

# Kommunale Demographietypen: Typisierung der Städte und Gemeinden durch eine Clusteranalyse

*Bernd Behrenschorf, Institut für Entwicklungsplanung und Strukturforschung GmbH an der Universität Hannover*

Eine Hilfestellung bei der Analyse des demographischen Wandels auf der Ebene der Gemeinden bietet der „Wegweiser Demographischer Wandel“ mit der Entwicklung von Demographietypen. Im Folgenden wird die Erstellung der Demographietypen durch eine Typisierung von Gemeinden in der Bundesrepublik Deutschland nach demographischen, wirtschaftlichen und sozialen Merkmalen vorgestellt. Die Typisierung der Gemeinden wurde zu dem Zweck vorgenommen:

- Transparenz über die demographische Entwicklung herzustellen,
- Betroffenheit, Perspektiven und Potenziale zu vermitteln und
- Handlungsempfehlungen für Kommunen zu formulieren.

Damit die auf Basis der kommunalen Demographietypen erstellten Handlungsempfehlungen weiterhin als Hilfestellung und Anregung genutzt werden können, wurde die Typisierung mit neuen Indikatorenwerten aktualisiert. Absicht dieser Aktualisierung ist es, die Stabilität der Demographietypen zu gewährleisten, nicht die Beständigkeit der Einordnung der Gemeinden zu überprüfen. Das methodische Vorgehen hierzu ist ebenfalls im folgenden Text nachzulesen.

## 1. Clusteranalyse – Ziel und Vorgehensweise

Die Clusteranalyse ist ein Typisierungsverfahren, dessen Ziel es ist, die untersuchten Raumeinheiten so zu gruppieren, dass die Unterschiede zwischen den Raumeinheiten innerhalb einer Gruppe („Cluster“) möglichst gering und die Unterschiede zwischen den Gruppen möglichst groß sind.

Das Grundprinzip der Clusteranalyse geht von der Positionierung jedes einzelnen Falles in einem mehrdimensionalen Raum aus, dessen Achsen die der Analyse zugrunde liegenden Variablen sind. In diesem Koordinatensystem können die Abstände zwischen den positionierten Fällen anhand verschiedener Verfahren gemessen werden.

Die einzelnen Fälle werden danach anhand der gemessenen Abstände, d. h. gemäß ihrer Ähnlichkeit zu Typen zusammengefasst. Das Ergebnis dieser Gruppierung (Clusterbildung) wird anhand weiterführender Verfahren hinsichtlich seiner Plausibilität und möglichen Korrekturen überprüft.

## 2. Vorbereitende Schritte

Auf der Basis inhaltlich-theoretischer Überlegungen wurden die folgenden acht Variablen als relevant für die Untersuchung ausgewählt:

- Bevölkerungsentwicklung 2003 bis 2020 (Index; 2003 = 100)
- Medianalter 2020 (Jahre)
- Arbeitsplatzzentralität (Verhältnis: Anzahl sozialversicherungspflichtig Beschäftigte am Arbeitsort / Anzahl sozialversicherungspflichtig Beschäftigte am Wohnort, 2003)
- Arbeitsplatzentwicklung 1998 bis 2003 (Index; 1998 = 100)
- Arbeitslosenquote (in %, 2003)
- Kommunale Steuereinnahmen gemittelt über vier Jahre (pro Einwohner, 2000 bis 2003)
- Anteil Hochqualifizierter (in % aller sozialversicherungspflichtig Beschäftigter, 2003)
- Anteil Mehrpersonenhaushalte mit Kindern (in % aller Haushalte, 2003)

### 2.1 Sichtung der Verteilung der Werte im Datensatz

In einem ersten Schritt wird die Verteilung der Werte im Datensatz für jede einzelne Variable mithilfe von Lage- und Streuungsmaßen analysiert. Mithilfe der SPSS-Prozedur DESCRIPTIVES ergibt sich die folgende zusammenfassende deskriptive Statistik:

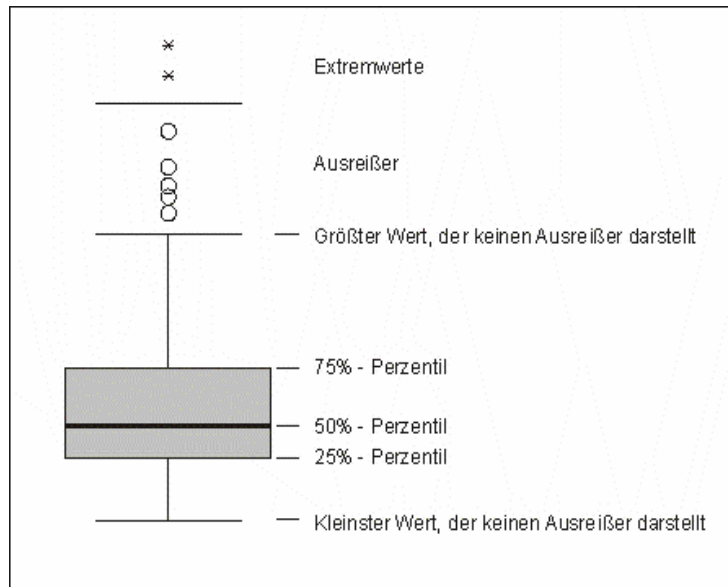
**Tabelle 1: Deskriptive Statistik**

Deskriptive Statistik					
	N	Minimum	Maximum	Mittelwert	Standardabweichung
Bevölkerungsentwicklung 2003 bis 2020	2959	52,93	156,86	99,3031	8,75072
Medianalter 2020	2959	37,73	60,18	48,2618	3,06916
Arbeitsplatzzentralität	2959	,11	4,43	,8486	,42922
Arbeitsplatzentwicklung 1998 - 2003	2959	41,65	302,67	100,0647	14,46026
Arbeitslosenquote	2959	4,31	36,66	12,0846	5,91949
Steuereinnahmekraft über 4 Jahre	2959	149,09	6748,70	616,8655	343,16103
Anteil hochqual. Beschäftigter (WO)	2959	1,60	26,76	7,5573	3,70238
Anteil Mehrpersonenhaushalte mit Kindern	2959	12,11	68,96	38,7823	8,66630
Gültige Werte (Listenweise)	2959				

