universität wien

MAGISTERARBEIT

Titel der Magisterarbeit

„Multivariate Analyse von Nutzungsdaten von Mobiltelefonen“

Verfasserin

Bettina Vala, Bakk.rer.soc.oec.

angestrebter akademischer Grad

Magistra der Sozial- und Wirtschaftswissenschaften (Mag.rer.soc.oec.)

Wien, Oktober 2008

Studienkennzahl lt. Studienblatt:	A 066 951
Studienrichtung lt. Studienblatt:	Magisterstudium Statistik UniStG
Betreuer:	Ao. Univ. Prof. Dr. Marcus Hudec

Danksagung:

Ich danke Herrn Prof. Marcus Hudec für die Betreuung dieser Arbeit.

Ich danke meinen Eltern dafür, dass sie mir dieses Studium ermöglicht haben.

Ich danke Nina Huber dafür, dass sie mich immer wieder in meinem Studium motiviert hat und sie eine gute Freundin für mich ist.

INHALTSVERZEICHNIS

1 EINLEITUNG

Für Unternehmen in der Wirtschaft spielt der Bereich des Marketings eine wichtige Rolle. Unternehmen versuchen durch Werbekampagnen das Bewusstsein des Produktes und der Marke zu erhöhen. Um viele Kunden durch Werbekampagnen anzusprechen, werden soziale Netzwerke verwendet. In dieser Arbeit handelt es sich um ein konkretes Netzwerk, nämlich aus dem Bereich der Telekommunikation. Dieses soziale Netzwerk entsteht, indem die Kunden miteinander telefonieren beziehungsweise SMS austauschen. Dadurch stehen die Kunden zueinander in einer für den Telekommunikationsbetrieb messbaren Beziehung. Mittels Kennzahlen kann man nun untersuchen, ob Kunden eine wichtige oder eher unwichtige Rolle in diesem Netzwerk haben. In erster Linie ist der Degree von Bedeutung. Er gibt an, mit wie vielen Personen der zu untersuchende Kunde in Beziehung steht.
Das soziale Netzwerk aus dem Bereich der Telekommunikation wurde anhand von diesen Kennzahlen untersucht. Die Thematik der Analyse von sozialen Netzwerken wird in dieser Arbeit nur theoretisch erörtert.

Durch die Analyse der Netzwerkdaten erhält man nützliche Eigenschaften der einzelnen Kunden. Anhand der Eigenschaften wird versucht die Kunden durch ihr Telefonierverhalten zu segmentieren. Als zentrale Methode bietet sich die Clusteranalyse an. Da dieses Netzwerk aus dem Bereich der Telekommunikation eine hohe Anzahl an Kunden enthält, wird in erster Linie auf die Two-Step Clusteranalyse und den K-Means Algorithmus zur Segmentierung zurückgegriffen. In dieser Arbeit wird diese Thematik nicht nur theoretisch betrachtet, sondern auch praktisch durchgeführt. Andere Verfahren sind für so große Datenmengen weniger geeignet. Eine weitere Problematik bei großen Datensätzen stellt die Stabilität der Ergebnisse dar. Ein möglicher Lösungsansatz zur Behandlung großer Datenmengen ist die Ziehung von Stichproben. Die Stichproben werden anhand von der Clusteranalyse untersucht und anschließend deren Ergebnisse

miteinander verglichen, um zu überprüfen, ob die Ergebnisse auf Basis von Stichproben Stabilität aufweisen.
In einer weiteren Analyse wird versucht die Fragestellung, ob es einen Zusammenhang zwischen den telefonierten Sekunden und den Degree gibt, zu beantworten.
Im letzten Schritt wird versucht durch die Clusteranalyse eine vollständige Segmentierung des gesamten Datensatzes durchzuführen.
Die praktische Durchführung erfolgt durch die statistische Software Clementine von SPSS.

2 SOZIALE NETZWERKANALYSE

2.1 WAS IST EIN SOZIALES NETZWERK?

In diesem Kapitel wird in erster Linie das Buch „Einführung in die Netzwerkanalyse“ von Dorothea Jansen herangezogen.
Ein soziales Netzwerk besteht aus Akteuren, die zueinander in Verbindung stehen. Verbindungen können freundschaftlicher oder wirtschaftlicher Natur sein. Ein Beispiel für freundschaftliche Verbindungen könnte eine Schulklasse sein. In diesem Fall sind die Schüler die Akteure dieses Netzwerkes. Dieses soziale Netzwerk entsteht dadurch, dass jene Schüler miteinander verbunden werden, die eine Freundschaft miteinander haben. Diese Beziehungen können stark beziehungsweise schwach sein. Bei starken Beziehungen spricht man im Fachjargon von strong ties und bei schwachen Beziehungen spricht man von weak ties. Weiters kann man nicht davon ausgehen, dass die Verbindungen auf Gegenseitigkeit beruhen. Das heißt, wenn Schüler A Schüler B als Freund wählt, muss nicht notwendigerweise Schüler B Schüler A als Freund bestimmen. Daraus folgt, dass die Beziehungen selten komplett ausbalanciert beziehungsweise symmetrisch sind.

Wirtschaftlich gesehen, gewinnen soziale Netzwerke immer mehr an Bedeutung. Vor allem im Bereich des Marketings spielt die Netzwerktheorie eine wichtige Rolle.
„Im Gebiet des Marketings wird vor allem versucht durch Werbekampagnen das Markenbewusstsein eines Produktes bei Kunden zu erhöhen (Apte et al. 2002)“. „Diese Kampagnen basieren typischerweise auf Informationen über die individuellen Charakteristiken der Kunden, wobei mittels Klassifikationsbäumen oder logistischer Regression die Wahrscheinlichkeit, dass ein Kunde auf die Kampagne anspricht oder zu einem anderen Anbieter wechselt, abgeschätzt wird (Piatetsky-Shapiro and Masand 1999, Rosset et a. 2001, Bichler and Kiss 2004)“.

Um das Markenbewusstsein der Kunden zu erhöhen, wird gerne auf die Strukturen real existierender sozialer Netzwerke zurückgegriffen. Eine wichtige Rolle spielt hier ist der Begriff Mundwerbung (Word-of-Mouth, kurz WOM). „WOM ist eine mündliche Kommunikationsart bei der Erfahrungen bezüglich Marken, Produkten oder Services von Person zu Person weiter gegeben werden (Arndt 1967)“. Vor allem im Internet ist diese Kommunikationsart weit verbreitet. Beispiele dafür sind Meinungsplattformen, Kundenchats, oder auch Diskussionsforen. Diese Beispiele wurden 2004 von Dale Dougherty und Craig Cline unter dem Schlagwort Web 2.0 zusammengefasst. Im Marketing wird diese Kommunikation der Kunden mit Hilfe von virales Marketing ausgenutzt. „Virales Marketing (auch Viral-Marketing oder manchmal Virus-Marketing, kurz VM) ist eine Marketingform, die existierende soziale Netzwerke ausnutzt, um Aufmerksamkeit auf Marken, Produkte oder Kampagnen zu lenken, indem Nachrichten sich analog zur Verbreitung eines Virus ausbreiten (Wikipedia 2008)“. Dadurch kann schnell eine Fülle an Kunden erreicht werden und deren Verhalten analysiert werden.
Soziale Netzwerke spielen insbesondere in der Telekommunikationsindustrie eine wichtige Rolle. Auch in diesem Bereich ist es von Wichtigkeit, wie sich die einzelnen Kunden im Netzwerk verhalten. Eine bedeutende Fragestellung der Telekommunikationsindustrie ist, wie wichtig einzelne Kunden in der Netzwerkstruktur sind. Wichtigkeit bedeutet in diesem Zusammenhang, wie zentral liegt diese Person im Netzwerk. Hat diese Person einen großen Einfluss auf andere Personen, beziehungsweise steht diese wichtige Person auch mit anderen wichtigen Personen in Verbindung. Das heißt wie ist die Person in Bezug auf sein Kommunikationsverhalten in diesem Netzwerk eingebettet.

In der Netzwerkanalyse gibt es drei Analyseebenen, die auf unterschiedliche Fragestellungen untersucht werden können.

- Die erste Analyseebene entspricht der Untersuchung des einzelnen Akteurs. Hierbei wird untersucht, wie der Akteur im Netzwerk eingebettet ist. Dabei stellt sich die Frage, wie zentral er ist und ob er

Einfluss auf andere Akteure im Netzwerk hat. Wichtig dabei ist auch, ob jener Akteur der sehr zentral im Netzwerk liegt, auch großen Einfluss auf andere Akteure hat.

- Die zweite Analyseebene besteht aus der Erforschung ganzer Gruppen in einem Netzwerk. Gruppen werden in diesem Zusammenhang auch als Cliquen bezeichnet. Hier wird in erster Linie versucht Cliquen zu finden, deren Akteure sehr eng miteinander vernetzt sind und zu andern Akteuren beziehungsweise Cliquen sehr weitläufig miteinander verbunden sind. Weiters wird auch hier wieder untersucht wo sich die Cliquen im Netzwerk befinden.
- Auf der dritten Analyseebene wird das ganze Netzwerk als Ganzes betrachtet. In diesem Fall werden ganze Gesellschaften beobachtet. Man kann dann verschiedene Gesellschaften miteinander vergleichen. Diese Vergleiche ruhen auf den Ergebnissen der vorherigen Analyseebenen.

2.2 GRAPHENTHEORIE

In diesem Kapitel wird in erster Linie das Buch „Einführung in die Netzwerkanalyse“ von Dorothea Jansen herangezogen.
Ein wichtiges Hilfsmittel der sozialen Netzwerkanalyse bildet die Graphentheorie, welche zur Definition von Maßzahlen aber auch zur Visualisierung herangezogen werden kann. Ein soziales Netzwerk wird als ein Graph, oder in der Netzwerkanalyse genannt Soziogramm, dargestellt. Ein Graph setzt sich aus V einer Menge von Knoten (vertex) und E einer Menge von Kanten (edge) zusammen. Eine Kante wird durch das zugehörige Knotenpaar definiert.

Ein Graph heißt vollständig, falls alle möglichen Kanten existieren, das heißt alle Knoten sind direkt miteinander verbunden.

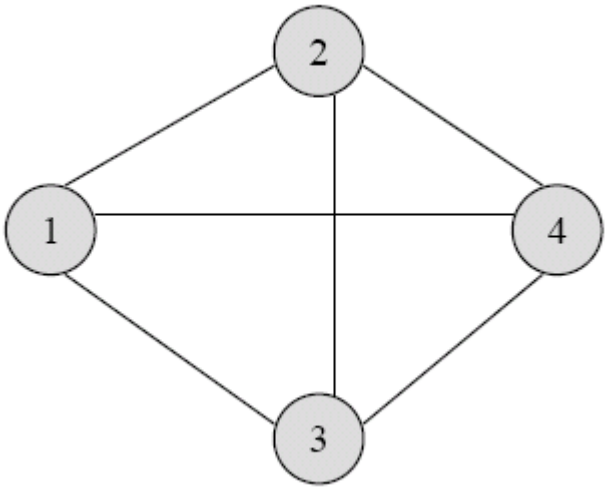

Abbildung 2.1: Beispiel eines vollständigen Graphen

Der Knoten repräsentiert den Akteur (z. B. Kunde eines Telekommunikationsunternehmens) und die Kante entspricht der Beziehung oder Relation (z. B. Telefonieren oder SMS schreiben). Diese Relation liegt in den meisten Fällen nur in binärer Form da. Das heißt, ist die Relation vorhanden (1) oder nicht (0). Die Beziehung kann aber auch eine Gewichtung haben, zum Beispiel Kontakthäufigkeit (wie oft haben zwei Kunden miteinander telefoniert), Dauer der Kontakte (wie lange haben zwei Kunden miteinander telefoniert). Dies macht dann die Stärke der Beziehung aus (strong ties oder weak ties).

Ein Graph kann sowohl ungerichtete, als auch gerichtete Beziehungen haben. Ist eine Richtung der Relation vorhanden, dann spricht man von einem gerichteten Graphen (Diagraph), zum Beispiel Kunde 1 ruft Kunde 2 an. Daraus folgt, dass die Kanten eines Diagraphs durch geordnete Paare definiert sind. Die Relation wird nicht mehr als Linie dargestellt, sondern als Pfeil.

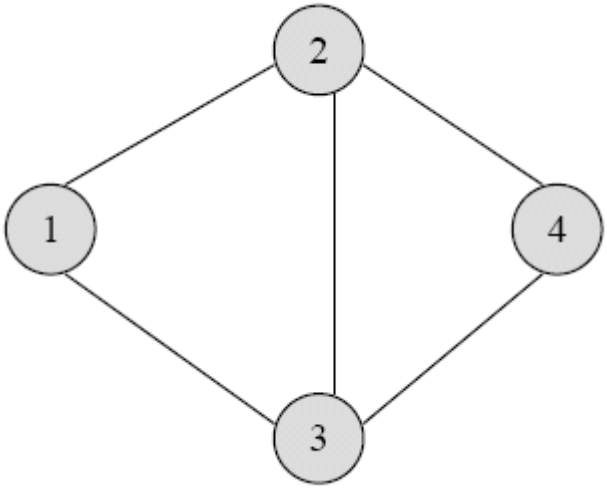

Abbildung 2.2: Beispiel eines ungerichteten Graphen

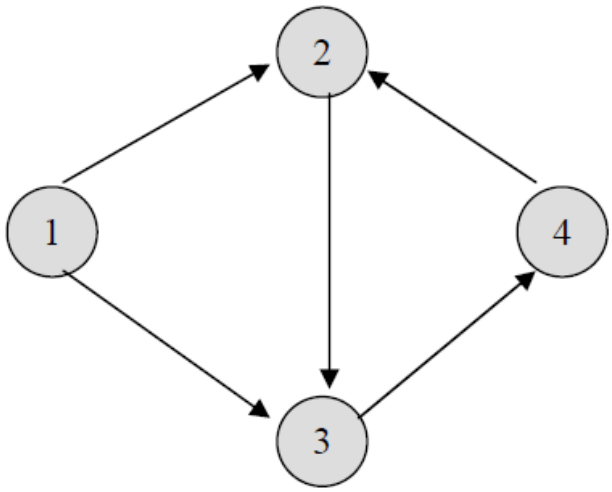

Abbildung 2.3: Beispiel eines gerichteten Graphen

Es gibt vier mögliche Ausprägungen der gerichteten Beziehung. Um diese Ausprägungen näher zu beschreiben, wird im folgenden ein Graph mit genau zwei Akteuren betrachte, nämlich Akteur i und Akteur j. Diese Form des Graphen nennt man Dyade. Sie entspricht dem kleinst möglichem Gefüge eines Netzwerkes. Eine erste Möglichkeit der Ausprägung der Kante ist, dass zwischen den Akteuren keine Relation besteht ($D_{ij} = (0,0)$, Null Dyad). Angenommen es gibt nun eine Beziehung zwischen den beiden Akteuren. In diesem Fall kann die Beziehung in die eine

Richtung($D_{ij} = (1,0)$, Asymmetric Dyad), in die andere Richtung ($D_{ij} = (0,1)$, Asymmetric Dyad) oder in beide Richtung gehen ($D_{ij} = (1,1)$, Symmetric Dyad). Ein Netzwerk mit n Akteuren besitzt $\binom{n}{2} = \frac{n!}{(n-2)!\ 2!}$ Dyaden.

Weiters kann man in einem Graphen die Pfade zwischen Akteur i und Akteur j betrachten. Ein Pfad wird berechnet durch die Anzahl der Kanten, die von Akteur i zu Akteur j führen. In der Netzwerkanalyse ist die kürzeste Pfaddistanz von Bedeutung. Die kürzeste Pfaddistanz $d(i, j)$ entspricht der kürzesten Länge (Pfad) zwischen Akteur i und Akteur j. Existieret keine Pfad zwischen zwei Akteuren, so beträgt die Pfaddistanz ∞.

Graphen können auch anhand von Matrizen dargestellt werden. Diese Matrizen nennen sich in der Graphentheorie Adjazenzmatrizen. Zwei Knoten heißen adjazent, wenn sie durch eine Kante miteinander verbunden sind. In der Netzwerkanalyse spricht man allgemein von Soziomatrizen. Im Folgenden werden die beiden Termini Adjazenzmatrix und Soziomatrix synonym verwendet. Die Zeilen und Spalten der Matrix stehen für die Akteure im Netzwerk. Die Matrix ist somit eine quadratische Matrix. In den Zeilen stehen die Akteur, von denen die Beziehung ausgeht und in den Spalten die Akteure, die die Beziehung annehmen. Eine 1 in der Matrix bedeutet, dass eine Beziehung vorhanden ist und eine 0, dass keine Beziehung besteht. Solch eine Soziomatrix sieht folgendermaßen aus:

		Eingehend Beziehung			
		Akteur 1	Akteur 2	Akteur 3	Akteur 4
Ausgehende Beziehung	Akteur 1	0	1	0	1
	Akteur 2	0	0	1	0
	Akteur 3	1	0	0	1
	Akteur 4	0	1	0	0

Tabelle 2.1:Soziomatrix

Die Diagonale der Matrix besteht nur aus Nullen. Dies bedeutet, dass hier keine Beziehung zu dem Akteur selbst möglich ist.
Eine weitere Eigenschaft der Soziomatrix stellt die Symmetrie der Matrix bei ungerichteten Netzwerken dar. Aufgrund dessen reicht es, die obere oder untere Diagonalmatrix zu betrachten, da diese identisch sind. Im Falle gerichteter Netzwerken unterscheiden sich die obere und untere Diagonalmatrix. Die Soziomatrix ist im Allgemeinen asymmetrisch.
In der Adjazenzmatrix lassen sich nur die direkten Beziehungen zu den einzelnen Akteuren ablesen. Eine weitere Besonderheit der Adjazenzmatrix ergibt sich, wenn man die Matrix potenziert. Aufgrund der Potenz der Adjazenzmatrix kann man auch die indirekten Beziehungen ablesen. Man nennt diese Matrix Erreichbarkeitsmatrix. Zum Beispiel die zweite Potenz der Adjazenzmatrix bedeutet, dass man nun die Anzahl der Wege der Länge zwei von Akteur i und Akteur j ablesen kann.
Mit Hilfe dieser Matrizen lassen sich nun drei wesentliche Merkmale berechnen, die den gesamten Graphen beschreiben.
Das wichtigste Merkmal zur Beschreibung des Graphen entspricht der Dichte. Die Dichte errechnet sich aus dem Quotienten von den tatsächlichen Kanten und allen möglichen Kanten.

$$\Delta = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}}{n(n-1)} \quad \text{für } i \neq j \tag{2.1}$$

Ein weiteres Merkmal stellt die Multiplexität dar. Multiplexität bedeutet, dass zwischen zwei Knoten mehr als nur eine Kante existiert. Sie wird folgendermaßen berechnet:

$$M = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij(m)}}{n(n-1)} \quad \text{für } i \neq j\,, \tag{2.2}$$

wobei m ein selbst gewählter Grenzwert ist, der angibt ab wann eine Beziehung als multiplex gilt.

Zur Beschreiben eines Graphen kann man auch die Kohäsion heranziehen. Dieses Merkmal lässt sich nur für gerichtete Graphen berechnen. Die Kohäsion entspricht dem Verhältnis aus der Anzahl der Beziehungen zu der Anzahl aller möglichen Dyaden.

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_{ij} + x_{ji})}{[(n(n-1))/2]} \tag{2.3}$$

für $i \neq j$, $i < j$ und $(x_{ij} + x_{ji}) = 1$, falls beide Werte 1, sonst 0

Im folgenden Kapitel wird nun auf die Charakterisierung des einzelnen Akteurs im Netzwerk eingegangen.

2.3 Kennzahlen zur Charakterisierung von Akteuren im Netzwerk

Zur literarischen Unterstützung wird sowohl das Buch „Einführung in die Netzwerkanalyse“ von Dorothea Jansen, als auch das Paper „Centrality in Social Networks“ von Linto C. Freeman (1979) verwendet.

Zur Charakterisierung der Akteure von Netzwerken werden die Kennzahlen Zentralität und Prestige herangezogen. Anhand dieser Kennzahlen wird versucht herauszufinden, welche Rolle die einzelnen Akteure im Netzwerk spielen. Das Konzept der Zentralität wird so definiert, dass derjenige Akteur prominent im Netzwerk ist, der an vielen Beziehungen im Netzwerk beteiligt ist (Knoke, Burt 1983). Um die Kennzahlen der Zentralität berechnen zu können, werden nur ungerichtete Beziehungen benötigt. Beim Konzept des Prestiges wird der Akteur dahingehend untersucht, ob der Akteur von vielen anderen Akteuren im Netzwerk ausgewählt wird. Es

kann aber durchaus vorkommen, dass ein Akteur, der sehr zentral ist, nur wenig Prestige genießt oder umgekehrt. Ein Beispiel hierfür kann in der Telekommunikation gefunden werden. Eine Telefonistin in einer Firma wird zwar von vielen Kunden angerufen, das heißt sie besitzt Prestige, aber ruft nur selten selber die Kunden an. Ihre Zentralität ist daher niedrig. Zur Berechnung der Kennzahl Prestige braucht man allerdings gerichtete Beziehungen.

