identifikation von Personengruppen mit ähnlichem Verzehrsverhalten. Dabei werden hier nur Kinder im Alter von zwei bis achtzehn Jahren betrachtet, die an mindestens zehn Untersuchungen teilgenommen haben. Es ergeben sich Verlaufskurven mit mehr als zehn Messzeitpunkten. Um Kinder mit ähnlichem Verzehrsverhalten zu finden, soll eine Clusteranalyse durchgeführt werden.

Probleme bei der geplanten Umsetzung dieser Clusteranalyse mit SAS sind die unterschiedlich langen Verlaufskurven (abhängig vom Alter des Kindes) und die verschiedenen Startpunkte (abhängig vom Zeitpunkt des Eintritts in die Studie).

Das vorliegende SAS-Programm liefert nach Definition eines geeigneten Distanzmaßes die zugehörige Distanzmatrix und führt die Clusteranalyse anschließend durch. Die Ergebnisse werden diskutiert und durch Beispiele ergänzt.

Keywords: Verlaufskurven, Distanzmatrix, Clusteranalyse.

1 Die DONALD-Studie

In der DONALD-Studie werden seit 1985 gesunde Säuglinge, Kinder und Jugendliche im Alter von 3 Monaten bis 18 Jahren zu 29 Zeitpunkten (s. Abbildung 1) untersucht. Ein Anlass bzw. eine Untersuchung besteht in der Regel aus einem Interview, einer medizinischen und einer anthropometrischen Untersuchung. Im Anschluss daran wird zu Hause ein Drei-Tage-Wiege-Ernährungsprotokoll erstellt. Bei Probanden ab einem Alter von rund drei Jahren wird am dritten Protokolltag ein 24-Stunden-Urin gesammelt.

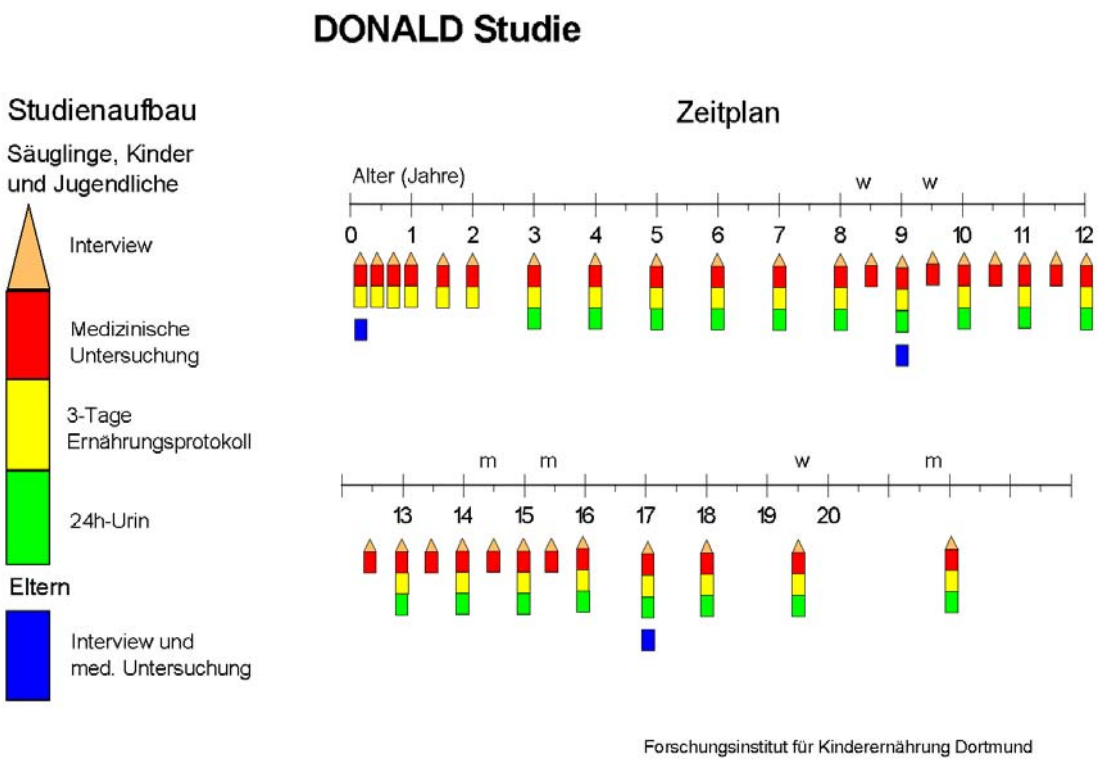


Abbildung 1: Studiendesign der DONALD-Studie: Komponenten der Untersuchung bei Säuglingen, Kindern, Jugendlichen und Eltern sowie Zeitplan der altersabhängigen Untersuchungen

Ein Ziel dieser Studie ist, Kinder mit ähnlichem Verzehrsverhalten in Gruppen zusammenzufassen. Dabei ist der Fettverzehr bedingt durch die wachsende Anzahl adipöser Kinder (siehe [1]) von besonderem Interesse. Hier wird er als prozentualer Anteil (%) an der Gesamtenergiezufuhr berechnet, damit der Fettverzehr zu verschiedenen Alterszeitpunkten verglichen werden kann.

