7 Geomarketing-Prozess II: Data Mining-Methoden - Modell zur Vergleichbarkeit und Übertragung von Filialerfolgsfaktoren in verschiedenen Regionen

Für die Erstellung eines Modells für die Vergleichbarkeit und Übertragung von Filialerfolgsfaktoren in verschiedenen Regionen für das gesamte Bundesgebiet Deutschlands wird eine zwei-phasige Segmentation vorgenommen.

Erster Schritt: Die Grundidee ist, Deutschland auf Basis von PLZ-gebietsbezogenen Angaben zu Merkmalen aus dem soziodemographischen Umfeld sowie auf Basis von Wirtschaftskennzahlen in mehrere, in sich homogene Großräume (Cluster) zu untergliedern. Dabei können die Cluster durchaus in einem räumlichen Kontinuum zusammenhängen, müssen dies aber nicht zwingend, da sich beispielsweise Ballungsräume in West- und Süddeutschland trotz ihrer räumlichen Distanz durchaus in soziodemographischen und wirtschaftlichen Faktoren ähneln können. Die Wahl von PLZ-Gebieten ist aufgrund der verfügbaren umfangreichen Datenlage auf dieser Aggregationsebene getroffen worden. Gleichzeitig werden durch diese Wahl insbesondere die städtischen Gebiete detaillierter unterschieden als bei einer Durchführung der Clusterung auf Gemeindeebene[1]. Durch das unterschiedliche Skalenniveau der vorliegenden Segmentierungsmerkmale ist die Verwendung der *Two-Step-Clusteranalyse* angezeigt (*Kapitel 7.1*).

Zweiter Schritt: Zur Beurteilung der Frage, welche Merkmale als Erfolgsfaktoren einer Filiale angesehen werden können, werden für ausgewählte, im ersten Schritt erzeugte Cluster auf Grundlage von Unternehmens- und sozioökonomischen Daten aus dem mikrogeographischen Filialumfeld (500 m Isodistanz[2]) – und damit unter Berücksichtigung von Nachbarschaftseffekten – Entscheidungsbaum-Modelle erstellt. Unter der Prämisse der im ersten Schritt getroffenen Zusammenfassung von PLZ-Gebieten zu homogenen Räumen, die sich durch eine bestimmte makrogeographische Konstellation definieren, sollten diese Entscheidungsbaum-Modelle grundsätzlich im

[1] Für die Betrachtung von ländlichen Gebieten ist der Bezug auf Gemeindeebene sinnvoll. Da aber Städte wie Berlin oder München als eine Gemeinde gelten, ist bei der Analyse der städtischen Gebiete der Bezug auf die PLZ-Gebiete aussagekräftiger (Berlin hat etwa 190, München 75 PLZ-Gebiete).

[2] Wenn die Annahme aus dem Verkehrsbereich in Bezug auf den ÖPNV zugrunde gelegt wird, dann ist ein überwiegender Teil der Fußgänger bereit, 300 -500 m zur nächsten Haltestelle zu gehen. Bei einer weiteren Entfernung nimmt die Bereitschaft diese zurückzulegen, linear ab (*siehe z. B. Gutachten IVU 1998, 1993, IVU & SNV 1994*).

gesamten ausgewählten Cluster gültig sein und sich somit auf vergleichbare andere Gebiete derselben Clusterzugehörigkeit in Deutschland übertragen lassen (*Kapitel 7.2*).

Durch die Anwendung der *Two-Step-Clusteranalyse* und durch die Berücksichtigung zahlreicher Typisierungsmerkmale erhalten wir ein mehrfarbiges Deutschlandbild.

Dieses zweistufige Vorgehen rechtfertigt sich durch die folgende Hypothese: Die Erfolgsfaktoren einer Filiale unterscheiden sich in ihrer Ausprägung und in ihrer Gewichtung in den verschiedenen Gebieten Deutschlands (Clustern). Daher werden die Entscheidungsbaum-Analysen clusterspezifisch durchgeführt. Beispielsweise wird für alle Filialen eines Clusters (z. B. Cluster 13, 14 und 15, in dem u.a. Gebiete aus Berlin, Köln und München enthalten sind) ein Entscheidungsbaum-Modell zunächst nur für die Berliner Filialen erstellt und mit den Ergebnissen der Entscheidungsbaum-Modelle aus Köln und München verglichen. Sind die Modelle dieser drei Filialräume aus den Clustern untereinander vergleichbar, kann von einer statistisch hochwertigen Übertragbarkeit der gefundenen Entscheidungsbaum-Analyseergebnisse auf andere Filialen in demselben Cluster ausgegangen werden. Die regionale Übertragbarkeit wird in *Kapitel 9.3* dargestellt.

7.1 Ermittlung strukturähnlicher Regionen mittels der Two-Step-Clusteranalyse

Im Folgenden wird eine zusammenfassende Beschreibung der *Two-Step-Clusteranalyse* nach JANSSEN & LAATZ (2005) vorgenommen. Die Anwendung auf den Untersuchungsfall wird beschrieben, und die Ergebnisse werden in Form von Tabellen und exemplarischen Karten einiger Regionen dargestellt.

7.1.1 Two-Step-Clusteranalyse

Clustern ist ein exploratives, d. h. strukturentdeckendes Verfahren zur Gruppierung von Fällen. Ziel der Clusteranalyse ist es, Fälle bzw. Untersuchungsobjekte in Gruppen zusammenzufassen. Die Mitglieder einer Gruppe sollen dabei eine weitgehend verwandte Eigenschaftsstruktur aufweisen. Zwei Untersuchungsobjekte sind sich dann ähnlich, wenn sie sich in der Gesamtheit der berücksichtigten Eigenschaften ähneln (Homogenität). Die Zusammenfassung von Objekten zu homogenen Gruppen erfolgt

auf Basis des Ähnlichkeits- bzw. Distanzmaßes[3]. Zwischen den Gruppen sollten dagegen möglichst wenig Gemeinsamkeiten bestehen (Heterogenität).

Die *Two-Step-Clusteranalyse* stellt ein junges Klassifikationsverfahren in SPSS[4] dar, das die gleichzeitige Verarbeitung von metrischen und kategorialen Variablen erlaubt und schon aufgrund der hohen Datenmenge für den Untersuchungsfall hervorragend geeignet ist. Bei der *Two-Step-Clusteranalyse* vollzieht sich das Clustern der Fälle in zwei Stufen, einer Vorcluster- und einer Clusterstufe. Ausgehend vom Distanz- bzw. Ähnlichkeitsmaß[5] werden in der ersten Stufe die Fälle sequentiell abgearbeitet und verschiedene Pre-Cluster aus jeweils sehr ähnlichen Fällen gebildet. In der zweiten Stufe werden die Pre-Cluster mittels eines Verfahrens der hierarchischen Clusteranalyse[6] zu den Endclustern fusioniert. Die hierarchische Clusteranalyse kann für diese Stufe verwendet werden, weil die Anzahl der Pre-Cluster (als Eingangsinformationen für den zweiten Schritt) im Vergleich zu der Anzahl der ursprünglichen Datenfälle nur noch sehr klein ist und damit eine paarweise Beurteilung der Ähnlichkeiten leicht möglich wird. Die Bildung der endgültigen Cluster wird so lange fortgeführt, bis entweder eine vom Anwender vordefinierte Clusteranzahl k erreicht wird oder aber automatisch eine optimale, gut separierbare Clusteranzahl innerhalb einer vom Anwender vorgegebenen maximal möglichen Clusteranzahl k gefunden wird. Als Ähnlichkeitsmaß fungieren verschiedene Modellauswahlkriterien (z. B. Schwarz-Bayes Kritierium), die auf der Basis der Log-Likelihood-Funktion automatisch gebildet werden.

Clusteranalysen fordern grundsätzlich die Unabhängigkeit aller einfließenden Klassifikationsmerkmale, was meist durch eine vorgeschaltete Faktorenanalyse[7] hergestellt werden kann. Aus folgenden Gründen wird jedoch auf die Verwendung der Ergebnisse dieser Voranalyse verzichtet.

---

[3] Distanz- oder Proximitätsmaß: Berechnung eines Wertes, der die Unterschiede zweier Untersuchungsobjekte hinsichtlich der untersuchten Merkmale symbolisiert, z. B. quadrierte euklidische Distanz.

[4] SPSS und Clementine ist die hier angewandte Statistik- und Data Mining-Software *'Statistical Package for Social Sciences'*.

[5] Bei metrischen Variablen beruht das Distanzmaß auf der euklidischen Distanz, bei kategorialen Variabeln auf dem Log-Likelihood Verfahren.

[6] Die hierarchische Clusteranalyse benötigt als Modellvoraussetzung: metrische Variablen, die Vergleichbarkeit der Variablen-Einheiten, die statistische Unabhängigkeit der Eigenschaftsmerkmale, die Konsistenz der Daten und lässt sich nur auf eine verhältnismäßig kleine Datenmenge anwenden (vgl. JANSSEN & LAATZ 2005).

[7] Die Ergebnisse der Faktorenanalyse sind im Anhang dargestellt.

Die Faktorenanalyse ist ein dimensionsreduzierendes Verfahren. Bei der Eingrenzung der Variablen auf eine kleine Anzahl von Faktoren gehen grundsätzlich Informationen verloren, und zwar sowohl von der Gesamtheit aller Variablen als auch für jede Einzelne (Kommunalitäten < 1). Beispielsweise würden bei einer Extraktion von zehn Faktoren nur noch 70 % der Gesamtvarianz aller ursprünglichen Variablen abgebildet werden. Hinzukommt, dass mit zunehmender Anzahl der extrahierten Faktoren auch jeder Faktor selbst einen immer geringeren Anteil der Varianz aller ursprünglichen Variablen erklärt, was sich in den immer kleiner werdenden Eigenwerten ausdrückt. Damit würde jeder extrahierte Faktor mit sehr unterschiedlichem Beitrag als Variable in die Clusteranalyse einfließen.

Weiterhin ergibt sich zusätzlich ein rechentechnisches Problem bei der Verwendung von standardisierten, metrischen Werten für die Faktoren. Bei der Beurteilung von Ähnlichkeiten in der *Two-Step-Clusteranalyse* mittels der Likelihood-Funktion verlängert das Vorliegen von kontinuierlichen, fein abgestuften Variablen die Rechenzeit überproportional mit der Anzahl der berücksichtigten Faktoren. Jede Rundung wiederum würde den bei der Faktoranalyse ohnehin schon auftretenden Informationsverlust in einem nur schwer quantifizierbaren Ausmaß weiter erhöhen.