Zusätzlich zu den angegebenen Lage- und Streuungsparametern kann die Verteilung der

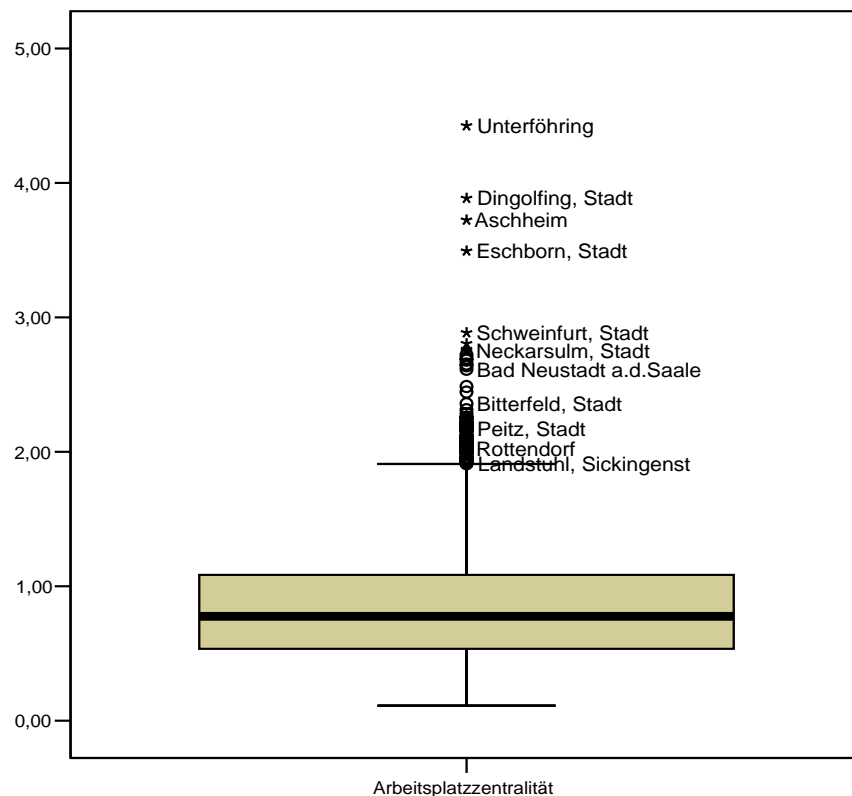
Werte für jede einzelne Variable anhand eines „boxplots“ grafisch dargestellt werden. Dies ermöglicht es, die Existenz und Lage von Extremwerten sowie Ausreißern zu erkennen.

**Abbildung 1: Boxplot – Legende**



**Quelle:** Brosius, Felix (1998): SPSS 8.0 Professionelle Statistik unter Windows. Bonn: MITP. 381

Anhand der beispielhaft ausgewählten Variable „Arbeitsplatzzentralität“ (vgl. Abb. 2) lässt sich eine typische Konstellation für soziodemografische und ökonomische raumbezogene Daten erkennen: Es existieren zahlreiche Extremwerte und Ausreißer, die sich sehr weit von dem 50 %-Perzentil (dem Median) entfernen, während die untere Hälfte der Werte in einem relativ engen Wertebereich konzentriert ist.

**Abbildung 2: Boxplot, Beispiel Variable Arbeitsplatzzentralität**

## 2.2 Korrelationsprüfung

Im zweiten Schritt ist zu prüfen, ob die ausgewiesenen Variablen nicht in hohem Maße miteinander korrelieren. Durch die **Korrelationsprüfung** gilt es zu vermeiden, dass Variablen, deren Aussagen nahezu deckungsgleich sind und die daher hoch miteinander korrelieren, gemeinsam in die Clusteranalyse eingehen. Gingen sie gemeinsam in die Analyse ein, so würde dies bedeuten, dass die in ihnen enthaltene Information im Vergleich zu einer anderen Aussage (die nur von einer Variablen repräsentiert wird) mehrfach gewichtet wird. Für diesen Test reicht ein einfacher linearer Korrelationstest für metrisch skalierte Variablen anhand des Korrelationskoeffizienten nach Pearson. Korrelationskoeffizienten mit Absolutwerten größer als 0,6 bedeuten eine hohe bis sehr hohe Korrelation. Korrelieren zwei Variablen höher, so sollte eine der beiden betroffenen Variablen aus den weiteren Analysen ausgeschlossen werden.

Die Korrelationsanalyse weist für die folgenden Variablenpaare jeweils mittlere signifikante Korrelationswerte auf (vgl. Tab. 2):

- Bevölkerungsentwicklung & Medianalter 2020 ( $R = -0,504$ )
- Bevölkerungsentwicklung & Arbeitslosenquote ( $R = -0,496$ )
- Medianalter 2020 & Arbeitslosenquote ( $R = 0,589$ )
- Arbeitsplatzzentralität & Steuereinnahmekraft ( $R = 0,480$ )

- Arbeitsplatzzentralität & Mehrpersonenhaushalte (R= -0,483)

Da keine dieser Korrelationen als „stark“ bezeichnet werden kann, ist der Forderung, dass die einfließenden Variablen nicht hoch miteinander korrelieren sollen, Genüge getan und es kann auf den Ausschluss einzelner Variablen aus den folgenden Analysen verzichtet werden.