Diese Kennzahlen sind von Bedeutung, damit die Akteure im sozialen Netzwerk genauer beschrieben werden können. Akteure die Zentralität oder Prestige genießen, werden dahingehend charakterisiert, dass diese Akteure eine gute Position im Netzwerk besitzen. Eine gute Position im Netzwerk bedeutet, dass die Akteure soziales Kapital daraus schlagen können. Soziales Kapital sind Ressourcen, die für Akteure von großer Wichtigkeit sind. Beispiele für soziales Kapital sind Anerkennung in einer Gesellschaft, ein gewisser Informationsvorsprung anderen gegenüber und Vertrauen von anderen.
Das Konzept der Zentralität und Prestige kann nicht nur auf einzelne Akteure, sondern auch für die vergleichende Analyse gesamter Netzwerke angewendet werden. In diesem Fall werden gesamte Netzwerke miteinander verglichen, bei denen unterschiedliche zentrale Strukturen existieren. Es stellt sich die Frage, ob der Informationsfluss bei Netzwerken mit mehr zentraleren Akteuren höher ist als im Vergleich zu anderen stärker ausgewogenen strukturierten Netzwerken.
In den folgenden Kapiteln wird nun auf die einzelnen Berechnungen der Maßzahlen eingegangen.

2.3.1 ZENTRALITÄT

2.3.1.1 DEGREE-BASIERTE ZENTRALITÄT

Wie schon zuvor beschrieben, wird die Zentralität daran gemessen wie viele Beziehungen der Akteur besitzt. Es wird hier untersucht wie gut der

Akteur im Netzwerk eingebettet beziehungsweise vernetzt ist. Um dies nun messen zu können, wird die degree-basierte Zentralität berechnet. Die Maßzahl setzt zunächst nur ein ungerichtetes, symmetrisches Netzwerk voraus.
Die degree-basierte Zentralität wird folgendermaßen berechnet:

$$C_D(n_i) = d_i = \sum_{j=1}^{n} x_{ij} = \sum_{j=1}^{n} x_{ji} \quad \text{für } i \neq j \tag{2.4}$$

Diese Formel entspricht der Aufsummierung der Zeile oder Spalte der Adjazenzmatrix des Akteurs.

Liegt ein gerichtetes, asymmetrisches Netzwerk vor, wird in diesem Fall die degree-basierte Zentralität, oder auch Outdegree genannt, folgendermaßen berechnet:

$$C_D(n_i) = od_i = \sum_{j} x_{ij} \quad \text{für } i \neq j \tag{2.5}$$

Der Outdegree entspricht der Summe aller Beziehungen, die von dem untersuchten Akteur wegführen. In der Adjazenzmatrix addiert man die Zeile des zu untersuchenden Akteurs.
Um die degree-basierte Zentralität des einen Netzwerkes mit einer degree-basierte Zentralität aus einem anderen Netzwerk vergleichen zu können, muss die Kennzahl standardisiert werden. Das heißt die Größe des Netzwerkes muss in der Berechnung der Maßzahl miteinbezogen werden. Dies erfolgt, indem man die berechnete degree-basierte Zentralität durch die Anzahl aller möglichen Beziehungen, die ein Akteur besitzen kann, dividiert. Die Anzahl aller möglichen Beziehungen beträgt $(n-1)$, wenn die Beziehung zu sich selbst ausgeschlossen ist. Die degree-basierte Zentralität nimmt dann nur mehr Werte zwischen 0 und 1 an. Ein Wert nahe bei 1 bedeutet, dass der Akteur stark im Netzwerk verbunden ist und ein Wert nahe bei 0 sagt aus, dass der Akteur sehr wenige Beziehungen zu anderen Akteuren hat.

2.3.1.2 CLOSENESS-BASIERTE ZENTRALITÄT

Dieses Zentralitätsmaß betrachtet nicht die Anzahl der wegführenden, oder ankommenden Beziehungen eines Akteurs, sondern die Nähe des Akteurs zu allen anderen Akteuren im Netzwerk. Zur Berechnung wird ausgehend von Akteur i die Pfaddistanzen zu allen anderen Akteuren j herangezogen. Das heißt hier werden sowohl direkte, als auch indirekte Beziehungen berücksichtigt. Das closeness-basierte Zentralitätsmaß ist definiert als

$$C_C(n_i) = \frac{1}{\sum_{j=1}^{n} d(n_i, n_j)} \quad \text{für } i \neq j \tag{2.6}$$

Um diese Maßzahl von unterschiedlichen Netzwerken vergleichbar zu machen, muss diese wieder auf den Wertebereich zwischen 0 und 1 normiert werden. Dies erfolgt mit dem Größe $\frac{1}{(n-1)}$ (reflexive Beziehung ausgeschlossen). Das heißt $C_c(n_i)$ wird durch diesen Größe dividiert. Daraus ergibt sich folgende Formel:

$$C'_C(n_i) = \frac{n-1}{\sum_{j=1}^{n} d(n_i, n_j)} \quad \text{für } i \neq j \tag{2.7}$$

Ein Problem kann entstehen, falls nicht alle Akteure miteinander verbunden sind. Sobald ein Akteur unverbunden ist, so ist die Distanz zu dem untersuchenden Akteur ∞. Dies führt wiederum dazu, dass das Zentralitätsmaß nicht definiert ist. Um dieses Problem zu umgehen, kann man unverbundene Akteure in der Berechnung weglassen.

2.3.1.3 BETWEENNESS-BASIERTE ZENTRALITÄT

Diese Maßzahl betrachtet nicht mehr die einzelnen Dyaden, sondern zerlegt das Netzwerk in Triaden. Triaden sind Netzwerke mit jeweils drei Akteuren. Man bildet den Quotienten von der Anzahl der kürzesten Wege

(geodesic) von j nach k, die über i laufen und der Anzahl der kürzesten Wege von j nach k. Die Formel ist definiert als

$$C_B(n_i) = \sum_{j<}^{n} \sum_{k}^{n} \frac{1}{g_{jk}} g_{jk}(n_i) \quad \text{für } i \neq j \neq k, \tag{2.8}$$

wobei $g_{jk}(n_i)$ ...für die Anzahl der kürzesten Wege von j nach k, die über i führen

g_{jk} ...Anzahl der kürzesten Wege von j nach k

Um dieses Zentralitätsmaß auf den Wertebereich zwischen 0 und 1 zu normieren, wird $C_B(n_i)$ durch die theoretisch größtmögliche Betweenness von $(n^2 - 3n + 2)/2$ dividiert.

$$C'_B(n_i) = \frac{2C_B(n_i)}{(n^2 - 3n + 2)} \tag{2.9}$$

Die betweenness-basierte Zentralität lässt sich sowohl auf ungerichtete, als auch auf gerichtete Netzwerke anwenden. Es gibt aber einen kleinen Unterschied in der Berechnung. Bei ungerichteten Graphen gibt es keinen Unterschied zwischen dem Akteurpaar (j,k) und (k,j). Jedoch bei gerichteten Graphen wird sowohl das Paar (j,k), als auch das Paar (k,j) in die Berechnung miteinbezogen.

2.3.2 PRESTIGE

Mit Prestige wird gemessen, wie viel Autorität und Achtung ein Akteur im sozialen Netzwerk genießt. Die Maßzahlen können nur auf gerichtete Netzwerke angewendet werden. Hier steht nicht die Aktivität vom beobachteten Akteur im Vordergrund, sondern es wird untersucht von wie vielen Akteuren der untersuchte Akteur gewählt worden ist.

2.3.2.1 INDEGREE-BASIERTES PRESTIGE

Das indegree-basierte Prestige ist die einfachste Kennzahl um Prestige messen zu können. In diesem Fall werden alle direkten Beziehungen, die zum Akteurs i führen, summiert.
Dies lässt sich mathematisch folgendermaßen darstellen:

$$P_D(n_i) = id_i = \sum_{j=1}^{n} x_{ji} \quad \text{für } i \neq j \tag{2.10}$$

Natürlich muss wieder beachtet werden, dass man die Netzwerkgröße in das Prestigemaß mit einbezieht, um es mit Maßen aus anderen Netzwerken mit unterschiedlicher Größe vergleichbar zu machen. Ein Akteur kann maximal $(n-1)$ eingehende Beziehungen haben. Daher wird $P_D(n_i)$ durch diese Größe dividiert. Dadurch erhält man ein Prestigemaß, das Werte zwischen 0 und 1 annimmt.

2.3.2.2 PROXIMITY-PRESTIGE

Das Proximity-Prestige ist dem closeness-basierten Zentralitätsmaß sehr ähnlich. Es werden hier wieder die Pfaddistanzen von Akteur i zu allen andern Akteuren j betrachtet. Der Unterschied zur closeness-basierten Zentralität ist, dass Probleme bei unverbundenen Akteuren nicht auftreten. Hierbei wird folgendes Maß betrachtet:

$$I_i = \frac{Zahl\ \ der\ \ Akteure,\ \ die\ \ i\ \ erreichen\ \ können}{(n-1)} \tag{2.11}$$

Dieses Maß nennt sich Einflusssphäre. Die Einflusssphäre wurde hier mit dem Wert $(n-1)$ auf die Größe des Netzwerkes normiert.
Das Problem des unverbundenen Akteurs wird hier vermieden, da die Einflusssphäre bei unverbundenen Akteuren 0 beträgt. Im Unterschied zur closeness-basierten Zentralität, die bei unverbundenen Akteuren ∞ ergibt.

Das Proximity-Prestigemaß lässt sich folgendermaßen berechen:

$$P_P(n_i) = \frac{I_i^2}{(n-1)\sum_{j=1}^{n} d_{ij}} \quad \text{für } i \neq j \tag{2.12}$$

2.3.2.3 RANGPRESTIGE

Das Rangprestigemaß ist dem Proximity-Prestigemaß sehr ähnlich. Das Rangprestigemaß berücksichtigt zusätzlich zu den direkten und indirekten Beziehungen jeweils noch das Prestige von den direkten und indirekten Wählern. Berechnet wird das Rangprestigemaß folgendermaßen:

$$P_R(n_i) = x_{1i}P_R(n_1) + x_{2i}P_R(n_2) + \ldots + x_{ni}P_R(n_n) = \sum_{j=1}^{n} x_{ji}P_R(n_j) \tag{2.13}$$

In dieser Formel stehen die x-Werte für Beziehung vorhanden (hat Wert 1) oder nicht vorhanden (hat Wert 0) und werden mit den entsprechenden Prestiges gewichtet.

2.4 VERGLEICH DER VERSCHIEDEN ZENTRALITÄTSMAßE

Die degree-basierte Zentralität betrachtet nur die direkten Beziehungen von einem Akteur zu allen anderen Akteuren. Dieses Zentralitätsmaß misst, ob die Kommunikation im Netzwerk gut oder schlecht ist. Im Gegensatz zur degree-basierten Zentralität betrachtet die closeness-basierte Zentralität nicht nur die direkten sondern auch die indirekten Beziehungen, indem die Pfaddistanzen von dem untersuchten Akteur zu allen anderen Akteuren berechnet werden. Der Nachteil der closeness-basierten Zentralität ist, dass für unverbundene Akteure das Maß nicht definiert ist. Auch die betweenness-basierte Zentralität bezieht die indirekten Beziehungen mit ein. Aber im Unterschied zur closeness-basierten Zentralität werden hier alle möglichen Triaden untersucht. Die Triaden setzen sich aus zwei

Akteuren und dem untersuchten Akteur zusammen. Zur Berechnung werden die kürzesten Pfaddistanzen von zwei Akteuren, die durch den beobachteten Akteur gehen herangezogen.
Freemann (1979) vergleicht die drei Zentralitätsmaße anhand aller möglicher Graphen, die sich aus fünf Akteuren bilden lassen. Er fand heraus, dass das betweenness-basierte Zentralitätsmaß die wichtigsten Akteure am besten charakterisiert. Weiters zeigte er, dass das degree-basierte Zentralitätsmaß und die betweenness-basierte Zentralität wichtige Indikatoren für das Gruppenverhalten sind.
Bolland (1988) untersuchte ebenfalls die Zentralitätsmaße. Er fand heraus, dass die degree-basierte Zentralität und die closeness-basierte Zentralität sehr anfällig auf Veränderungen in der Netzwerkstruktur sind. Die betweenness-basierte Zentralität ist im Stande kleine Veränderungen in der Netzwerkstruktur zu erfassen, aber ist trotzdem fehleranfällig.

3 CLUSTERANALYSE

3.1 EINLEITUNG

In diesem Kapitel wird in erster Linie das Buch „Multivariate Analysemethoden. Eine anwendungsorientierte Einführung.“ von Backhaus et al. 1996 verwendet.
Die Clusteranalyse ist ein explorativer Vorgang, bei dem die Daten in Gruppen eingeteilt werden. Diese Methode wird oft in den Bereichen des Marketings verwendet, um Kunden nach bestimmten Eigenschaften zu segmentieren. Beispiele dafür sind das Kaufverhalten für bestimmte Produkte, das Telefonierverhalten und noch viele weitere Fragestellungen, die die Wirtschaft bietet. Die Clusteranalyse wird aber unter anderem auch in den Bereichen der Medizin, Psychologie und Soziologie angewandt.
Ziel der Clusteranalyse ist es, Gruppen beziehungsweise Cluster zu finden, wobei die Personen beziehungsweise Objekte innerhalb eines Clusters sehr ähnliche Eigenschaften aufweisen sollen (homogen) und die Cluster untereinander aber sehr verschieden sein sollen (heterogen).

Diese Ähnlichkeit beziehungsweise Unähnlichkeit wird anhand eines Proximitätsmaß gemessen. Die distanzbasierenden Clusterverfahren können in zwei Schritten kurz zusammengefasst werden.

Es wird von einer Rohdatenmatrix der Dimension $N \times M$ ausgegangen, die folgendermaßen aussieht:

	Variable 1	Variable 2	...	Variable M
Objekt 1 Objekt 2 . . Objekt N				

Nun wird im ersten Schritt das Ähnlichkeits- oder Distanzmaß der Variablen von allen möglichen Objektpaaren berechnet.
Dadurch ergibt sich die Ähnlichkeits- oder Distanzmatrix der Dimension $N \times N$.

	Objekt 1	Objekt 2	...	Objekt N
Objekt 1 Objekt 2 . . Objekt N				

Das Ähnlichkeitsmaß und das Distanzmaß haben eine unterschiedliche Bedeutung und auch unterschiedliche Eigenschaften.

Das Ähnlichkeitsmaß wird größer, je ähnlicher zwei Objekte bezüglich der Variablen sind. Es gelten folgende Eigenschaften für das Maß:

$$s_{nm} = s_{mn} \qquad n, m = 1,2,..., N \tag{3.1}$$
$$s_{nm} \leq s_{nn}$$

In den meisten Fällen wird noch vorausgesetzt, dass $s_{nm} \geq 0$ und $s_{nn} = 1$ gilt. Das heißt das Ähnlichkeitsmaß kann nur Werte im Bereich 0 und 1 annehmen.

Das Distanzmaß wird kleiner, je ähnlicher zwei Objekte bezüglich der Variablen sind. Es gelten folgende Eigenschaften für das Maß:

$$d_{nn} = 0 \text{ und } d_{nm} \geq 0$$
$$d_{nm} = d_{mn} \qquad n, m = 1,2,..., N \tag{3.2}$$

Es gibt eine Fülle von verschiedenen Ähnlichkeits- und Distanzmaßen. Aufgrund verschiedener Skalenniveaus der Variablen werden unterschiedliche Maße angewandt. Unter anderem werden für metrischen Skalen die L_2-Norm oder die Mahalanobis-Distanz und für Nominal-Skalen der Tanimoto-Koeffizient oder der M-Koeffizient verwendet. Die richtige Anwendung und Berechnung der Maße können unter anderem im Buch „Multivariate Analysemethoden. Eine anwendungsorientierte Einführung.“ von Backhaus et al. 1996 nachgelesen werden.

Im zweiten Schritt der Clusteranalyse werden dann jene Objekte zu Gruppen zusammengefasst, die eine hohe Übereinstimmung bezüglich des Maßes besitzen. Für die Zusammenfassung von Objekten zu Gruppen stehen verschiedene Clustermethoden zu Verfügung. In den folgenden Kapiteln wird auf die hierarchische Clusteranalyse, die Two-Step Clusteranalyse und auf den K-Means Algorithmus näher eingegangen.

3.2 HIERARCHISCHE CLUSTERANALYSE

Man unterscheidet zwischen agglomerativen und divisiven hierarchischen Clusteranalysen. Agglomerativ bedeutet, dass ausgehen von der kompletten Objektmenge (kleinste Einheit) eine Bildung von Clustern erfolgt. Im Unterschied zum agglomerativen Verfahren wird im divisiven Algorithmus von einem gesamten Cluster ausgegangen, der schrittweise in kleiner Cluster zerlegt wird.

Die hierarchische Clusteranalyse erschafft eine Hierarchie von möglichen Clustern. Unterschiedliche Ebenen der Hierarchie stellen jeweils eine Lösung im Sinne einer Partition dar. Je tiefer die Ebene desto feiner ist die Granularität der Partition.
Dies wird graphisch durch ein sogenanntes Dendrogramm dargestellt.

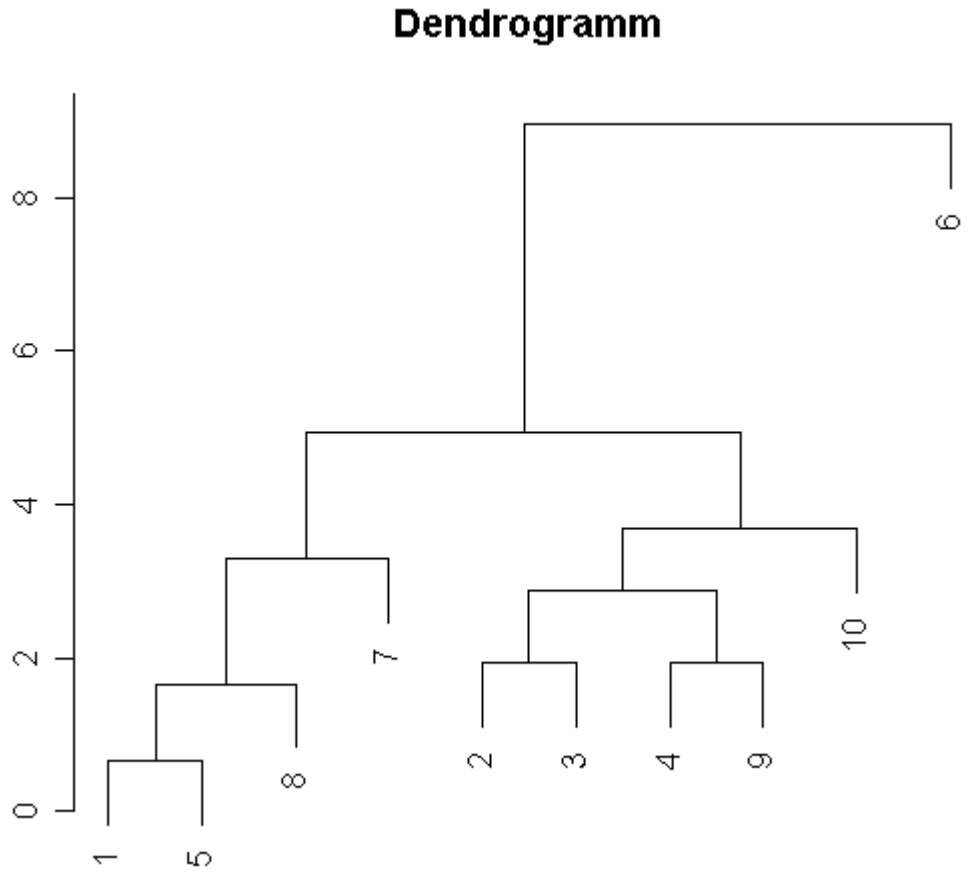

Abbildung 3.1: Dendrogramm

Der Ablauf der agglomerativen hierarchischen Clusteranalyse erfolgt in fünf Schritten (vgl. Backhaus et al. 1996):

1. Man beginnt mit der feinsten Partition. Das heißt N Objekte bilden N Cluster.
2. Berechnung des gewählten Distanzmaßes für alle Objektpaare. Dies entspricht $\binom{N}{2}$ Distanzen.
3. Fusionierung jene Cluster, die die geringste Distanz aufweisen. Die Anzahl der Cluster reduziert sich um 1.
4. Berechnung der neuen Distanzmatrix und gehe zu Schritt 2.
5. Abbruch des Algorithmus, wenn nur noch ein Cluster vorhanden ist.