7.1.2 Anwendung Two-Step-Clusteranalyse

Werden statt der Faktorwerte die gesamten Variablen in ihren ursprünglichen Ausprägungen in die Clusteranalyse einbezogen, kann ein ausgeglichenes Verhältnis zwischen der Anzahl der demographischen und der wirtschaftlichen Variablen beobachtet werden. Da für die hier genannten Fragestellungen die Berücksichtigung beider Bereiche gleich relevant ist, kann bei der Verwendung der Originalvariablen die Gleichgewichtung von demographischen Faktoren einerseits und wirtschaftlichen Faktoren andererseits sichergestellt werden.

Da es in der vorliegenden Arbeit primär nicht um die Bewertung der Wichtigkeit der einzelnen Faktoren bzw. Variablen geht, kann auch die Forderung nach der statistischen Unabhängigkeit der Clustermerkmale als eher untergeordnet angesehen werden. Mit dem Ziel der Ausgrenzung homogener Cluster sollen gerade auch bestimmte Konstellationen von Merkmalen mit gegebenenfalls einander verstärkenden Effekten, Überlagerungen oder Wirkungsabhängigkeiten berücksichtigt werden, welche die Ausgrenzung eines typischen Clusters überhaupt erst bedingen.

In die Clusteranalyse gehen somit die folgenden soziodemographischen und ökonomischen Faktoren ein: Bevölkerungsanzahl, Bevölkerungsdichte, Fläche, Ausländeranteil, Anteil der Sozialversicherungspflichtigen, Anteil der unter 18jährigen, Anteil der 18-29jährigen, Anteil der 30-39jährigen, Anteil der 40-49jährigen, Anteil der 50-59jährigen, Anteil der Personen über 59 Jahre, Kaufkraft je Einwohner, einzelhandelsrelevanter Umsatz je Einwohner, Zuzüge minus Wegzüge, sozialversicherungspflichtige Beschäftige am Arbeitsort insgesamt, Anteil der sozialversicherungspflichtigen Beschäftigen am Arbeitsort unterteilt in 13 Wirtschaftsbereiche[8] und die Zentralitätskennziffer.

Bei Betrachtung der Entwicklung der Modellauswahlkriterien nach Durchführung der *Two-Step-Clusteranalyse* (operationalisiert durch die schrittweise Änderung der Verhältnisse der Distanzmaße, *siehe Abb. 7-1*) ist unter den vorgegebenen maximal 20 möglichen Clusterlösungen eine 4er, 7er, 14er oder 16er Lösung[9] inhaltlich denkbar.

[8] 13 Wirtschaftsbereiche: 1. Land- und Forstwirtschaft, Fischerei 2. Bergbau, Gewerbe von Steine und Erden, Energie- und Wasserversorgung 3. Verarbeitendes Gewerbe 4. Baugewerbe 5. Handel, Instandhaltung, Reparatur 6. Verkehr und Nachrichtenübermittlung 7. Kredit- und Versicherungsgewerbe 8. Gastgewerbe 9. Grundstücks- und Wohnungswesen 10. Erziehung und Unterricht 11. Gesundheits- , Veterinär-, Sozialwesen 12. Erbringung sonstiger Dienstleistungen 13. Verwaltung, Verteidigung, Sozialversorgung

[9] Weitere Ergebnistabellen Two-Step-Clusteranalyse siehe Anhang

Automatische Clusterbildung

Anzahl der Cluster	Bayes-Kriterium nach Schwarz (BIC)	BIC-Änderung[a]	Verhältnis der BIC-Änderungen[b]	Verhältnis der Distanzmaße[c]
1	177730,025			
2	150515,135	-27214,890	1,000	2,611
3	140437,339	-10077,796	,370	1,374
4	133254,824	-7182,515	,264	1,701
5	129263,899	-3990,925	,147	1,145
6	125849,506	-3414,393	,125	1,316
7	123389,807	-2459,699	,090	1,618
8	122083,321	-1306,486	,048	1,008
9	120790,910	-1292,411	,047	1,004
10	119506,502	-1284,407	,047	1,085
11	118366,912	-1139,590	,042	1,042
12	117295,929	-1070,983	,039	1,082
13	116348,999	-946,931	,035	1,040
14	115459,967	-889,031	,033	1,226
15	114837,925	-622,042	,023	1,120
16	114342,182	-495,743	,018	1,370
17	114131,425	-210,757	,008	1,009
18	113927,801	-203,625	,007	1,024
19	113742,182	-185,618	,007	1,006
20	113561,114	-181,068	,007	1,119

a. Die Änderungen wurden von der vorherigen Anzahl an Clustern in der Tabelle übernommen.

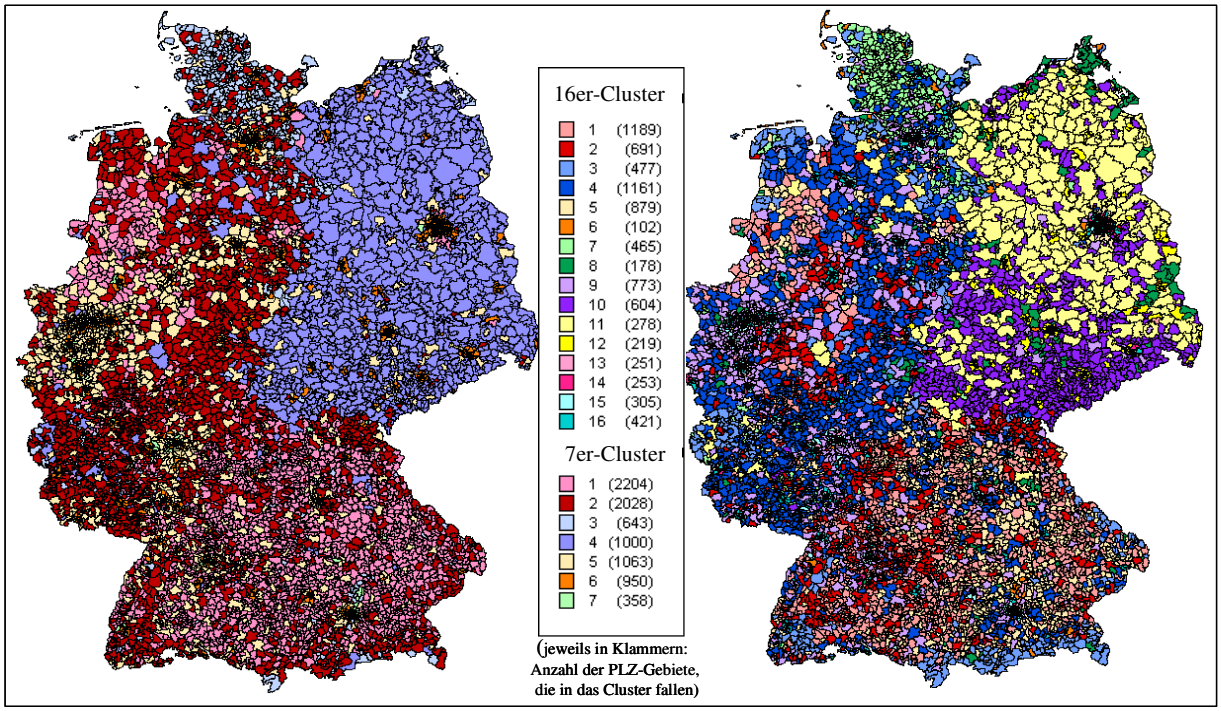
b. Die Änderungsquoten sind relativ zu der Änderung an den beiden Cluster-Lösungen.

c. Die Quoten für die Distanzmaße beruhen auf der aktuellen Anzahl der Cluster im Vergleich zur vorherigen Anzahl der Cluster.

Quelle: eigene Berechnung (SPSS)

Abb. 7-1: Clusterbildung: Two-Step-Clusteranalyse

Für jede dieser Lösungen findet eine bundesweite Visualisierung statt, um so neben den statistischen Maßzahlen eine weitere, inhaltliche Evaluationsgrundlage für die geeignete Clusteranzahl zu erhalten. Dabei impliziert der grundsätzlich explorative Charakter des Verfahrens eine analytisch offene, ergebnisorientierte Interpretation, wobei die Anzahl der Cluster in gewisser Weise immer eine Gratwanderung zwischen einer stark generalisierten ‘schwarz-weiß-Darstellung’ (hier z. B. die einfache Unterscheidung zwischen Ost und West) und damit einem maximalen Informationsverlust einerseits und der ‘feinsten Partition’ andererseits ist (z. B. wenn jedes PLZ-Gebiet als ein Unikat eingeht und ein eigenes Cluster darstellt). Je gröber, desto inhomogener werden die Cluster; je feiner, desto homogener sind sie. Im Hinblick auf die Fragestellungen der nachfolgenden Entscheidungsbaum-Analysen ermöglicht eine hinreichende feine Clusterung der PLZ-Gebiete zudem, dass die Baummodelle, Grafiken und Entscheidungsregeln übersichtlicher werden.



Quelle: eigene Berechnung (Darstellung Filialinfo)

Abb. 7-2: Two-Step Clusteranalyse: 7er Cluster und 16er Cluster

Beispielhaft wird hier die Clustereinteilung für eine 7er und eine 16er Clusterlösung dargestellt[10].

Die 7er Clusterlösung wird aufgrund der zu geringen Differenzierungen, insbesondere der Regionen innerhalb der Großstädte und Metropolen, nicht gewählt. Dahingegen wird die 16er Lösung gegenüber der ebenfalls denkbaren 14er Lösung bevorzugt, weil gerade in den städtischen Gebieten nochmals die feinen Unterscheidungen einzelner Stadtteile deutlich werden. Der Ansatz geht zwar von einer makroräumlichen Ebene aus, der dann aber im weiteren Schritt mit mikroräumlichen Analysen verfeinert wird. Insofern wird im Folgenden mit der 16er Unterteilung gearbeitet. Die Verteilung der PLZ-Gebiete auf die 16 Cluster zeigt die folgende Abbildung (*Abb. 7-3*).