**Tabelle 2: Korrelationskoeffizienten (bivariate Korrelation, nach Pearson)**

		Korrelationen							
		Bevölkerungsentwicklung 2003 bis 2020	Medianalter 2020	Arbeitsplatzzentralität	Arbeitsplatzentwicklung 1998 - 2003	Arbeitslosenquote	Steuereinnahmekraft über 4 Jahre	Anteil hochqual. Beschäftigter (WO)	Anteil Mehrpersonenhaushalte mit Kindern
Bevölkerungsentwicklung 2003 bis 2020	Korrelation nach Pearson	1	-,504	-,097	,356	-,496	,145	,140	,208
	Signifikanz (2-seitig)		,000	,000	,000	,000	,000	,000	,000
	N	2959	2959	2959	2959	2959	2959	2959	2959
Medianalter 2020	Korrelation nach Pearson	-,504	1	-,015	-,359	,589	-,329	-,001	-,227
	Signifikanz (2-seitig)	,000		,428	,000	,000	,000	,941	,000
	N	2959	2959	2959	2959	2959	2959	2959	2959
Arbeitsplatzzentralität	Korrelation nach Pearson	-,097	-,015	1	,103	,193	,480	,106	-,483
	Signifikanz (2-seitig)	,000	,428		,000	,000	,000	,000	,000
	N	2959	2959	2959	2959	2959	2959	2959	2959
Arbeitsplatzentwicklung 1998 - 2003	Korrelation nach Pearson	,356	-,359	,103	1	-,442	,294	,063	,133
	Signifikanz (2-seitig)	,000	,000	,000		,000	,000	,001	,000
	N	2959	2959	2959	2959	2959	2959	2959	2959
Arbeitslosenquote	Korrelation nach Pearson	-,496	,589	,193	-,442	1	-,388	-,037	-,366
	Signifikanz (2-seitig)	,000	,000	,000	,000		,000	,046	,000
	N	2959	2959	2959	2959	2959	2959	2959	2959
Steuereinnahmekraft über 4 Jahre	Korrelation nach Pearson	,145	-,329	,480	,294	-,388	1	,233	-,121
	Signifikanz (2-seitig)	,000	,000	,000	,000	,000		,000	,000
	N	2959	2959	2959	2959	2959	2959	2959	2959
Anteil hochqual. Beschäftigter (WO)	Korrelation nach Pearson	,140	-,001	,106	,063	-,037	,233	1	-,183
	Signifikanz (2-seitig)	,000	,941	,000	,001	,046	,000		,000
	N	2959	2959	2959	2959	2959	2959	2959	2959
Anteil Mehrpersonenhaushalte mit Kindern	Korrelation nach Pearson	,208	-,227	-,483	,133	-,366	-,121	-,183	1
	Signifikanz (2-seitig)	,000	,000	,000	,000	,000	,000	,000	
	N	2959	2959	2959	2959	2959	2959	2959	2959

### 2.3 Test auf Normalverteilung

Regionalökonomische Zahlen ebenso wie sozio-demographische Raumdaten weisen niemals eine echte Normalverteilung auf.

Sie weisen häufig einige Ausreißer und Extremwerte (meist Städte oder suburbane Gemeinden) sowie insgesamt eine linkssteile Verteilung auf.

Die Überprüfung der annähernden Normalverteilung erfolgt anhand der Normalverteilungsdiagramme (Q-Q) und des trendbereinigten Q-Q-Diagramms jeder einzelnen Variablen. Im vorliegenden Datensatz zeigen die folgenden Variablen eine sehr gute Anpassung an die Normalverteilungskurve:

- Anteil der Mehrpersonenhaushalte mit Kindern
- Medianalter
- Bevölkerungsentwicklung

Die anderen Variablen weisen die in sozio-ökonomischen, raumbezogenen Datensätzen typischen Abweichungen von der Normalverteilung auf, die für die Durchführung der Clusteranalyse akzeptabel erscheinen. Somit wird der bereits definierte und anhand

explorativer Datenanalyse vorgestellte Variablen in die Clusteranalyse einfließen.

## **2.4 Aufteilung des Datensatzes in Großstädte und kleinere Städte und Gemeinden**

Die durch die Clusteranalyse zu erreichende Typisierung der betrachteten Städte und Gemeinden soll als Grundlage für zu entwickelnde Handlungsempfehlungen zur Gestaltung des demographischen Wandels dienen. Bei vorab testweise durchgeführten Analysen ergaben sich stets Klassifizierungen, bei denen Großstädte und kleinere Gemeinden gemeinsamen Gruppen zugeordnet wurden. Bei Betrachtung der eingehenden Variablen erwies sich diese Zuordnung durchaus als korrekt und plausibel. Im Hinblick auf die aus der Gruppenzugehörigkeit abzuleitenden Handlungsempfehlungen ist diese gemeinsame Klassifizierung eher hinderlich, da großen Städten – auch bei ähnlichem demographischen Profil – in der Regel andere Maßnahmen empfohlen werden müssen als kleineren Gemeinden. Aus pragmatischen Erwägungen wurde daher beschlossen, für Städte mit 100.000 Einwohnern und mehr ( $n = 82$ ) und für Städte und Gemeinden mit weniger als 100.000 Einwohnern ( $n = 2877$ ) jeweils eine gesonderte Clusteranalyse durchzuführen.

Durch die Aufteilung in zwei Analysegruppen ergeben sich in den Teildatensätzen Verschiebungen bei den Korrelationswerten. Bei der kleineren Gruppe der Großstädte ergeben sich nun bei den Variablenpaaren „Bevölkerungsentwicklung 2003-2020“ & „Medianalter 2020“ ( $R = -,739$ ) und „Steuereinnahmekraft über 4 Jahre“ & „Arbeitslosenquote“ ( $R = -,650$ ) stärkere Korrelationen. Diese Variablen korrelieren im Teildatensatz der kleineren Städte und Gemeinden nur schwach und können daher dort bedenkenlos in die Analyse eingehen. Aus Gründen der Vergleichbarkeit der beiden Analysegruppen werden diese Variablen auch für die Gruppe der Großstädte verwendet.