Weiterhin werden nur Kinder zwischen zwei und achtzehn Jahren betrachtet, da erst nach dem zweiten Geburtstag die Umstellung von der Säuglingsernährung auf die Familienernährung abgeschlossen ist. Es ergeben sich Verlaufskurven mit maximal 17 Beobachtungen für jedes Kind, da einmal pro Jahr ein 3-Tage-Ernährungsprotokoll erstellt wird. Verlaufskurven mit weniger als zehn Beobachtungen werden jedoch als zu kurz angesehen, um das längerfristige Verzehrsverhalten des jeweiligen Kindes zu charakterisieren, und in der nachfolgenden statistischen Analyse nicht mehr berücksichtigt. Diese Überlegungen führen zu einer Stichprobe mit 228 Kindern, die mit Hilfe von SAS ausgewertet wird.

2 Statistische Problemstellung

Das übliche statistische Instrument zur Erkennung von Strukturen in beliebigen Objektmengen ist die Clusteranalyse. Die Voraussetzung für eine Clusteranalyse ist die Definition einer geeigneten Distanzmatrix. Das bedeutet, es muss geklärt werden, wie die Verlaufskurven zweier Kinder aussehen müssen, um als „ähnlich“ angesehen zu werden.

Gerade im Zusammenhang mit der Entstehung der Adipositas ist wichtig, wie hoch der durchschnittliche Fettverzehr eines Kindes während Kindheit und Jugend gewesen ist. Des Weiteren spielt es eine Rolle, wie homogen ein Kind gegessen hat, ob die Kost jedes Jahr ähnlich zusammengesetzt war oder ob es starke Unterschiede gibt. Diese beiden Überlegungen führen zum

Vergleich von Gesamtmittelwerten m(i) und Gesamtstreuungen s(i) für jedes der 228 Kinder ($i \in \{1,...,228\}$). Diese deskriptiven Kenngrößen werden bei jedem Kind über die Anzahl der verfügbaren Anlässe y(i,t) mit $1 \leq t \leq 17$ gebildet.

Durch die Berechnung der „Standard Deviation Scores" (SDS)

$$x(i,t) = [y(i,t) - m(i)] / s(i)$$

werden die Messgrößen eines einzelnen Kindes von der Höhe und Homogenität der Fettzufuhr unabhängig (siehe auch [2]). So können auch Kinder mit ähnlichen Mustern in ihrer Verlaufskurve in Gruppen zusammengefasst werden. Die Berücksichtigung der drei vorgestellten Kriterien:

1.) ähnlicher durchschnittlicher Fettverzehr von Kind i und Kind j
2.) ähnliche Streuung bei Kind i und Kind j
3.) ähnlicher Verlauf des Fettverzehrs bei Kind i und Kind j

mit $i, j \in \{1,...,228\}$ führt zu folgendem Distanzmaß d bzw. Distanzmatrix D = (d(i, j)) :

$$d(i,j) = | m(i) - m(j) | + | s(i) - s(j) | + 1/k \sum | x(i,t) - x(j,t) | \text{ und } k,t \in \{1,...,17\} \text{ mit } t \leq k.$$

Dabei sind t die verschiedenen Anlässe, zu denen beide Kinder (Kind i und Kind j) untersucht wurden, und k ist die Anzahl der berechneten Distanzen der einzelnen Meßpunkte.

Basierend auf der Distanzmatrix D wird anschließend eine Clusteranalyse durchgeführt. Als Maß für die Heterogenität wird „Two-stage Linkage" mit der „kth-Nearest-Neighbour Method" gewählt. Diese Methode ist von W. S. Sarle entwickelt und empfohlen worden (s. [3] und [4]).

3 Umsetzung mit SAS

Die Hauptaufgabe von SAS ist die Bereitstellung der Distanzmatrix. Anschließend wird mit PROC CLUSTER die Clusteranalyse durchgeführt und die Ergebnisse interpretiert. Diese Aufgabe wird in vier Teilschritte zerlegt, für die jeweils SAS-Programme geschrieben werden.

1.) Berechnung des durchschnittlichen prozentualen Fettanteils an der Gesamtenergiezufuhr pro Tag für jedes 3-Tage-Ernährungsprotokoll und Extraktion aller Kinder mit mehr als zehn Anlässen
→ datzehn.sas.

Input für datzehn.sas ist eine permanente SAS-Datei, die die Kindnummern, die Anlässe und alle wesentlichen aufgenommenen Nährstoffe aus den Ernährungsprotokollen enthält. Output ist eine temporäre SAS-Datei mit den Kindnummern, den Anlässen, den prozentualen Fettgehalten und der Anzahl der insgesamt wahrgenommenen Anlässe für jedes einzelne Kind. Weiterhin werden die Anlässe für jedes Kind durchnummeriert.

