Individuelle Clusterung, oder: Womit kann man Bayern vergleichen?

Martin Westphal
Bauer Systems KG – Bauer Media Group
Burchardstraße 11
Hamburg
martin.westphal@bauerverlag.de

Nicolai Brauns
Bauer Systems KG – Bauer Media Group
Burchardstraße 11
Hamburg
nicolai.brauns@bauerverlag.de

Sergej Steinberg
Bauer Systems KG – Bauer Media Group
Burchardstraße 11
Hamburg
sergej.steinberg@bauerverlag.de

Zusammenfassung

Innerhalb unserer Logistikstrukturen werden regelmäßig strukturelle und organisatorische Veränderungen vorgenommen, um die Qualität der Zustellung hauseigener und externer Printprodukte zu erhöhen. Um die Wirksamkeit dieser Maßnahmen beurteilen zu können, besteht die Notwendigkeit, den Gebieten in denen diese Strukturtests durchgeführt werden, vergleichbare Gebiete zuzuordnen. In diesem Beitrag wird ein Verfahren vorgestellt, das diesen Anforderungen gerecht wird.

Schlüsselwörter: Data Mining, Clusteranalyse, Business Intelligence, GIS

1 Einleitung

1.1 Bauer Media Group

Die Bauer Media Group ist mit über 300 Zeitschriften in 15 Ländern eines der größten Medienhäuser in Europa. Dabei ist Bauer in den reichweitenstarken Segmenten der Programm-, Frauen- und Jugendzeitschriften Marktführer in Deutschland. Nach einer Studie der Arbeitgemeinschaft Media-Analyse e.V. erreicht die Bauer Media Group mit ihren Zeitschriften jeden zweiten Deutschen (vgl. [1]).

1.2 Das Logistiknetzwerk der Bauer Media Group

Für die Zustellung von Zeitschriften und anderen Printprodukten unterhält die Bauer Media Group seit ca. 40 Jahren ein eigenes postalternatives Logistiknetzwerk. Über dieses Netzwerk werden wöchentlich ca. 1,8 Mio. Zeitschriften-Abonnements und bis zu 5 Mio. Sendungen externer Kunden zugestellt. Um diesen Service zu ermöglichen, wird

eine leistungsstarke und komplexe Logistikstruktur mit ca. 40.000 Zustellern vorgehalten (vgl. Abbildung 1).

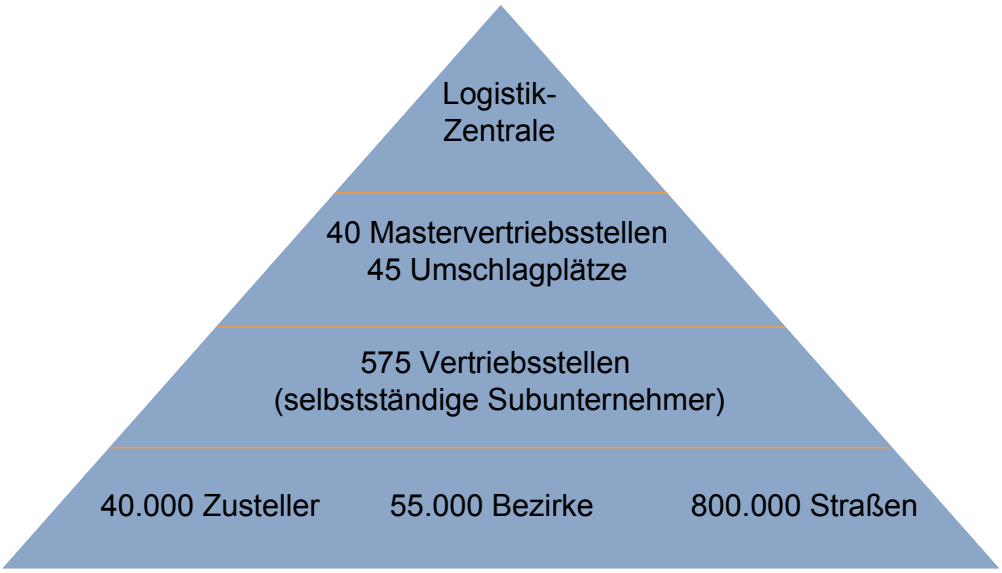

Abbildung 1: Struktur des Logistiknetzwerks der Bauer Media Group

In den Strukturen werden regelmäßig strukturelle und organisatorische Tests vorgenommen, um die Zustellqualität zu messen und zu steigern. Hieraus entsteht die Notwendigkeit den Gebieten, in denen Tests durchgeführt werden, ähnliche Gebiete zuzuordnen. Diese werden dann zum Vergleich herangezogen, um die Wirkung der Maßnahmen beurteilen zu können.