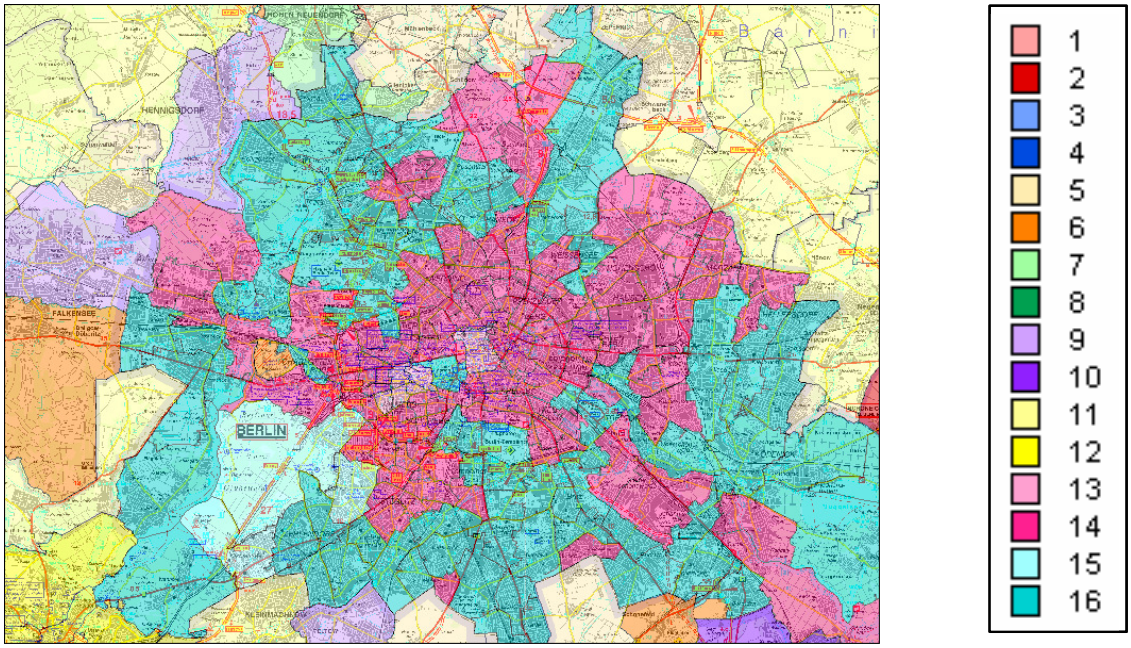
[10] Die Farbauswahl ist hier anhand der automatischen Voreinstellung erfolgt. Es ist bewusst auf eine inhaltliche Farbwahl der Cluster verzichtet worden, da es sich um unabhängige, gleichwertige Gruppen handelt und mögliche Ähnlichkeiten von Clustern durch eine Farbstufung nicht vorgetäuscht werden sollen.

Clusterverteilung

		N	% der Kombination	% der Gesamtsumme
Cluster	1	1189	14,4%	14,4%
	2	691	8,4%	8,4%
	3	477	5,8%	5,8%
	4	1161	14,1%	14,1%
	5	879	10,7%	10,7%
	6	102	1,2%	1,2%
	7	465	5,6%	5,6%
	8	178	2,2%	2,2%
	9	773	9,4%	9,4%
	10	604	7,3%	7,3%
	11	278	3,4%	3,4%
	12	219	2,7%	2,7%
	13	251	3,0%	3,0%
	14	253	3,1%	3,1%
	15	305	3,7%	3,7%
	16	421	5,1%	5,1%
	Kombiniert	8246	100,0%	100,0%
Gesamtwert		8246		100,0%

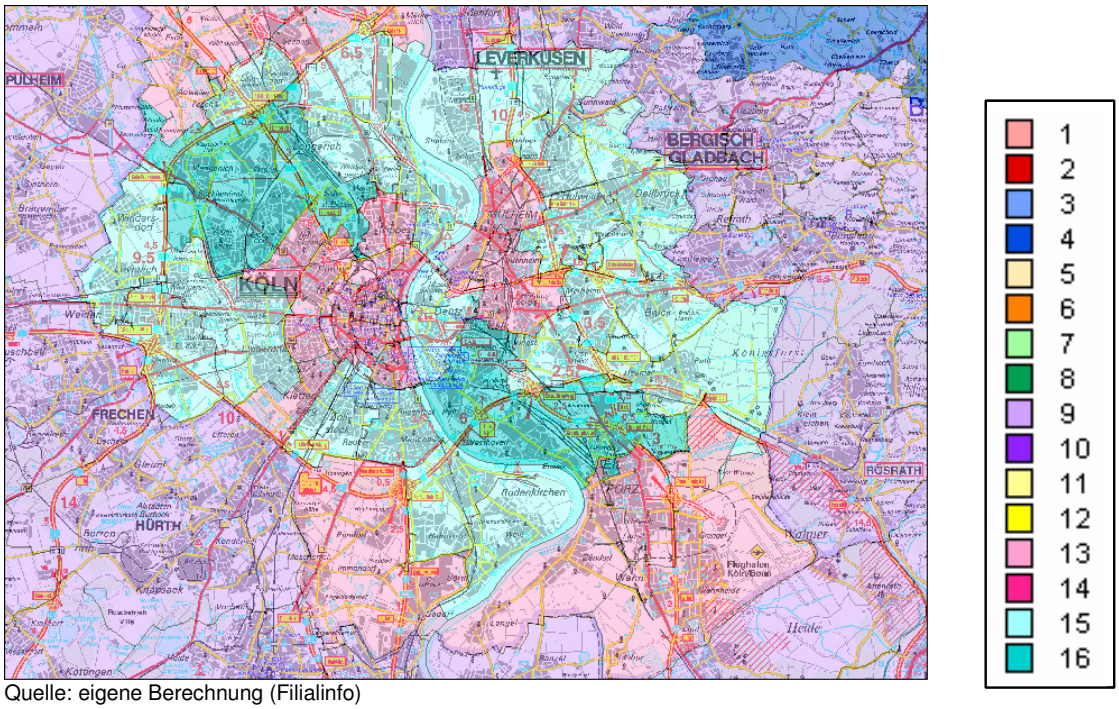
Quelle: eigene Berechnung (SPSS)

Abb. 7-3: Clusterverteilung: Two-Step-Clusteranalyse



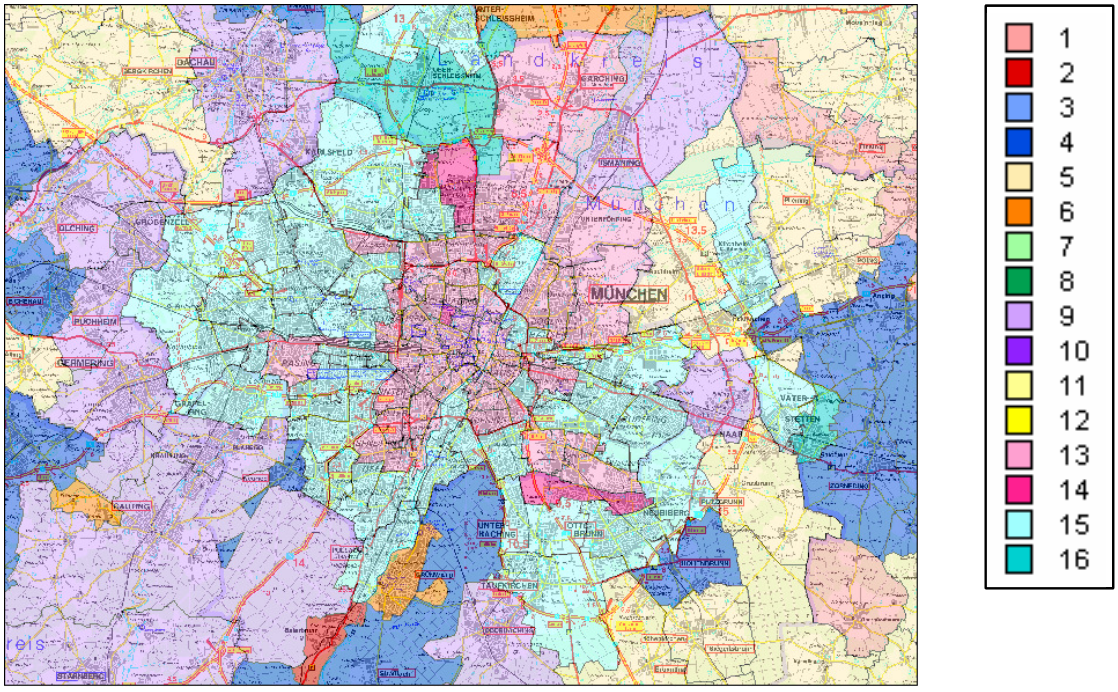
Quelle: eigene Berechnung (Filialinfo)

Abb. 7-4: Two-Step-Clusteranalyse - 16er Cluster, Berlin



Quelle: eigene Berechnung (Filialinfo)

Abb. 7-5:Two-Step-Clusteranalyse - 16er Cluster, Köln



Quelle: eigene Berechnung (Filialinfo)

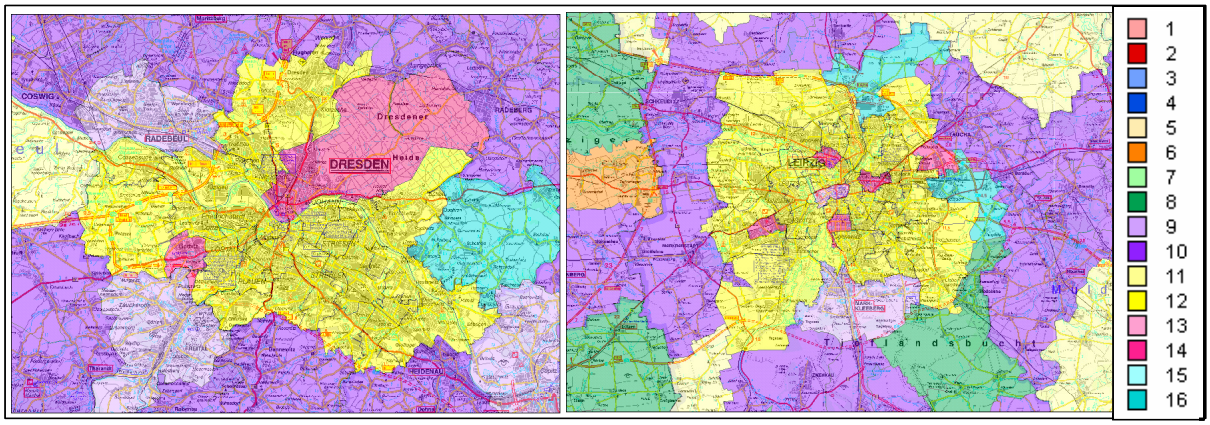
Abb. 7-6:Two-Step-Clusteranalyse - 16er Cluster, München

Die Cluster lassen sich anhand der sozioökonomischen Merkmale beschreiben und sind hier nach der Bedeutung für städtische Gebiete mit hoher Bevölkerungsdichte bis hin zu den landwirtschaftlichen Gebieten und den Mittelgebirgen Deutschlands geordnet:

Die **Cluster 13 bis 16** umfassen vorwiegend flächenarme PLZ-Gebiete in städtischen Gebieten mit hohem Ausländeranteil (über 13,5 %), hoher Bevölkerungsdichte (über 2.200 EW/km^2) und hoher einzelhandelsrelevanter Kaufkraft je Einwohner (über 5.500).