Die hierarchischen Verfahren unterscheiden sich in der Distanzberechnung von Objektmengen. Es gibt unter anderem das Single-Linkage-Verfahren, das Complete-Linkage Verfahren, das Centroid-Verfahren, das Ward-Verfahren und noch einige andere. Es ist zu beachten, dass nicht alle Proximitätsmaße für die einzelnen Verfahren geeignet sind. Zum Beispiel sind für das Single-Linkage-Verfahren und das Complete-Linkage-Verfahren alle Proximitätsmaße geeignet, jedoch für das Centroid-Verfahren und das Ward-Verfahren sind nur die Distanzmaße für die Berechnung geeignet. Auf die Vorgangsweisen bei diesen Verfahren wird hier jedoch nicht näher eingegangen (vgl. Backhaus et al. 1996).

Die Eigenschaften der hierarchischen Clusteranalyse sind:

- Die Analyse ist nur für kleine Datensätze geeignet.
- Das Ergebnis der Clusteranalyse hängt von dem Verfahren ab.
- Die Anzahl der Cluster muss im Vorhinein nicht bekannt sein.

3.3 TWO-STEP CLUSTERANALYSE

Die Two-Step Clusteranalyse ist im Unterschied zu der hierarchischen Clusteranalyse ein Algorithmus für den Umgang mit sehr großen Datenmengen. Es können sowohl metrische als auch kategorielle Variablen oder Merkmale verwendet werden. Der Vorteil dieser Clusteranalyse ist, dass der Algorithmus nur einen Datendurchlauf benötigt. Der Algorithmus ist in zwei Stufen aufgeteilt. In der ersten Stufe werden die Fälle in viele kleine Sub-Cluster vorgruppiert. In der zweiten Stufe werden dann diese kleinen Sub-Cluster zu der gewünschten Anzahl von Clustern zusammengefasst. Der Algorithmus kann wie auch die hierarchische Clusteranalyse die Anzahl der Cluster automatisch bestimmen. Im Folgenden werden nun diese zwei Stufen genauer betrachtet. Zur literarischen Unterstützung wird der „Clementine 9.0 Algorithmus Guide“ von SPSS verwendet.

Stufe 1: Pre-Cluster

In dieser Stufe wird eine sequentielle Methode angewendet. Es wird ein Fall nach dem anderen betrachtet und entschieden, ob der aktuelle Fall in den zuvor gebildeten Cluster fällt, oder ob ein neuer Cluster gebildet werden soll, der nur diesen Fall enthält. Diese Entscheidung basiert auf einem Distanzkriterium. Als Distanzkriterium kann die Log-Likelihood-Distanz herangezogen werden. Dieses Distanzmaß kann sowohl für metrische als auch für kategorielle Variablen verwendet werden. Es handelt sich hierbei um ein Maß, das auf einer Wahrscheinlichkeit beruht. Der Abstand zwischen zwei Clustern steht im Zusammenhang mit der Abnahme im Log-Likelihood, wenn die zwei Cluster zu einem Cluster zusammengefasst werden. Um die Log-Likelihood zu berechnen, wird vorausgesetzt, dass die metrischen Variablen einer Normalverteilung folgen und die kategoriellen Variablen einer Multinomialverteilung folgen. Weiters wird vorausgesetzt, dass die Variablen und somit auch die Fälle unabhängig voneinander sind.

Die Distanz zwischen Cluster j und s ist definiert als

$$d(j,s) = \xi_j + \xi_s - \xi_{<j,s>}, \tag{3.3}$$

wobei

$$\xi_v = -N_v\left(\sum_{k=1}^{K^A}\frac{1}{2}\log(\hat{\sigma}_k^2 + \hat{\sigma}_{vk}^2) + \sum_{k=1}^{K^A}\hat{E}_{vk}\right) \tag{3.4}$$

und

$$\hat{E}_{vk} = -\sum_{l=1}^{L_k}\frac{N_{vkl}}{N_v}\log\frac{N_{vkl}}{N_v} \tag{3.5}$$

Die Notation ist wie folgt:

K^A .....Anzahl der metrischen Variablen

K^B .....Anzahl der kategoriellen Variablen

L_k .....Anzahl der Kategorien der k-ten kategoriellen Variable

N .....Anzahl der Fälle

N_v .....Anzahl der Fälle in Cluster v

$\hat{\sigma}_k^2$ .....geschätzte Varianz der k-ten metrischen Variable

$\hat{\sigma}_{vk}^2$ .....geschätzte Varianz der k-ten metrischen Variable in Cluster v

N_{vkl} .....Anzahl der Fälle in Cluster v, deren k-te kategorielle Variable die l-te Kategorie annimmt

Ein weiteres Distanzkriterium stellt die euklidische Distanz dar. Diese Distanz kann allerdings nur dann verwendet werden, wenn alle verwendeten Variablen metrisch sind. Die euklidische Distanz zwischen zwei Vektoren $x = (x_1,...,x_n)$ und $y = (y_1,...,y_n)$ ist definiert als

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3.6}$$

Die Distanz zwischen zwei Clustern entspricht dann der euklidischen Distanz zwischen den Clusterzentren. Ein Clusterzentrum ist definiert als ein Vektor, der die Mittelwerte der Variablen des Zentrums enthält.

Diese nun zuvor beschriebene Prozedur, bei der entschieden wird, ob ein Fall in ein vorhandenes Cluster fällt, oder ob ein neues Cluster gebildet wird, wird durch den Bau eines sogenannten Cluster Feature Tree (CF) nach dem BIRCH-Verfahren (Balanced Iterative Reducing and Clustering using Hierarchies) umgesetzt (vgl Zhang et al. 1996). Der CF-Baum besteht aus Ebenen von Knoten. Der CF-Baum besitzt eine Wurzel, die alle Fälle repräsentiert. Die unterste Ebene des Baumes besteht aus einer Reihe von Blättern, die die Sub-Cluster darstellen. Die Wurzel und die Blätter sind durch Ebenen von Knoten, den sogenannten Ästen, miteinander verbunden. Die Anzahl der Ebenen muss im Vorhinein festgelegt werden. Jeder Knoten besitzt folgende Anzahl von Einträgen, nämlich die Anzahl der Fälle, die in den nachfolgenden Knoten vorhanden sind, der Mittelwert und die Varianz der metrisch skalierten Variablen und die Anzahl der Kategorien für jede kategorielle Variable.
Die Zuordnung jedes Falls startet bei der Wurzel und wird rekursiv durch die für den Fall passenden Einträge der Knoten in das passende Blatt geleitet und einem Cluster innerhalb des Blattes zugewiesen. Nun wird überprüft, ob durch die Zuordnung des Falles der Schwellwert für die Heterogenität innerhalb des Blattes nicht überschritten wird. Ist dies der Fall, bleibt der Fall in diesem Blatt und die Einträge der übergeordneten Knoten werden neu berechnet. Wird der Schwellwert überschritten, wird innerhalb des Blattes ein neues Cluster gebildet, welches den Fall beinhaltet. Sollte dadurch die maximale Anzahl an zugeordneten Fällen innerhalb des Blattes überschritten werden, wird das Blatt in zwei Blätter geteilt. Dies erfolgt, indem die zwei heterogensten Cluster ermittelt werden, die den zwei neuen Blättern entsprechen. Die übrigen Cluster, die in dem Blatt vorhanden waren, werden durch ein Ähnlichkeitskriterium den zwei neuen Blättern zugewiesen. Nachdem die Größe des CF-Baumes im Vorhinein festgelegt werden muss, können nur so viele neue Blätter erzeugt werden bis die

maximale Größe des Baumes erreicht wird. Wird die maximale Größe des Baumes erreicht, wird der Baum restrukturiert, indem der Schwellwert für die Heterogenität innerhalb eines Knoten erhöht wird. Dadurch wird der rekonstruierte CF-Baum kleiner und hat mehr Platz für neue Fälle. Dieser Prozess wird solange durchgeführt bis alle Fälle den Baum durchlaufen haben (vgl. Zhang et al. 1996).

Stufe 2: Clustern

Nun werden die zuvor gebildeten Sub-Cluster zu der gewünschten Anzahl von Cluster gruppiert. Nachdem die Anzahl der Sub-Cluster erheblich geringer ist als die Anzahl der Fälle, lassen sich herkömmliche Clustermethoden effektiv anwenden. Das Statistikprogramm SPSS verwendet hier die zuvor beschriebene hierarchische Clusteranalyse. Je höher jedoch die Anzahl der Sub-Cluster, die in der Stufe 1 erzeugt wurden, ist, desto genauer wird das endgültige Ergebnis. Dies führt allerdings wiederum dazu, dass sich die Laufzeit der zweiten Stufe erhöht.

Nachdem SPSS für die Aggregation der Sub-Cluter die hierarchische Clusteranalyse (siehe Kapitel 3.2) verwendet, muss die Anzahl der endgültigen Cluster nicht bekannt sein. Man kann die Anzahl auch automatisch ermitteln. Im ersten Schritt wird das BIC (Baysian Information Criterion) oder das AIC (Akaike Information Criterion) für jede Anzahl von Cluster innerhalb eines festgelegten Bereichs berechnet. Es wird für die anfänglich geschätzte Anzahl von Cluster verwendet. Im zweiten Schritt wird die anfänglich geschätzte Anzahl dadurch modifiziert, indem der größte Anstieg in der Distanz zwischen den zwei nahe liegensten Clustern in jeder hierarchischen Clusterebene gesucht wird. Das BIC und AIC für J Cluster ist definiert als

$$BIC(J) = -2\sum_{j=1}^{J} \xi_j + m_J \log(N) \tag{3.7}$$

$$AIC(J) = -2\sum_{j=1}^{J} \xi_j + 2m_J \ , \tag{3.8}$$

wobei

$$m_J = J\left\{2K^A + \sum_{k=1}^{K^B}(L_k - 1)\right\} \tag{3.9}$$

Im Allgemeinen liefert das BIC eine geringere Anzahl an Cluster, da für $N > 8$ der Strafterm für die Zahl der Parameter stärker wirkt.

Einerseits ist die Two-Step Clusteranalyse sehr gut geeignet, um große Datensätze zu segmentieren, da diese Prozedur nur einen Durchlauf aller Fälle hat. Andererseits muss man beachten, dass in der Stufe 1: Pre-Cluster die Konstruktion des CF-Baumes von der Reihenfolge der Fälle abhängig ist. Je nach dem in welcher Reihenfolge die Daten den CF-Baum durchlaufen, ist die Konstruktion des Baumes verschieden. Daher sollte man in der Praxis die Fälle zufällig ordnen, um die Stabilität der Ergebnisse zu überprüfen.

3.4 K-MEANS CLUSTERANALYSE

Zur literarischen Unterstützung wird der „Clementine 9.0 Algorithmus Guide“ von SPSS verwendet.
Der K-Means Algorithmus wird auch Clusterzentrenanalyse genannt. K-Means ist ein iterativer Algorithmus und funktioniert daher für große Datenmengen sehr gut. Allerdings müssen die verwendeten Variablen metrisch skaliert werden.

Der Algorithmus funktioniert folgendermaßen:

1. Wähle k Anfangs-Cluster-Zentren.
2. Berechne für jeden Fall die quadrierte euklidische Distanz zu den Cluster-Zentren und weise es dem nahesten Zentrum zu.
3. Berechne die Cluster-Zentren neu.
4. Wiederhole Schritt 2 und 3 bis:
   - es keinen Unterschied mehr in den Cluster-Zentren gibt.
   - die angegebene Anzahl der Iterationen durchgeführt wurde.

Ein Cluster-Zentrum ist ein Vektor, der aus den Mittelwerten der Fälle der einzelnen Variablen besteht, die dem Cluster zugewiesen wurden.
In Schritt 2 wird für die Berechnung der Distanz die quadrierte euklidische Distanz ermittelt. Die Distanz ist definiert als

$$d(i,j) = \left\| X_i - C_j \right\|^2 = \sum_{q=1}^{Q} (x_{qi} - c_{qj})^2 \,, \tag{3.10}$$

wobei

X_i .....Vektor der Variablen des Falls i

C_j .....Vektor der Cluster-Zentren für Cluster j

Q .....Anzahl der Variablen

x_{qi} .....Wert der q-ten Variable für den i-ten Fall

c_{qj} .....Wert der q-ten Variable für den j-ten Fall

Für jeden Fall wird die Distanz zwischen dem Fall und jedem Cluster-Zentrum berechnet und jenem Cluster-Zentrum zugewiesen, welche die kürzeste Distanz hatte. Nachdem alle Fälle zugewiesen wurden, werden die Cluster-Zentren neu berechnet. Das Cluster-Zentrum entspricht einem Mittelwertsvektor:

$$C_j = \overline{X}_j , \tag{3.11}$$

wobei die einzelnen Komponenten des Mittelwertsvektors folgendermaßen berechnet werden:

$$\overline{x}_{qj} = \frac{\sum_{i=1}^{n_j} x_{qi}(j)}{n_j} \tag{3.12}$$

wobei

n_j .....Anzahl der Fälle in Cluster j

$x_{qi}(j)$ ..Wert der q-ten Variable für den i-ten Fall, welches in dem Cluster j zugewiesen wurde

Die Problematik des K-Means Algorithmus liegt darin, dass auch hier bei unterschiedlichen Anfangswerten unterschiedliche Ergebnisse entstehen können. Daher sollte man sich genau überlegen, mit welchen Anfangswerten begonnen werden soll. Eine Möglichkeit besteht darin, dass eine kleine Stichprobe gezogen wird. Auf diese Stichprobe wird die hierarchische Clusteranalyse angewandt. Dadurch erhält man zunächst die Anzahl der Cluster als geschätzten Anfangswert. Im nächsten Schritt müssen nun die k Anfangs-Cluster-Zentren bestimmt werden. In der Praxis wird empfoh-

len den Algorithmus für unterschiedliche Werte von k Anfangs-Cluster-Zentren durchlaufen zu lassen und die beste Lösung zu wählen.
Ein weiteres Problem stellt die Reihenfolge der Daten dar. Wird die Reihenfolge der Daten geändert, kann es zu anderen Startwerten und somit zu anderen Ergebnissen kommen. Daher sollten auch hier die Ergebnisse auf Stabilität geprüft werden.

4 ANWENDUNG ANHAND EINES DATENSATZES

4.1 BESCHREIBUNG DER DATEN

Der zu untersuchende Datensatz besteht aus 43 Variablen. Die Variablen beschreiben das Telefonierverhalten der 677.173 untersuchten Personen. Der Untersuchungszeitraum für die Ermittlung des Telefonierverhaltens beträgt ein Monat.
Da es sich hier um ein gerichtetes soziales Netzwerk handelt, können die Variablen in ausgehende und eingehende Kontakte unterteilt werden. Ausgehende Kontakte werden mit „out“ und eingehende Kontakte werden mit „in“ bezeichnet. Die Kontakte werden gemessen in Degree (Anzahl der Personen, siehe Kapitel 2), Gesprächsdauer gemessen in Sekunden, Häufigkeit der Anrufe und Anzahl der SMS. Diese verschieden gemessenen Kontakte werden noch dahingehend unterteilt, indem das technische Kommunikationsnetzwerk (Provider) berücksichtig wird. Es werden insgesamt fünf unterschiedliche Kommunikationsnetzwerke unterschieden, nämlich Mobilfunknetz Intern, Mobilfunknetz Extern, Ausland, Festnetz und Unbekannt. Die Variablen sind alle metrisch skaliert.
Weiters enthält der Datensatz eine Variable Haushalts-ID. Das heißt, dass Personen, die im selben Haushalt leben die gleiche Haushalts-ID besitzen.
Die Variable Filter ist die einzige binäre Variable und bedeutet, dass eine Person, die einen Degree größer gleich 300 hat oder die Sekunden größer gleich 300.000 sind, wird angenommen, dass es sich hierbei um eine Hotline oder ähnliches handelt und bekommt daher den Wert 1. Ist dies nicht der Fall beträgt der Wert 0.

In der folgenden Tabelle werden die Variablen durch wichtige deskriptive Kennzahlen beschrieben:

Variable	Min	Max	Mittelwert	Std.Abw.
in-degree-intern	0	258	8,040	6,709
in-degree-extern	0	362	7,196	7,328
in-degree-ausland	0	1737	1,796	7,366
in-degree-festnetz	0	89	1,027	1,488
in-degree-unbekannt	0	230	3,169	3,264
in-sec-intern	0	447523	6684,342	9562,087
in-sec-extern	0	293011	3252,161	5745,093
in-sec-ausland	0	161946	241,936	1356,828
in-sec-festnetz	0	103535	350,953	1185,897
in-sec-unbekannt	0	92712	661,859	1429,819
in-calls-intern	0	14146	55,888	61,944
in-calls-extern	0	904	23,009	31,572
in-calls-ausland	0	442	1,152	4,796
in-call-festnetz	0	607	2,745	7,552
in-call-unbekannt	0	822	4,624	7,680
in-sms-intern	0	4832	14,340	57,122
in-sms-extern	0	3303	16,506	68,149
in-sms-ausland	0	4865	5,190	20,790
in-sms-festnetz	0	846	0,132	4,128
in-sms-unbekannt	0	7359	3,378	15,047
out-degree-intern	0	893	8,182	7,223
out-degree-extern	0	442	8,498	8,580
out-degree-ausland	0	159	1,303	2,902
out-degree-festnetz	0	156	0,964	1,447
out-degree-unbekannt	0	807	6,060	7,900
out-sec-intern	0	286437	7124,139	10627,272
out-sec-extern	0	296078	4306,327	8080,710
out-sec-ausland	0	170329	383,872	1970,336
out-sec-festnetz	0	169748	518,071	2219,101

Variable	Min	Max	Mittelwert	Std.Abw.
out-sec-unbekannt	0	248182	2048,933	4524,235
out-calls-intern	0	2834	57,355	65,722
out-calls-extern	0	1679	36,807	54,688
out-calls-ausland	0	855	2,545	10,237
out-calls-festnetz	0	4099	3,746	12,411
out-calls-unbekannt	0	2236	14,418	21,727
out-sms-intern	0	4653	14,839	61,015
out-sms-extern	0	3178	15,774	69,795
out-sms-ausland	0	2592	2,669	17,406
out-sms-festnetz	0	923	0,128	3,979
out-sms-unbekannt	0	970	0,233	3,899
HAUSHALTS-ID	0	4201788	3407174,587	978430,727
FILTER	0	1	--	--

Tabelle 4.1: dekriptive Statistik

Nun können ausgehend von diesen Variablen weitere für die Analyse nützliche Variablen berechnet werden. Einerseits kann man die Struktur der einzelnen Provider als auch die Struktur der Provider insgesamt betrachten.

Um die Struktur der einzelnen Provider zu beschreiben, können nun folgende Variablen gebildet werden.
Es gibt $X_1,...,X_i,...,X_N$ Personen. Für jedes X_i kann folgendes berechnet werden, wobei IS für in_Sekunden und OS für out_Sekunden steht:

- Summe der eingehenden und ausgehenden Sekunden pro Provider: beschreibt die gesamte Nutzung pro Provider
  $IS_i^j + OS_i^j$, für $j = 1,...,5$ (Provider)

- Differenz der eingehenden und ausgehenden Sekunden pro Provider: beschreibt die Dauer pro Provider
  $OS_i^j - IS_i^j$, für $j = 1,...,5$ (Provider)

- Summe der Indegree und Outdegree pro Provider:
  beschreibt die Anzahl der Personen, zu denen man Kontakt hat, gesamt pro Provider
  $ID_i^j + OD_i^j$, für $j = 1,...,5$ (Provider)

- $\frac{OS_i^j}{\sum_{j=1}^{5} OS_i^j}$, für $j = 1,...,5$ (Provider)

Um die Struktur der Provider insgesamt zu beschreiben, können unter anderem diese Variablen gebildet werden:

- Summe der ausgehenden Sekunden aller Provider:
  $\sum_{j=1}^{5} OS_i^j$
- Summe der Outdegree aller Provider:
  $\sum_{j=1}^{5} OD_i^j$

Berechnungen dieser Art können auch mit den Variablen Degree, Call und SMS durchgeführt werden.