2.) Nullsetzen der nicht oder noch nicht erhobenen Anlässe. Um die Summe aus den Differenzen der Standard Deviation Scores zu bilden, sollten auch die nicht erhobenen Anlässe in die Datei mit aufgenommen und gesondert gekennzeichnet werden. Hier geschieht das durch Nullsetzen der fehlenden Anlässe
→ null.sas.

null.sas benutzt den Output von datzehn.sas und gibt eine temporäre Datei mit der Kindnummer, den Anlässen und dem prozentualen Fettgehalt für alle Anlässe aus. Bei nicht oder noch nicht erhobenen Anlässen ist der prozentuale Fettgehalt gleich null. Unterstützt wird dieses Programm durch vier SAS-Macros.

3.) Bildung von Mittelwert und Streuung für jedes Kind und Übertragung aller Einzelwerte, Mittelwerte und Streuungen in ein Feld
→ array.sas.

array.sas benutzt temporäre Dateien aus null.sas und gibt wiederum eine temporäre SAS-Datei aus, die aus einem einzigen großen Array besteht, das alle oben genannten Messgrößen enthält.

4.) Bildung der Distanzmatrix D als Basis für die Clusteranalyse
→ basisclu.sas.

Ausgehend von einer temporären SAS-Datei array.sas wird eine permante SAS-Datei gebildet, die die oben definierte Distanzmatrix enthält. Dieses Programm enthält zur Vereinfachung der Bildung der Distanzmatrix D zwei SAS-Macros.

Wie bereits erwähnt, wird anschließend die Clusteranalyse mit PROC CLUSTER und der Option METHOD=TWOSTAGE durchgeführt.

4. Ergebnisse

Das Ergebnis der Clusteranalyse sind vier Cluster. Cluster eins besteht aus 35 Kindern, Cluster zwei aus 81 Kindern, Cluster drei enthält 57 Kinder und Cluster vier 55 Kinder. Ein Problem, welches häufig bei Clusteranalysen auftritt, ist die korrekte Interpretation der Cluster.
Zuerst wird kurz dargestellt, welche der gemessenen Parameter die Zugehörigkeit zu einem Cluster nicht beeinflussen. Das Geschlecht scheint in allen Clustern nahezu gleichverteilt zu sein. Jedes Cluster besteht aus kurzen und langen Verläufen, einige Verläufe geben ein Bild aus der frühen Kindheit wieder, andere beginnen erst am Anfang der Pubertät. Das Alter der Kinder scheint also ebensowenig die Zugehörigkeit zu einem Cluster zu beeinflussen.
Die Cluster werden in erster Linie durch das unterschiedliche Niveau des prozentualen Fettverzehrs geprägt. Sei M(*i*) der Mittelwert des Fettverzehrs in Cluster i. Dann gilt:

$$M(3) > M(1) > M(2) > M(4).$$

Diese Unterschiede im Fettverzehr finden sich nicht nur allgemein, sondern auch bei jedem Anlass mit einer einzigen Ausnahme. Diese betrifft fünfzehnjährige Jugendliche. Hier ist der Mittelwert des prozentualen Fettverzehrs in Cluster eins höher als in Cluster drei.

Für die Standardabweichungen S(*i*) der vier Cluster gilt:

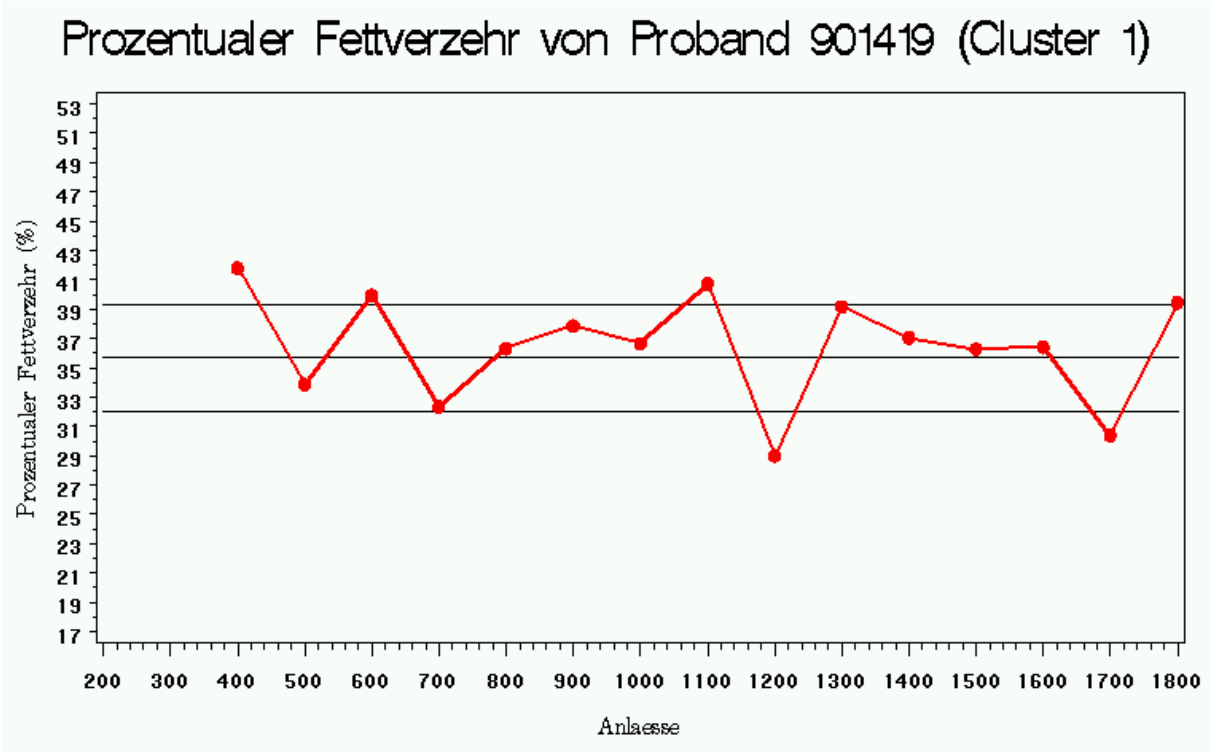
$$S(2) > S(4) > S(3) > S(1).$$

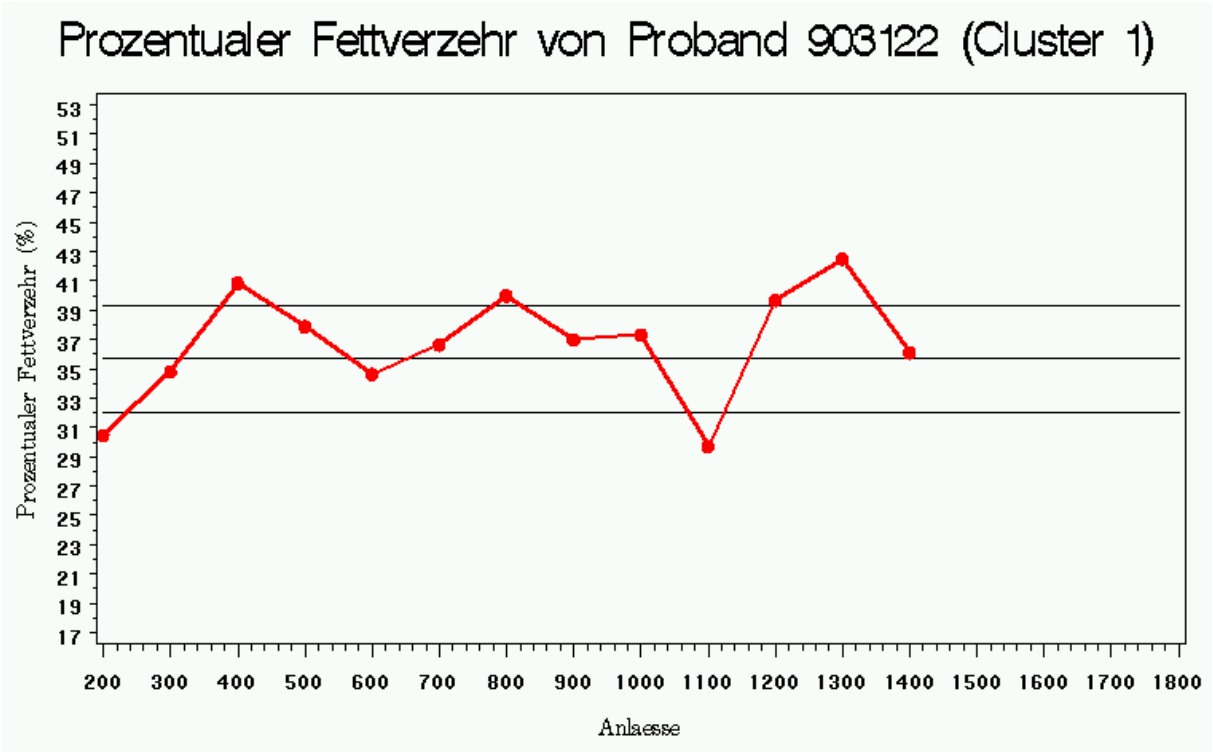
Die Unterschiede sind jedoch nur zwischen Cluster eins und den übrigen drei Clustern ausgeprägter.