Cluster 9 folgt hinsichtlich der demographischen Merkmale unmittelbar hinter den Clustern 13 bis 16. Die Gebiete im Cluster 9 haben aber neben einer eher durchschnittlichen Bevölkerungdichte (950 EW/km^2) und viel Fläche eine sehr hohe - ähnlich wie Cluster 13 - Gesamtzahl an Branchen und einen überdurchschnittlich hohen Anteil des verarbeitenden Gewerbes am Gesamtgewerbe (30,2 %).

Cluster 11 setzt sich aus vornehmlich flächenreichen PLZ-Gebieten in Randgebieten von Großstädten in den neuen Bundesländern zusammen. Es weist die größte negative Bevölkerungsentwicklung (mehr Fortzüge als Zuzüge) und einen geringen Ausländeranteil (1,7 %) auf. Außerdem haben die Gebiete eine sehr geringe einzelhandelsrelevante Kaufkraft je Einwohner, die nur knapp von **Cluster 12** übertroffen wird. Cluster 12 umfasst vornehmlich die Innenräume größerer Städte in den neuen Bundesländern mit dem geringstem Anteil an unter 18-Jährigen, einem sehr hohen Anteil am Bereich Erziehung und Unterricht (9,5 %) sowie Verwaltung (mit 10 % nahezu doppelt so hoch wie der Durchschnitt), was sich vermutlich auf die vergangene Funktion als DDR-Bezirkshauptstädte zurückführen lässt.



Quelle: eigene Berechnung (Darstellung Filialinfo)

Abb. 7-7: Two-Step-Clusteranalyse - 16er Cluster, Dresden, Leipzig

Cluster 10 findet sich häufig in unmittelbarem Umfeld der Großstädte Ost- und Westdeutschlands (mit Einzugsbereich um 30 km), mit niedrigstem Ausländeranteil (1,4 %), viel Fläche, überdurchschnittlich hohem Anteil sozialversicherungspflichtiger Beschäftigter an der Bevölkerung und niedrigem Anteil der Personen unter 18 Jahre.

Die **Cluster 1, 3 sowie 5 bis 7** weisen eine sehr geringe Bevölkerungsdichte (unter 175 EW/km^2) auf, in **Cluster 6** ist sie am geringsten. Weiterhin sind diese Cluster gekennzeichnet durch einen unterdurchschnittlichen Ausländeranteil (unter 5,9 %), einen sehr hohen Anteil der unter 18 Jährigen in Haushalten (über 18 %), viele Haushalte mit drei und mehr Personen (über 35 %) sowie wenige Single-Haushalte, wenig Branchen. **Cluster 3** hat zudem den geringsten Anteil der sozialversicherungspflichtig Beschäftigten an der Bevölkerung.

Cluster 7 liegt überwiegend in Schleswig-Holstein, mit hohem Anteil an Land- und Forstwirtschaft/Fischerei. **Cluster 8** befindet sich vornehmlich in den Grenzgebieten Ost zu Polen. Beide Cluster haben eine geringe Bevölkerungsdichte, einen niedrigen Ausländeranteil, einen hohen Anteil an Zwei-Personen-Haushalten.

Cluster 2 mit dem höchsten Anteil an verarbeitendem Gewerbe (47 % am Gesamtgewerbe) ist vorwiegend in Süddeutschland, Westfalen und im Bergischen Land vorzufinden.

Cluster 4 weist einen überdurchschnittlichen Anteil der Branche Gesundheits-, Veterinär- und Sozialwesen am Gesamtgewerbe (14,8 %) auf, der nur noch im ‘Stadtcluster’ 14 übertroffen wird. Cluster 4 umfasst überwiegend die Regionen der Mittelgebirge sowie nördliches und östliches Niedersachsen mit einem dichten Angebot an Kur- und Reha-Einrichtungen.

Durch die Anwendung der *Two-Step-Clusteranalyse* als ein strukturentdeckendes Verfahren ist es gelungen, auf der Grundlage von soziodemographischen sowie von Wirtschaftskennzahlen, Deutschland in mehrere, in sich homogene Großräume (Cluster) zu gliedern. Ein solches Vorgehen ermöglicht es auch, eine Alternative zu der oft umstrittenen Einteilung von Stadt und Land, deren Grenzen nicht eindeutig definiert sind, gegenüberzustellen. Denn die Stadt-Land Untergliederung lässt beispielsweise die ‘Speckgürtel’ unmittelbar um die Städte außer Acht oder Agglomerationen, wie das Ruhrgebiet, können bei der Unterteilung nicht adäquat statistisch erfasst und bewertet werden. Gerade die Veränderung, das Wachstum von Gebieten sind unabhängig von administrativen Grenzen interessant.

Da hier die bevölkerungsstarken Gebiete im Mittelpunkt der Untersuchung stehen, sind vor allem folgende Cluster von Interesse: Cluster 13 bis 16 (höchste Bevölkerungsdichte) sowie Cluster 9 (hohe Bevölkerungdichte und sehr hohe Anzahl an Gewerbebetrieben), Cluster 11 und 12 (Randgebiete und Innenräume der vornehmlich ostdeutschen Großstädte, hoher Verwaltungsgrad) und Cluster 10 (Umfeld von Großstädten).[11]

Insgesamt macht das Ergebnis deutlich, dass die Clusteranalyse eine gute Unterscheidung von räumlichen Strukturen ermöglicht und damit in jedem Falle die fundiertere Alternative zu einer im Vornherein angenommenen Einteilung nach Alten und Neuen Bundesländern und Städtegrößen darstellt, wie sie oft in der Praxis angewendet wird (*siehe Anhang*).

7.2 Ermittlung von Erfolgsfaktoren mittels CHAID

Zur Beurteilung der Frage, welche Merkmale als Erfolgsfaktoren einer Filiale gelten, werden für ausgewählte Cluster auf der Grundlage von Unternehmensdaten und sozioökonomischen Daten aus dem mikroräumlichen Filialumfeld (500 m Isodistanz) Entscheidungsbaum-Modelle erstellt. Es wird im Weiteren (*Kapitel 9.3*) aufgezeigt, dass die Modelle grundsätzlich im gesamten ausgewählten Cluster gültig und somit auf vergleichbare andere Gebiete derselben Clusterzugehörigkeit in Deutschland übertragbar sind.

7.2.1 Entscheidungsbäume

Klassifikations- und Entscheidungsbäume dienen prinzipiell der Identifizierung von Segmenten, Untergruppen und Mustern durch Baumdiagramme. Entscheidungsbäume sind somit statistische Segmentierungsmodelle. Es handelt sich hier um vergleichsweise junge Methoden, die in ihrer Bedeutung aufgrund der Notwendigkeit von schnellen effektiven Mustererkennungen in Massendaten zunehmen.

Entscheidungsbäume unterteilen einen Datenbestand auf der Basis von selbsttätig entdeckten Zusammenhängen zwischen einer Zielvariablen und mehreren Prediktorvariablen in immer feinere Subsegmente (Knoten). Sie identifizieren auf

[11] Weitere Abbildungen sowie eine Übersichtstabelle zu den Kennziffern zum Vergleich der Cluster befinden sich im Anhang.

mehreren Ebenen (Ästen) Interaktionen und grenzen Gruppenzugehörigkeiten ab. Außerdem ist es möglich, aus einem Baummodell Klassifikations- und Vorhersageregeln abzuleiten und diese auf bestehende und neue Daten anzuwenden.

Es gibt unterschiedliche Algorithmen der Entscheidungsbäume, wobei im Folgenden die beiden wichtigsten und in SPSS[12] enthaltenen Verfahren beschrieben werden:

- C&RT Classification and Regression Trees
  (*BREIMAN, FRIEDMAN, OLSHEN, STONE 1984)*
- CHAID Chi-squared Automatic Interaction Detector (*KASS 1980)*

Bezüglich eines vorgegebenen Kriteriums (eine Zielvariable: z. B. Vertriebserfolg oder Vertriebserfolg pro Mitarbeiter) werden alle Beobachtungen im Datensatz derart in Untergruppen aufgespalten, dass sich die einzelnen, durch die Teilung entstehenden Gruppen, hinsichtlich des Zielkriteriums jeweils in sich homogener sind. Die Splits erfolgen auf Basis der Verringerung des Gini-Koeffizienten (als Maß der zu reduzierenden Inhomogenität) auf mehreren Ebenen, wobei auf den einzelnen Ebenen ein Split immer genau durch einen Parameter vorgenommen wird, der mit dem Zielkriterium am stärksten zusammenhängt. Bei C&RT ist ein binärer Algorithmus zum Baumaufbau hinterlegt, d. h. jeder übergeordnete Knoten wird immer genau in zwei Teilstichproben gesplittet. Es handelt sich um einen rekursiven Prozess, bis eines der vorab zu definierenden Abbruchkriterien erreicht ist. C&RT verarbeitet innerhalb der Prediktoren sowie für das Zielkriterium sowohl kategoriale als auch metrische Größen. Bei Verwendung von metrischen Zielkriterien werden bei C&RT Mittelwerte und Standardabweichungen zur Berechnung des Gini-Koeffizienten berücksichtigt.

Das Verfahren von CHAID unterteilt den Datenbestand auf der Basis von statistischen Zusammenhangsmaßen und deren Signifikanzeinschätzungen. Somit kann das CHAID-Verfahren aus statistischer Sicht als wesentlich leistungsfähiger eingeschätzt werden als der C&RT-Algorithmus mit dem Gini-Koeffizienten als ein rein deskriptives Zusammenhangsmaß. Auf der Grundlage von den jeweils vorliegenden Signifikanzen werden optimale Trennungen in Gruppen vorgenommen, wobei der Baum nicht wie bei C&RT binär, sondern mit x-Unterknoten (entsprechend der Anzahl der Kategorien der

[12] Weiterhin sind in SPSS die Verfahren QUEST und Exhaustive CHAID implementiert: QUEST (Quick, Unbiased, Efficient Statistical Tree). Es handelt sich um ein schnelles Verfahren, das die in anderen Verfahren auftretende Verzerrung zugunsten von Prädiktorvariablen mit vielen Kategorien vermeidet. QUEST kann nur ausgewählt werden, wenn die abhängige Variable nominal ist; Exhaustive CHAID ist lediglich eine Abwandlung von CHAID, die für jede Prädiktorvariable alle möglichen Aufteilungen untersucht. Sie ist von daher sehr rechenintensiv.

Prediktoren) entsteht. Im Gegensatz zum Homogenitätsgedanken bei C&RT ist es „*allgemein (...) das Ziel einer Chaid-Analyse, eine Menge von Objekten derart in Gruppen aufzuteilen, dass sich die einzelnen Gruppen bezüglich eines vorgegebenen Kriteriums möglichst deutlich voneinander unterscheiden"* (BROSIUS 1997, S. 127).