Als Lösung dieses Problems bietet es sich an, Gebiete zu Clustern von jeweils einander ähnlichen Gebieten zu gruppieren.

2 Motivation

Die Bildung von Clustern setzt voraus, dass die zu betrachtenden Objekte eine oder mehrere Eigenschaften besitzen, die das Objekt hinlänglich beschreiben. Die Eigenschaftsausprägungen sollen dabei in Form von Vektoren dargestellt werden können. Bei Betrachtung geografischer Gebiete kann dies z.B. die Fläche, die Bevölkerungsdichte oder das Durchschnittseinkommen der Einwohner des Gebietes sein. Der erste Schritt stellt somit die Auswahl von auf den Anwendungsfall bezogenen, unabhängigen Variablen dar.

Die Vektoren der Eigenschaftsausprägungen können in einem Vektorraum als Punktwolke dargestellt werden. Die Ähnlichkeit wird dabei bei vielen Clusterverfahren (z.B. k-Means-Clustering) über ein Distanzmaß bestimmt, welches die Zuordnung der Objekte zu Clustern bestimmt. Ähnliche Objekte werden so zu Gruppen zusammengefasst.

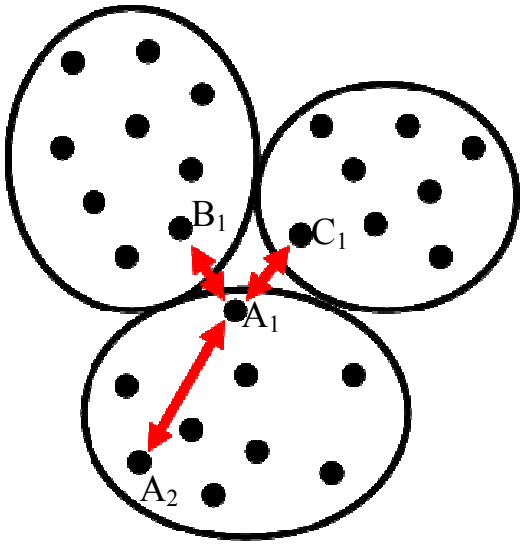

Abbildung 2: Zuordnung von Punkten zu Clustern

Somit sind die Objekte in Cluster A einander aufgrund ihrer Eigenschaften ähnlich und damit vergleichbar. Jedoch zeigt Abbildung 2, dass das Objekt A_1 den Objekten B_1 und C_1 näher liegt als Objekt A_2 im gleichen Cluster A.

Dies ist auf die grundsätzliche Eigenschaft von Clusterverfahren zurückzuführen, dass sie alle Objekte gleichberechtigt behandeln und jedes Objekt genau einem Cluster zugeordnet wird. In unserem Fall ist es wichtig, jeweils für einzelne Punkte eigene Umkreise zu bilden (daher sprechen wir von individueller Clusterung). Es ist auch durchaus möglich (je nach Testart), dass die Umkreise auch nach unterschiedlichen Regeln gebildet werden müssen.

Es gibt eine weitere Besonderheit in der Aufgabenstellung: Die Gebiete sind i.d.R. nicht homogen, das heißt, es handelt sich um „Makrogebiete“, die sich in „Mikrogebiete“ (z.B. PLZ-Gebiete) unterteilen lassen. Diese Zusammensetzung sollte auch berücksichtigt werden, da die „Mikrogebiete“ sowohl unterschiedliche Gewichtungen als auch jeweils eigene Charakteristiken besitzen.

3 Bildung individueller Gebietscluster

Eine Lösung dieses Problems ist, individuelle Gebietscluster mit jeweils einem Element als Mittelpunkt zu bilden. Für diese werden jeweils spezifische Distanzen zu allen anderen Elementen errechnet. Das Vorgehen lässt sich wie folgt beschreiben:

Es gibt N Variablen x_i, die L Mikrogebiete (z.B. PLZ-Gebiete) beschreiben. Damit liegen für alle Mikrogebiete die Variablen x_{il} mit $i=1...N$, $l=1...L$ vor. Ein Makrogebiet setzt sich dabei aus h_l Mikrogebieten zusammen. Demnach lassen sich die L Mikrogebiete M Makrogebieten $m_1...m_M$ zuordnen.