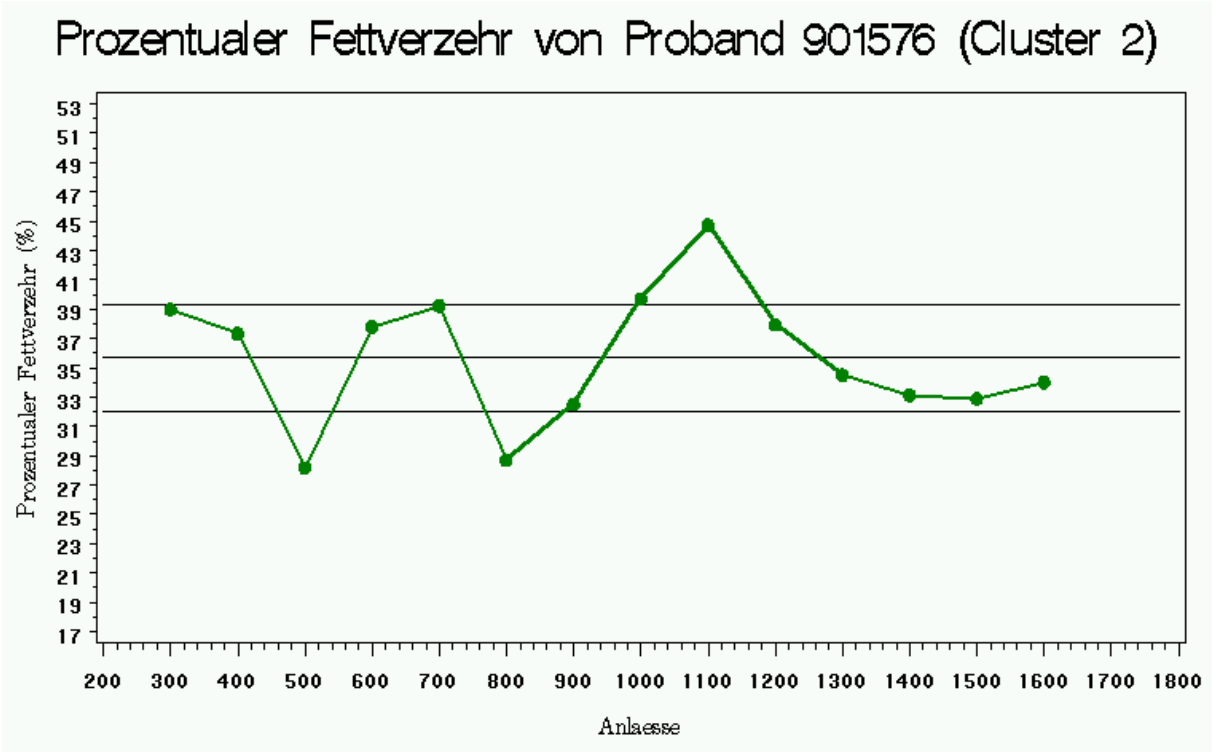
5. Beispiele

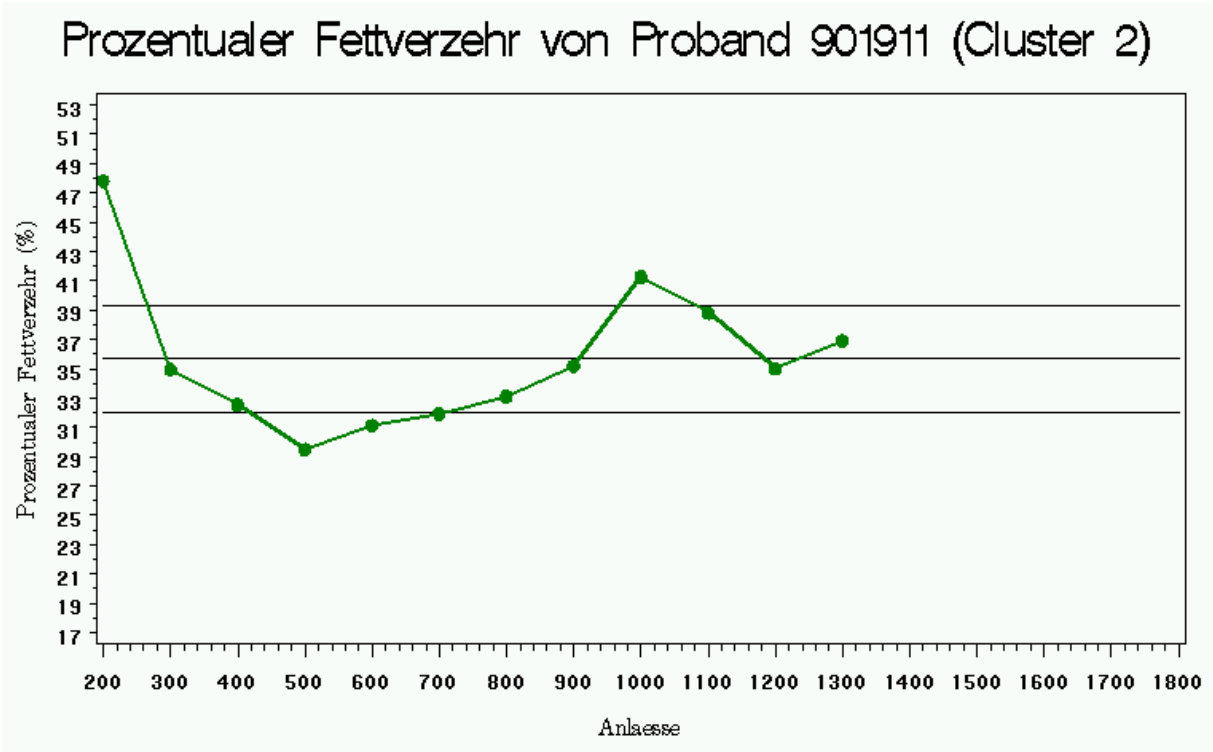
Zur Ergänzung der aufgeführten Ergebnisse der Clusteranalyse werden einige typische Verläufe präsentiert, um die Cluster detaillierter charakterisieren zu können. Für eine bessere Vergleichbarkeit wurden das erste Quartil (Q1 = 32.02 %), der Median (Median = 35.66 %) und das dritte Quartil (Q3 = 39.27 %) der gesamten Beobachtungen in die Graphen mit aufgenommen. Diese deskriptiven Maße wurden aus allen 228 Verläufen berechnet.