Signifikante Unterschiede zur Zielvariablen werden über verschiedene Tests berechnet. Im Falle einer metrischen Zielvariable kommt der F-Test zur Beurteilung von Unterschieden von Gruppenmittelwerten zum Einsatz. Im Falle einer kategorialen Zielvariable (ordinal oder nominal) werden Chi-Quadrat Statistiken (analog zu Kreuztabellen) berechnet. Bei nominalen Zielkriterien wird die Chi-Quadrat Methode nach Pearson, bei ordinalen Zielvariablen hingegen der Likelihood-Quotient (LR) zugrundegelegt[13]. Im Weiteren wird hier der Pearson´sche Chi²-Wert genutzt, zumal sich vor allem bei großen Stichproben beide Maßzahlen immer mehr annähern.

$$CHi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(HKbij - HKeij)^2}{HKeij}$$

$$LR = \sum_{i=1}^{I} \sum_{j=1}^{J} HKbij \cdot \ln\left(\frac{HKbij}{HKeij}\right)$$

i...I = Zahl der Zeilen im Kreuztabellenfeld
j...J = Zahl der Spalten im Kreuztabellenfeld
HKb = beobachtete Häufigkeiten für das Kreuztabellenfeld in Zeile i und Spalte j
HKe = erwartete Häufigkeiten für das Kreuztabellenfeld in Zeile i und Spalte j

Ebenso wie bei C&RT ist auch bei CHAID von praktischer Bedeutung, dass die Methodik auch auf Seiten der Prediktoren mit allen Variablentypen funktioniert, d. h. Variablen können nominal, ordinal oder kontinuierlich sein, wobei letztere zur Berechnung der Zusammenhangsmaße per Voreinstellung in Dezile kategorisiert werden. Fehlende Werte werden als eigene Kategorie behandelt.

Im Vergleich beider Entscheidungsbaum-Verfahren haben neben den bereits erwähnten Unterschieden eigene Analysen gezeigt, dass der CHAID-Algorithmus bei der gegebenen Fragestellung sehr gute Ergebnisse mit interpretierbaren Baumstrukturen liefert. Hinzukommt, dass bei CHAID jede Prediktorvariable innerhalb eines Astes nur

[13] Die Chi-Quadrat Methode berechnet die Summe der quadrierten Abweichungen zwischen den beobachteten und den per Zufall zu erwartenden Häufigkeiten innerhalb der Kreuztabelle als residuale Differenzen. Hingegen berechnet der Likelihood-Quotient die Summe der logarithmierten Verhältnisse zwischen beobachteten und erwarteten Häufigkeiten innerhalb der einzelnen, von den Variablenkategorien aufgespannten Zellen in der Kreuztabelle.

einmal zur Baumaufteilung verwendet wird, wohingegen bei C&RT durch den binären Baumaufbau ein Merkmal mehrfach genutzt und damit die baumübergreifende Relevanz als Splitkriterium überbetont werden kann.

Zudem ist CHAID als die flexibelste Methode in Bezug auf die Verwendung unterschiedlich skalierter Attribute bekannt (siehe BAGOZZI 1994). Gerade das ist ein außerordentlicher Vorteil bei der Nutzung vieler unterschiedlich skalierter Marktdaten; Transformationen von einem Skalenniveau in ein anderes sind nicht mehr notwendig.

7.2.2 Anwendung CHAID: Filialerfolg nach Two-Step-Clustern

In diesem Kapitel wird der Algorithmus der *CHAID-Analyse* vorgestellt. Hierzu werden die zugrunde liegenden statistischen Methoden sowie die wesentlichen Schritte einer *CHAID-Analyse* anhand der praktischen Beispiele aus dem Anwendungsfall der Deutschen Post nachvollzogen. Das Verständnis der Methode ist unabdingbar, um die Ergebnisse der Analyse unter Berücksichtigung von Einschränkungen in der Aussagekraft richtig zu interpretieren, zu optimieren und erneute weitere partielle oder auch abgewandelte Analysen zu initiieren.

So vielfältig die Anwendungsgebiete von CHAID zu sein scheinen, sind die Vorteile des Verfahrens in der Praxis vielfach noch nicht ausreichend erkannt. Gründe dafür sind im möglichen Mangel an entsprechend aussagekräftigen Marktdaten zu sehen, in deren Kosten bei der Beschaffung oder aber in einem enormen Zeitaufwand, der sich bei Datengrundlagen aus verschiedenen Quellen, Formaten und Zeiträumen ergibt, die für die Analysen zusammengeführt werden müssen.

Weiterhin ist aber auch die zugrunde liegende Methodik nicht von jedem ungeschulten Mitarbeiter ohne weiteres anwendbar. Hier gilt es zu unterscheiden zwischen einer Art *'Basismodellierung'*, wo ein einmaliger Modelldurchlauf schnell erste grobe Aussagen zu Analyseergebnissen liefert, und dem *'Advanced Modelling'*, wo bei mehreren Modellaufbauten gezielt bestimmte Prediktoren bezüglich ihrer Relevanz und ihrer Splits im Baum genauer untersucht, Prediktoren ausgetauscht bzw. neu kombiniert oder die Schwellenwerte beim Aufbau der Baumstruktur (Signifikanzniveaus zur Trennung, Abbruchskriterien zum Baumaufbau) verfeinernd eingesetzt werden.

Zur Beantwortung der Fragestellung, welche Faktoren mit dem Vertriebserfolg einer Filiale in Zusammenhang stehen, fließen in die CHAID-Bäume sowohl räumliche als

auch nicht-räumliche Faktoren aus dem unmittelbaren, fußläufigen Filialumfeld ein (Isodistanz 500 m)[14]. So sind verschiedene Unternehmensdaten aller postbetriebenen Filialen herangezogen sowie für jede einzelne dieser Filialen über GIS aggregierte Attribute aus dem Bereich der ökonomischen sowie soziodemographischen Marktdaten (Marktzellen) hinzugenommen worden. Als abhängige Zielvariable fungiert der gesamte Vertriebserfolg eines bestimmten Zeitraums, dessen Variabilität von Standort zu Standort eine Konsequenz verschiedener Umfeldbedingungen sein kann. Sind die Kritierien für den Filialerfolg hinreichend bekannt, so können diese Regeln unterstützend für die Standortüberprüfung bzw. für die Entscheidung eines an den Standort angepassten Filialtyps herangezogen werden.

Die Entscheidungsbäume wurden für jedes der in *Kapitel 7.1.2* ausgewiesenen Cluster separat erstellt und besitzen somit eine clusterspezifische Gültigkeit, d. h. die Baummodelle (= Kriterien für den Vertriebserfolg einer Filiale) gelten unter den Rahmenbedingungen des soziodemographischen Umfelds sowie der Wirtschaftskennzahlen des betreffenden Clusters.

Eine umfassende Aufstellung von möglichen Standortfaktoren, die hier von Relevanz sind, sind in *Kapitel 4* im Zusammenhang mit der Location Decision Scorecard erörtert.

Für den clusterspezifischen CHAID-Prozess musste im Vorfeld eine Auswahl an Prediktoren getroffen werden, um hier einige Ergebnisse beispielhaft und nachvollziehbar darstellen zu können[15].

- Lage (1A-Lage, mittlere Lage etc.)
- Anzahl Gewerbebetriebe
- Anzahl Privathaushalte
- Fluktuation
- Kennzahl Zahlungsverhalten
- Kennzahl Kundenqualität
- Anzahl hochwertige Einzelhändler
- Anzahl Kauf- und Warenhäuser

---

[14] Die Isodistanz von 500 m wurde hier als Optimum gewählt, da zum einen eine kleinräumigere Aggregation (z. B. auf 300 m) eine zu hohe Lageungenauigkeit und damit fehlerhafte Informationen bedingt, eine gröbere Aggregation (z. B. 1.500 m) zum anderen dem Anspruch der Fußläufigkeit im Filialumfeld widerspricht.

[15] Anzahl Haltestellen, Frequenzdaten und einige weitere relevante Marktdaten sind an dieser Stelle noch nicht in die Auswertung mit CHAID eingegangen, da sie aus unternehmensrechtlichen Gründen (Datenschutz) nicht verfügbar waren, erst zu einem späteren Zeitpunkt in der Datenbank vorlagen und/oder aufgrund der nicht geeigneten Form für die Voranalyse im GIS und die Weiterverarbeitung zur Verfügung standen.

- Anzahl Nahversorger
- Anzahl Händler Bücher und Zeitungen
- Anzahl Baumärkte
- Anzahl sonstige Einzelhändler
- Anzahl Teleshops
- Anzahl Verbrauchermärkte
- Anzahl Kreditbanken CASHGroup
- Anzahl Bausparkassen
- Anzahl Sparkassen
- Anzahl Kreditgenossenschaften
- Anzahl Landesbanken
- Anzahl Wettbewerber (Hermes, GLS, Pickpoints)

Am Beispiel des Clusters 9[16], für den die betreffenden Informationen für insgesamt 612 Filialen vorliegen, werden einige, nach mehrfachen Durchläufen gewonnene Ergebnisse vorgestellt. Der Filial- oder Vertriebserfolg ist die abhängige Variable, die es zu erklären gilt. Als begrenzende Baumaufbaukriterien (auch für die anderen Cluster) wurden eine maximale Anzahl von Ebenen unterhalb des Stamm(= Haupt)-Knotens von zehn festgesetzt, was der Hälfte der Prediktorvariablen entspricht. Ein weiteres Abbruchkriterium ist die Mindestanzahl der Fälle in den Knoten des Baumes, die sich nach der Eingangsfallzahl richtet. Wenn im Zuge des Baumaufbaus die Anzahl der Fälle in einem übergeordneten Knoten beispielsweise im Cluster 9 unter 40 sinkt, wird dieser nicht mehr verzweigt. Auch wenn beim Verzweigen wenigstens ein neues, untergeordnetes Segment mit weniger als 20 Fällen entsteht, wird der übergeordnete Knoten nicht weiter verzweigt. Als Schwellenwert für die automatische Selektion eines Prediktors wird wegen der generell geringen Fallzahl der Unterschied der Mittelwerte des Vertriebserfolgs in den Subgruppen auf einem Signifikanzniveau von $p < 0,1$ angenommen.

