Autor: Helmut Schels, Stadt Ingolstadt, Stadtplanungsamt

Vereinfachte Clusteranalyse mit Excel

Clusteranalyse mit Excel nach einer der hierarchischen Methoden (Single-Linkage) – Kurzbeschreibung

Hintergrund

Die Clusteranalyse ein statistisches Analyseinstrument, das auf einer umfangreichen theoretischen Grundlage basiert. Dazu gibt es noch eine ganze Reihe von verschiedenen Methoden, wie man die Cluster bilden kann.

Oft wird für das „Clustern“ das Softwareprogramm SPSS (Statistical Package for the Social Sciences) verwendet, das hierzu hervorragende Tools anbietet. Der Nachteil: der sinnvolle Umgang mit SPSS muss erst gelernt werden und die Software ist alles andere als billig. Für größere Städte, die sich eine oder einige Lizenzen kaufen und in denen sich ein oder wenige Mitarbeiter/innen auf SPSS spezialisiert haben, ist das kein Problem. Kleinere Städte und auch kleinere Großstädte haben oft keine SPSS-Spezialisten und der finanzielle Aufwand für die Beschaffung des Programms sowie für die Einarbeitung steht meist nicht in günstiger Relation zu den Ergebnissen. Aus diesem Grund wird nachfolgend eine einfache und in vielen Fällen (nicht in allen) praktikable Methode der Clusteranalyse mit Excel beschrieben.

Übersicht über die Vorgehensweise

1. Auswahl der Daten
   - Wissenschaftliche Begründung
   - Tabellarische Darstellung
   - Berechnung von Mittelwert und Standardabweichung
2. Standardisierung der Daten
   - Standardisierung der Daten
   - Daten standardisiert auf Mittelwert = 0 und Standardabweichung = 1
3. Berechnung der Euklidschen Distanz
   - Jeder Wert jeden Gebietes mit jedem Wert jeden anderen Gebietes
   - Bildung der quadrierten Differenz
   - Summe der quadrierten Differenzen = Euklidsche Distanz
4. Clusterbildung
   - Sortierung der Euklidschen Distanzen aufsteigend
   - Clusterbildung mit hierarchischem Verfahren

1. Auswahl der Daten

Bei der Auswahl der Daten sollten wissenschaftliche Begründung und Kenntnisse der empirischen Forschungsmethoden Voraussetzungen sein. Andernfalls können die Ergebnisse der Clusteranalyse zwar richtig berechnet und dargestellt, in ihrer Aussage jedoch falsch bzw. nichtssagend sein und zu falschen Schlüssen verleiten (was bei wissenschaftlicher Vorgehensweise natürlich auch passieren kann!).
Im nächsten Schritt werden die Daten in Excel tabellarisch dargestellt, in der Zeilenbeschriftung die Nummern oder Bezeichnungen der Gebietseinheiten oder Teilbereiche des Untersuchungsgegenstandes, in der Spaltenbeschriftungen die Bezeichnungen der Merkmale bzw. Variablen. Diese Tabelle kann man dann z.B. umbenennen in „Daten“.

Anschließend werden Mittelwert und Standardabweichung für jede Variable berechnet.
Formel für Mittelwert: =MITTELWERT(Wertereihe), z.B. =MITTELWERT (C2:C12)
Formel für Standardabweichung: =STABWN(Wertereihe), z.B. =STABWN (C2:C12)

Beispiel für Schritt 1 und 2: Daten zur Sozialraumtypisierung nach 11 Gebietseinheiten

SBZ	Stadtbezirk	Bev. mit Migrations-hintergrund	Anteil Geschie-dene	Anteil unter 18 Jahre	Anteil über 65 Jahre	Einwohner-entwicklung 1996-2006	Besiedlungs-dichte	Arbeits-losen-quote	Wohn-fläche je Ein-wohner
1	**Mitte**	32,4	8,3	14,2	18,6	8,6	1 132 E/km²	3,5	39,3 m²
2	**Nordwest**	68,0	7,5	18,5	19,4	-4,1	3 795 E/km²	6,0	29,8 m²
3	**Nordost**	51,8	7,5	17,4	19,0	2,3	4 004 E/km²	4,9	34,5 m²
4	**Südost**	38,1	7,0	18,0	19,3	3,9	1 105 E/km²	3,5	39,1 m²
5	**Südwest**	28,0	6,4	17,4	18,5	11,9	2 032 E/km²	2,6	41,9 m²
6	**West**	18,0	3,9	22,6	14,8	22,8	181 E/km²	1,4	42,6 m²
7	**Etting**	31,4	4,0	22,5	14,3	29,3	447 E/km²	2,0	40,3 m²
8	**Oberhaunstadt**	33,2	5,8	16,7	21,5	0,1	847 E/km²	2,4	39,8 m²
9	**Mailing**	29,5	5,0	19,1	17,7	15,2	571 E/km²	2,2	41,4 m²
10	**Süd**	19,7	4,4	20,4	14,5	26,1	291 E/km²	1,7	42,4 m²
11	**Friedr.-Hollerst.**	40,9	6,6	15,6	15,2	17,8	1 624 E/km²	3,1	39,7 m²
	Mittelwert	35,6	6,0	18,4	17,5	12,2	1 457 E/km²	3,0	39,2
	Standardabweichung	14,3	1,5	2,6	2,4	11,1	1 331 E/km²	1,4	3,8