Für den Fall, dass ein Mikrogebiet zwei oder mehr Makrogebieten zugeordnet werden kann, wird dieses Gebiet und dessen quantitative Eigenschaften (z.B. Anzahl der Haushalte) nach einer zu definierenden Regel geteilt. So würde die in Abbildung 3 markierte PLZ sowohl dem Makrogebiet C als auch dem Makrogebiet D zugeordnet werden.

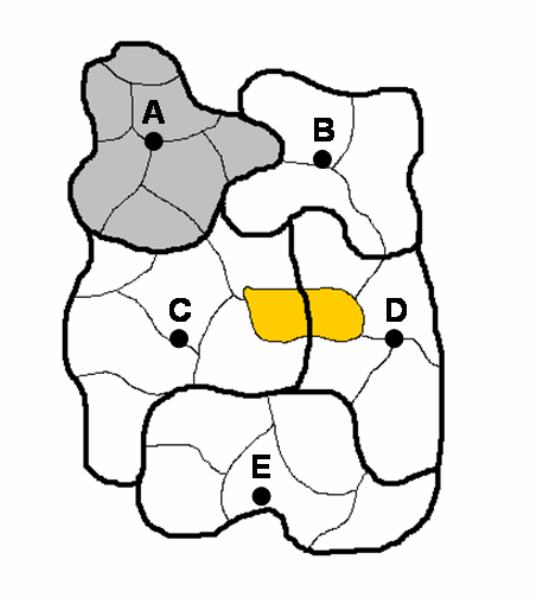

Abbildung 3: Zusammensetzung der Makrogebiete aus Mikrogebieten und Teilung eines Mikrogebietes

Es werden nun die Mittelwerte und Standardabweichungen je Variable und Makrogebiet berechnet.

$$X_i^m = \frac{\sum_{l \ni m} x_{il} * h_i}{\sum_{l \ni m} h_l} \quad m = 1...M \qquad s_i^m = \sqrt{\frac{\sum_{l \ni m} (x_{il} - X_i^m)^2 * h_i}{\sum_{l \ni m} h_l}} \quad m = 1...M$$

Diese Werte werden anschließend mittels Minimums- und Maximumswert der Variablen normiert.

$$X_i^m = (X_i^m - xx_i)/(XX_i - xx_i) \qquad \text{wobei} \qquad xx_i = \min_{l=1...L} x_i \qquad XX_i = \max_{l=1...L} x_i$$

Über eine Distanzfunktion sowohl für die Mittelwerte als auch die Standardabweichungen wird für alle Gebietspaare die individuelle Distanz zueinander ermittelt.

$$D1_{kp} = \sqrt{\frac{\sum_{i=1}^{N} (X_i^k - X_i^p)^2}{N}} \qquad D2_{kp} = \sqrt{\frac{\sum_{i=1}^{N} (s_i^k - s_i^p)^2}{N}} \quad k, p = 1...M$$

Abschließend werden die Distanzen durch Gewichtungsvariablen α, β und λ zu einem einheitlichen Distanzmaß zusammengeführt.

$$D_{kp} = \sqrt[1+\lambda]{\frac{\alpha * D1_{kp} + (1-\alpha) D2_{kp}}{\beta}}$$

Das Verfahren wurde in SAS realisiert und brachte für den angeforderten Test die in Tabelle 1 dargestellten Ergebnisse. Es wurden jedem Testgebiet jeweils diejenigen drei Gebiete zugeordnet, welche die geringste Distanz zum jeweiligen Testgebiet aufweisen.

Tabelle 1: Anwendung des Verfahrens in der Logistikstruktur

Testgebiet	Top 3 Vergleichsgebiete	Distanz Top 3	Distanz zum Rest
VS…1	VS…1a, VS…1b, VS…1c	0,16	0,57
VS…2	VS…2a, VS…2b, VS…2b	0,20	0,63
VS…3	VS…3a, VS…3b, VS…3c	0,13	0,65

Abbildung 4 zeigt die Ergebnisse in der Kartendarstellung. Dabei sind die Testgebiete jeweils mit einem Kreis markiert und zugehörige Vergleichsgebiete jeweils gleich eingefärbt. Wie man erkennen kann, sind „ähnliche“ Gebiete nicht unbedingt auch immer „physische“ Nachbarn.