- Der erste Split (*Abb. 7-9*) erfolgt auf Basis des Merkmals *'Anzahl Gewerbebetriebe'* und führt zu fünf Untergruppen mit signifikant verschiedenen Mittelwerten ($p < 0,0001$, siehe Knotenstatistik am Split). Damit ist in Cluster 9 der Einfluss der Anzahl der Gewerbetriebe im 500 m-Filialumfeld wichtiger als der üblicherweise angenommene Einfluss der Wettbewerber. Vier dieser Segmente werden aufgrund der Merkmale *'Anzahl Wettbewerber (Hermes, GLS, Pickpoints)', 'Anzahl*

[16] Cluster 9: Hohe Bevölkerungsdichte, hoher Anteil Gewerbe, überdurchschnittlicher Ausländeranteil etc. *(siehe Kapitel 7.1)*

- *Sparkassen', 'Anzahl Händler Bücher/Zeitungen', 'Summe Privathaushalte', 'Anzahl Teleshops' bzw. 'Anzahl Kreditbanken CASHGroup*[17]' z.T. auf mehreren Stufen weiter ausdifferenziert; die übrigen Prediktoren spielen keine signifikante Rolle.
- Nach Greifen eines der angegebenen Abbruchkriterien lassen sich die 612 Filialen insgesamt in zwölf Endknoten ausdifferenzieren. So schließt Knoten 13 insgesamt 5,6 % der Filialen mit einem überdurchschnittlich hohen Vertriebserfolg ein (91.327), was statistisch mit dem Vorliegen von mehr als 424 Gewerbebetrieben und mehr als drei Teleshops im 500 m-Filialumfeld einhergeht. Dahingegen weisen die Filialen in Knoten 18 (7,8 % der 612) den niedrigsten mittleren Vertriebserfolg auf (3.928), der statistisch mit einer Anzahl von Gewerbebetrieben unter 169, mindestens einem Wettbewerber (*Hermes, GLS, Pickpoints*), mehr als einer Sparkasse und mit höchstens einem Händler Bücher/Zeitungen im 500 m-Umfeld der Filiale korrespondiert.

Die Charakteristik der übrigen Endknoten ist der folgenden Abbildungen *(Abb. 7-8, Abb.* 7-9*)* zu entnehmen.

Gewinnzusammenfassung für Knoten

Ranking	Knoten	N	Prozent	Mittelwert
1	13	34	5,6%	91327,349
2	12	28	4,6%	68365,942
3	11	20	3,3%	66601,570
4	4	60	9,8%	57834,722
5	17	21	3,4%	57521,045
	8	67	10,9%	42831,941
	16	82	13,4%	37032,831
	6	77	12,6%	20256,933
	14	29	4,7%	20069,158
	9	116	19,0%	18285,450
-	19	30	4,9%	9282,89963
-	18	48	7,8%	3928,68242

Aufbaumethode: CHAID
Abhängige Variable: Vertriebserfolg im Zeitraum x bis y

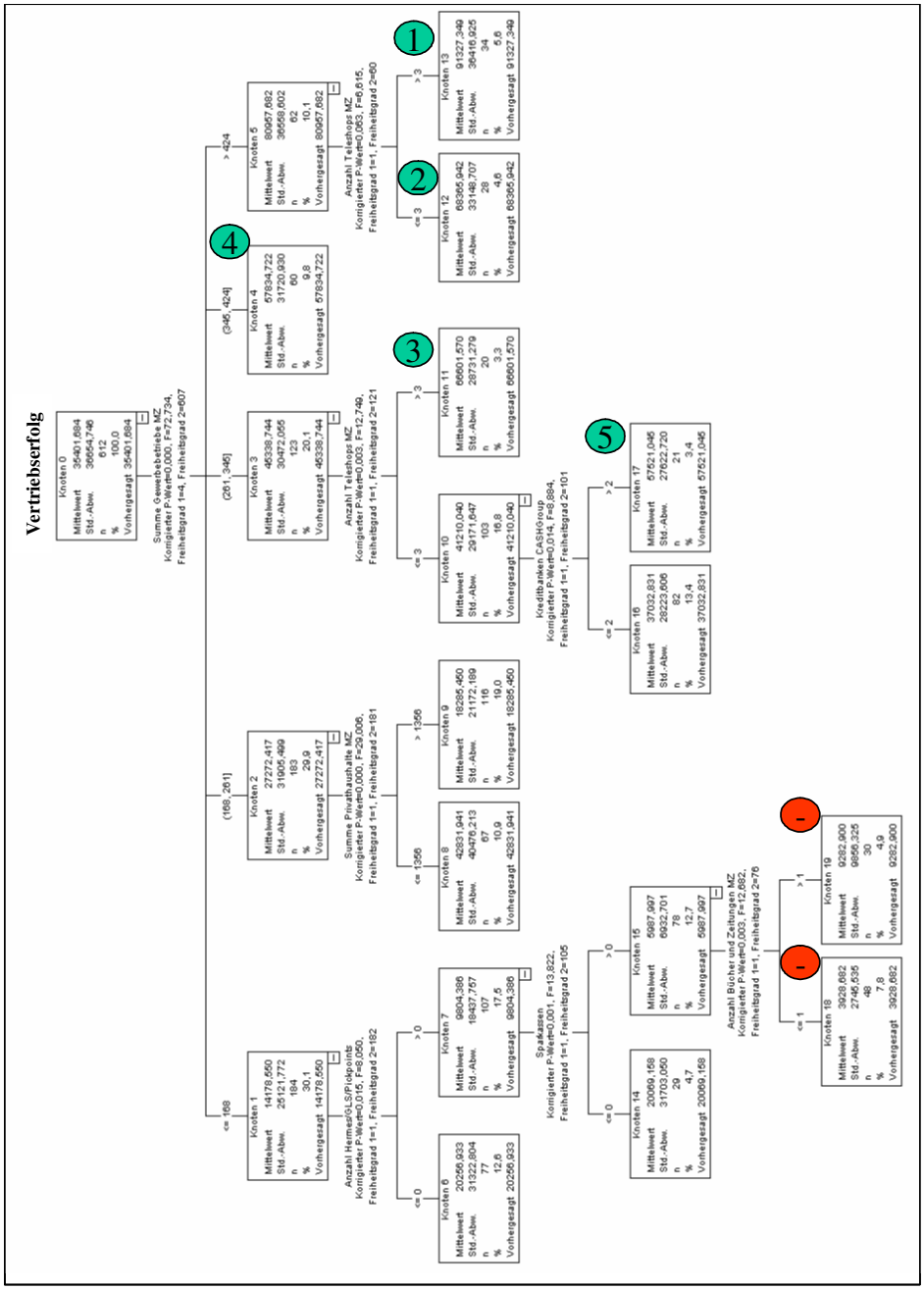
Quelle: eigene Berechnung (SPSS) [18]

Abb. 7-8: Gewinnzusammenfassung Endknoten CHAID-Modell

Filialen in Cluster 9

[17] Zur CASH-Group gehören die Commerzbank, Deutsche Bank, Dresdner Bank, HypoVereinsbank und die Postbank. Bei der Berechnung ist die Postbank nicht eingegangen, da diese für die Kannibalisierung bzw. den Eigenwettbewerb eine Rolle spielt.

[18] Der Vertriebserfolg wurde hier aufgrund des Unternehmensgeheimnisses in Punktewerte abgewandelt und bezieht sich auf den Zeitraum x bis y.



Quelle: eigene Berechnung (SPSS)

Abb. 7-9: Baumstruktur für CHAID-Modell der Filialen in Cluster 9

Die Baumstrukturen für die übrigen Cluster unterscheiden sich von den Ergebnissen in Cluster 9 hinsichtlich der Baumtiefe bzw. der Anzahl der Endknoten (als Konsequenz der Ausgangsfallzahl im Stammknoten) sowie der Zahl und Relevanz der aufgenommenen Prediktoren. Beispielsweise ist in Cluster 13 die Anzahl der hochwertigen Einzelhändler in 500 m-Isodistanz maßgebliches Splitkriterium, die Gewerbebetriebe hingegen spielen keine Rolle. Für die Cluster 14 und 15 hat die Anzahl der Banken einen größeren Einfluß auf den gesamten Vertriebserfolg als die der Gewerbebetriebe. In dem zweiten Split unterscheiden sich die beiden Cluster wiederum dahingegend, dass der Vertriebserfolg der Filialen in Cluster 14 insbesondere von Teleshops in der Umgebung beeinflusst wird, in Cluster 15 spielen jedoch die Anzahl von Privathaushalten im Umfeld eine größere Rolle.

Wird die Zielvariable insofern verändert, dass beispielsweise der Vertriebserfolg ohne bankbezogene Vertriebswerte berechnet wird, so ergeben sich veränderte Abhängigkeiten. Für Cluster 14 und 15 (bevölkerungsstarke städtische Gebiete - *siehe Beschreibung in Kapitel 7.1.2*) sind Kriterien wie hohe Bevölkerungsdichte, hohes Gewerbeaufkommen, Nähe zu Verbrauchermärkten und überdurchschnittlicher Anteil der Single- und Zwei-Personen-Haushalte relevant.

Die Ergebnisse von diesen Klassifikations- und Vorhersageregeln können auf bestehende und neue Daten, basierend auf den gefundenen Segmenten und Mustern, angewendet werden. Ziel ist es, die Ergebnisse der *CHAID-Analysen* zu nutzen, um auf deren Basis weitere Auswertungen mit der Grid-Methodik durchzuführen. Dabei werden die gefundenen Muster auf Gebiete mit ähnlicher Struktur innerhalb des jeweiligen Clusters übertragen (Greenfield-Analyse[19]). Dies wird beispielhaft für ausgewählte Gebiete aus den Clustern 14 und 15, welche die zusätzlichen Kriterien hohe Wettbewerberdichte und hohe Fußgängerfrequenz aufweisen durch Anwendung des GIS-basierten *GBI-Tools* berechnet (*Kapitel 9*)[20].

[19] Greenfield-Analyse wird hier definiert als Bestimmung von Gebieten, welche die zuvor berechneten Erfolgsfaktoren aufweisen und unbeeinflusst von bestehenden Filialen als Zielgebiete ausweisen.

[20] Aufgrund der Geheimhaltung der Unternehmensdaten werden hier nur beispielhaft Daten und Ergebnisse dargestellt.

7.2.3 Anwendung CHAID: Produktabsatz, Kunden- und Vertriebserfolgsanalysen

Ein weiteres Anwendungsfeld von CHAID für den vertrieblichen Bereich ist z. B. den Absatz von Produkten in direktem Zusammenhang mit soziodemographischen Variablen der Kunden zu berechnen.

Die Anwendung von CHAID ist weiterhin für die Beantwortung von Fragestellungen zu vielen anderen Vertriebsthemen geeignet: ob die Nähe von Bankenstandorten oder Einzelhandel Auswirkungen auf den Vertriebserfolg einer Filiale hat, ob zusätzlich der Erfolg mit zunehmender Einwohneranzahl der Gemeinde kontinuierlich steigt, ob sich der Erfolg signifikant in jedem Einwohner-Cluster unterscheidet in Abhängigkeit der Lage der Filiale in Alten oder Neuen Bundesländern. Es lässt sich bis zu einem bestimmten Grad feststellen, ob der Beratereinsatz in bestimmten Gebieten erfolgreicher ist als in anderen[21]. Entscheidend dabei ist immer die Datenqualität und -menge, die dem Entscheidungsbaum zur Verfügung gestellt werden kann. Hier gilt, umso mehr Daten, umso aussagekräftiger ist das Ergebnis, umso signifikanter sind die Gruppenunterteilungen.