2. Standardisierung der Daten

Die Daten liegen normalerweise in sehr unterschiedlichen Dimensionen vor, entsprechend groß sind die Unterschiede in Mittelwerten und Standardabweichungen. Im obigen Beispiel ist die Standardabweichung der Arbeitslosenquote nur 1,4, bei der Besiedlungsdichte 1 331 E/km².
Um jeder Variable das gleiche Gewicht in den folgenden Berechnungen zu geben, werden alle Zahlen standardisiert, d. h. alle Daten werden so umgerechnet, dass eine standardisierte Verteilung mit dem Mittelwert 0 und der Standardabweichung 1 entsteht.

Zu diesem Zweck kopiert man entweder das ganze Tabellenblatt von Excel oder kopiert die Datentabelle mit den Berechnungen von Mittelwert und Standardabweichung in ein leeres Tabellenblatt und benennt es praktischerweise um, z.B. in „Standardisierung“. Daraufhin löscht man alle Daten, nicht aber die Berechnungen von Mittelwert und Standardabweichung.

In die erste Zelle, in der Daten standen gibt man dann die Formel für die Standardisierung ein, die grundsätzlich lautet:
=STANDARDISIERUNG(Variable;Mittelwert;Standardabweichung)

Im obigen Beispiel würde die Formel so aussehen, wenn die Datentabelle „Daten“ heißt:
=STANDARDISIERUNG('Daten'!C2;'Daten'!C$14;'Daten'!C$15)

Der Bezug 'Daten'!C2 verweist auf den Originalwert der Bevölkerung mit Migrationshintergrund des Stadtbezirks 1-Mitte mit 32,4%.
Der Bezug 'Daten'!C$14 verweist auf den berechneten Mittelwert von 35,6.
Der Bezug 'Daten'!C$15 verweist auf die berechnete Standardabweichung von 14,3.
Das $-Zeichen vor der Zahl bewirkt, dass sich der Bezug zu Mittelwert und Standardabweichung immer auf die Zeilen 14 bzw. 15 bezieht, auch wenn die Formel nach unten kopiert wird. Beim Kopieren der Formel nach rechts in die nächsten Spalten passt Excel dagegen den Bezug auf den jeweiligen Mittelwert bzw. auf die jeweilige Standardabweichung in der Spalte an (da kein $ vor Spaltenbuchstaben).
Die Formel in dieser Zelle wird in alle Zellen des Wertebereichs kopiert (im Beispiel bis zum 11. Stadtbezirk und bis zur Variable „Wohnfläche je Einwohner“).
Wenn alle Formeln korrekt eingegeben wurden, muss beim Mittelwert jeweils eine 0 erscheinen, bei der berechneten Standardabweichung eine 1.

Beispiel für Schritt 3: Tabelle mit den standardisierten Daten zu obiger Beispieltabelle