Anhand dieser Daten werden die Methoden Two-Step Clusteranalyse (siehe Kapitel 3.3) und K-Means Clusteranalyse (siehe Kapitel 3.4) in den folgenden Kapiteln angewandt. Zur Berechnung wird das Statistikprogramm Clementine von SPSS verwendet.

Bevor jedoch mit der Analyse begonnen werden kann, werden die Daten anhand von Histogrammen graphisch betrachtet.

Als Beispiel wird hier das Histogramm für die Variable in_sec_intern:

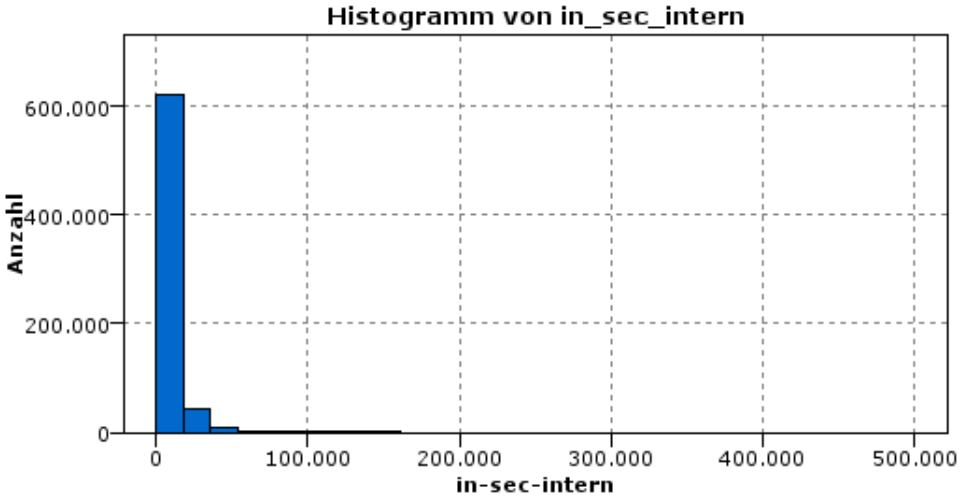

Abbildung 4.1: Histogramm der Variable in_sec_intern

Wie man erkennen kann, sind die Daten extrem rechtsschief. Die Histogramme aller anderen Variablen sind sehr ähnlich und deuten ebenfalls auf eine extrem rechtsschiefe Verteilung der Daten. Die Daten sollten daher transformiert werden, um einer Normalverteilung näher zu kommen. Eine mögliche Transformation stellt das Logarithmieren der Daten dar. Da die Ausprägung Null in allen Variablen häufig vorkommt, wird die Transformation $\log(x+1)$ verwendet.

Nach der Transformation der Variable in_sec_intern sieht das Histogramm folgendermaßen aus:

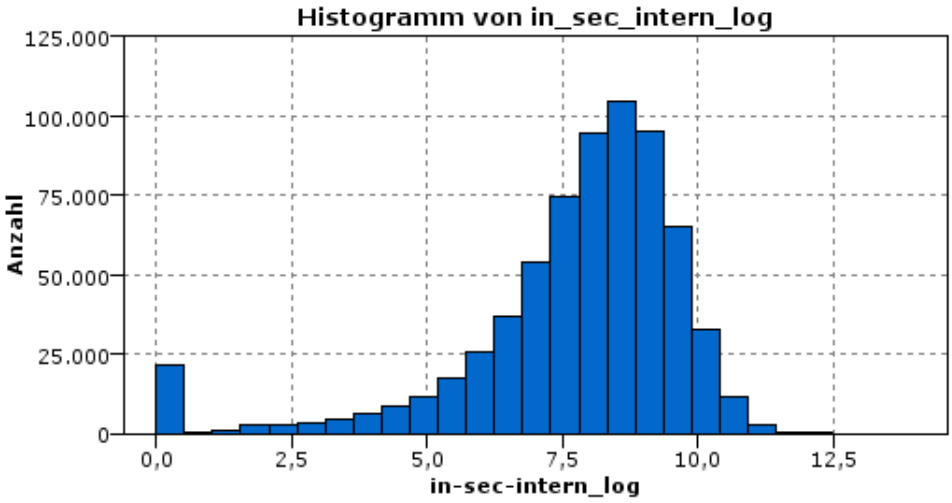

Abbildung 4.2: Histogramm der logarithmierten Variable in_sec_intern_log

Wie man anhand der Graphik erkennen kann, stellt die Transformation für die Variable in_sec_intern eine gute Lösung dar. Die Variable in_sec_extern verhält sich ähnlich wie die abgebildete Variable. Bei den Variablen der Provider Ausland, Festnetz und Unbekannt dominiert der Nullwert, da ein Großteil der Kunden diesen Providerdienst nicht nutzen.

Hier ein Beispiel für die Variable in_sec_ausland_log:

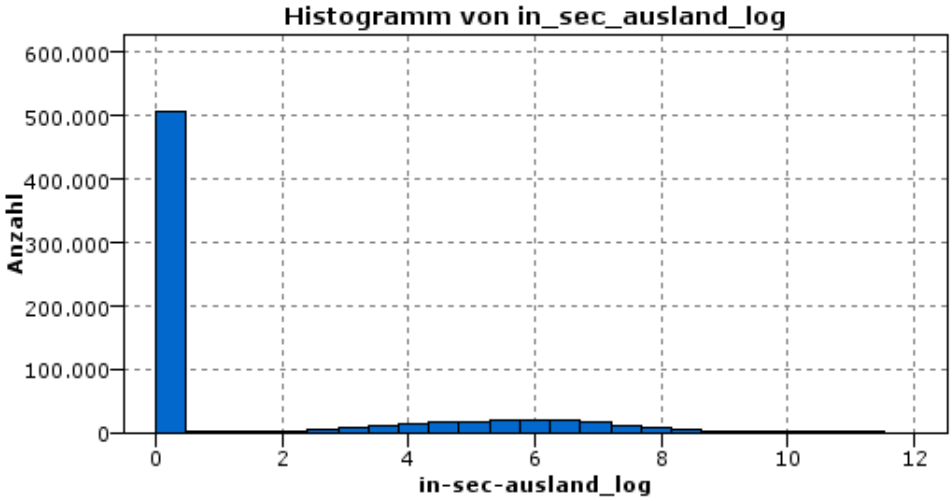

Abbildung 4.3: Histogramm der logarithmierten Variable in_sec_ausland_log

Die anderen Variablen verhalten sich sehr ähnlich wie die jetzt abgebildeten Variablen. Es werden aber auch hier mit den logarithmierten Daten die weitern Analysen durchgeführt, um stabilere Ergebnis zu erhalten.

4.2 PRÜFUNG DER STABILITÄT

Sowohl bei der Anwendung der Two-Step Clusteranalyse als auch bei der Anwendung des K-Means Alogrithmus muss, wie in der Beschreibung der Verfahren ausgeführt wurde, beachtet werden, dass die Ergebnisse von der Reihenfolge der Daten abhängig sind. Um diesen Effekt zu umgehen, sollten die Daten bei praktischen Berechnungen mehrmals zufällig permutiert werden und die resultierenden Ergebnisse miteinander verglichen werden. Bei sehr großen Datenmengen würde dies zu einem großen Aufwand führen. Daher werden bei sehr großen Datenmengen gerne zufällig Stichproben gezogen.

Anhand des nun vorher beschriebenen Datensatzes wird nun versucht die gesamte Nutzung der Kunden zu segmentieren. Das heißt, dass die Summen der hineinkommenden und hinausgehenden Sekunden pro Provider (intern, extern, Ausland, Festnetz und Unbekannt) gebildet werden. Dies entspricht fünf neu berechneten Variablen, die folgendermaßen deskriptiv beschrieben werden können:

Feld	Min	Max	Mittelwert	Std.Abw.
sum_in_out_sec_intern	0	502817	13808,481	18238,088
sum_in_out_sec_extern	0	363834	7558,488	12050,133
sum_in_out_sec_ausland	0	184186	625,808	2721,501
sum_in_out_sec_festnetz	0	190672	869,024	2801,807
sum_in_out_sec_unbekannt	0	314197	2710,792	5127,873

Tabelle 4.2: deskriptive Statistik

Mithilfe dieser Variablen, die ebenfalls log-transformiert werden, wird nun anhand von Stichproben eine Two-Step Clusteranalyse und eine K-Means Clusteranalyse durchgeführt und die Ergebnisse auf ihre Stabilität geprüft.

4.2.1 EINE 10%IGE STICHPROBE

Es wird nun eine Stichprobe von 10% gezogen. Diese Stichprobe wird vier Mal permutiert. Es wird überprüft, ob die Ergebnisse einer Two-Step Clusteranalyse innerhalb dieser Stichprobe stabil sind. Eine erste Analyse mit automatischer Wahl der Anzahl der Cluster wurde durchgeführt.

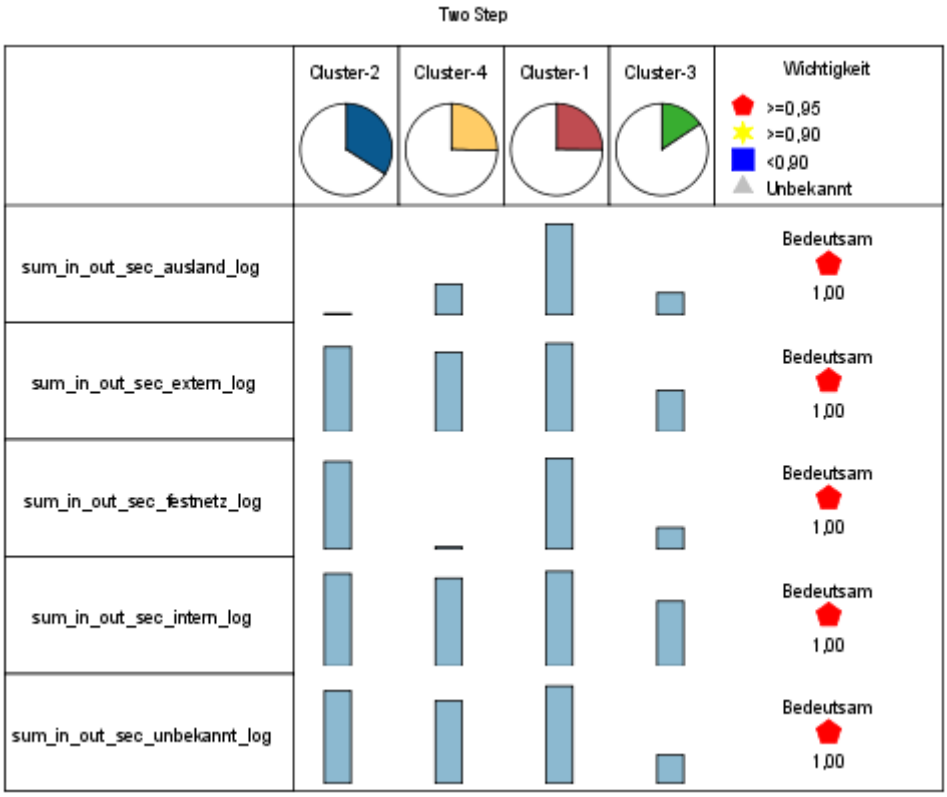

Abbildung 4.4: Two-Step Clusteranalyse einer 10%igen Stichhprobe

Anhand dieser Graphik erkennt man wie sich die Variablen auf die einzelnen Cluster aufteilen. Diese Abbildung kann auch in Zahlen ausgedrückt werden.

Two Step

	Cluster-2 Anzahl: 22914 (33%)	Cluster-4 Anzahl: 17063 (25%)	Cluster-1 Anzahl: 16917 (25%)	Cluster-3 Anzahl: 10517 (15%)	Wichtigkeit >=0,95 >=0,90 <0,90 Unbekannt
sum_in_out_sec_ausland_log	0,09 (0,45)	2,09 (3,00)	6,23 (1,57)	1,52 (2,72)	Bedeutsam 1,00
sum_in_out_sec_extern_log	8,34 (1,34)	7,80 (1,45)	8,68 (1,29)	4,06 (3,30)	Bedeutsam 1,00
sum_in_out_sec_festnetz_log	6,00 (1,53)	0,16 (0,64)	6,23 (1,57)	1,47 (2,40)	Bedeutsam 1,00
sum_in_out_sec_intern_log	9,08 (1,20)	8,63 (1,33)	9,29 (1,15)	6,38 (3,02)	Bedeutsam 1,00
sum_in_out_sec_unbekannt_log	7,25 (1,38)	6,48 (1,48)	7,63 (1,40)	2,23 (2,82)	Bedeutsam 1,00

Abbildung 4.5: Two-Step Clusteranalyse einer 10%igen Stichprobe mit Angabe des Mittelwerts und der Standardabweichung

Die Interpretation dieser Graphiken ist sehr einfach. Zum Beispiel ist die Variable sum_in_out_sec_ausland_log in Cluster 1 sehr hoch ausgeprägt, während sie in den anderen Clustern sehr niedrig ausgeprägt ist. In Zahlen bedeutet das, dass die Kunden im Mittel 6,23 Sekunden ins Ausland telefonieren. Wenn man den Mittelwert nun zurücktransformiert, dann ergibt sich ein Wert von 506,75 Sekunden. Die Zahl innerhalb der Klammer entspricht der Standardabweichung. Es lässt darauf deuten, dass jene Kunden in Cluster 1 sind, die viel ins Ausland telefonieren. Wenn man den Cluster 1 gesamt betrachtet, befinden sich all jene Kunden, die alle Provider sehr frequentieren. Cluster 2 besteht aus jenen Kunden, die die Provider Extern, Festnetz, Intern und Unbekannt, jedoch nicht das Ausland benutzen. In Cluster 4 benutzen die Kunden den Provider Festnetz nicht. Cluster 3 entspricht im Vergleich zu den anderen Clustern dem „Wenig-Telefonierer“.

Diese Analyse wurde nun noch mit 3 verschiedenen zufälligen Permutationen dieser Stichprobe durchgeführt. Um einen Überblick über die Größen

der einzelnen Cluster der Permutationen zu bekommen, wurden diese in der folgenden Tabelle gegenübergestellt.

	Permutation 1	Permutation 2	Permutation 3	Permutation 4
Cluster 1	241.526	241.602	241.428	241.601
Cluster 2	85.497	85.549	85.448	85.550
Cluster 3	154.354	155.020	153.531	155.016
Cluster 4	195.796	195.002	196.766	195.006

Tabelle 4.3: Anzahl der Kunden der einzelnen Cluster der Permutationen

Zum Vergleich der Ergebnisse der Permutationen wurde für jeden Cluster eine Graphik erstellt. Diese Graphik stellt die zurücktransformierten Mittelwerte der Provider der einzelnen Permutationen dar. Die Mittelwerte wurden zurücktransformiert, um eine bessere Interpretation zu ermöglichen.

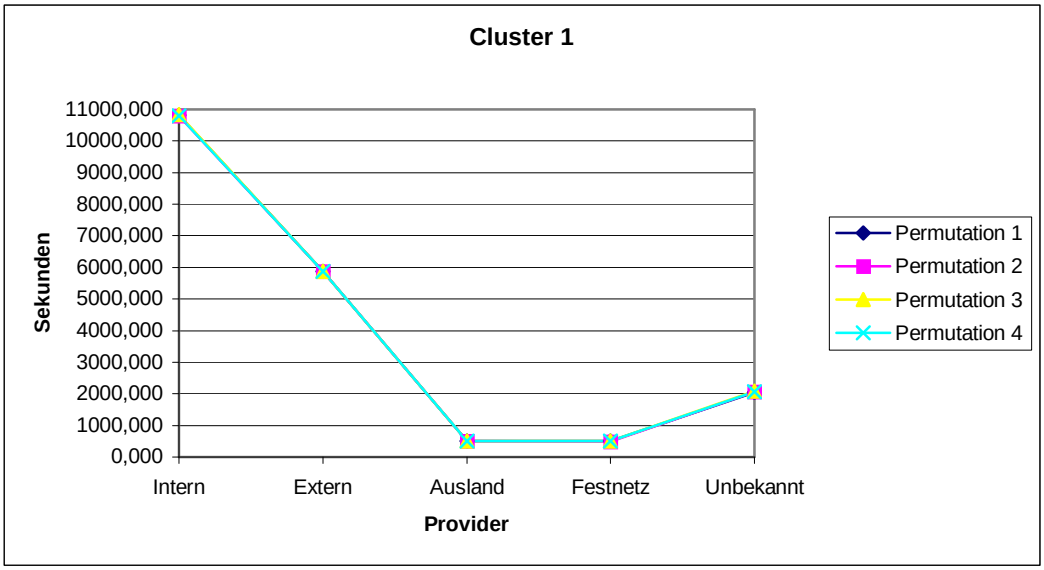

Abbildung 4.6: Vergleich der Ergebnisse der Permutationen in Cluster 1 einer 10%igen Stichprobe

Cluster 1 hat in allen vier Permutationen eine Größe von 35%.
Wie man nun erkennen kann, zeigen die Analysen der Permutationen in Cluster 1 eine Übereinstimmung. Cluster 1 enthält jene Kunden, die im

Mittel bis zu 11.000 Sekunden den Provider Intern benutzen. Der Provider Extern wird im Mittel bis zu 6.000 Sekunden benutzt. Die Provider Ausland, Festnetz und Unbekannt werden nur gering verwendet. Dieser Cluster entspricht trotzdem dem „Viel-Telefonierer“.

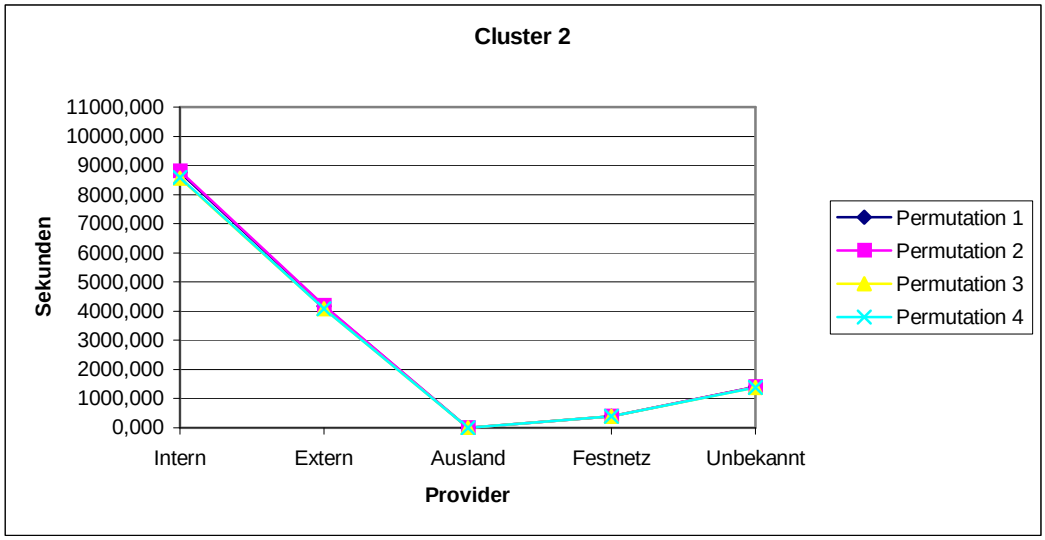

Abbildung 4.7: Vergleich der Ergebnisse der Permutationen in Cluster 2 einer 10%igen Stichprobe

Cluster 2 entspricht einer Größe von 12% in allen Permutationen. Cluster 2 ist dem Cluster 1 sehr ähnlich. Der Unterschied liegt lediglich darin, dass die Kunden weniger telefonieren, aber auch hier herrscht wieder eine Übereinstimmung der Ergebnisse der Permutationen.