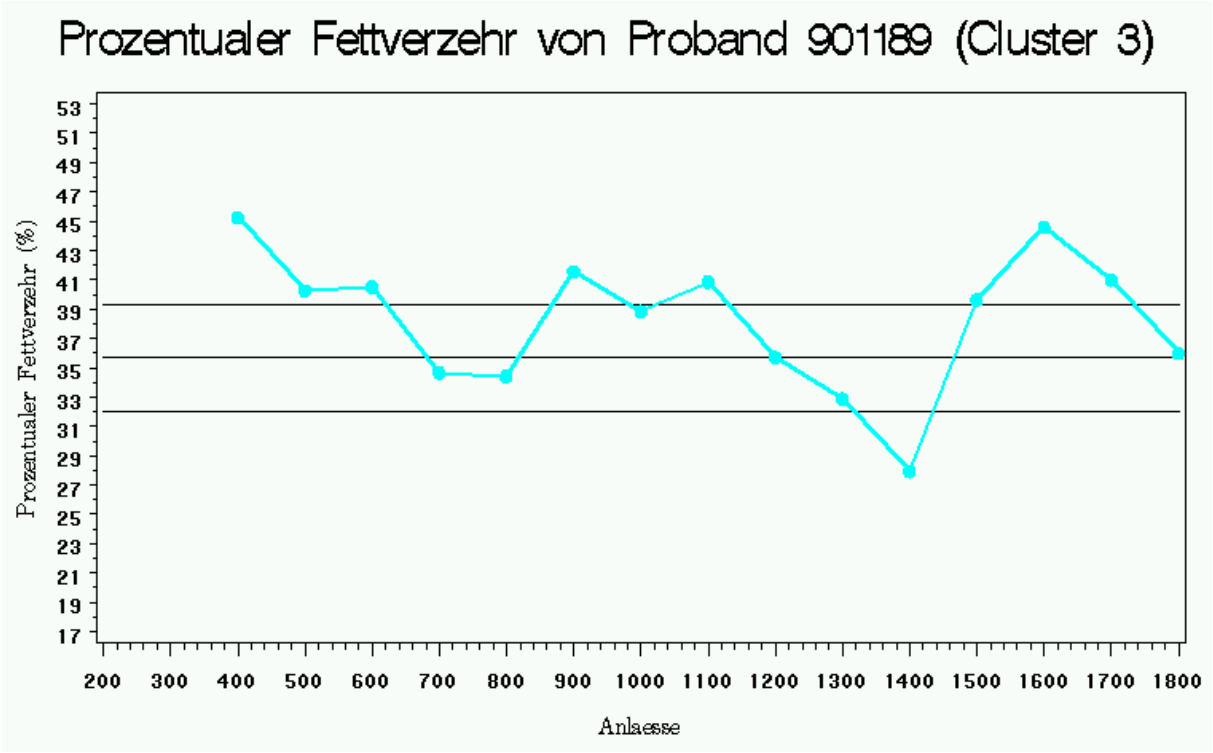
Hier bitte einfügen:
Graphik-Dateien: c11.gif, c12.gif, c21.gif, c22.gif, c31.gif, c32.gif, c41.gif, c42.gif