SBZ	Stadtbezirk	Bev. mit Migrations-hintergrund	Anteil Geschie-dene	Anteil unter 18 Jahre	Anteil über 65 Jahre	Einwohner-entwicklung 1996-2006	Besied-lungs-dichte	Arbeits-losen-quote	Wohn-fläche je Ein-wohner
1	**Mitte**	-0,2	1,5	-1,7	0,5	-0,3	-0,3	0,4	0,0
2	**Nordwest**	2,4	1,0	0,1	0,8	-1,5	1,8	2,2	-2,6
3	**Nordost**	1,2	1,0	-0,4	0,6	-0,9	2,0	1,4	-1,3
4	**Südost**	0,2	0,7	-0,2	0,8	-0,8	-0,3	0,4	0,0
5	**Südwest**	-0,6	0,3	-0,4	0,4	0,0	0,5	-0,3	0,8
6	**West**	-1,3	-1,5	1,7	-1,2	1,0	-1,0	-1,3	0,9
7	**Etting**	-0,3	-1,4	1,6	-1,4	1,6	-0,8	-0,8	0,3
8	**Oberhaunstadt**	-0,2	-0,2	-0,7	1,7	-1,1	-0,5	-0,5	0,2
9	**Mailing**	-0,4	-0,7	0,3	0,1	0,3	-0,7	-0,7	0,6
10	**Süd**	-1,2	-1,1	0,8	-1,3	1,3	-0,9	-1,0	0,9
11	**Friedr.-Hollerst.**	0,4	0,4	-1,1	-1,0	0,5	0,1	0,0	0,2
	Mittelwert	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
	Standardabweichung	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0

3. Berechnung der Euklidschen Distanz

Die Euklidische Distanz ist die Summe aller quadrierten Distanzen für jede Variable zwischen zwei Gebietseinheiten. Um beispielsweise die Euklidische Distanz der Stadtbezirke 1 und 11 zu berechnen, muss für jede Variable der standardisierte Wert des Stadtbezirks 1 von demjenigen des Stadtbezirks 11 abgezogen werden und anschließend die Differenz quadriert werden (dadurch erhält man nur positive Werte). Anschließend werden diese quadrierten Differenzen (im Beispiel 8 Variablen, also 8 Werte) summiert. Im Beispiel der Stadtbezirke 1 und 11 ergibt sich dann eine Zahl von aufgerundet 4,8 als Euklidische Distanz, für die Stadtbezirke 2 und 11 sind es 25,9. Die Stadtbezirke 1 und 11 sind sich von den Daten der Sozialraum-typisierung her gesehen also sehr viel ähnlicher als die Stadtbezirke 2 und 11, deren Euklidische Distanz wesentlich höher ausfällt.

Das folgende Beispiel zeigt die Berechnung:

Berechnung der Euklidschen Distanz

(W1G1-W1G2)^2 + (W2G1-W2G2)^2 ….. + (W5G4-W5G11^)^2 … + (W8G10-W8G11)^2

W = Wert oder Variable (standardisiert); G = Gebiet (im Bsp. Stadtbezirk)

Standardisierte Werte Gebiet 1

Stadtbezirk (Gebiet)	Bev. mit Migrations-hintergrund	Anteil Geschie-dene	Anteil unter 18 Jahre	Anteil über 65 Jahre	Einwohner-entwicklung 1996-	Besiedlungsdichte	Arbeitslosenquote	Wohn-fläche je Ein-wohner
1	-0,2	1,5	-1,6	0,5	-0,3	-0,2	0,4	0,0
2	2,3	0,9	0,1	0,8	-1,5	1,8	2,1	-2,4
3	1,1	1,0	-0,4	0,6	-0,9	1,9	1,3	-1,2
4	0,2	0,7	-0,2	0,7	-0,7	-0,3	0,4	0,0
5	-0,5	0,3	-0,4	0,4	0,0	0,4	-0,3	0,7
6	-1,2	-1,4	1,6	-1,1	1,0	-1,0	-1,2	0,9
7	-0,3	-1,3	1,6	-1,3	1,5	-0,8	-0,8	0,3
8	-0,2	-0,2	-0,7	1,6	-1,1	-0,5	-0,5	0,2
9	-0,4	-0,7	0,3	0,1	0,3	-0,7	-0,6	0,6
10	-1,1	-1,0	0,8	-1,2	1,3	-0,9	-0,9	0,8
11	0,4	0,4	-1,1	-1,0	0,5	0,1	0,0	0,2
1	-0,2	1,5	-1,6	0,5	-0,3	-0,2	0,4	0,0
2	2,3	0,9	0,1	0,8	-1,5	1,8	2,1	-2,4
3	1,1	1,0	-0,4	0,6	-0,9	1,9	1,3	-1,2