## **3. Durchführung der Clusteranalyse**

### **3.1 Standardisierung**

Da die Variablen teilweise unterschiedliche Dimensionen aufweisen (z. B. Arbeitslosenquote und Kaufkraft), müssen sie standardisiert werden, um inhaltliche Verzerrungen zu vermeiden. Als Standardisierungsverfahren kam die z-Transformation zur Anwendung. Durch die z-Transformation werden die Variablen solchermaßen standardisiert, dass jede einen Mittelwert von null und eine Standardabweichung von eins aufweist.

### **3.2 Hierarchische Clusteranalyse**

Hierarchische Clusterverfahren sind schrittweise aufgebaut: Zu Beginn der Clusteranalyse stellt jede einzelne Raumeinheit einen Cluster dar, die Zahl der Cluster beträgt also  $n$  (= Anzahl der Fälle im Datensatz). Im ersten Schritt werden zwischen allen Raumeinheiten die Unterschiede hinsichtlich der eingehenden Variablen paarweise gemessen und die beiden Raumeinheiten, welche die geringste Distanz aufweisen, d. h. die sich am ähnlichsten sind,

werden zu einem Cluster zusammengefasst. Nach dem ersten Schritt existieren daher  $n-1$  Cluster, für die wiederum die Distanzen paarweise gemessen werden um danach die beiden ähnlichsten Einheiten zusammenzufassen. So wird fortgefahren, bis nach  $n-1$  Schritten nur noch ein Cluster (hier: alle betrachteten Kommunen) übrig bleibt.

Neben der hierarchischen Clusteranalyse bietet sich noch die Alternative „Clusterzentrenanalyse“, die jedoch bei der vorliegenden Aufgabenstellung nicht als primäres Analyseverfahren zum Einsatz kommen kann, da zwei Voraussetzungen der Clusterzentrenanalyse nicht erfüllt sind: 1. Die Anzahl der Cluster und 2. deren „Zentrum“ (arithmetisches Mittel aller enthaltenen Elemente) müssen vorab bekannt sein.

Hinsichtlich der Durchführung der hierarchischen Clusteranalyse sind die folgenden drei Fragen vorab zu klären:

- Wie wird die Distanz zwischen zwei Raumeinheiten gemessen?
- Nach welchem Verfahren werden Raumeinheiten zu Clustern zusammengefasst?
- In welchem Schritt wird der Prozess der Clusterbildung angehalten?

### 3.3 Verwendetes Distanzmaß

Zunächst ist das zum Einsatz kommende Ähnlichkeitsmaß festzulegen, das als Grundlage für die Clusterbildung dienen soll, da es verschiedene mathematische Möglichkeiten gibt, in einem mehrdimensionalen Raum Abstände zu messen. Wir verwenden die **quadrierte euklidische Distanz**, die sich für raumwirtschaftliche Fragestellungen als besonders sinnvoll erwiesen hat. Hierfür werden für jede einzelne Variable die Unterschiede zwischen zwei Raumeinheiten gemessen, dann quadriert und anschließend aufsummiert. Somit ergibt sich für jedes Raumeinheiten-Paar die Gesamtdistanz (über alle Variablen) und damit ein Kriterium für deren Ähnlichkeit.

Es bieten sich darüber hinaus folgende Alternativen:

- **Euklidischer Abstand:** Er ist als Wurzel aus der quadrierten euklidischen Distanz definiert. Im zweidimensionalen Raum repräsentiert er die „Luftlinie“ zwischen zwei Punkten und ist somit das wohl anschaulichste Distanzmaß.
- **Block:** Hierfür werden die Differenzen zwischen den beiden Werten für jede Variable gemessen und dann deren Beträge (Absolutwerte) aufsummiert. Sie wird auch als Manhattan-Distanz bezeichnet, da sie im zweidimensionalen Raum die Strecke abbildet, die in einem rechtwinkligen Straßensystem zurückgelegt werden muss, um von A nach B zu kommen.
- **Tschebyscheff:** Die Distanz zwischen zwei Raumeinheiten wird für jede Variable einzeln gemessen. Der hierbei auftretende Maximalwert wird als Distanz zwischen den beiden Raumeinheiten betrachtet. Für die vorliegende Fragestellung ist dieses Verfahren nicht geeignet, da das Ergebnis zu sehr von einzelnen Variablen und den hierbei auftretenden extremen Unterschieden geprägt wird.

Die Wahl fiel auf die „quadrierte euklidische Distanz“, da sie zum einen das am besten etablierte Distanzmaß ist, und zum anderen, da die im Folgenden vorgestellte Methode der Clusterbildung den Einsatz dieses Distanzmaßes vorschreibt.

### 3.4 Methode der Clusterbildung

Danach gilt es, die Frage zu klären, auf der Basis welchen Verfahrens die Raumeinheiten – bei gegebener Distanz – zu Typen zusammengefasst werden sollen. Auch hierfür stehen verschiedene mathematische Methoden zur Verfügung.

Von den üblichen Methoden weist das **Ward-Verfahren** die stärkste Tendenz zur Bildung von Gruppen mit ähnlicher Fallanzahl auf. Die Anwendung des **Single-Linkage-Verfahrens** führt eher zur Bildung einer großen Gruppe und mehrerer singulär besetzter „Ausreißercluster“, während das **Zentroid-** sowie das **Average-Linkage-Verfahren** einen mittleren Platz hinsichtlich dieser Effekte einnehmen.