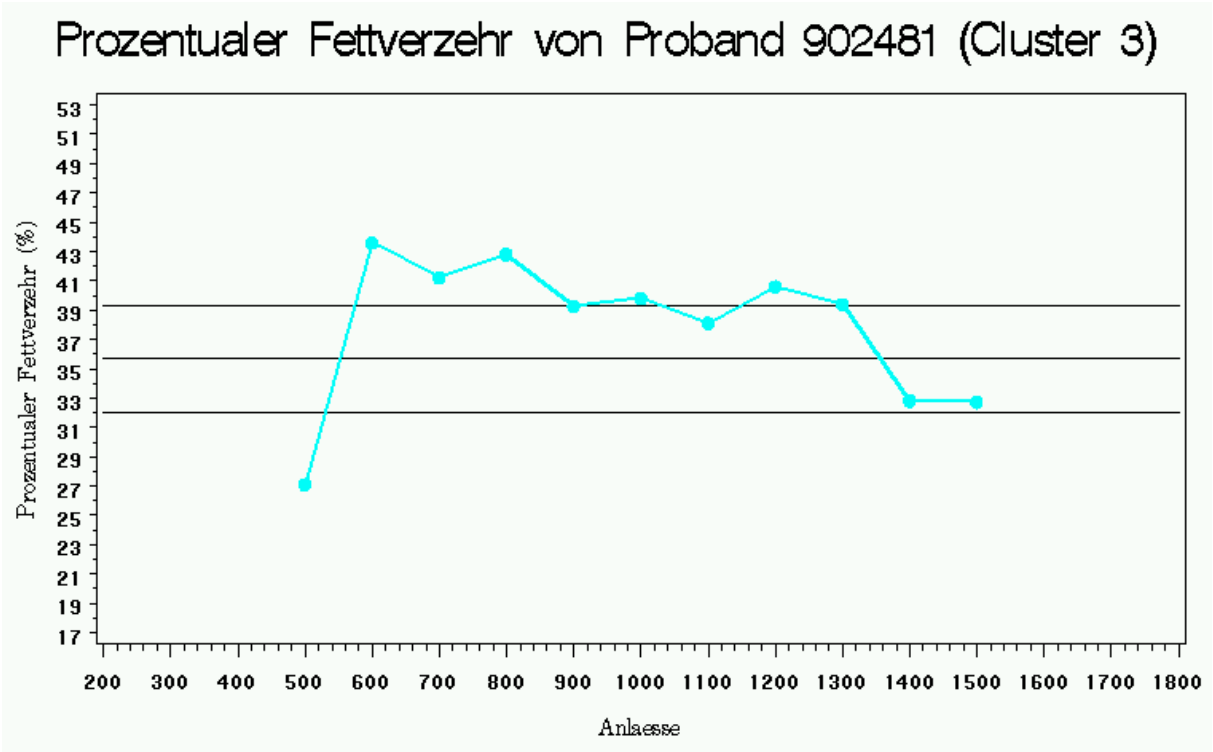

Prozentualer Fettverzehr von Proband 901419 (Cluster 1)
Prozentualer Fettverzehr (%)
53
51
49
47
45
43
41
39
37
35
33
31
29
27
25
23
21
19
17
200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800
Anlaesse


Prozentualer Fettverzehr von Proband 903122 (Cluster 1)
Prozentualer Fettverzehr (%)
53
51
49
47
45
43
41
39
37
35
33
31
29
27
25
23
21
19
17
200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800
Anlaesse


Prozentualer Fettverzehr von Proband 901576 (Cluster 2)
Prozentualer Fettverzehr (%)
53
51
49
47
45
43
41
39
37
35
33
31
29
27
25
23
21
19
17
200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800
Anlaesse


Prozentualer Fettverzehr von Proband 901911 (Cluster 2)
Prozentualer Fettverzehr (%)
53
51
49
47
45
43
41
39
37
35
33
31
29
27
25
23
21
19
17
200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800
Anlaesse


Prozentualer Fettverzehr von Proband 901189 (Cluster 3)
Prozentualer Fettverzehr (%)
53
51
49
47
45
43
41
39
37
35
33
31
29
27
25
23
21
19
17
200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800
Anlaesse


Prozentualer Fettverzehr von Proband 902481 (Cluster 3)
Prozentualer Fettverzehr (%)
53
51
49
47
45
43
41
39
37
35
33
31
29
27
25
23
21
19
17
200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800
Anlaesse