Standardisierte Werte Gebiet 1

Stadtbezirk (Gebiet)	Bev. mit Migrations-hintergrund	Anteil Geschie-dene	Anteil unter 18 Jahre	Anteil über 65 Jahre	Einwohner-entwicklung 1996-	Besiedlungsdichte	Arbeitslosenquote	Wohn-fläche je Ein-wohner
11	0,4	0,4	-1,1	-1,0	0,5	0,1	0,0	0,2
11	0,4	0,4	-1,1	-1,0	0,5	0,1	0,0	0,2
11	0,4	0,4	-1,1	-1,0	0,5	0,1	0,0	0,2
11	0,4	0,4	-1,1	-1,0	0,5	0,1	0,0	0,2
11	0,4	0,4	-1,1	-1,0	0,5	0,1	0,0	0,2
11	0,4	0,4	-1,1	-1,0	0,5	0,1	0,0	0,2
11	0,4	0,4	-1,1	-1,0	0,5	0,1	0,0	0,2
11	0,4	0,4	-1,1	-1,0	0,5	0,1	0,0	0,2
11	0,4	0,4	-1,1	-1,0	0,5	0,1	0,0	0,2
11	0,4	0,4	-1,1	-1,0	0,5	0,1	0,0	0,2
11	0,4	0,4	-1,1	-1,0	0,5	0,1	0,0	0,2
1	-0,2	1,5	-1,6	0,5	-0,3	-0,2	0,4	0,0
1	-0,2	1,5	-1,6	0,5	-0,3	-0,2	0,4	0,0
1	-0,2	1,5	-1,6	0,5	-0,3	-0,2	0,4	0,0

Quadrierte Differenzen

Bev. mit Migrations-hintergrund	Anteil Geschie-dene	Anteil unter 18 Jahre	Anteil über 65 Jahre	Einwohner-entwicklung 1996-2006	Besiedlungsdichte	Arbeitslosenquote	Wohn-fläche je Ein-wohner
0,36	1,18	0,28	2,04	0,68	0,14	0,11	0,02
3,60	0,34	1,25	3,00	3,90	2,66	4,37	6,75
0,58	0,37	0,45	2,52	1,96	3,20	1,63	1,90
0,04	0,08	0,82	2,93	1,58	0,15	0,11	0,03
0,81	0,01	0,47	1,91	0,29	0,09	0,09	0,33
2,58	3,19	6,99	0,02	0,20	1,18	1,50	0,56
0,44	2,88	6,84	0,14	1,07	0,78	0,61	0,02
0,29	0,28	0,16	6,64	2,55	0,34	0,23	0,00
0,64	1,15	1,79	1,04	0,05	0,63	0,43	0,19
2,22	2,00	3,37	0,08	0,56	1,00	0,91	0,47
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
6,22	0,26	2,71	0,09	1,32	4,01	3,07	6,11
1,85	0,23	1,43	0,03	0,33	4,66	0,88	1,57

Summe der Quadrierten Differenzen

= Euklidsche Distanz

Stadtbezirk (Gebiet)	Stadtbezirk (Gebiet)	Euklidsche Distanzen
1	11	**4,8**
2	11	**25,9**
3	11	**12,6**
4	11	**5,7**
5	11	**4,0**
6	11	**16,2**
7	11	**12,8**
8	11	**10,5**
9	11	**5,9**
10	11	**10,6**
11	11	**0,0**
1	1	**0,0**
2	1	**23,8**
3	1	**11,0**

Um für alle Gebiete und für alle Variablen Euklidische Distanzen zu berechnen wird zunächst ein neues Arbeitsblatt gewählt. Es empfiehlt sich beispielsweise wie oben gezeigt, in der linken Hälfte alle standardisierten Werte für alle 8 Variablen nach Stadtbezirken aufsteigend so oft untereinander zu schreiben wie man Gebietseinheiten (Stadtbezirke) hat. In diesem Beispiel stehen links die Daten der 11 Stadtbezirke 11-mal untereinander, also 121 Zeilen. Damit jede Gebietseinheit sicher mit jeder anderen Gebietseinheit verglichen wird, wurden im obigen Beispiel die Werte eines Stadtbezirks 11-mal untereinander geschrieben (im Beispiel zuerst der Stadtbezirk 11). Das wird für alle 11 Gebietseinheiten gemacht.
Dadurch werden zwar alle Gebietseinheiten zweimal miteinander verglichen, aber es wird auch keine Gebietseinheit vergessen.

Lesebeispiel für die Tabelle von Seite 3

Vom Wert1 (Bevölkerung mit Migrationshintergrund) des Gebietes1 (-0,2) wird der Wert1 des Gebietes 11 (0,4) subtrahiert und das Ergebnis quadriert $(-0{,}2-0{,}4)^2 = -0{,}6^2 = 0{,}36$.
So wird mit allen Variablen der ersten Zeile verfahren (Vergleich Gebiet 1 mit 11). Die Summe dieser acht quadrierten Differenzen ergibt gerundet 4,8 und wird als Euklidische Distanz bezeichnet (siehe letzte Spalte der Tabelle oben).