Auf der Basis dieser Beurteilung wird in der vorliegenden Analyse dem Ward-Verfahren der Vorzug gegeben. Dies ist in raumbezogenen sozio-demographischen und ökonomischen Daten durchaus üblich.

### 3.5 Bestimmung der Anzahl zu betrachtender Cluster

Da das mathematische Verfahren der Clusteranalyse als Endergebnis immer nur einen einzigen Cluster hat, ist die Entscheidung zu treffen, an welcher Stelle der Prozess der Zusammenfassung anzuhalten ist.

Da es Ziel der Clusteranalyse ist, Gruppen zu bilden, die in sich möglichst homogen sind, untereinander aber möglichst große Unterschiede aufweisen, gleichzeitig aber aus Gründen der Übersichtlichkeit eine möglichst geringe Clusteranzahl anzustreben ist, wird der Prozess an der Stelle angehalten, an der erstmals sehr unterschiedliche Gruppen zu einem einzigen Cluster zusammengefasst werden.

Das Kriterium dafür ist der Anstieg der Fehlerquadratsumme, die in der Zuordnungsübersicht unter „Koeffizienten“ abgebildet ist. Beginnen die Koeffizienten sprunghaft anzusteigen, bedeutet dies, dass ab diesem Schritt sehr unähnliche Cluster zusammengeführt werden. Folglich ist die Clusteranalyse vor diesem ersten sprunghaften Anstieg anzuhalten. In der vorliegenden Analyse lassen sich jedoch keine markanten sprunghaften Anstiege erkennen.

Bei der Betrachtung der prozentualen Zuwächse ist jedoch erkennbar, dass im Fall der Gruppe der kleineren Städte und Gemeinden eine Lösung mit **neun** und im Fall der Großstädte eine Lösung mit **sechs** extrahierten Clustern sinnvoll ist.

## 4. Analyse der identifizierten Gruppen: Mittelwertunterschiede und Varianzanalyse

Die identifizierte Lösung ist im Folgenden hinsichtlich ihrer inhaltlichen Interpretierbarkeit und statistischen Zuverlässigkeit zu überprüfen. Dies wird anhand einer Analyse der Mittelwertunterschiede und einer ANOVA (Analysis of Variance) erfolgen.

**Tabelle 3<sup>1</sup>: Mittelwertvergleich – Städte >= 100.000 EW – sechs Cluster**

		Bericht							
6 Cluster		Bevölkerungsentwicklung 2003 bis 2020	Medianalter 2020	Arbeitsplatzkonzentration	Arbeitsplatzentwicklung 1998 - 2003	Arbeitslosenquote	Steuereinnahmekraft über 4 Jahre	Anteil hochqual. Beschäftigter (WO)	Anteil Mehrpersonenhaushalte mit Kindern
1	Mittelwert	100,0979	43,7144	1,4809	102,1387	15,8264	914,6085	11,3429	22,0549
	N	17	17	17	17	17	17	17	17
	Standardabweichung	2,97171	1,76573	,23578	2,08612	2,12257	163,24720	1,79352	3,46656
2	Mittelwert	93,0176	47,7163	1,1023	98,2181	16,9517	736,6798	7,5867	25,2320
	N	19	19	19	19	19	19	19	19
	Standardabweichung	2,93129	1,15126	,16550	6,25533	2,76086	122,62970	2,03070	2,88985
3	Mittelwert	102,3746	42,2439	1,7205	105,5464	11,5061	1190,9913	16,6530	23,0083
	N	21	21	21	21	21	21	21	21
	Standardabweichung	3,92332	1,90683	,25196	4,96949	2,03440	356,62372	4,27827	5,45243
4	Mittelwert	97,9183	46,3965	1,2234	97,7926	12,7347	908,2301	7,9653	33,6226
	N	13	13	13	13	13	13	13	13
	Standardabweichung	3,30276	1,74708	,31627	3,63019	1,83763	201,87600	2,42824	4,27937
5	Mittelwert	98,8576	44,2039	1,2669	92,3902	21,3326	461,1478	16,2526	24,2178
	N	9	9	9	9	9	9	9	9
	Standardabweichung	8,66261	2,99380	,10915	4,38214	3,58381	107,30559	3,97205	1,65161
6	Mittelwert	81,1374	52,9530	1,2457	86,5887	22,9779	368,3263	14,1261	25,0779
	N	3	3	3	3	3	3	3	3
	Standardabweichung	2,47219	2,03902	,16159	2,31790	1,04968	63,71839	2,70844	1,72690
Insgesamt	Mittelwert	97,8650	45,0820	1,3816	99,7751	15,3565	873,3948	11,9377	25,2171
	N	82	82	82	82	82	82	82	82
	Standardabweichung	6,23993	3,16456	,32511	6,56822	4,13659	326,65118	4,87261	5,42593

<sup>1</sup> Die Reihenfolge der Cluster findet sich in einer geänderten, aktuelleren Variante im Online-Wegweiser unter [www.wegweiserdemographie.de](http://www.wegweiserdemographie.de).