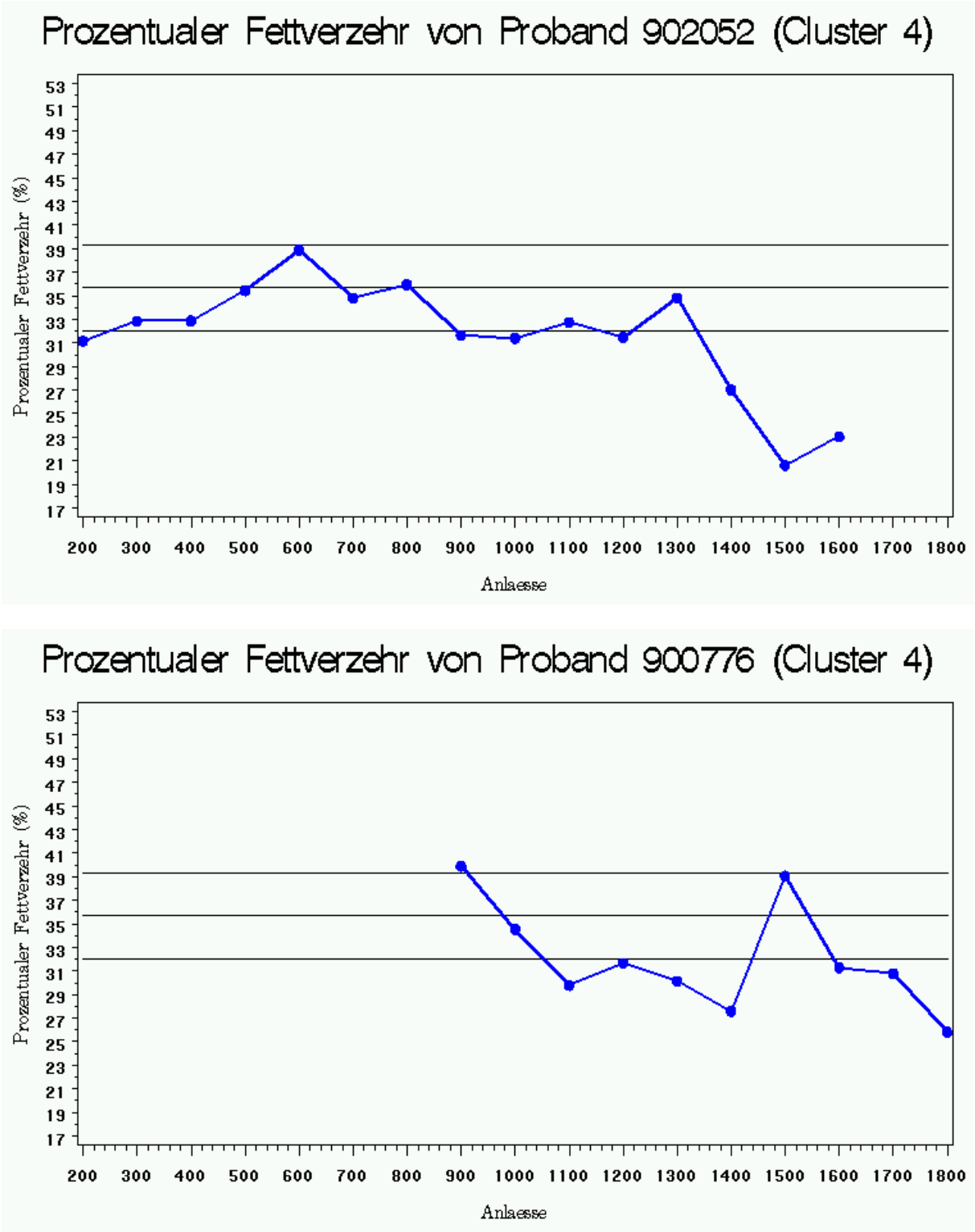


Abbildung 2: Verlaufskurven von acht Kindern aus den vier Clustern

Cluster eins besteht aus Kindern, deren relativer Fettverzehr bei den meisten Anlässen höher als der Median und niedriger als das dritte Quartil ist. Nur wenige Beobachtungen sind niedriger als der Median oder höher als das dritte Quartil, was die im Gegensatz zu den anderen Clustern geringere Standardabweichung erklärt. Zu Beginn der Pubertät, das heißt im Alter von elf oder zwölf Jahren, nimmt der prozentuale Fettverzehr bei den meisten Kindern stark ab und fällt unter das erste Quartil. Diese Veränderung der Ernährungsgewohnheiten wird jedoch nicht lange durchgehalten, weshalb der

prozentuale Fettverzehr bei Elf- bis Zwölfjährigen in diesem Cluster ein lokales Minimum bildet. Der Mittelwert aller achtzehnjähriger Kinder ist 36.16 %, also etwas höher als der Median.

In Cluster zwei ist der prozentuale Fettverzehr in der Kindheit sehr inhomogen. Deshalb besitzt dieses Cluster auch die größte geschätzte Standardabweichung. In allen Quartilsregionen liegen infolgedessen Beobachtungen. Zu Beginn der Pubertät steigt in einigen Fällen der prozentuale Fettverzehr stark an und ist dann höher als das dritte Quartil. Später fällt er wieder und der Mittelwert aller Achtzehnjährigen beträgt 34.28 %, ist also etwas niedriger als der Median.

Cluster drei besteht aus Kindern, die sehr viel Fett verzehren. In den meisten Verläufen ist der prozentuale Fettverzehr in mehr als 50 % aller Anlässe größer als das dritte Quartil. Jedoch - ähnlich wie in Cluster eins - gibt es auch hier einen Altersbereich, wo der prozentuale Fettverzehr abnimmt. Dieser ist jedoch zeitlich später am Ende der Pubertät bei den Fünfzehnjährigen zu beobachten. Der Mittelwert der Achtzehnjährigen beträgt 38.35 %, ist also etwas niedriger als das dritte Quartil.

In Cluster vier sind Kinder, die sehr wenig Fett verzehren. In den meisten Verläufen ist der prozentuale Fettverzehr in mehr als 50 % aller Anlässe niedriger als das erste Quartil. Ähnlich wie in Cluster drei nimmt der Fettverzehr am Ende der Pubertät ab, aber diese Abnahme bleibt in diesem Cluster im Gegensatz zu den anderen bestehen. Der Mittelwert aller Achtzehnjährigen ist 27.99 %, also niedriger als das erste Quartil.

Insgesamt zeigt sich, dass der prozentuale Fettanteil an der Energiezufuhr höher ist, als allgemein empfohlen wird (30-35%). Die vier Cluster führen zu folgenden Schlußfolgerungen:

1.) Ernährungsgewohnheiten scheinen sich schon in der frühen Kindheit herauszubilden. Kinder, die zu diesem Zeitpunkt wenig oder viel Fett essen, tun das vielfach später ebenso.

2.) Cluster eins und Cluster drei zeigen, dass es Kinder gibt, die insbesondere in der Pubertät versuchen, ihre bisherigen Ernährungsgewohnheiten zu durchbrechen, aber in der Regel misslingen diese Versuche. Nur in Cluster vier, wo ohnehin schon wenig Fett gegessen wird, sind Bemühungen dieser Art erfolgreich.

Literatur

1. Freedman, D. S., Srinivan S. R., Valdez, R. A., Williamson D. F. und Berenson, G. S. (1999). Secular increases in rel. weight and adiposity among children over two decades: The Bogalusa Heart Study. *Pediatrics*, **99**, 420-426.
2. Hermanussen, M., Lange, S. und Grasedyk, L. (2001). Growth tracks in early childhood. *Acta Paediatrica*, **90**, 381-386.
3. Kaufman, L. und Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.
4. SAS Institute Inc. (1990). *SAS/STAT User's Guide; Vol. 1, Version 6, 4th ed.*, SAS Institute Inc.