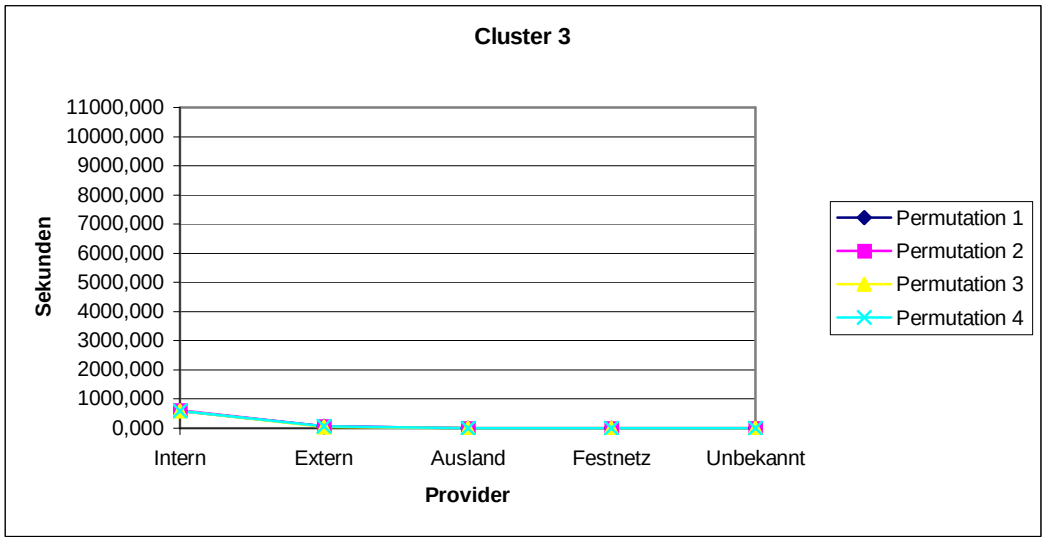

Abbildung 4.8: Vergleich der Ergebnisse der Permutationen in Cluster 3 einer 10%igen Stichprobe

Die Größe des Cluster 3 liegt bei 22% in allen Permutationen. Cluster 3 unterscheidet sich hinsichtlich Cluster 1 und 2 sehr. Der Cluster enthält jene Kunden die gesamt gesehen sehr wenig telefonieren. Die Mittelwerte der Provider Ausland, Festnetz und Unbekannt liegen unter 9 Sekunden. Im Provider Intern wird im Mittel bis zu 600 Sekunden telefoniert und im Provider extern bis zu 60 Sekunden telefoniert. Dieser Cluster entspricht somit dem „Wenig-Telefonierer“. Die Ergebnisse der Permutationen stimmen auch hier wieder überein.

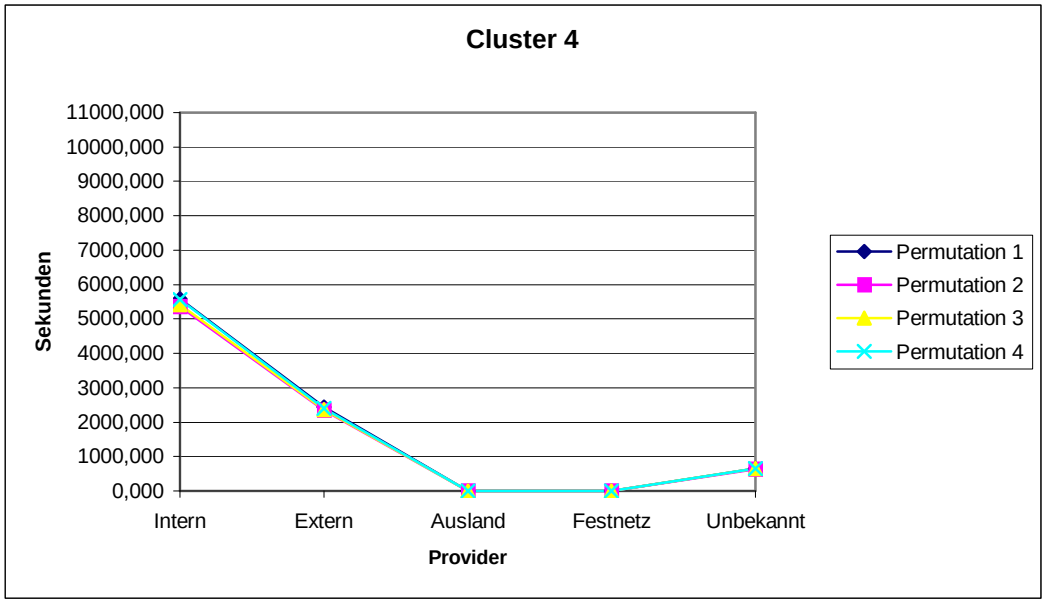

Abbildung 4.9: Vergleich der Ergebnisse der Permutationen in Cluster 4 einer 10%igen Stichprobe

Cluster 4 umfasst eine Größe von 28% in allen Permutationen.
Vom Telefonierverhalten sind die Kunden dem Cluster 1 und 2 ähnlich, jedoch ist die Dauer noch geringer als in Cluster 2. Die Ergebnisse der Permutationen stimmen abermals überein.

Zusammengefasst kann man sagen, dass die einzelnen Ergebnisse eine hohe Stabilität aufweisen. Man kann kaum Unterschiede in den Ergebnissen der Two-Step Clusteranalyse der einzelnen Permutationen der 10%igen Stichprobe erkennen. Man kann daher davon ausgehen, dass Stabilität vorliegt.

Die K-Means Clusteranalyse hängt ebenfalls davon ab, in welcher Reihenfolge die Daten den Algorithmus durchlaufen. Auch in diesem Fall wird die gleiche Prozedur wie bei der Two-Step Clusteranalyse durchgeführt. Es wird auch hier die gleiche 10%ige Stichprobe verwendet. Ein erstes Ergebnis des K-Means Algorithmus mit einer vorgegebenen Anzahl von vier Clustern sieht folgendermaßen aus:

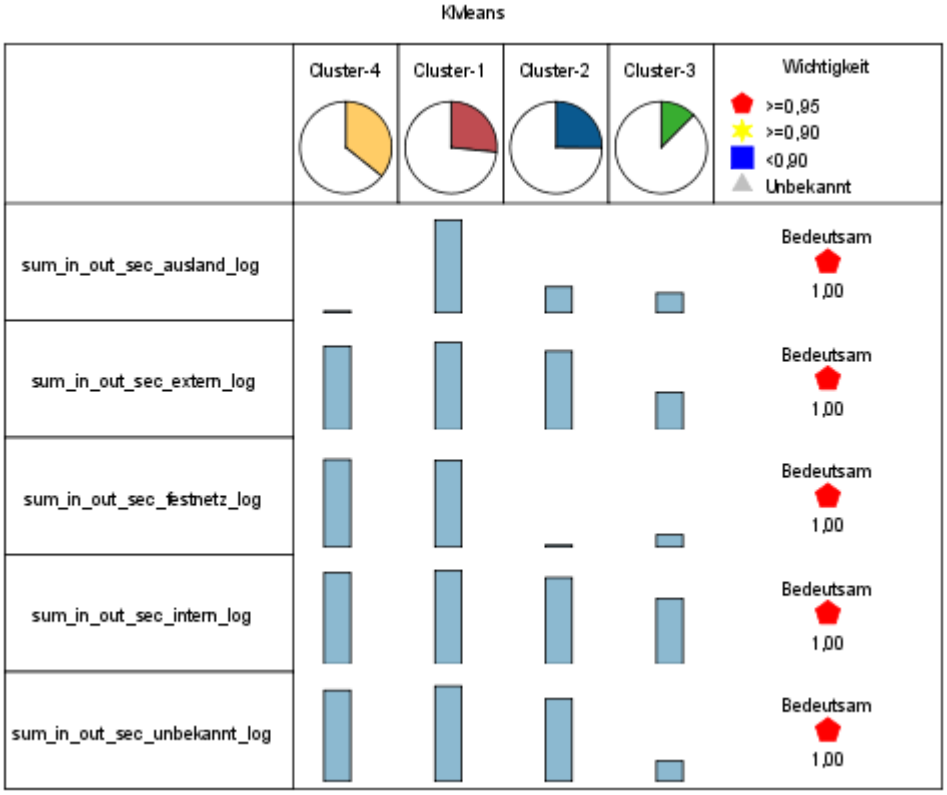

Abbildung 4.10: K-Means Clusteranalyse einer 10%igen Stichprobe

KMeans

	Cluster-4 Anzahl: 24108 (35%)	Cluster-1 Anzahl: 17935 (26%)	Cluster-2 Anzahl: 16910 (25%)	Cluster-3 Anzahl: 8458 (12%)	Wichtigkeit >=0,95 >=0,90 <0,90 Unbekannt
sum_in_out_sec_ausland_log	0,12 (0,55)	6,36 (1,57)	1,79 (2,76)	1,36 (2,58)	Bedeutsam 1,00
sum_in_out_sec_extern_log	8,21 (1,59)	8,56 (1,49)	7,73 (1,58)	3,70 (3,20)	Bedeutsam 1,00
sum_in_out_sec_festnetz_log	6,01 (1,51)	5,95 (1,92)	0,14 (0,57)	0,85 (1,81)	Bedeutsam 1,00
sum_in_out_sec_intern_log	8,98 (1,41)	9,20 (1,31)	8,49 (1,84)	6,44 (2,94)	Bedeutsam 1,00
sum_in_out_sec_unbekannt_log	7,15 (1,56)	7,47 (1,64)	6,47 (1,52)	1,60 (2,36)	Bedeutsam 1,00

Abbildung 4.11: K-Means Clusteranalyse einer 10%igen Stichprobe mit Angabe des Mittelwerts und der Standardabweichung

Wenn man nun das erste Ergebnis der Two-Step Clusteranalyse mit diesem Ergebnis der K-Means Clusteranalyse vergleicht, kann man erken-

nen, dass es sich sehr ähnlich ist. Cluster 1 entspricht Cluster 1 in der Two-Step Analyse. Cluster 2 stimmt mit Cluster 4 überein. Cluster 3 entspricht Cluster 3 und Cluster 4 entspricht Cluster 2 in der Two-Step Clusteranalyse. Zur Veranschaulichung wird dies in einer Kreuztabelle dargestellt:

		K-Means Clusteranalyse			
		Cluster 1	Cluster 2	Cluster 3	Cluster 4
Two-Step Cluster-analyse	Cluster 1	16.651	14	0	252
	Cluster 2	0	122	0	22.792
	Cluster 3	506	683	8.374	954
	Cluster 4	778	16.091	83	111

Tabelle 4.4:Kreuztabelle der Two-Step Clusteranalyse und der K-Means Clusteranalyse

94,8% der Beobachtungen wurde in den beiden Verfahren identisch klassifiziert.

Nach Durchführung der restlichen Permutationen und Segmentierung wurden die Ergebnisse miteinander verglichen. Es zeigt sich ebenfalls, dass die Resultate Stabilität aufweisen. Von der Interpretation her gibt es kaum Unterschiede zu den Ergebnissen der Two-Step Clusteranalyse.

4.2.2 VERSCHIEDENE 10%IGE STICHPROBEN

Im vorigen Kapitel wurde eine 10%ige Stichprobe vier mal permutiert und segmentiert. Nun werden vier verschieden zufällige 10%ige Stichproben gezogen und anhand derselben Variablen segmentiert. Um einen Vergleich der vorigen und jetzigen Analysen zu gewährleisten, wurden die Anzahl der Cluster wieder auf vier gesetzt. Sowohl die Ergebnisse der Two-Step Clusteranalyse als auch die der K-Means Clusteranalyse wei-

sen die selben Eigenschaften wie im Kapitel 4.2.1 auf. Betrachtet man die Resultate der verschiedenen Stichproben, kann man darauf schließen, dass sie sehr homogen sind.

4.2.3 STICHPROBENERHÖHUNG

Um wirklich jeden Zufall der Ergebnisse auszuschließen, wird nun die Größe der Stichprobe schrittweise angehoben und deren Ergebnisse miteinander verglichen. Ausgehend von der 10%igen Stichprobe wird diese in Zehnerschritten bis zu einer 50%igen Stichprobe angehoben. Es wird wieder versucht die gesamte Nutzung der Kunden anhand des Two-Step Algorithmus und des K-Means Algorithmus zu segmentieren.

Es ergeben sich bei einer Durchführung der Two-Step Clusteranalyse folgende Ergebnisse:

Cluster 1			Intern		Extern	
	Größe	Größe in %	Mittelwert	Stabw.	Mittelwert	Stabw.
Stichprobe 10%	10257	15%	585,399	21,021	50,676	26,633
Stichprobe 20%	20937	15%	606,286	19,635	55,092	26,358
Stichprobe 30%	31653	15%	601,447	19,968	55,261	25,924
Stichprobe 40%	41718	15%	609,941	19,656	55,261	26,249
Stichprobe 50%	51733	15%	629,177	19,656	54,757	26,249

Cluster 1	Ausland		Festenetz		Unbekannt	
	Mittelwert	Stabw.	Mittelwert	Stabw.	Mittelwert	Stabw.
Stichprobe 10%	3,874	14,975	3,280	9,773	7,273	15,232
Stichprobe 20%	3,855	14,895	3,459	10,235	7,390	15,103
Stichprobe 30%	4,099	15,511	3,345	10,012	7,750	15,103
Stichprobe 40%	3,707	14,580	3,263	9,827	7,688	15,232
Stichprobe 50%	3,874	14,991	3,267	9,881	7,174	14,784

Tabelle 4.5: Vergleich der Ergebnisse der Stichproben einer Two-Step Clusteranalyse in Cluster 1

Wie man erkennen kann, bleiben die Mittelwerte annähernd gleich. Der einzige Unterschied liegt in der Größe der Cluster. In der 10%igen Stichprobe liegt die Anzahl der Kunden bei 10.257. Je höher die Stichprobe wird, umso größer wird die Anzahl der Kunden in dem Cluster. In der Größe in % gibt es jedoch keine Unterschiede.

Die Ergebnisse der Stichproben der anderen drei Clustern verhalten sich entsprechend jenen im ersten Cluster. Von der Interpretation her hat sich zu den vorherigen Stabilitätsprüfungen nichts geändert.

Bei der Anwendung der K-Means Clusteranalyse wurde das selbe Verfahren wie bei Two-Step Clusteranalyse angewandt. Zusammengefasst kann man sagen, dass die einzelnen Ergebnisse ebenfalls sehr homogen sind. Anhand dieses Beispiels kann man erkennen, dass die Stichprobenerhöhung in Cluster 1 keine Veränderungen der Resultate verursacht.

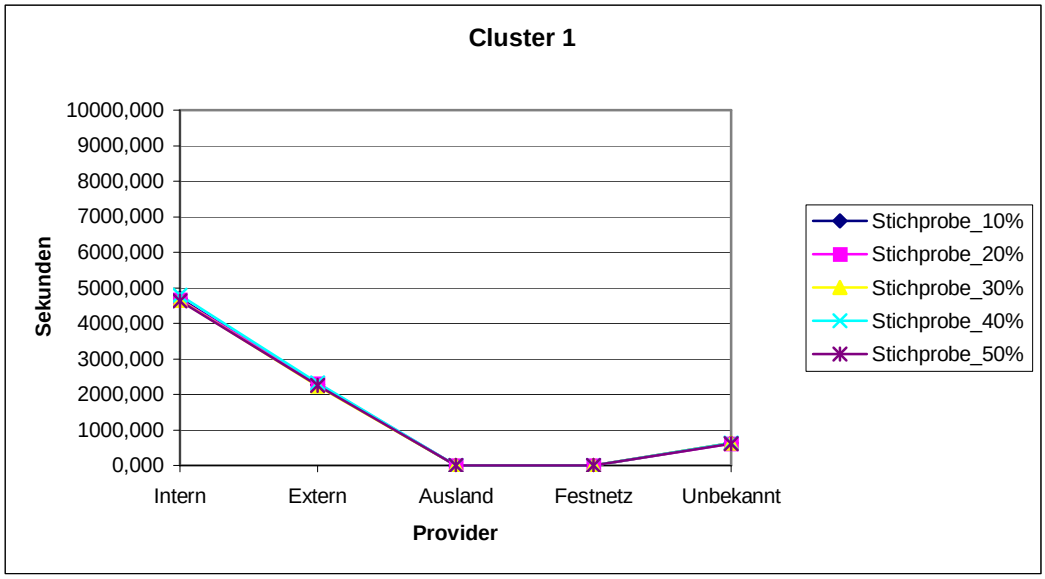

Abbildung 4.12: Vergleich der Ergebnisse der Stichproben in Cluster 1

4.2.4 GESAMTER DATENSATZ

Die letzte Prüfung zur Stabilität erfolgt dadurch, dass der gesamte Datensatz vier Mal permutiert wird. Es wird wieder hinsichtlich der gesamten Nutzung der einzelnen Provider versucht mittels der Two-Step Clusteranalyse und dem K-Means Algorithmus Gruppen zu finden.

Im ersten Schritt wird die Two-Step Clusteranalyse mit automatischer Wahl der Anzahl der Cluster durchgeführt. Das Ergebnis sieht folgendermaßen aus:

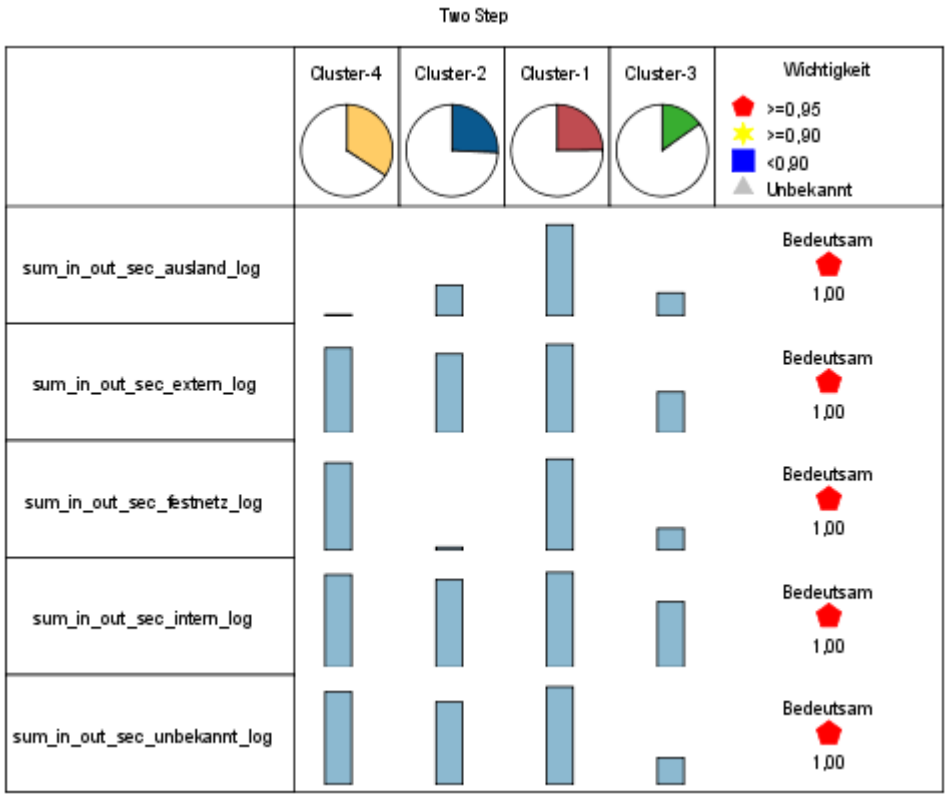

Abbildung 4.13: Two-Step Clusteranalyse des gesamten Datensatzes

Two Step

	Cluster-4 Anzahl: 231398 (34%)	Cluster-2 Anzahl: 174582 (25%)	Cluster-1 Anzahl: 168476 (24%)	Cluster-3 Anzahl: 102717 (15%)	Wichtigkeit >=0,95 >=0,90 <0,90 Unbekannt
sum_in_out_sec_ausland_log	0,08 (0,44)	2,09 (3,00)	6,25 (1,57)	1,57 (2,77)	Bedeutsam 1,00
sum_in_out_sec_extern_log	8,33 (1,35)	7,75 (1,49)	8,89 (1,29)	4,01 (3,32)	Bedeutsam 1,00
sum_in_out_sec_festnetz_log	5,99 (1,53)	0,21 (0,76)	6,24 (1,56)	1,50 (2,42)	Bedeutsam 1,00
sum_in_out_sec_intern_log	9,08 (1,20)	8,57 (1,37)	9,29 (1,14)	6,42 (3,05)	Bedeutsam 1,00
sum_in_out_sec_unbekannt_log	7,26 (1,38)	6,44 (1,49)	7,85 (1,36)	2,10 (2,78)	Bedeutsam 1,00

Abbildung 4.14: Two-Step Clusteranalyse des gesamten Datensatzes mit Angabe des Mittelwerts und der Standardabweichung

Wenn man die zwei Graphiken betrachtet, erkannt man, dass dieses Ergebnis, welches den gesamten Datensatz beinhaltet, dem Ergebnis der 10%igen Stichprobe von der Interpretation her sehr ähnlich ist.
In den folgenden Abbildungen werden nun die Analysen der vier Permutationen gegenübergestellt.