**Tabelle 4<sup>2</sup>: Mittelwertvergleich – Gemeinden < 100.000 EW – neun Cluster**

Bericht									
9 Cluster		Bevölkerungsentwicklung 2003 bis 2020	Medianalter 2020	Arbeitsplatzzentralität	Arbeitsplatzentwicklung 1998 - 2003	Arbeitslosenquote	Steuereinnahmekraft über 4 Jahre	Anteil hochqual. Beschäftigter (WO)	Anteil Mehrpersonenhaushalte mit Kindern
1	Mittelwert	98,4687	48,9239	1,3146	100,1169	12,8467	755,5567	6,2178	30,2499
	N	459	459	459	459	459	459	459	459
	Standardabweichung	6,03701	3,01626	,41674	8,34310	3,78067	288,93198	2,39393	5,93052
2	Mittelwert	114,2490	49,9101	,6512	89,5619	16,3134	290,8031	12,0016	43,1988
	N	96	96	96	96	96	96	96	96
	Standardabweichung	16,12395	1,87625	,25853	12,99116	3,31724	100,67818	4,48153	6,23553
3	Mittelwert	101,6532	46,8584	,6659	100,6958	9,2397	582,9779	6,1606	44,3024
	N	1156	1156	1156	1156	1156	1156	1156	1156
	Standardabweichung	5,82550	2,13618	,29028	8,63031	2,30845	151,72848	2,22288	6,59696
4	Mittelwert	95,9363	49,5475	,5993	92,4992	12,2113	490,9746	5,0713	36,9831
	N	260	260	260	260	260	260	260	260
	Standardabweichung	6,00674	1,69221	,19609	10,66141	2,81727	110,34559	1,68104	4,56190
5	Mittelwert	101,0509	47,8768	,7718	103,7790	8,4771	760,5493	12,5036	35,5242
	N	307	307	307	307	307	307	307	307
	Standardabweichung	4,93616	1,92283	,31416	9,83112	2,03358	202,35537	3,63858	6,81033
6	Mittelwert	102,7383	46,5329	,7906	124,8246	8,1766	736,6813	7,6428	44,6961
	N	220	220	220	220	220	220	220	220
	Standardabweichung	6,66322	2,26327	,34108	22,09674	2,46524	292,36387	3,31770	5,49007
7	Mittelwert	104,9978	47,1563	1,7248	113,4261	6,9688	1818,9856	16,5226	34,6455
	N	33	33	33	33	33	33	33	33
	Standardabweichung	11,44069	2,39952	,64201	21,86526	1,66708	486,55483	4,72533	5,66725
8	Mittelwert	103,9140	46,2454	2,7778	133,3612	7,3229	5096,5743	16,9222	33,0313
	N	5	5	5	5	5	5	5	5
	Standardabweichung	,39685	2,75888	1,24920	27,72467	1,36141	1048,76266	3,72962	10,94218
9	Mittelwert	86,7556	53,0469	,9518	85,5504	24,9636	282,4785	8,1730	34,5464
	N	341	341	341	341	341	341	341	341
	Standardabweichung	8,11530	2,10167	,32777	11,93657	4,42359	76,55059	2,12298	7,22971
Insgesamt	Mittelwert	99,3441	48,3524	,8334	100,0729	11,9914	609,5539	7,4324	39,1689
	N	2877	2877	2877	2877	2877	2877	2877	2877
	Standardabweichung	8,80915	3,01824	,42210	14,62339	5,93663	340,85499	3,58708	8,42739

Die Tabellen zeigen für jeden Cluster die Anzahl der darin enthaltenen Raumeinheiten (N) und für jede der einfließenden Variablen das arithmetische Mittel sowie die Standardabweichung als Indikator der Unterschiedlichkeit der im Cluster enthaltenen Fälle. Auf der Basis dieser Lage- und Streuungsparameter für die eingehenden Variablen kann jedes einzelne Cluster nun inhaltlich beschrieben, interpretiert und benannt werden.

Davor sollte aber noch getestet werden, inwiefern sich diese Cluster signifikant voneinander unterscheiden. Die Unterschiede werden zwar bereits durch die Differenzen der Mittelwerte und der Standardabweichungen in deskriptiver Art deutlich. Zusätzlich überprüft die einfaktorielle Varianzanalyse (ANOVA), inwiefern die beobachteten Unterschiede signifikant sind und sich somit die Cluster deutlich voneinander abgrenzen lassen. Die Varianzanalyse setzt hierfür die Streuung der Variablen innerhalb der Gruppen (und damit die Homogenität der Gruppen) in Beziehung zu der Streuung zwischen den Gruppen (und damit die Unterschiedlichkeit der Gruppen) und errechnet daraus den F-Wert. Entscheidend ist, inwiefern sich ein F-Wert in gegebener Höhe auch zufällig ergeben kann. Da in beiden Fällen bei allen eingehenden Variablen sehr hohe Signifikanz (Niveau 0,01; d. h. die Irrtumswahrscheinlichkeit ist kleiner als 1 %) ausgewiesen wird, ist auszuschließen, dass die Unterschiede zufällig aufgetreten sind. Die identifizierten sechs bzw. neun Cluster weisen

<sup>2</sup> Die Reihenfolge der Cluster findet sich in einer geänderten, aktuelleren Variante im Online-Wegweiser unter [www.wegweiserdemographie.de](http://www.wegweiserdemographie.de).

somit zuverlässig markante Unterschiede auf. Dennoch zeigt die immer noch recht hohe Streuung innerhalb der Cluster, dass versucht werden sollte, eine trennschärfere Abgrenzung zwischen den Gruppen zu erzielen.

## 5. Überprüfung / Korrektur der Lösungen

Ein wichtiger Nachteil aller vorgestellten Varianten der hierarchischen Clusteranalyse liegt im schrittweisen Vorgehen begründet. Sobald ein Fall einem Cluster zugeordnet ist, kann er nicht mehr entnommen und einem anderen Cluster zugeschlagen werden. Da sich durch die schrittweise Erweiterung der Cluster aber erhebliche Verschiebungen der Cluster-Mittelwerte ergeben können, ist es nicht nur möglich, sondern sogar wahrscheinlich, dass einzelne Fälle zum Schluss einem Cluster zugehören, obwohl sie näher an der Mitte eines anderen Clusters liegen. Deswegen sollte das Ergebnis der hierarchischen Clusteranalyse überprüft bzw. einer Korrektur unterzogen werden. Auch hierfür stehen verschiedene Verfahren zur Auswahl, von denen bei der vorliegenden Analyse zwei zum Einsatz kamen:

- die Clusterzentrenanalyse und
- die Diskriminanzanalyse.