4. Clusterbildung

Summe der Quadrierten Differenzen = Euklidsche Distanz

Stadtbezirk (Gebiet)	Stadtbezirk (Gebiet)	Euklidsche Distanzen
1	11	**4,8**
2	11	**25,9**
3	11	**12,6**
4	11	**5,7**
5	11	**4,0**
6	11	**16,2**

Als Vorbereitung zur Clusterbildung wird der Teil der Tabelle mit den berechneten Euklidischen Distanzen (siehe links) kopiert und mit dem Befehl „Bearbeiten-Inhalte einfügen-Werte“ in ein neues Arbeitsblatt eingefügt. Rechts von diesen Werten auf dem neuen Arbeitsblatt schreibt man quer als Überschrift Cluster 1, in die nächste Zelle Cluster 2 usw.

Bei der eigentlichen Clusterbildung wird nun wie folgt verfahren:

1. Sortieren der Euklidischen Distanzen aufsteigend vom niedrigsten Wert.
2. Die Zeilen mit dem Werte 0, normalerweise diejenigen, bei denen eine Gebietseinheit mit sich selbst verglichen wird, können gelöscht werden.
3. Die beiden Gebiete mit der niedrigsten Euklidischen Distanz werden zum Cluster 1 zusammengefasst (im Bsp. SBZ 10 und 6).
4. Bei der nächst niedrigen Euklidschen Distanz wird geprüft, ob bereits eines der beiden verglichenen Gebiete zu einem Cluster gehört:
   wenn ja, dann wird das noch nicht zugeordnete Gebiet (im Bsp. SBZ 7) dem Cluster (im Bsp. Cluster 1) des bereits zugeordneten Gebiets (im Bsp. SBZ 10) zugeschlagen
   wenn nein (im Bsp. SBZ 8 und 4), dann wird ein neues Cluster (im Bsp. Cluster 2) gebildet.
5. Schritt 3 wird so lange wiederholt, bis alle Gebiete einem Cluster zugeordnet sind.

Die nachfolgende Übersicht veranschaulicht das Vorgehen. Der Einfachheit und Übersichtlichkeit halber wurden hier die doppelten Vergleichswerte (z.B. 6 mit 10 ist das Gleiche wie 10 mit 6) gelöscht, so dass jeder Vergleich nur einmal vorkommt.

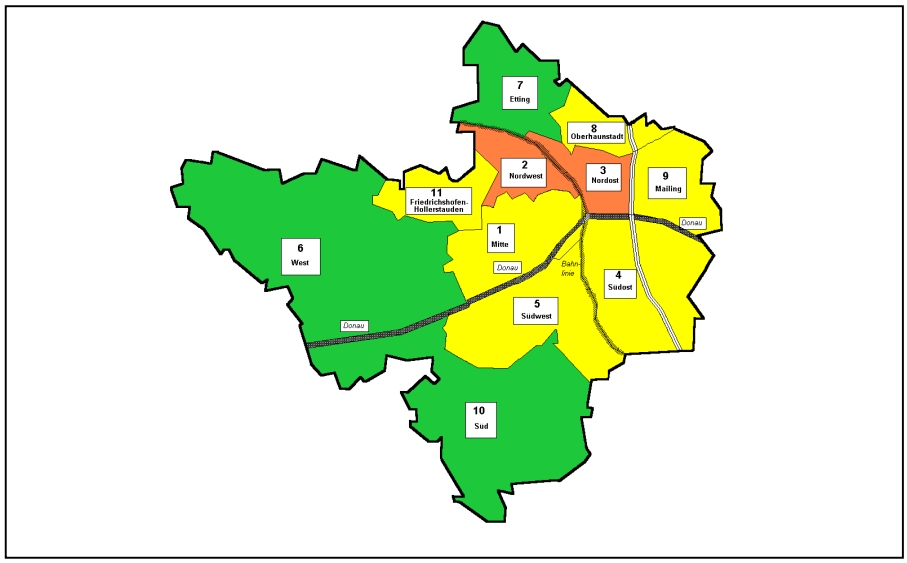
Clusterbildung (Hierarchische Methode)