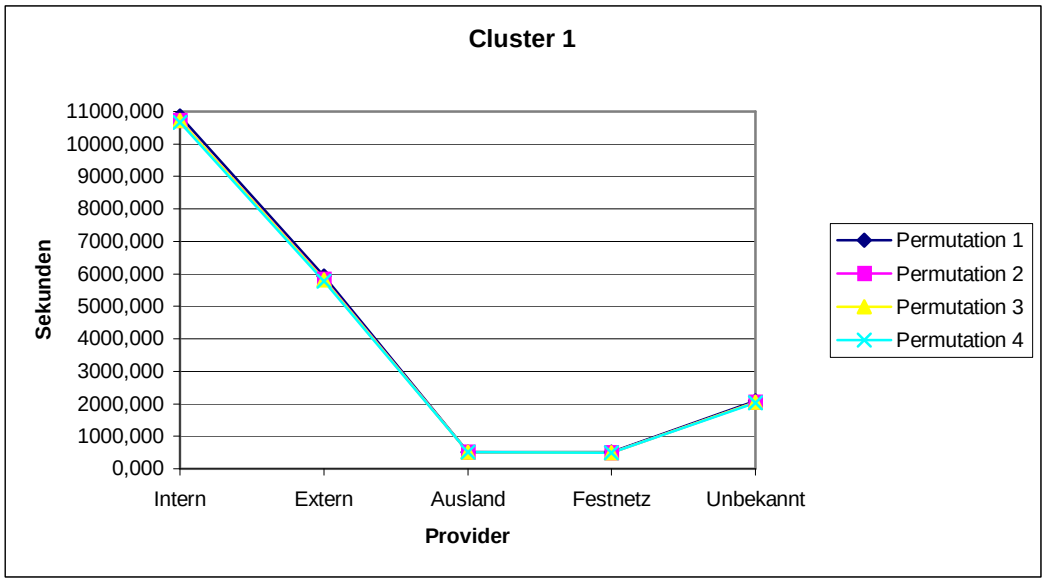

Abbildung 4.15: Vergleich der Ergebnisse des gesamten Datensatzes in Cluster 1

Wie man erkennen kann, gibt es kaum einen Unterschied zwischen den Resultaten der vier Permutationen. Cluster 1 enthält jene Kunden, die die Provider Intern und Extern am meisten nutzen. Die Provider Ausland und Festnetz wurden weniger genutzt, aber im Vergleich zu den folgenden Cluster ist die Nutzung jedoch groß. Dieser Cluster entspricht dem „Viel-Telefonierer“.

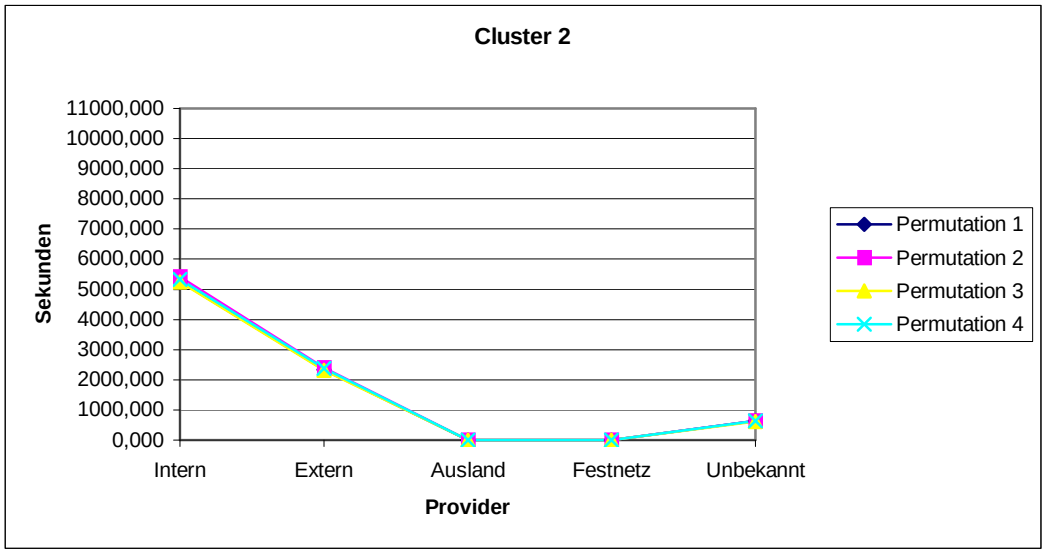

Abbildung 4.16: Vergleich der Ergebnisse des gesamten Datensatzes in Cluster 2

Cluster 2 enthält jene Kunden, die den Provider Intern und Extern mittelmäßig benutzten. Diese Kunden verwenden so gut wie gar nicht die Provider Ausland und Festnetz. Die Nutzung Unbekannt liegt zirka bei einem Mittelwert 630 Sekunden.

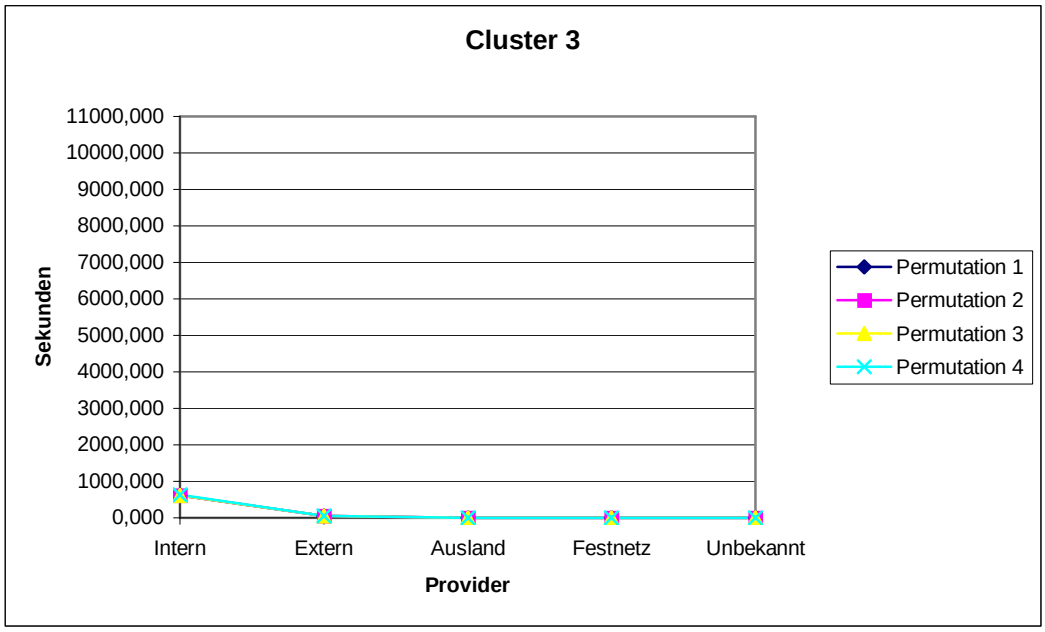

Abbildung 4.17: Vergleich der Ergebnisse des gesamten Datensatzes in Cluster 3

Cluster 3 entspricht eindeutig dem „Wenig-Telefonierer“. Die Nutzung der fünf Provider liegt im Mittel konstant unter 1000 Sekunden.

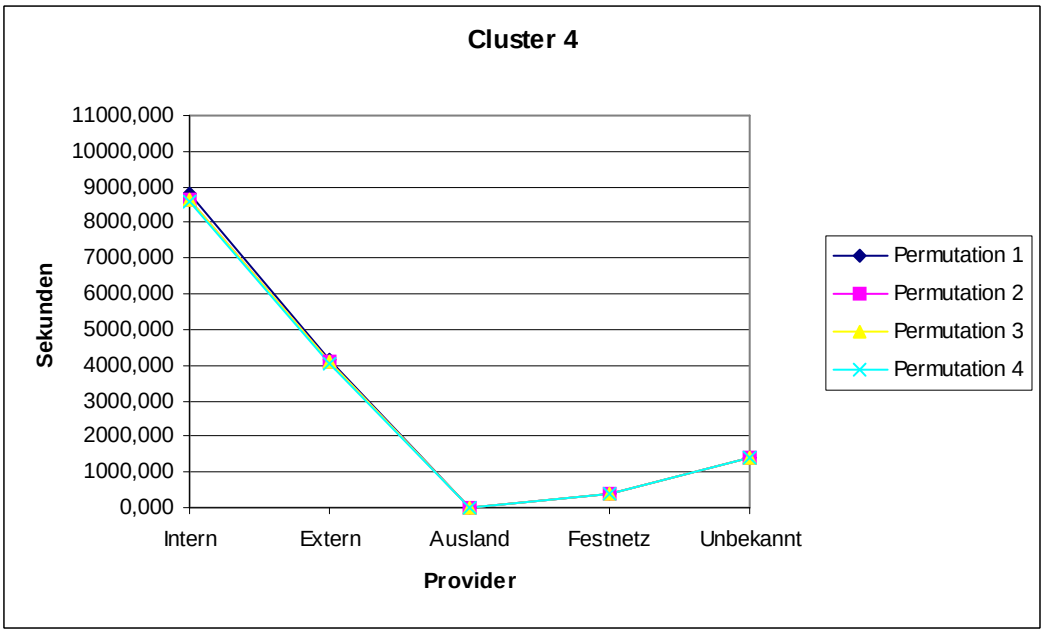

Abbildung 4.18: Vergleich der Ergebnisse des gesamten Datensatzes in Cluster 4

Der vierte Cluster beinhaltet die Kunden, die den Provider Intern stark nutzen. Der Mittelwert liegt bei allen Permutationen bei 8700 Sekunden. Der Mittelwert des Provider Extern liegt bei der Hälfte des Mittelwertes des Providers Intern. Der Provider Ausland wird gar nicht genutzt. Der Mittelwert beträgt 0,09 Sekunden. Das Festnetz wird im Mittel 400 Sekunden lang genutzt. Der Provider Unbekannt wird im Vergleich zu dem Provider Festnetz stärker genutzt.

Zusammengefasst betrachtet sind die Ergebnisse der Permutationen des gesamten Datensatzes sehr homogen. Von der Interpretation her, kann man sagen, dass sie zu den vorhergehenden Analysen sehr ähnlich sind.

Diese Prozedur wird mit dem K-Means Verfahren ebenfalls durchgeführt. Man kommt zu denselben Resultaten wie bei der Two-Step Clusteranalyse.

Nach dem Vergleich von Permutationen und verschiedenen Stichproben, stellt sich heraus, dass sowohl die Two-Step Clusteranalyse als auch der K-Means Algorithmus sehr stabile Ergebnisse liefern. Nachdem dies überprüft wurde, werden nun nächste Fragestellungen bezüglich der Kundensegmentierung analysiert.

4.3 Zusammenhang zwischen Sekunden und Degree

Der Degree entspricht der Anzahl der Personen zu denen man telefonischen Kontakt hat. Man unterscheidet zwischen dem Indegree (Anzahl der Personen von denen man kontaktiert wird) und dem Outdegree (Anzahl der Personen die man kontaktiert). Wie sich der Degree berechnet ist nachzulesen in Kapitel 2.3. Die Erhebung der Variablen des Degree ist für einen Provider in jedem Fall aufwendiger, als die Erhebung der Telefonierdauer gemessen in Sekunden. Daher stellt sich nun die Frage, ob es reicht nur die Variablen der Sekunden zu analysieren, oder ob eine Analy-

se der Variablen des Degree ebenfalls sinnvoll ist, da sie gegebenenfalls andere Ergebnisse liefert. Um nun dies zu untersuchen, wird im ersten Schritt eine Clusteranalyse über die Telefonierdauer gemessen in Sekunden durchgeführt und im zweiten Schritt werden dann die Variablen des Degree innerhalb der einzelnen Cluster analysiert.

Es wird mit einer Two-Step Clusteranalyse über die Telefonierdauer gemessen in Sekunden mit automatischer Wahl der Anzahl der Cluster begonnen. In diesem Fall wurden nur zwei Cluster gefunden, die eine Interpretation des viel Telefonieren und des wenig Telefonieren haben. Da dies keine gute Segmentierung ist, wird die Anzahl der Cluster selbst gewählt, nämlich sieben Cluster. Der Grund warum gerade sieben Cluster gewählt werden, liegt darin, dass zuvor eine Clusteranalyse mit automatischer Wahl der Anzahl der Cluster nur mit den Variablen der eingehenden Sekunden durchgeführt wurde. Als Ergebnis erhielt man eine gute Segmentierung durch sieben Cluster.

Um zu überprüfen, ob die Wahl der Anzahl der Cluster sinnvoll ist, wird eine weitere Analyse mit acht Cluster durchgeführt. In der folgenden Kreuztabelle kann man nachlesen wie sich die Cluster zusammensetzen.

$T1-TwoStep_in_out_sec_log	Cluster-1	Cluster-2	Cluster-3	Cluster-4	Cluster-5	Cluster-6	Cluster-7
Cluster-1	138230	505	0	0	0	0	0
Cluster-2	1006	65017	501	0	2080	3993	391
Cluster-3	0	401	68281	0	0	0	0
Cluster-4	0	157	0	103949	0	0	0
Cluster-5	0	178	0	0	98929	0	0
Cluster-6	20	61536	777	3419	0	0	0
Cluster-7	0	44	0	0	0	63819	0
Cluster-8	0	44	0	0	0	0	63896

Tabelle 4.6: Kreuztabelle zweier Two-Step Clusteranalysen

In der Analyse mit acht Clustern ist der Cluster 2 jener Cluster, der in der Analyse mit sieben Clustern nicht vorkommt. Die anderen Cluster unterscheiden sich in beiden Analysen kaum von einander. Da der Cluster 2 der Analyse mit acht Clustern zum Großteil in Cluster 2 der Analyse mit sieben Clustern steckt, bringt eine Clusteranalyse mit acht Clustern keine bessere Segmentierung.

Um einen weitern Vergleich zu der Analyse mit sieben Clustern durchzuführen, wurde eine Clusteranalyse mit sechs Cluster berechnet. Das Resultat sieht folgendermaßen aus:

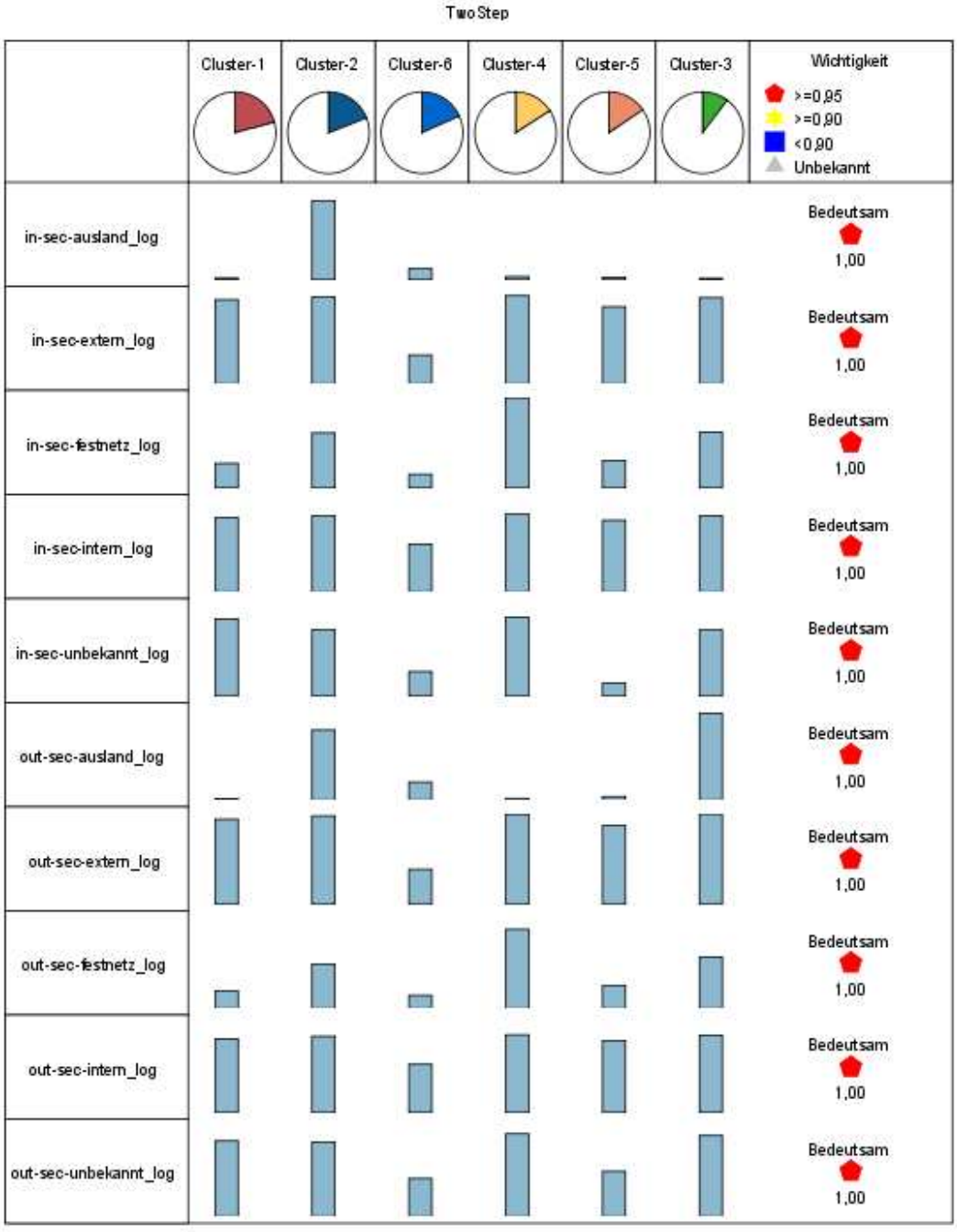

Abbildung 4.19: Two-Step Clusteranalyse mit den Variablen der Sekunden

Two Step

	Cluster-1 Anzahl: 141265 (20%)	Cluster-2 Anzahl: 128986 (19%)	Cluster-6 Anzahl: 123920 (18%)	Cluster-4 Anzahl: 108112 (15%)	Cluster-5 Anzahl: 104972 (15%)	Cluster-3 Anzahl: 69918 (10%)	Wichtigkeit >=0,95 >=0,90 <0,90 Unbekannt
in-sec-ausland_log	0,12 (0,63)	6,08 (1,36)	0,85 (2,03)	0,25 (0,92)	0,18 (0,83)	0,10 (0,48)	Bedeutsam 1,00
in-sec-extern_log	7,40 (1,37)	7,64 (1,52)	2,54 (2,91)	7,77 (1,40)	6,79 (1,49)	7,60 (1,50)	Bedeutsam 1,00
in-sec-festnetz_log	1,62 (2,38)	3,83 (2,93)	0,90 (1,95)	5,94 (1,44)	1,79 (2,49)	3,88 (2,91)	Bedeutsam 1,00
in-sec-intern_log	8,19 (1,28)	8,40 (1,28)	5,27 (3,06)	8,57 (1,18)	7,88 (1,32)	8,41 (1,28)	Bedeutsam 1,00
in-sec-unbekannt_log	5,95 (1,30)	5,12 (2,64)	1,88 (2,60)	6,07 (1,73)	1,00 (1,89)	5,09 (2,59)	Bedeutsam 1,00
out-sec-ausland_log	0,09 (0,50)	4,63 (3,13)	1,19 (2,42)	0,12 (0,58)	0,23 (0,98)	5,74 (1,30)	Bedeutsam 1,00
out-sec-extern_log	7,50 (1,52)	7,77 (1,58)	3,08 (2,97)	7,93 (1,51)	6,92 (1,63)	7,95 (1,48)	Bedeutsam 1,00
out-sec-festnetz_log	1,33 (2,25)	3,39 (3,06)	0,99 (2,11)	6,10 (1,57)	1,75 (2,55)	3,96 (3,03)	Bedeutsam 1,00
out-sec-intern_log	8,12 (1,36)	8,38 (1,37)	5,29 (3,13)	8,55 (1,26)	7,90 (1,35)	8,51 (1,27)	Bedeutsam 1,00
out-sec-unbekannt_log	6,67 (1,41)	6,57 (2,22)	3,36 (3,05)	7,30 (1,40)	3,99 (2,84)	7,14 (1,67)	Bedeutsam 1,00

Abbildung 4.20: Two-Step Clusteranalyse mit den Variablen der Sekunden mit Angabe des Mittelwerts und der Standardabweichung

Trotzdem es jetzt einen Cluster weniger gibt, kann man die Cluster gut voneinander unterscheiden und interpretieren. Nachdem wieder eine Kreuztabelle erzeugt wurde, wurde ersichtlich, dass der 7 Cluster nicht notwendig ist, da er hauptsächlich dem Cluster 6 entsprach. Daher wird nun mit diesem Ergebnis überprüft, ob es nun einen Zusammenhang zwischen dem Degree und den Sekunden vorhanden ist. Dafür werden nun Boxplots der Variablen des Degree gruppiert nach den Clustern untersucht.