### 5.1 Clusterzentrenanalyse

Wenn die Clusterzentrenanalyse auch als primäres Analysewerkzeug nicht infrage kommt, da zu Beginn weder Clusterzentren noch die Clusteranzahl bekannt sind, ist sie doch als nachgeschaltetes Überprüfungs- und Korrekturinstrument durchaus geeignet.

In diesem Fall werden die durch die hierarchische Clusteranalyse ermittelten Clusterzentren (d. h. hier: die arithmetischen Mittelwerte der standardisierten beteiligten Variablen innerhalb der einzelnen Cluster) und die Clusteranzahl als Vorgabe für die Clusterzentrenanalyse (SPSS-Prozedur QUICK CLUSTER) verwendet.

Im einfachsten Fall wird die Clusterzentrenanalyse einfach dazu verwendet, die Fälle auf Basis der bestehenden Clusterzentren neu zuzuordnen (METHOD = CLASSIFY). Hierbei werden viele Gemeinden wieder dem gleichen Cluster zugeordnet, dem sie auch nach der hierarchischen Clusteranalyse angehörten.

Es kommt aber auch zu zahlreichen Neuordnungen (im Fall der 2877 kleineren Gemeinden ergeben sich etwa 600 Anderszuordnungen). Einen Schritt weiter geht die Clusterzentrenanalyse, die nach der Neuklassifizierung die Clusterzentren neu berechnet, danach eine weitere Neuklassifizierung vornimmt, diese wieder als Ausgangspunkt für eine Clusterzentrenberechnung nimmt etc. Dieses erfolgt so lange, bis das Neuordnen keine oder nur minimale Änderungen gegenüber dem vorigen Ergebnis ergibt oder eine einstellbare Maximalzahl von Neuberechnungen erfolgt ist (METHOD = KMEANS).

Diese Variante, die für das weitere Vorgehen zum Einsatz kam, produziert bei der Gruppe der kleineren Gemeinden etwa 800 Anderszuordnungen. Eine zusätzliche Methode zur

Kontrolle einer gefundenen Clusterlösung besteht darin, zu überprüfen, inwiefern beim Einsatz eines völlig andersartigen Rechenverfahrens ein weitgehend deckungsgleiches Ergebnis erzielt wird und somit die errechneten Cluster mit einem anderen statistischen Verfahren reproduziert werden können.

Eine hierfür geeignete Vorgehensweise zur Überprüfung der Zuordnung jedes einzelnen Falles zu den identifizierten Clustern und ggf. deren Neuordnung stellt die Diskriminanzanalyse dar.

## 5.2 Diskriminanzanalyse

Die Diskriminanzanalyse ist ein strukturprüfendes Verfahren, welches die Ausprägung einer nominal skalierten Variablen in Abhängigkeit von einem Set unabhängiger, metrisch skalierten Variablen schätzt. Im vorliegenden Fall bedeutet dies, dass die Diskriminanzfunktion die Zugehörigkeit der einzelnen Fälle (Raumeinheiten) zu den identifizierten Clustern schätzt.

Das Verfahren beruht auf der Berechnung einer Diskriminanzfunktion:

$$D = b_0 + b_1 * X_1 + \dots + b_n * X_n$$

$X_1$  bis  $X_n$ : unabhängige (erklärende Variablen)  
b Koeffizienten

Der solchermaßen für jeden einzelnen Fall (jede Raumeinheit) berechnete Diskriminanzwert stellt im zweiten Schritt die Grundlage der wahrscheinlichkeits-theoretisch begründeten Zuordnung der einzelnen Fälle zu den verschiedenen Ausprägungen der abhängigen, nominal skalierten Variablen (hier: Clusterzugehörigkeit).

Im Fall der 82 Großstädte wurden von der Diskriminanzanalyse 100 % und bei der Gruppe der 2877 kleineren Gemeinden 91,5 % der Fälle als korrekt klassifiziert ausgewiesen. Dies ist ein höchst zufriedenstellendes Ergebnis und bestätigt die Gültigkeit der gefundenen Clusterlösungen. Man könnte nun noch einige von der Diskriminanzanalyse unterschiedlich klassifizierte Fälle wiederum umgruppieren. Hierauf wurde aber wegen des bereits sehr guten Ergebnisses der Clusterzentrenanalyse verzichtet.

## 6. Endgültige Clusterlösungen

Die endgültigen Clusterlösungen sind in folgenden Schritten entstanden:

- Aufteilung des Datensatzes in zwei Teildatensätze: 82 Großstädte (100.000 Einwohner) und 2877 restliche Städte und Gemeinden
- Durchführung einer hierarchischen Clusteranalyse für jeden Teildatensatz mit den acht Indikatoren: Bevölkerungsentwicklung 2003 bis 2020, Medianalter 2020, Arbeits-

platzzentralität, Arbeitsplatzentwicklung, Arbeitslosenquote, Steuereinnahmekraft, Anteil Hochqualifizierte, Anteil Mehrpersonenhaushalte mit Kindern

- Wahl der Clusteranzahl: sechs Cluster beim Teildatensatz „Großstädte“, neun Cluster beim Teildatensatz „Städte und Gemeinden < 100.000 EW“
- Korrektur der gefundenen Lösungen durch eine nachgeschaltete Clusterzentrenanalyse
- Überprüfung der Lösungen mit einer Diskriminanzanalyse

Die endgültigen Lösungen weisen folgende Verteilung auf:

**Tabelle 5<sup>3</sup>: Großstädte >= 100.000 EW – sechs Cluster**

	Anzahl	Prozent
Cluster 1	21	25,6
2	19	23,2
3	19	23,2
4	11	13,4
5	7	8,5
6	5	6,1
Gesamt	82	100,0

**Tabelle 6<sup>4</sup>: Städte und Gemeinden < 100.000 EW – neun Cluster**

	Anzahl	Prozent
Cluster 1	514	17,9
2	90	3,1
3	740	25,7
4	579	20,1
5	361	12,5
6	165	5,7
7	71	2,5
8	5	,2
9	352	12,2
Gesamt	2877	100,0

<sup>3</sup> Die Reihenfolge der Cluster findet sich in einer geänderten, aktuelleren Variante im Online-Wegweiser unter [www.wegweiserdemographie.de](http://www.wegweiserdemographie.de).