Der Vertriebserfolg wird hier als abhängiger Parameter eingesetzt. Aus der Vielzahl der Marktdaten werden Daten im Umfeld einer Filiale, innerhalb einer 500 m Isodistanz oder Isochrone über GIS berechnet: Anzahl Banken, hochwertiger Einzelhandel, Einwohner, Altersstruktur.

Der CHAID-Prozess ist mehrfach durchlaufen worden unter Hinzunahme und im Austausch von Parametern:

Durch die Anwendung verschiedener Verfahren ist im Vorfeld eine Auswahl an Parametern getroffen worden, um hier einige Ergebnisse beispielhaft und nachvollziehbar darstellen zu können.

Die Parameter, die in die clusterspezifischen Entscheidungsbaum-Analysen der Erfolgsfaktoren in unterschiedlichen Iterationsstufen eingehen, sind:

- **Unternehmensdaten:** Vertriebserfolg einzelner Produktlinien und gesamter Vertriebserfolg, Anzahl Berater, Anzahl Mitarbeiter, offene Schalterstunden, Filialtyp

[21] Verständlich ist natürlich auch, dass viele interne Faktoren eine Rolle für den Vertriebserfolg spielen. Viele dieser Faktoren, die digital vorliegen, lassen sich auch in das Modell einbinden, können hier aber aus Geheimhaltungsgründen nicht veröffentlicht werden.

- **Soziodemographische Marktdaten:** Gemeindeklassifizierung (Metropole, Großstadt, Kleinstadt etc.) oder Einwohnerzahl, Altersstruktur, Anteil Ausländer, Mobilität, vorwiegender Soziotyp
- **Ökonomische Marktdaten:** Anzahl der Einzelhändler gesamt, Nahversorger, hochwertige Einzelhändler, Anzahl der Banken gesamt im Umfeld von 500 m und 300 m , weitere Wettbewerber, Alte und Neue Bundesländer, Typ der Filiale, Lage (1-A Lage, mittlere Lage etc.) und einzelhandelsrelevante Kaufkraft je Einwohner, Umsatzkennziffer, Bebauungstyp
- **Kundendaten:** Einkommen, Einlagevolumen, Depotvolumen, Alter, Bonität, Wohnort, Geschlecht, Beruf, Ausbildung, Familienstand, Produktnutzung, Dauer der Kundenbeziehung, Internetnutzung u.w.

Die Durchführung der *CHAID-Analyse* zeigt die Zusammenhänge zwischen Parametern der Kunden und den Produktabsätzen in bestimmten Regionen innerhalb eines definierten Zeitraumes auf.

Diese Methodik ist auch anwendbar, wenn nur die Adresse des Kunden vorliegt. Über die Adresszuordnung können die Kennzahlen, die ebenfalls auf die Adresse zutreffen, genutzt werden, um Zusammenhänge zwischen Absatz und Gebietsfaktoren aufzudecken. Über die Kundenadresse ist die räumliche Verbindung zu den übrigen Daten gegeben. So wird festgestellt, welcher vorwiegende Soziotyp (Lifestyletyp, Alterstruktur etc.) in den Regionen herrscht und wo der Produktabsatz besonders hoch ist. Auch hier können anschließend durch eine Greenfield-Analyse weitere Gebiete oder Straßenabschnitte gefunden werden, die potentielle Kunden für das Produkt aufzeigen. Diese Methodik ist für das Direktmarketing sehr von Nutzen, da hier durch adressgezieltere Mailings hohe Streuverluste vermieden werden können und damit Kosten eingespart werden.

Bezüglich einer klar definierten Zielgruppe wird das Marketing auf deren Bedürfnisse, Anforderungen und Eigenheiten fokussiert. Gruppen wie Singles mit hohem Kaufkraftniveau werden von vornherein andere Waren angeboten als Familien. Diese Daten werden z. B. für Cross-Selling-Mailaktionen genutzt. Gleichzeitig können abwanderungswillige Kunden anhand des Umsatzrückgangs festgestellt und speziell in

Aktionen beworben werden. Dies ist nur ein Ausschnitt des zielgruppenorientierten *'Database Marketing'*[22].

Für den Bereich des Vertriebserfolgs kann nur exemplarisch[23] auf die Erkenntnisse, welche die CHAID-Anwendung erbracht hat, eingegangen werden - hier zum Thema Rolle von Banken und Einzelhandel und Rolle von Alten und Neuen Bundesländern:

- Der Vertriebserfolg ist in den Filialen mit dem höchsten Beraterstamm am größten.
- Hier teilen sich die Gruppen nochmals in diejenigen auf, in deren 500 m Umfeld besonders viele Banken liegen. Davon wiederum haben die Filialen mit mehr als zehn Bankstandorten im Umfeld den höchsten Vertriebserfolg. Wenn keine Bankstandorte in der Umgebung vorhanden sind, dann sind die Filialen am erfolgreichsten, die in Großstädten liegen und im Umfeld besonders viel Einzelhandel verzeichnen können.
- Bei den Filialen, die über keine Berater verfügen, ist der Faktor Einzelhandel im Umkreis am signifikantesten. Am schlechtesten schneiden die Filialen ab, die keine Berater haben und in Großstädten liegen oder die keine Berater haben und sehr wenig Einzelhandel im Umkreis aufweisen können.
- Filialen in den Metropolen und Kleinstädten sind signifikant erfolgreicher als die in Mittelstädten.
- Eine Differenzierung zwischen alten und neuen Bundesländern ist hier nicht relevant.
- Filialen, die zwar keine Berater haben, aber in den Neuen Bundesländern liegen, insbesondere in Klein- oder Mittelstädten schneiden signifikant besser ab als vergleichsweise die Filialen in den Alten Bundesländern ohne Berater.

Diese Ergebnisse aus den *CHAID-Analysen* sind wiederum Grundlage für weitere Untersuchungen und Entscheidungen.

---

[22] Weitere Ansätze finden sich bei SCHÜSSLER (2000, S. 165-171) und zu unterschiedlichen Anwendungsmöglichkeiten von Scoring-Modellen siehe NITSCHE (1998, S. 98ff).

[23] CHAID-Tabellen siehe Anhang

7.2.4 Anwendung CHAID: Wettbewerberanalysen

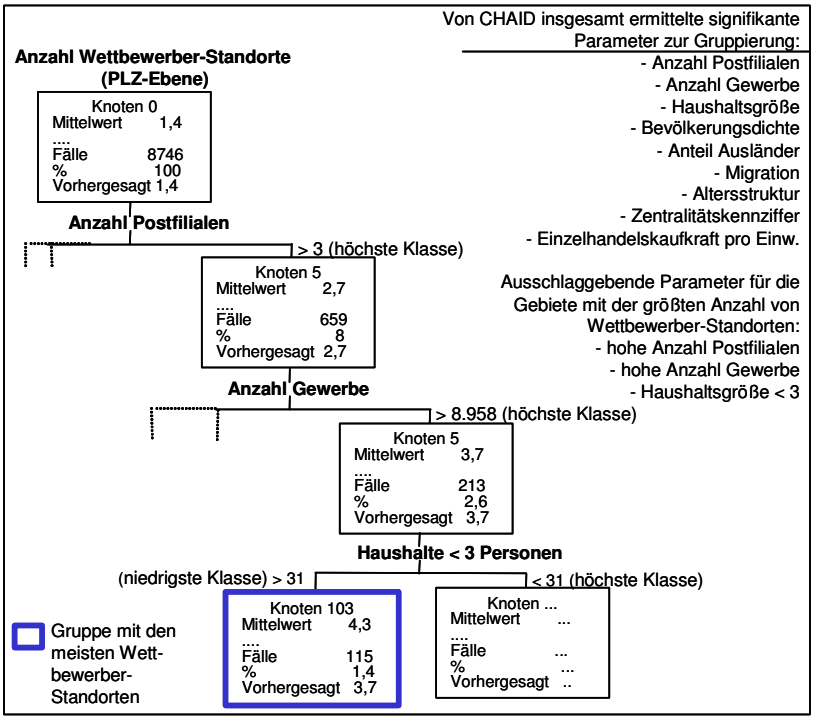
Weiterhin bietet die Methode im Zusammenhang mit Wettbewerberanalysen einen relevanten Mehrwert. Der Zusammenhang zwischen Vertriebserfolg und Wettbewerberdichte kann hier hergestellt werden. Eine weitere Betrachtung, kann der Blick auf die Strategie des Wettberwerbers in der Verteilung seiner Filialen sein. Sind Zusammenhänge erkennbar, die darauf schließen lassen wie sich die Ausdehnung weiter vollzieht und wohin der Wettbewerb weiter wachsen wird?

Als Beispiel soll hier der Wettbewerber *Hermes Logistik Group*[24] mit derzeit 11.000 *Hermes PaketShops* betrachtet werden. Die Analyse auf makroräumlicher Ebene mit den Zeitreihen von 2002 bis 2006 ergibt als Ergebnis, dass *Hermes* insbesondere in den Metropolen und Großstädten eine hohe Standortdichte hat. In etwa 80 % der Städte hat *Hermes* mehr Standorte als die Deutsche Post. Dies betrifft vor allem die Metropolen wie Berlin, München, Hamburg und Großstädte über 200.000 Einwohner wie Köln, Dortmund und weitere. Das größte Delta an Standorten weist Berlin auf: In Berlin unterhält *Hermes* mit rund 600 Standorten die dreifache Anzahl im Vergleich zu Postfilialen. Wogegen gerade in den Klein-, Landstädten und Landgemeinden *Hermes* unterdurchschnittlich vertreten ist. Hier hat die Post den Infrastrukturauftrag zu erfüllen, *Hermes* nicht.

Eine Analyse des Umfeldes der Filialen und der Wettbewerber mit möglichst umfangreichen Marktdaten wird im vornhinein im GIS durchgeführt. Mit diesem ersten Ergebnis wird eine *CHAID-Analyse* durchgeführt. Die Anzahl der Wettbewerber-Standorte ist die abhängige Größe. Als Ergebnis erhält man eine anhand von Merkmalen getroffene Charakterisierung von Gebieten, in denen sich vornehmlich Filialen des Wettbewerbers befinden.