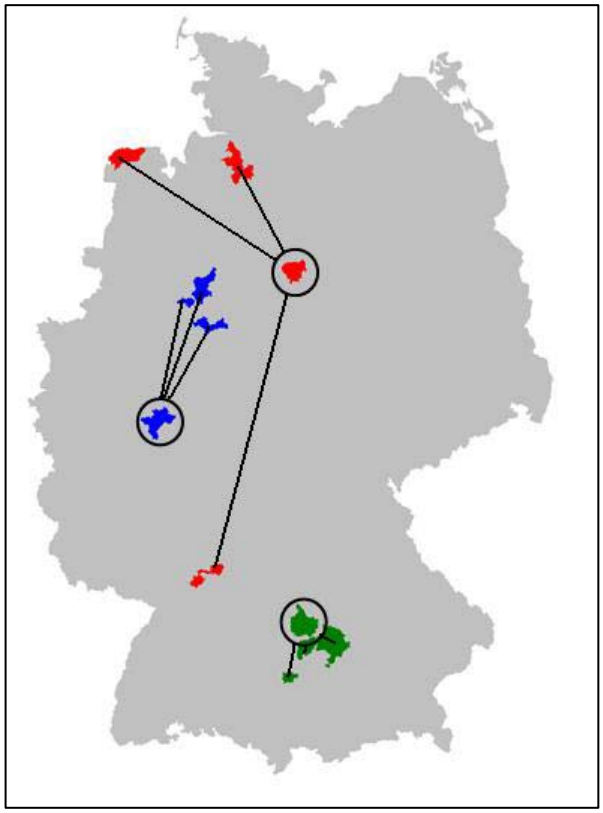

Abbildung 4: Kartendarstellung der Anwendung des Verfahrens in der Logistikstruktur

4 Validierung

Zur Validierung des Verfahrens wurde auf bereits bestehende VDZ-Clusteraufteilungen (vgl. Abbildung 5) für die Grossogebiete des Einzelhandels zurückgegriffen (vgl. [3]). Das Verfahren wurde auf die einzelnen Grossogebiete angewandt und die Distanz jedes Gebietes zu jedem anderen berechnet. Tabelle 2 zeigt, dass die Gebiete aus gleichen Clustern deutlich näher beieinander liegen als die der restlichen Cluster.

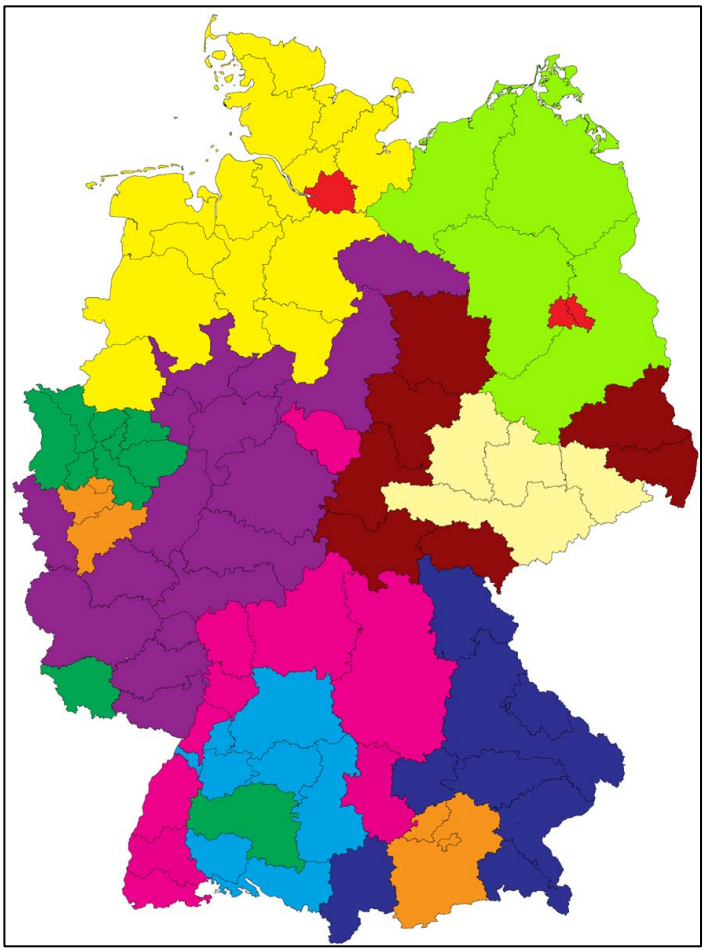

Abbildung 5: Karten der Grossogebietscluster

Tabelle 2: Validierung über Grossogebietscluster

Clustername	Anzahl Grossisten	Distanz innerhalb	Distanz zu anderen
Urban 1	3	0,63	0,75
Urban 2	6	0,45	0,62
Nord	14	0,45	0,59
Mitte 1	9	0,43	0,60
Mitte 2	14	0,43	0,58
Südwest 1	8	0,38	0,57
Südwest 2	8	0,43	0,58
Südost	8	0,36	0,60
Ost Nord	4	0,40	0,72
Ost Süd 1	4	0,32	0,71
Ost Süd 2	7	0,39	0,72