<sup>4</sup> Die Reihenfolge der Cluster findet sich in einer geänderten, aktuelleren Variante im Online-Wegweiser unter [www.wegweiserdemographie.de](http://www.wegweiserdemographie.de).

## 7. Erneute Überprüfung der Zuordnung mit Daten aus 2005

Nach Vorliegen der Daten aus dem Jahr 2005 zu den Indikatoren der Clusteranalyse (vgl. Seite 2) sind die Zuordnungen der Gemeinden zu Clustertypen erneut überprüft worden. Die Gemeinden, bei denen sich die Indikatoren von 2003 auf 2005 wesentlich geändert hatten, sollten im Zuge der Überprüfung auf die Cluster neu verteilt werden. Anschließend ist es möglich, die Gemeinden zu identifizieren, die inzwischen einem anderen Cluster angehören und ggf. die dafür verantwortlichen Indikatoren zu analysieren.

Da es nicht um die komplette Neuerstellung von Demographietypen gehen sollte, baute die Analyse auf den bestehenden Ergebnissen auf. Die Clusteranalyse wurde mit der SPSS-Prozedur QUICK CLUSTER durchgeführt. Dabei wurden die bei der Version 2003 ermittelten Clusterzentren (d. h. hier: die arithmetischen Mittelwerte der standardisierten beteiligten Variablen innerhalb der einzelnen Cluster) und die Clusteranzahl als Vorgabe verwendet. Die aktuell vorliegenden Indikatoren wurden zuvor standardisiert, sodass jeder einen Mittelwert von null und eine Standardabweichung von eins aufweist (z-Transformation).

Generell konnten alle Gemeinden in die Überprüfung einbezogen werden, wenn die acht Indikatoren der Clusteranalyse vollständig vorlagen. Alle Gemeinden, die Ziel einer Gebietsänderung waren, wurden wegen nicht mehr gültiger bzw. nicht vorhandener Prognosezahlen von der Clusterzentrenanalyse ausgenommen. Die Bundesagentur für Arbeit kann ebenfalls keine Arbeitslosenzahlen für Gemeinden liefern, die von Gebietsänderungen betroffen waren. Diese fallen daher ebenfalls für die Analyse aus.

Bei der Clusterzentrenanalyse wurde die von SPSS zur Verfügung gestellte Variante KMEANS verwendet, die bereits bei der nachgeschalteten Analyse der 2003er Daten zum Einsatz kam. Hierbei werden die Clusterzentren nach der Neuordnung neu berechnet und eine weitere Zuordnung vorgenommen. Dieses Verfahren wird so lange wiederholt, bis das Neuordnen keine oder nur minimale Änderungen gegenüber dem vorigen Ergebnis ergibt. Dieses Verfahren berücksichtigt besonders die Tatsache, dass sich die Clusterzentren seit der letzten Berechnung (Datenstand 2003) vermutlich leicht geändert haben, und versucht, eine möglichst korrekte Lösung hinsichtlich der aktuellen Verteilung zu erreichen. Durch das Neuberechnen der Clusterzentren während der Analyse ergibt sich andererseits eine gewisse Abweichung der Clustermittelwerte gegenüber der bisherigen Fassung.

Wie bisher kam nach der Clusteranalyse als strukturprüfendes Verfahren wiederum die Diskriminanzanalyse zum Einsatz. Da diese auf anderen mathematischen Grundlagen beruht als die Clusteranalyse, kann sie für eine unabhängige Überprüfung bzw. Korrektur der Clusterergebnisse herangezogen werden.

Nach dem ersten Lauf der Diskriminanzanalyse wurde ein kleiner Teil der Gemeinden unter 100.000 Einwohner umklassifiziert. Hier wurden nur solche Fälle berücksichtigt, bei denen die Diskriminanzanalyse eine sehr hohe Wahrscheinlichkeit für die Korrektheit der Umgruppierung ausgab. Schließlich bescheinigte eine erneute Diskriminanzanalyse eine 100%ige Korrektheit der Zuordnung bei den Großstädten und eine 92%ige Korrektheit bei den Städten und Gemeinden mit unter 100.000 Einwohnern. Dieser Wert ist sogar

geringfügig besser als der bei der 2003er-Lösung (91,5 %). Man kann daher von der Gültigkeit der gefundenen Lösung ausgehen.

Die Verteilung nach Stand der Daten aus 2005 (Ausschluss Gemeinden ohne alle acht CA-Indikatoren) stellt sich wie folgt dar:

**Tabelle 7: Großstädte  $\geq$  100.000 EW – sechs Cluster**

	Anzahl	Prozent
Cluster 1	14	17,3
2	20	24,7
3	6	7,4
4	18	22,2
5	15	18,5
6	8	9,9
Gesamt	81	100,0

**Tabelle 8: Städte und Gemeinden  $<$  100.000 EW – neun Cluster**

	Anzahl	Prozent
Cluster 1	468	16,4
2	63	2,2
3	407	14,3
4	361	12,6
5	714	25,0
6	637	22,3
7	117	4,1
8	83	2,9
9	5	0,2
Gesamt	2.855	100,0

Januar 2007