---

[24] Die Hermes Logistik Gruppe ist einer der größten postunabhängigen Logistik-Dienstleister Deutschlands bei der Zustellung an Privatpersonen. Im Bereich Brief- und Infoservice arbeitet die Gruppe mit primeMail - einem 50:50-Joint Venture mit der Swiss Post International - und einer 29%-igen Beteiligung an der TNT Post (ehemalig Europost) zusammen (*www.hermes-logistik-gruppe.de; Stand: 10/2006).*



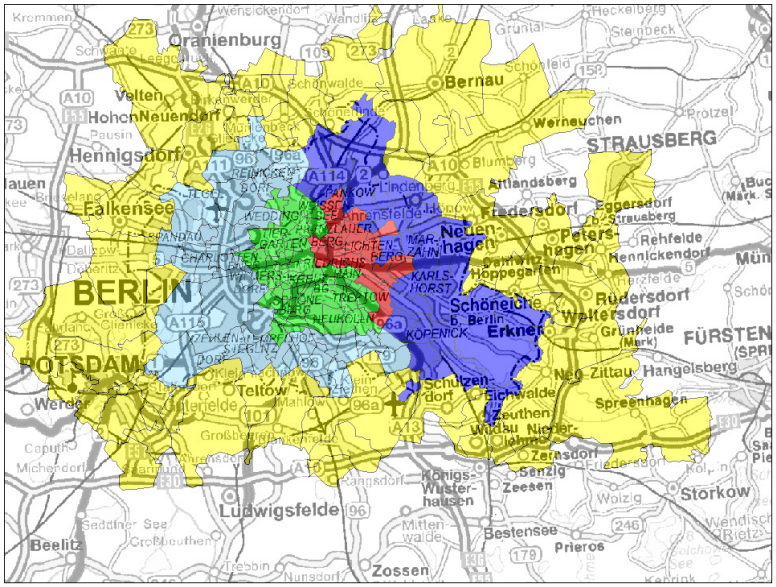
Quelle: eigene Berechnung (SPSS)

Abb. 7-10: CHAID-Analyse: Teilergebnisse Wettbewerberanalysen (vereinfacht)

Unter Berücksichtigung des Deltas der Zeitreihen im Wachstum der Wettbewerberstandorte von einem zum anderen Jahr ist festzuhalten, dass insbesondere Standorte wie *Hermes* dort verstärkt von 2003 bis 2005 zugenommen haben, wo bereits mehrere Postfilialen vorhanden sind, die Anzahl von Gewerbetreibenden hoch und die Haushaltsgröße eher klein ist. Die einzelhandelsrelevante Kaufkraft spielt hier - wie zu erwarten war - keine Rolle, denn die Produkte sind für jeden unabhängig vom Einkommen gleichermaßen relevant. Weiterhin sind die Zuzugsgebiete überdurchschnittlich von Filialeröffnungen betroffen. Der Ausländeranteil ist in der Umgebung der *Hermes*-Standorte eher hoch.

Werden weitere Informationen zu den einzelnen Wettbewerbern hinzugezogen, lässt sich von den bisherigen Ergebnissen ableiten, dass eine der Expansionsstrategien ist, in unmittelbarer Umgebung von bestehenden Clustern von Postfilialen eigene Dienstleistungen in bestehenden Einzelhandelsläden anzubieten. Des Weiteren lässt sich festhalten, dass zusätzlich vornehmlich in Metropolen und Großstädten, in denen die Post im Verhältnis zu *Hermes* unterdurchschnittlich vertreten ist, ein Netz von Standorten aufgebaut wird, um diese ungenutzten Potentiale abzuschöpfen.

Im Weiteren wird Berlin betrachtet, als die Metropole, in der die negative Abweichung der Standorte der Post im Gegensatz zu denen von *Hermes* am stärksten ist. Da Berlin ein sehr großes heterogenes Gebiet umfasst, ist eine Verfeinerung z. B. durch eine Clusterung für eine bessere Vergleichbarkeit der Analysen sinnvoll. Einerseits kann auf die 16 er Clusterung zurückgegriffen werden, diese bezieht sich aber auf das gesamte Bundesgebiet Deutschland. Bei Betrachtung eines spezifischen Raumes kann es durchaus sinnvoll sein, maßgebliche Kriteren für dieses Gebiet für eine Clusterung heranzuziehen. Eine Trennung erfolgt hier in bevölkerungsstarke Innenbezirke, die nochmals in Ost- und Westbezirke unterteilt werden, da sich diese Gebiete bis heute in vielen Kriterien unterscheiden (siehe BRAUN &TIEFELSDORF 1998).



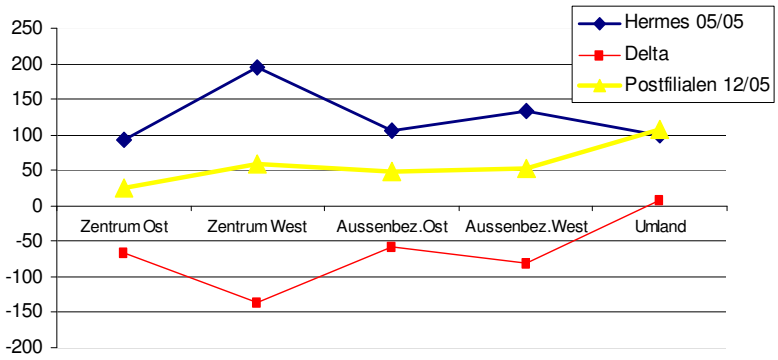
Quelle: eigene Berechnung (Filialinfo)

Abb. 7-11: Clusterung Berlins nach Bevölkerungsdichte und Zentrum Ost und West

Die PLZ-Gebiete Berlin und Umland werden nach innerstädtischen Bereichen (definiert anhand hoher Bevölkerungsdichte in zusammenhängenden PLZ-Gebieten) und Außenbezirken getrennt nach Ost und West klassifiziert. Die umliegenden PLZ-Gebiete Berlins bilden ein weiteres Cluster, so dass mögliche *'Speckgürtel'* Effekte berücksichtigt werden *(Abb. 7-11)*. Zur weiteren mikroräumlichen Analyse wird auf das 16er Cluster zurückgegriffen. Ein Vergleich zwischen den Clustern kann weitere Aufschlüsse über die Standortstrategie von *Hermes* ergeben.

Durch Umfeldanalysen (500 m Isodistanz) für *Hermes*-Standorte werden Indices gebildet, um einen Vergleich der Standorte untereinander und im nächsten Schritt auch Empfehlungen für Standorte für Postfilialen (z. B. *POSTPOINTS)* zu erhalten. Eine Verifizierung der Ergebnisse erfolgt über Standortbegehungen und durch Regionalkenntnisse.

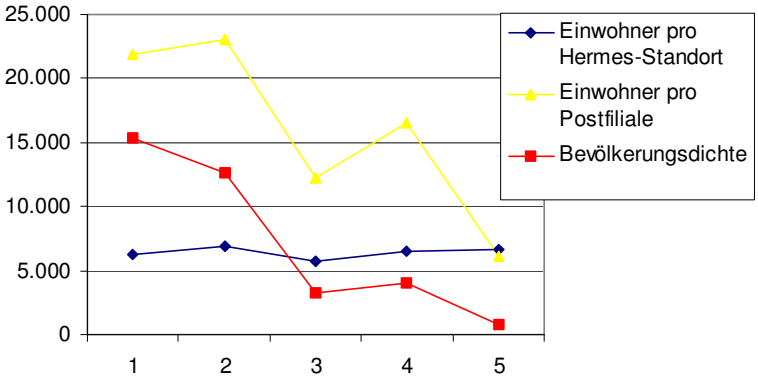
Die folgende Abbildung (*Abb. 7-12*) zeigt, dass *Hermes* in den westlichen Bezirken mit einer höheren Anzahl an Standorten vertreten ist als in den östlichen Teilen. Die Post hat eine nahezu identische Anzahl an Filialen in den Innenbezirken West und den Außenbezirken Ost und West. In den Innenbezirken Ost ist die Post mit geringerer Filialanzahl vertreten. Im Umland übertrifft die Post *Hermes* anhand ihrer Filialanzahl. Die Verteilung *Hermes*-Standorte und Postfilialen ist sehr unterschiedlich. Insbesondere in den Innenbezirken West (Zentrum West) ist die Differenz gravierend. Für weitere Analysen wird ein Gebiet aus dem Innenbezirk West ausgewählt, bei dem das Delta *Hermes* zu Postfilialen besonders hoch ist: der Neuköllner Raum. Dieses Gebiet fällt nach der 16 er Clusterung in das Cluster 14 (hohe Bevölkerungsdichte, hoher Ausländeranteil).



Quelle: eigene Berechnung

Abb. 7-12: Anzahl Standorte und Delta pro Cluster

In bevölkerungsstarken PLZ-Gebieten ist *Hermes* im Gegensatz zur Post entsprechend stark vertreten. Das Verhältnis Einwohner pro Standort ist bei *Hermes* innerhalb der Metropole sowohl in den Innen- als auch in den Außenbezirken kundenorientiert gestaltet. *Hermes* versucht mit seiner Standortpolitik möglichst die Anzahl der Einwohner pro Standort identisch zu halten (*Abb. 7-13*): 6.400 Einwohner pro Standort (Berlin und Umland). Die Post erreicht nur im Umland Berlins eine Abdeckung von etwa 6.100 Einwohnern pro Filiale. Innerhalb von Berlin liegt die Abdeckung bei rund 18.000 Einwohnern pro Filiale. Verfolgt die Post die Strategie standortmäßig mit dem Wettbewerb mitzuhalten, ist es wichtig, eine weitere Präsenz in Form bestimmter Filialtypen wie *POSTPOINTS* in den genannten Gebieten zu zeigen.



Quelle: eigene Berechnung

Abb. 7-13: Einwohner pro Standort pro Cluster

Hermes ist demnach sowohl im Cluster Innenbezirk West sehr stark vertreten als auch generell im Cluster 14. Eine Übertragung der Merkmale auf die Gebiete gleichen Clusters ist demnach gegeben. Die weiteren mikroräumlichen Analysen werden dieses in den folgenden Kapiteln vertiefend aufzeigen (*Kapitel 8 und 9*).

Die Erfolgsfaktoren für *Hermes*-Standorte, die bei der makroräumlichen Analyse über die *CHAID-Analysen* bestätigt wurden, fließen in die *Greenfield-Analyse* für die Findung für neue Standorte, in denen Postdienstleistungen angeboten werden (z. B. *POSTPOINTS),* ein. Hier wird die Grid-Methodik verwendet. Sie ermöglicht mathematische Kombination verschiedenster Faktoren, um so theoretisch beste Standorte ausfindig machen zu können (*siehe Kapitel 8.1 und Kapitel 9*).