5 Zusammenfassung

In diesem Beitrag wurde eine pragmatische Lösung des Problems erörtert, wie man zu spezifischen geografischen Gebieten vergleichbare Gebiete ermitteln kann.

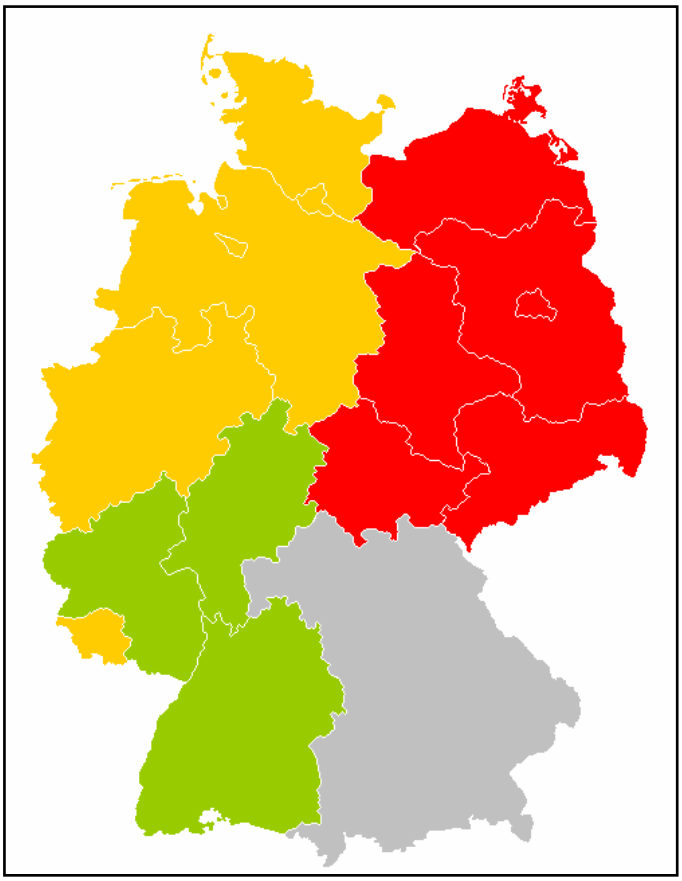

Abbildung 6: Womit kann man Bayern vergleichen?

Tabelle 3: Womit kann man Bayern vergleichen?

Bundesland	Distanz
Bayern	0,0
Baden-Württemberg	0,1
Rheinland-Pfalz	0,2
Hessen	0,2
Saarland	0,3
Nordrhein-Westfalen	0,3
Niedersachsen	0,3
Schleswig-Holstein	0,3
Bremen	0,4
Hamburg	0,4
Berlin	0,5
Sachsen	0,6
Thüringen	0,6
Mecklenburg-Vorpommern	0,7
Sachsen-Anhalt	0,7
Brandenburg	0,7

Das Verfahren zeichnet sich insbesondere dadurch aus, eine individuelle Lösung für jedes Makrogebiet zu bestimmen. Des weiteren können die Einflussparameter für jedes Gebiet jeweils neu ausgewählt werden. Das Verfahren berücksichtigt auch die Zusammensetzung der Makrogebiete aus Mikrogebieten und deren spezifische Eigenschaften.

Das Distanzverfahren lässt sich auf beliebige ähnliche Fragestellungen anwenden. So kann die individuelle Clusterung beispielsweise auch für die Bundesländer durchgeführt werden (vgl. Abbildung 6 und Tabelle 3).

Literatur

[1] ag.ma – Arbeitsgemeinschaft Media-Analyse e.V. (Hrsg.): ma 2009 Pressemedien I. Frankfurt a. M., 2009.

[2] R. Khattree, D.N. Naik: Multivariate Data Reduction and Discrimination with SAS Software. Cary, 2000.

[3] N. Brauns, S. Callsen, S. Steinberg: Methodik für die Schätzung der Distanz zwischen unterschiedlichen Clusterisierungen. Beitrag zu KSFE 2006.