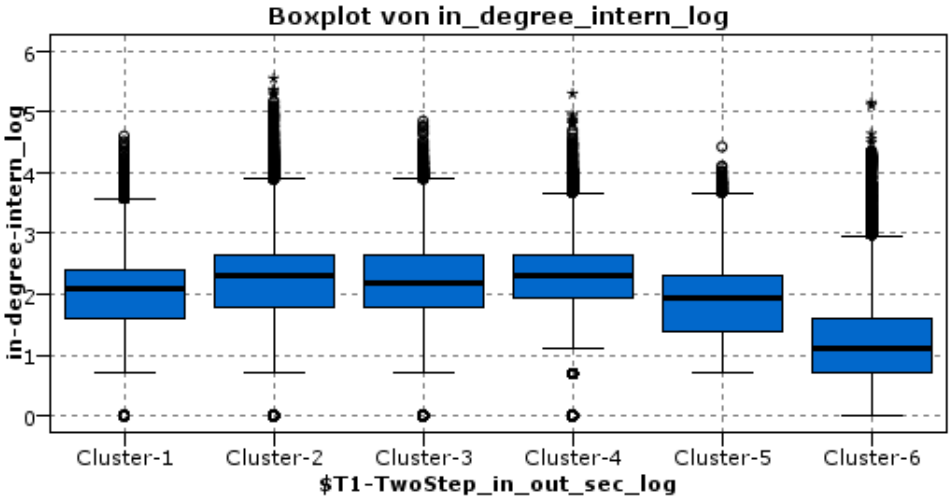

Abbildung 4.21: Boxplot von in_degree_intern_log gruppiert nach Cluster

Wenn man nun die Mittelwerte der Cluster der Two-Step Clusteranalyse der Variablen in_sec_intern_log mit dem Boxplot vergleicht, kann man erkennen, dass sich die Mittelwerte der Variable in_degree_intern_log sehr ähnlich verhalten. Zum Beispiel liegt der Mittelwert der Variable in_sec_intern_log in Cluster 6 bei 5,27 weit unter den anderen. Der Mittelwert der Variable in_degree_intern_log in Cluster 6 verhält sich genauso.

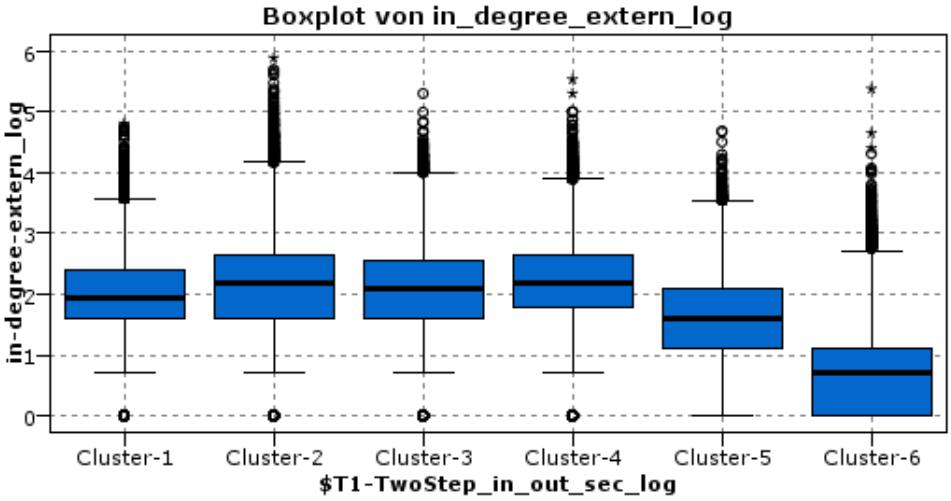

Abbildung 4.22: Boxplot von in_degree_extern_log gruppiert nach Cluster

Auch im Fall der Variable in_degree_extern_log ergibt sich eine Ähnlichkeit in den Mittelwerten zum Vergleich zu den Mittelwerten der geclusterten Variable in_sec_extern_log.

Um einen Überblick über alle Degree zu bekommen, werden die Mittelwerte gruppiert nach den Clustern in einer Tabelle zusammengefasst.

Feld	Cluster-1*	Cluster-2*	Cluster-3*	Cluster-4*	Cluster-5*	Cluster-6*	Wichtigkeit
in-degree-ausland_log	0.540	1.504	0.713	0.600	0.487	0.481	1.000 Bedeutsam
in-degree-extern_log	1.977	2.140	2.093	2.204	1.668	0.641	1.000 Bedeutsam
in-degree-festnetz_log	0.304	0.710	0.697	1.086	0.313	0.158	1.000 Bedeutsam
in-degree-intern_log	2.050	2.248	2.204	2.281	1.879	1.203	1.000 Bedeutsam
in-degree-unbekannt_log	1.432	1.445	1.410	1.605	0.667	0.676	1.000 Bedeutsam
out-degree-ausland_log	0.168	1.226	1.058	0.185	0.221	0.382	1.000 Bedeutsam
out-degree-extern_log	2.076	2.279	2.305	2.325	1.800	0.777	1.000 Bedeutsam
out-degree-festnetz_log	0.245	0.628	0.735	1.065	0.298	0.169	1.000 Bedeutsam
out-degree-intern_log	2.038	2.241	2.249	2.280	1.902	1.184	1.000 Bedeutsam
out-degree-unbekannt_...	1.713	1.854	2.085	2.134	0.951	0.750	1.000 Bedeutsam

Tabelle 4.7: Mittelwertstabelle der Variablen des Degree gruppiert nach Clustern einer Two-Step Clusteranalyse

Um einen möglichen Zusammenhang zwischen den Sekunden und dem Degree zu veranschaulichen, wird die Kundenzuordnung der einzelnen Cluster einer Two-Step Clusteranalyse der Sekunden mit einer der Two-Step Clusteranalyse des Degree anhand einer Kreuztabelle dargestellt.

		Two-Step Clusteranalyse mit Sekunden					
		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Two-Step Clusteranalyse mit Degree	Cluster 1	1.004	276	244	52	8.775	103.105
	Cluster 2	16.505	7.823	9.042	10.388	74.310	6.272
	Cluster 3	2.595	90.894	6.683	213	7.375	12.590
	Cluster 4	3.658	7.762	8.017	85.118	407	13
	Cluster 5	111.015	11.356	9.920	9.888	10.598	1.522
	Cluster 6	6.488	10.785	36.012	2.453	3.507	418

Tabelle 4.8: Kreuztabelle der Two-Step Clusteranalyse mit Sekunden und der Two-Step Clusteranalyse mit Degree

Es werden 73,9% der Beobachtungen in den beiden Verfahren identisch klassifiziert. Dies bedeutet, dass eine Clusteranalyse mit den Variablen des Degree kleine Unterschiede im Bezug auf die Clusteranalyse mit der Telefonierdauer gemessen in Sekunden hat. Damit können zusätzliche Erkenntnisse durch die Netzwerkanalyse gewonnen werden.

Im Folgenden wird nun der Zusammenhang der Sekunden und des Degree mittels der K-Means Clusteranalyse untersucht. Die Anzahl der Cluster wurde auf sechs gesetzt, da man schon durch die Durchführung der Two-Step Clusteranalyse diese Annahme treffen kann.
Dies führt zu folgendem Ergebnis:

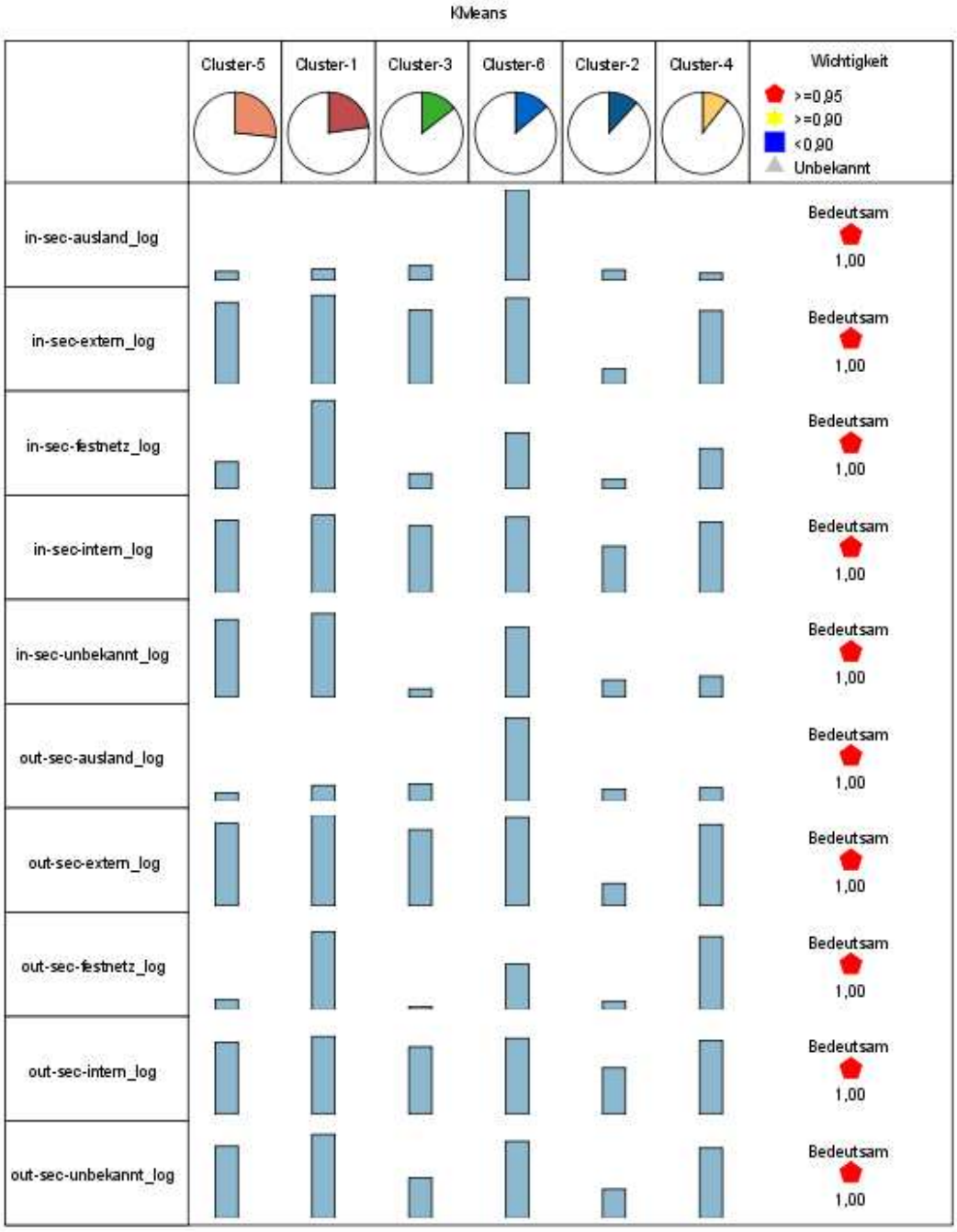

Abbildung 4.23: K-Means Clusteranalyse mit den Variablen der Sekunden

KMeans

	Cluster-5 Anzahl: 180098 (26%)	Cluster-1 Anzahl: 153623 (22%)	Cluster-3 Anzahl: 98858 (14%)	Cluster-6 Anzahl: 93700 (13%)	Cluster-2 Anzahl: 79943 (11%)	Cluster-4 Anzahl: 70951 (10%)	Wichtigkeit >=0,95 >=0,90 <0,90 Unbekannt
in-sec-ausland_log	0,57 (1,62)	0,75 (1,84)	0,96 (2,09)	5,97 (2,01)	0,67 (1,85)	0,47 (1,47)	Bedeutsam 1,00
in-sec-extern_log	7,21 (1,66)	7,84 (1,43)	6,56 (1,69)	7,58 (1,72)	1,35 (2,21)	6,49 (2,21)	Bedeutsam 1,00
in-sec-festnetz_log	1,76 (2,43)	5,82 (1,72)	1,00 (1,98)	3,70 (2,88)	0,83 (1,85)	2,65 (2,76)	Bedeutsam 1,00
in-sec-intern_log	8,02 (1,61)	8,56 (1,30)	7,37 (1,97)	8,33 (1,49)	5,15 (3,08)	7,83 (1,82)	Bedeutsam 1,00
in-sec-unbekannt_log	5,98 (1,29)	6,43 (1,29)	0,84 (1,45)	5,38 (2,40)	1,32 (2,23)	1,61 (2,15)	Bedeutsam 1,00
out-sec-ausland_log	0,67 (1,74)	1,24 (2,31)	1,35 (2,50)	6,43 (1,85)	0,94 (2,20)	1,04 (2,16)	Bedeutsam 1,00
out-sec-extern_log	7,23 (1,81)	7,99 (1,56)	6,69 (1,79)	7,79 (1,71)	1,95 (2,48)	7,13 (1,98)	Bedeutsam 1,00
out-sec-festnetz_log	0,79 (1,67)	6,00 (1,70)	0,24 (0,93)	3,52 (2,99)	0,85 (1,88)	5,65 (1,50)	Bedeutsam 1,00
out-sec-intern_log	7,91 (1,71)	8,53 (1,38)	7,41 (1,93)	8,34 (1,55)	5,10 (3,17)	8,10 (1,66)	Bedeutsam 1,00
out-sec-unbekannt_log	6,35 (1,83)	7,40 (1,51)	3,57 (2,85)	6,79 (2,01)	2,56 (2,84)	6,19 (2,05)	Bedeutsam 1,00

Abbildung 4.24: K-Means Clusteranalyse mit den Variablen der Sekunden mit Angabe des Mittelwerts und der Standardabweichung

Weiters werden die Mittelwerte der Variablen des Degree gruppiert nach den erzeugten Clustern ermittelt.

Feld	Cluster-1*	Cluster-2*	Cluster-3*	Cluster-4*	Cluster-5*	Cluster-6*	Wichtigkeit
in-degree-ausland_log	0.688	0.417	0.609	0.579	0.603	1.568	1.000 Bedeutsam
in-degree-extern_log	2.246	0.395	1.561	1.659	1.918	2.139	1.000 Bedeutsam
in-degree-festnetz_log	1.084	0.110	0.180	0.459	0.330	0.725	1.000 Bedeutsam
in-degree-intern_log	2.299	1.166	1.737	1.921	2.006	2.250	1.000 Bedeutsam
in-degree-unbekannt_log	1.681	0.580	0.592	0.784	1.426	1.485	1.000 Bedeutsam
out-degree-ausland_log	0.366	0.335	0.452	0.362	0.268	1.546	1.000 Bedeutsam
out-degree-extern_log	2.370	0.516	1.699	1.921	1.994	2.306	1.000 Bedeutsam
out-degree-festnetz_log	1.073	0.116	0.056	0.914	0.156	0.659	1.000 Bedeutsam
out-degree-intern_log	2.297	1.126	1.760	2.018	1.977	2.250	1.000 Bedeutsam
out-degree-unbekannt_...	2.208	0.546	0.809	1.636	1.612	1.934	1.000 Bedeutsam

Tabelle 4.9: Mittelwertstabelle der Variablen des Degree gruppiert nach Clustern einer K-Means Clusteranalyse

Auch in diesem Fall kann man davon ausgehen, dass eine Segmentierung durch die Sekunden im Vergleich zu einer Segmentierung durch den Degree zu keinen großen Unterschieden führt.

4.4 ENDERGEBNIS

Nachdem die Stabilität geprüft wurde und festgestellt wurde, ob es einen Zusammenhang zwischen den Sekunden und dem Degree gibt, wird zum Abschluss eine vollständige Segmentierung der Daten durchgeführt. Die Analyse erfolgt mit der Two-Step Clusteranalyse und dem K-Means Verfahren. Die Ergebnisse werden miteinander verglichen. Es werden die Variablen Sekunden und SMS für die Segmentierung verwendet.

4.4.1 ERGEBNIS DER TWO-STEP CLUSTERANALYSE

Im ersten Schritt wird eine Vorsegmentierung anhand der Variablen sum_sec_gesamt_log (entspricht der Summe aller Variablen der Sekunden) und sum_sms_gesamt_log (entspricht der Summe aller Variablen der SMS) vorgenommen. Die Anzahl der Cluster wird automatisch gewählt. Das Ergebnis sieht folgendermaßen aus:

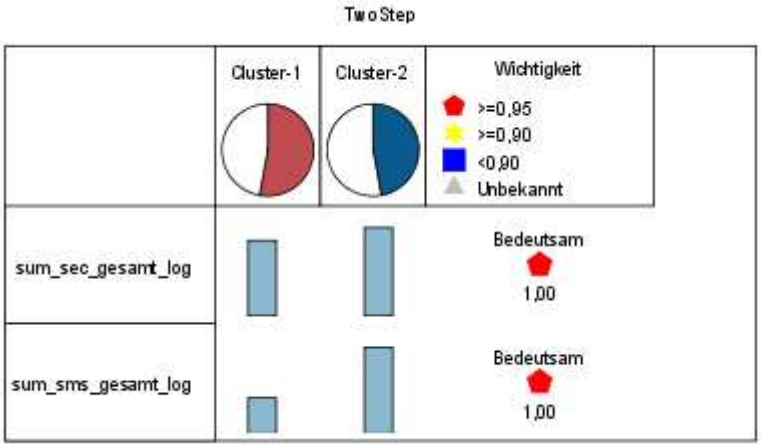

Abbildung 4.25: Two-Step Clusteranalyse

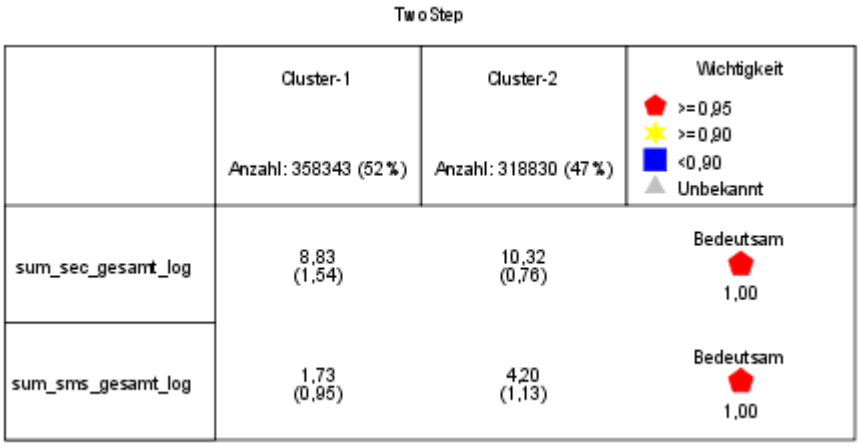

Abbildung 4.26: Two-Step Clusteranalyse mit Angabe des Mittelwerts und der Standardabweichung

Man erhält eine Vorgruppierung von zwei Clustern. Cluster 1 enthält 358.343 (52%) Personen, die gesamt im Mittel 6.835,29 Sekunden telefonieren und im Mittel 4,46 SMS schreiben. Cluster 2 enthält 318.830 (47%)

Personen, die gesamt im Mittel 30.332,26 Sekunden telefonieren und im Mittel 65,69 SMS schreiben. Der Interpretation zur Folge entspricht Cluster 1 dem „Wenig-Benutzer“ und Cluster 2 dem „Viel-Benutzer.

Um nun das Telefonierverhalten des „Wenig-Benutzer“ und des „Viel-Benutzer“ näher zu bestimmen, werden die einzelnen Cluster separat noch weiter analysiert.
Es wird im nächsten Schritt eine weitere Two-Step Clusteranalyse über die 358.343 „Wenig-Benutzer“, die zu Cluster 1 gruppiert wurden, durchgeführt. Die Anzahl der Cluster wird wieder automatisch gewählt. Es wird anhand der sowohl eingehenden als auch ausgehenden Sekunden und SMS pro Provider segmentiert. Dies entspricht insgesamt 20 Variablen. Als Ergebnis erhält man folgende Mittelwerte der fünf Cluster, wobei die Mittelwerte zurücktransformiert wurden:

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
in_sec_intern	1.684,81	2.564,7,	1.325,10	827,82	143,03
in_sec_extern	538,15	896,85	414,72	289,03	15,12
in_sec_ausland	45,99	0,21	0,36	2,03	0,38
in_sec_festnetz	13,15	26,66	6,46	15,28	0,80
in_sec_unbekannt	51,46	151,93	32,12	64,37	2,49
out_sec_intern	1.651,43	2.320,57	1.163,45	698,24	153,47
out_sec_extern	670,82	991,27	477,19	335,97	25,58
out_sec_ausland	99,48	0,34	0,54	3,01	0,84
out_sec_festnetz	11,94	25,05	4,75	13,73	0,99
out_sec_unbekannt	269,43	741,48	137,38	224,89	16,64
in_sms_intern	0,55	0,28	2,00	0,97	0,15
in_sms_extern	0,45	0,27	2,13	1,01	0,12
in_sms_ausland	1,59	0,90	0,75	1,32	0,34
in_sms_festnetz	0,00	0,00	0,00	0,27	0,00

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
in_sms_unbekannt	0,97	1,08	0,99	1,48	0,62
out_sms_intern	0,34	0,12	1,66	0,97	0,11
out_sms_extern	0,23	0,07	1,44	0,89	0,07
out_sms_ausland	0,68	0,04	0,15	0,51	0,14
out_sms_festnetz	0,00	0,00	0,00	0,27	0,00
out_sms_unbekannt	0,00	0,00	0,00	0,65	0,00

Tabelle 4.10: Two-Step Clusteranalyse zur Segmentierung der „Wenig-Benutzer“

Dem Cluster 1 werden 79.488 Kunden zugeordnet. Im Unterschied zu den anderen Clustern enthält Cluster 1 jene Kunden, die im Verhältnis zu den „Wenig-Benutzern“ noch recht viel telefonieren. Der größte Unterschied liegt vor allem in der Benutzung des Providers Ausland. Sowohl das Telefonieren als auch das Schreiben von SMS ist in diesem Cluster am größten.