SBZ 1	SBZ 2	Euklidsche Distanzen
10	6	**1,00**
10	7	**1,80**
7	6	**1,90**
8	4	**2,64**
5	4	**2,77**
9	5	**2,93**
4	1	**3,12**
10	9	**3,68**
5	11	**4,00**
3	2	**4,02**
8	5	**4,07**
5	1	**4,44**
1	11	**4,80**
9	6	**5,20**
9	4	**5,43**
9	7	**5,70**
4	11	**5,74**

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
10	**8**	**3**			
6	**4**	**2**			
7	**5**				
	9				
	1				
	11				

Das hier erreichte Ergebnis deckt sich gut mit einer Reihe weiterer Erkenntnisse zur Struktur der 11 Stadtbezirke in Ingolstadt. Grün sind die gehobenen Wohngebiete am Stadtrand, gelb die normalen Wohngebiete in der Kernstadt, rot die Gebiete, in denen soziale Probleme gehäuft auftreten. Die kartografische Darstellung zeigt dann auch die räumliche Verteilung:

Karte der Stadtbezirke der Stadt Ingolstadt



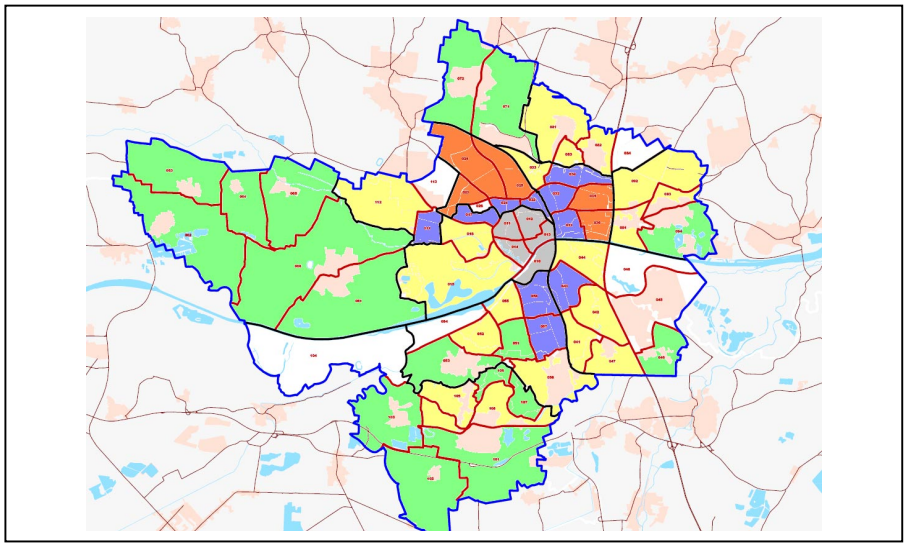
Zusammenfassung und praktische Umsetzbarkeit

Die hier gezeigte Vorgehensweise einer vereinfachten Cluster-Analyse mit Excel nach der Single-Linkage-Methode eignet sich vor allem für eine geringere Anzahl zu vergleichender Einheiten, z.B. Gebiete.

Da hier nur eines von vielen möglichen Verfahren zur Clusteranalyse verwendet wird, können auch kaum Feineinstellungen vorgenommen werden, z. B. ist es nicht möglich, wie es das Programm SPSS bietet, wenn keine sinnvolle Clusterung zustande kommt, auf eine andere Methode umzuschalten.

Allerdings können Clusteranalysen in vielen Fällen mit relativ geringem Aufwand und mit geringen Vorkenntnissen erstellt werden. Deshalb eignet sich die Clusteranalyse mit Excel vor allem für kleinere Großstädte und Städte.

Ein Beispiel einer Clusteranalyse, in der die 11 Stadtbezirke Ingolstadts in ihre 61 Unterbezirke aufgeteilt wurden mit den gleichen Variablen der Sozialraumtypisierung für jeden Unterbezirk zeigt nachfolgende Karte:



Hier sind aus den 3 Clustern der 11 Stadtbezirke 5 Cluster der 61 Unterbezirke schlüssig zu erstellen gewesen.

Ein ähnlicher Versuch mit rund 120 Gebietseinheiten (Stimmbezirken) wurde aufgrund der hohen Komplexität und des hohen Zeitaufwandes nicht weiter verfolgt. Hier lohnt sich dann der Einsatz einer professionellen Software wie SPSS in jedem Fall.

Autor:
Helmut Schels
Stadt Ingolstadt
Stadtplanungsamt, Sachgebiet Stadtentwicklung und Statistik
Spitalstr. 3
85049 Ingolstadt
Tel.: 0841 – 305 1056
E-mail: helmut.schels@ingolstadt.de