Cluster 2 enthält 110.962 Kunden. Cluster 2 unterscheidet sich hinsichtlich der anderen Cluster in der Benutzung der Provider Intern, Extern, Festnetz und Unbekannt. Diese Provider werden von den Kunden im Vergleich zu den anderen Clustern am meisten zum Telefonieren genutzt. Der Provider Ausland wird von den Kunden dieses Clusters am wenigsten genutzt. Das Schreiben von SMS wird von den Kunden dieses Cluster wenig genutzt.

Dem Cluster 3 wurden 73.756 Kunden zugeordnet. Die Kunden in Cluster 3 telefonieren hinsichtlich der Provider Intern und Extern noch weniger als die Kunden in Cluster 2. Im Unterschied zu den anderen Clustern ist hier das Schreiben von SMS unter der Benutzung der Provider Intern und Extern am größten.

Cluster 4 umfasst 7006 Kunden. Dies entspricht bei der gesamten Datengröße von 677.173 Kunden einem Prozentsatz von 1%. Daher sind die Kunden, die Cluster 4 zugeordnet wurden als Ausreißer zu betrachten.

Die Größe des Cluster 5 entspricht 87.131 Kunden. Unter den „Wenig-Benutzern“ enthält Cluster 5 jene Kunden, die am wenigsten telefonieren und SMS schreiben.

Im nächsten Schritt werden nun die „Viel-Benutzer“ ebenfalls anhand einer Two-Step Clusteranalyse mit den gleichen Variablen segmentiert. Die Anzahl der Cluster wird wiederum automatisch gewählt. Als Resultat erhält man folgende Mittelwerte der Variablen in den einzelnen Clustern, wobei die Ergebnisse wieder zurücktransformiert wurden:

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
in_sec_intern	3.903,95	3.676,54	6.835,29	7.630,20
in_sec_extern	3.497,19	657,52	3.603,72	3.903,95
in_sec_ausland	14,80	6,24	0,99	7,94
in_sec_festnetz	106,77	4,37	17,92	169,71
in_sec_unbekannt	326,01	28,37	102,54	588,93
out_sec_intern	6.903,99	3.865,09	7.114,28	7.784,36
out_sec_extern	4.358,01	795,32	4.581,50	4.864,87
out_sec_ausland	23,53	8,97	1,72	14,33
out_sec_festnetz	110,05	3,85	15,61	182,09
out_sec_unbekannt	1.175,15	191,48	486,85	2.343,90
in_sms_intern	20,54	5,89	29,57	3,48
in_sms_extern	25,31	3,57	39,85	4,00
In_sms_ausland	7,17	3,44	2,39	4,10
In_sms_festnetz	0,64	0,00	0,00	0,01
In_sms_unbekannt	6,24	2,03	2,86	2,56
out_sms_intern	24,03	5,17	31,79	2,74
out_sms_extern	25,31	2,78	37,47	2,39
out_sms_ausland	4,31	1,64	0,67	0,58

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
out_sms_festnetz	0,72	0,00	0,01	0,01
out_sms_unbekannt	1,77	0,03	0,05	0,03

Tabelle 4.11: Two-Step Clusteranalyse zur Segmentierung der „Viel-Benutzer“

Wenn man nun dieses Segmentierung der „Viel-Benutzer“ und die Segmentierung der „Wenig-Benutzer“ vergleicht, kann man erkennen, dass die Anzahl der Sekunden und SMS der „Viel-Benutzer“ deutlich höher sind als bei den „Wenig-Benutzern“. Trotzdem weisen die vier Cluster eine ähnliche Benutzung der Provider in verstärkter Form auf.
Cluster 1 enthält 22.145 Kunden. Kunden, die diesem Cluster zugeordnet wurden, weisen eine erhöhte Benutzung des Providers Ausland, sowohl zum Telefonieren als auch zum SMS Schreiben, auf. Dieser Cluster ist vergleichbar mit dem Cluster 1 aus der Gruppe der „Wenig-Benutzer“.
Dem Cluster 2 wurden 101.193 Kunden zugeordnet. Die Kunden in diesem Cluster weisen unter den „Viel-Benutzern“ eine geringe Nutzung des Telefonierens uns SMS Schreibens auf.
Cluster 3 besteht aus 78.141 Kunden. Diese Kunden verwenden die Provider Intern und Extern hinsichtlich des Telefonieren und SMS Schreiben sehr viel. Von den anderen Providern wird wenig Gebrauch gemacht.
Cluster 4 besitzt eine Größe von 117.351 Kunden. Dieser Cluster entspricht dem „Viel-Telefonierer“. Lediglich der Provider Ausland wird nur selten zum Telefonieren verwendet. Kunden in diesem Cluster machen jedoch keinen Gebrauch vom Schreiben von SMS. Dieser Cluster ist vergleich mit jenem Cluster 2 der „Wenig-Benutzer“.

Zusammengefasst kann man sagen, dass man die 677.173 Kunden zu acht Clustern segmentieren kann, die unterschiedliche Verhaltensweisen bezüglich des Telefonierens und des Schreiben von SMS haben. Sowohl bei den „Wenig-Benutzern“ als auch bei den „Viel-Benutzern“ kristallisieren sich ähnliche Gruppierungen heraus.

4.4.2 ERGEBNIS DES K-MEANS ALGORITHMUS

Das zuvor durchgeführte Szenario wird nun mit der K-Means Clusteranalyse ebenfalls untersucht. Im ersten Schritt wird wieder eine Analyse mit der vorgegeben Anzahl von zwei Clustern über die Variablen sum_sec_gesamt_log und sum_sms_gesamt_log durchgeführt. Es ergeben sich wieder zwei Cluster mit der Interpretation des „Wenig-Benutzer“ (Größe von 360.327 Kunden) und des „Viel-Benutzer (Größte von 316.846 Kunden). Nun werden diese Cluster einzeln genauer betrachtet. Zu Beginn wird der Cluster des „Wenig-Benutzer“ untersucht. Die Anzahl der Cluster wurde aufgrund von vorheriger Ergebnisse mit der Two-Step Clusteranalyse auf fünf gesetzt. Dies führt aber zu keiner guten Segmentierung. Aufgrund dessen wird die Anzahl der Cluster auf vier gesetzt. Das zurücktransformierte Ergebnis sieht folgendermaßen aus:

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
in_sec_intern	2.950,30	143,03	1.479,30	1.789,05
in_sec_extern	1.247,88	4,64	619,17	726,78
in_sec_ausland	1,12	0,67	0,58	67,72
in_sec_festnetz	169,72	0,86	1,66	10,59
in_sec_unbekannt	300,87	1,51	32,12	58,15
out_sec_intern	2.750,77	141,59	1.421,26	1.862,11
out_sec_extern	1.393,09	9,28	691,29	952,37
out_sec_ausland	1,23	1,23	0,34	496,70
out_sec_festnetz	168,02	1,05	1,48	9,80
out_sec_unbekannt	1.247,88	8,87	201,35	342,78
in_sms_intern	0,60	0,35	0,57	0,52
in_sms_extern	0,63	0,21	0,58	0,48
in_sms_ausland	1,08	0,45	0,72	1,36
in_sms_festnetz	0,01	0,00	0,00	0,00

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
in_sms_unbekannt	1,12	0,67	0,93	0,97
out_sms_intern	0,38	0,28	0,38	0,34
out_sms_extern	0,28	0,17	0,31	0,25
out_sms_ausland	0,11	0,23	0,14	0,40
out_sms_festnetz	0,01	0,00	0,00	0,00
out_sms_unbekannt	0,01	0,01	0,01	0,01

Tabelle 4.12: K-Means Clusteranalyse zur Segmentierung der „Wenig-Benutzer“

Cluster 1 hat eine Größe von 103.615 Kunden. Die Kunden, die diesem Cluster zugeordnet wurden, weisen eine hohe Benutzung der Provider Intern, Extern, Festnetz und Unbekannt. Unter den „Wenig-Benutzern“ gilt dieser Cluster trotzdem dem „Viel-Benutzer“.
Cluster 2 enthält 76.868 Kunden, die die alle Provider nur sehr wenig verwenden. Diese Kunden entsprechen dem „Wenig-Benutzern“ und den „Wenig-Benutzern“.
Cluster 3 umfasst 127.161 Kunden, die in erster Linie hauptsächlich die Provider Intern und Extern im größeren Maße nutzen.
Die Größe des Cluster 4 beträgt 52.682 Kunden. Diese Kunden unterscheiden sich zu den anderen dadurch, dass sie den Provider Ausland stark nutzen. Das gilt sowohl für das Telefonieren als auch für das SMS Schreiben. Die Provider Intern und Extern werden zum Telefonieren ebenfalls stärker genutzt.

Nun wird der Cluster der „Viel-Benutzer“ näher analysiert. Die Anzahl der Cluster wurde abermals auf vier gesetzt, da es bei der Two-Step Clusteranalyse eine gute Segemtierung ergab.

Als zurücktransformiertes Ergebnis erhält man folgendes:

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
in_sec_intern	6.633,24	3.164,29	5.270,13	5.942,18
in_sec_extern	3.070,74	571,49	2.464,13	3.497,19
in_sec_ausland	1,12	1,18	440,42	0,92
in_sec_festnetz	251,14	1,89	47,42	11,94
in_sec_unbekannt	436,03	20,33	209,61	114,58
out_sec_intern	6.767,26	3.164,29	5.430,66	6.246,89
out_sec_extern	3.788,54	684,39	3.132,79	4.490,76
out_sec_ausland	2,74	1,83	670,83	1,59
out_sec_festnetz	253,68	1,75	41,52	10,02
out_sec_unbekannt	1.668,03	138,77	1.001,25	449,34
in_sms_intern	5,55	6,46	5,49	34,52
in_sms_extern	5,75	4,05	5,36	58,74
In_sms_ausland	3,31	2,32	10,02	2,59
In_sms_festnetz	0,05	0,01	0,04	0,06
In_sms_unbekannt	2,74	2,13	2,56	3,35
out_sms_intern	4,75	5,55	5,23	39,04
out_sms_extern	3,85	3,01	4,31	56,97
out_sms_ausland	0,46	0,82	4,37	0,95
out_sms_festnetz	0,06	0,01	0,04	0,07
out_sms_unbekannt	0,09	0,05	0,12	0,23

Tabelle 4.13: K-Means Clusteranalyse zur Segmentierung der „Viel-Benutzer“

Cluster 1 entspricht eine Größe von 101.188 Kunden Diese Kunden benutzen in erster Linie die Provider Intern, Extern, Festnetz und Unbekannt zum Telefonieren. Dieser Cluster entspricht dem „Viel-Telefonierer“.

Cluster 2 enthält 90.799 Kunden. Im Vergleich zu den andern Clustern, enthält dieser jene Kunden, die am wenigsten telefonieren und SMS schreiben.
Cluster 3 umfasst 61.471 Kunden. Die Provider Intern und Extern werden vermehrt genutzt. Der Unterschied zu den anderen Clustern liegt in der Nutzung des Providers Ausland.
Cluster 4 hat eine Größe von 63.389 Kunden. Diese Kunden verwenden den Provider Extern am meisten. Ein großer Unterschied zu den anderen Clustern liegt auch darin, dass diese Kunden in alle Provider bis auf den Provider Ausland sehr viele SMS schreiben.

Aufgrund des K-Means Algorithmus erhält man eine gute Segmentierung durch acht Cluster.
Im Vergleich zu der Two-Step Clusteranalyse findet der K-Means Algorithmus ähnliche Cluster. Es wird in beiden Analysen sowohl ein „Viel-Benutzer“ als auch ein „Wenig-Benutzer gefunden. Allerdings kann die Benutzung der Provider etwas unterschiedlich sein. Der K-Means Algorithmus fand auch unter den „Viel-Benutzern“ einen Cluster, deren Kunden sehr viel SMS schreiben, welchen die Two-Step Clusteranalyse nicht fand. Das Fazit ist, dass unterschiedliche Methoden auch zu unterschiedlichen Ergebnissen führen.

5 ZUSAMMENFASSUNG

In Kapitel 1 ist nachzulesen, warum es in dieser Magisterarbeit einen Zusammenhang zwischen den Kapiteln Netzwerkanalyse und Clusteranalyse gibt. Die soziale Netzwerkanalyse bildet die theoretische Basis für die Indikatoren, welche im Bereich der Kundensegmentierung herangezogen werden.

Kapitel 2 beschäftigt sich mit der Frage was ein soziales Netzwerk ist. Weiters werden Kennzahlen vorgestellt, die dazu dienen, einen Akteur im Netzwerk zu analysieren. Das heißt es wird anhand der Kennzahlen untersucht, ob er zentral im Netzwerk liegt und ob er ein gewisses Prestige im Netzwerk besitzt.

In Kapitel 3 wird die Methodik der Clusteranalyse kurz vorgestellt. Da es sehr viele verschiedene Methoden gibt, Daten zu gruppieren, wird hier nur auf jene Analysen Bezug genommen, welche auch in der tatsächlichen Auswertung verwendet werden. Hierbei handelt es sich um die hierarchische Clusteranalyse, die Two-Step Clusteranalyse und die K-Means Clusteranalyse.

In Kapitel 4 werden die Two-Step Clusteranalyse und der K-Means Algorithmus anhand eines Datensatzes durchgeführt. Der Datensatz wird in diesem Kapitel näher beschrieben.
Weiters wird anhand von Stichprobenziehungen und Permutationen überprüft, ob deren Ergebnisse homogen sind. Das heißt es wird untersucht ob die Ergebnisse stabil sind. Gründe warum Ergebnisse nicht stabil sein können, sind in Kapitel 3 nachzulesen.
Um die Stabilität zu überprüfen, wird eine 10%-iger Stichprobe vier mal permutiert, verschiedene 10%-ige Stichhproben zufällig gezogen, eine Stichprobenerhöhung von 10% auf 50% durchgeführt und der gesamte Datensatz vier mal permutiert. In allen Fällen weisen sowohl die Ergebnis-

se der Two-Step Clusteranalyse als auch die Ergebnisse der K-Means Clusteranalyse Stabilität auf.
Im nächsten Schritt wird untersucht, ob es einen Zusammenhang zwischen den Sekunden und dem Degree gibt. Der Degree ist eine Kennzahl der Netzwerkanalyse und unter Umständen schwieriger für eine Provider zu erheben als die Sekunden. Daher wird überprüft, ob einer Clusteranalyse über die Telefonierdauer gemessen in Sekunden ausreicht, oder ob eine Clusteranalyse über die Variablen des Degree zu anderen Ergebnissen führt. Durch die Analyse mittels der Two-Step Clusteranalyse und der K-Means Clusteranalyse lässt sich, wie erwartet, ein Zusammenhang zwischen den Sekunden und dem Degree feststellen. Trotzdem stimmen die Cluster zu 73,9% überein, was darauf schließen lässt, dass durch die Netzwerkanalyse ein zusätzlicher Erkenntniswert gewonnen werden kann.
Zum Abschluss werden die Daten anhand der Telefonierdauer gemessen in Sekunden und der Anzahl an geschickten SMS segmentiert. Im ersten Schritt wird eine Vorsegmentierung durch die Summe der Sekunden gesamt und die Summe der SMS gesamt durchgeführt. Es ergeben sich dadurch zwei Cluster. Die zwei Cluster lassen sich interpretieren als „Viel-Benutzer“ und „Wenig-Benutzer“. Im nächsten Schritt werden die zwei entdeckten Cluster nochmals einzeln betrachtet und eine weitere Segmentierung durchgeführt. Sowohl die „Viel-Benutzer“ als auch die „Wenig-Benutzer“ lassen sich durch weitere vier Cluster segmentieren. Die Kunden werden sowohl durch die Two-Step Clusteranalyse als auch durch den K-Means Algorithmus in insgesamt acht Gruppen unterteilen.

6 LEBENSLAUF

Bettina Vala Bakk.rer.soc.oec.
Karl-Tornay-Gasse 45-47/5/6
1230 Wien

Persönliche Daten

Geburtsdatum:	14. Juli 1983
Geburtsort:	Mödling
Staatsbürgerschaft:	Österreich
Familienstand:	ledig

Schulischer und akademischer Werdegang

- Sommersemester 2007 – Wintersemester 2008/2009 Magisterstudium Statistik an der Universität Wien
- Sommersemester 2003 – Wintersemester 2006/2007 Bakkalaureatsstudium Statistik an der Universität Wien
- Wintersemester 2002 Studium Betriebswirtschaft an der Wirtschaftsuniversität Wien
- 1997 – 2002 Vienna Business School – Handelsakademie Mödling der Wiener Kaufmannschaft in 2340 Mödling,
  Maria-Theresien-Gasse 25
- 1993 – 1997 Realgymnasium in 1230 Wien,
  Anton Krieger-Gasse 25
- 1989 – 1993 Volksschule Biedermannsdorf in
  2362 Biedermannsdorf, Schulweg 7

7 ABBILDUNGSVERZEICHNIS

8 TABELLENVERZEICHNIS

9 REFERENZEN

Apte, C., B. Liu, E. Pednault, P. Smyth. 2002. *Business application of data mining*. CACM 45

Arndt, J. 1967. *Word of mouth advertising: A rieview of the literature*. Tech. Rep., Advertising Research Foundation Inc.

Backhaus, K., B. Erichson, W. Plinke, R. Weiber. 1996. *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung.* Springer, Berlin.

Bichler, M., C. Kiss. 2004. *A comparison of logistic regression, k-nearest neighbor, and decision tree induction for campaign management.* Tenth Americas Conference on Information Systems.

Deichsel G., H. J. Trampisch. 1985. *Clusteranalyse und Diskriminanzanalyse.* Gustav Fischer Verlag Stuttgart.

Fahrmeir, L., A. Hamerle. 1984. *Multivariate statistische Verfahren.* Walter de Gruyter, Berlin.

Freeman, L. C. 1979. *Centrality in social networks: conceptual clarification.* Social Networks 1 (215-239).

Jansen, D. 2006. *Einführung in die Netzwerkanalyse. Grundlagen, Methoden, Forschungsbeispiele.* 3. Auflage. VS Verlag für Sozialwissenschaften, Wiesbaden.

Kiss, C., M. Bichler. 2007. *Identification of influencers – measuring influence in customer networks.* Internet-based Information Systems.

Knoke, D., R. S. Burt. 1983*. Prominence*. Burt/Minor. (195-224)

Piatetsky-Shapiro, G., B. Masand. 1999. *Estimating campaign benefits and modeling lift.* KDD 1999. ACM, San Diego, CA, USA.

Rosset, S., E. Neumann, U. Eick, N. Vatnik, I. Idan. 2001. *Evaluation of prediction models for marketing campaigns.* KDD 01. ACM, San Francisco, CA, USA.

SPSS. 2004. *K-means Algorithm.* Clementine 9.0 Algorithmus Guide. (63-68).

SPSS. 2004. *TwoStep Cluster Algorithm*. Clementine 9.0 Algorithmus Guide. (69-73).

Steinhausen, D., K. Langer. 1977. *Clusteranalyse. Einführung in Methoden und Verfahren der automatischen Klassifikation.* Walter de Gruyter, Berlin.

Turan, V. 1996. *Algorithmische Graphentheorie.* Addison-Wesley, Bonn.

Wikipedia. Die freie Enzyklopadie. Virales Marketing. Stand: 17. 09. 2008. URL: http://de.wikipedia.org/wiki/Virales_Marketing. (Abruf: 17. 09. 2008)

Zhang, T., R. Ramakrishnan, M. Livny. 1996. *BIRCH: An efficient data clustering methode for very large databases.* Proceedings of the 1996 ACM SIGMOD international conference on knowledge discovery and data mining. (103-114).