Software-Beschreibung

Die Beschreibung erfolgt anhand eines Beispiels aus der Regional-Statistik der Datei (Demo-) *Regio2003.xls*. Von mathematisch-technischen Inhalten ist versucht möglichst zu abstrahieren. Abschließend sind die wesentichen Techniken anhand eines Beispiels zur Prüfung des Zusammenhanges zwischen der Bundestagswahlbeteiligungen 2002 und der Arbeitslosigkeit sowie deren Veränderungen gegenübergestellt.

Es sind nicht-parametrische Verfahren für die Erkenntnisgewinnung in Fragestellungen multivariater Natur realisiert, die insbesondere dem Problem der Unschärfen statistischer Erkenntnisgewinnung Rechnung tragen, wie diese beispielsweise aus der Berücksichtigung sowohl des Zählmasses wie auch der metrischen Lage von Daten für die Erkenntnisgewinnung resultieren. Wesentlich sind relationale Betrachtungen funktionalen vorgezogen und im Sinne einer möglichst unaufwendigen Verständnisbildung implementiert.

Die heute mit OLAP-Techniken zumeist einfach mögliche Verdichtung von Daten im Sinne von Vorgruppierung auf wesentliche Blickrichtungen lässt die Techniken auch für den Einsatz in betrieblichen Anwendungen einsetzbar erscheinen.

Struktur

Für die Anwendung sind das Programm Excel ab der Version `97 sowie Grundkenntnisse im Umgang mit Excel erforderlich. Von Bedeutung sind die offenen (im Folgenden „Blatt“ genannten) Tabellen:

- **Dialog**
- **Analyse**
- **Cluster**
- **Treiber**
- **Ranking**
- **Kennzahlen**
- **Beispiele**

Das Blatt ***Dialog*** dient der Steuerung der Anwendung, das Blatt ***Ranking*** der Darstellung von Filter- bzw. Ranking-Ergebnissen die zur Realisierung von Rekursionen auch als neue Kennzahlen übernommen werden können. Für Untersuchungen wesentlich ist das Blatt ***Analyse*** sowie die dem Blatt Analyse nachgelagerten Blätter ***Cluster*** und ***Treiber***, mit denen wesentliche Informationen zu ausgewählten Kennzahlkonstellationen insbesondere zu deren Zusammenwirken bereitgestellt sind bzw. gewonnen werden können. Auf dem Blatt ***Beispiel*** sind die ***log-files*** zu abgelegten Ergebnissen abrufbar, womit es auf einfache Weise möglich ist die Ergebnisse zu variieren und auch durch spielen

der Parameter Einblicke in die Funktionsweise der Techniken zu erhalten. Mit der Datei gegebene Kennzahlen sind über die Seite ***Kennzahlen*** dargestellt.

Darüber hinaus ist ein Blatt ***Kurzbeschreibung*** beigefügt, dass für technisch interessierte auch die im Folgenden nicht angesprochene Funktionalität beschreibt[1].

Mit den Techniken weniger vertraute Anwender sollten Änderungen bzw. Eingaben nur in den weiß unterlegten Parameter-Feldern vornehmen.

Dialog

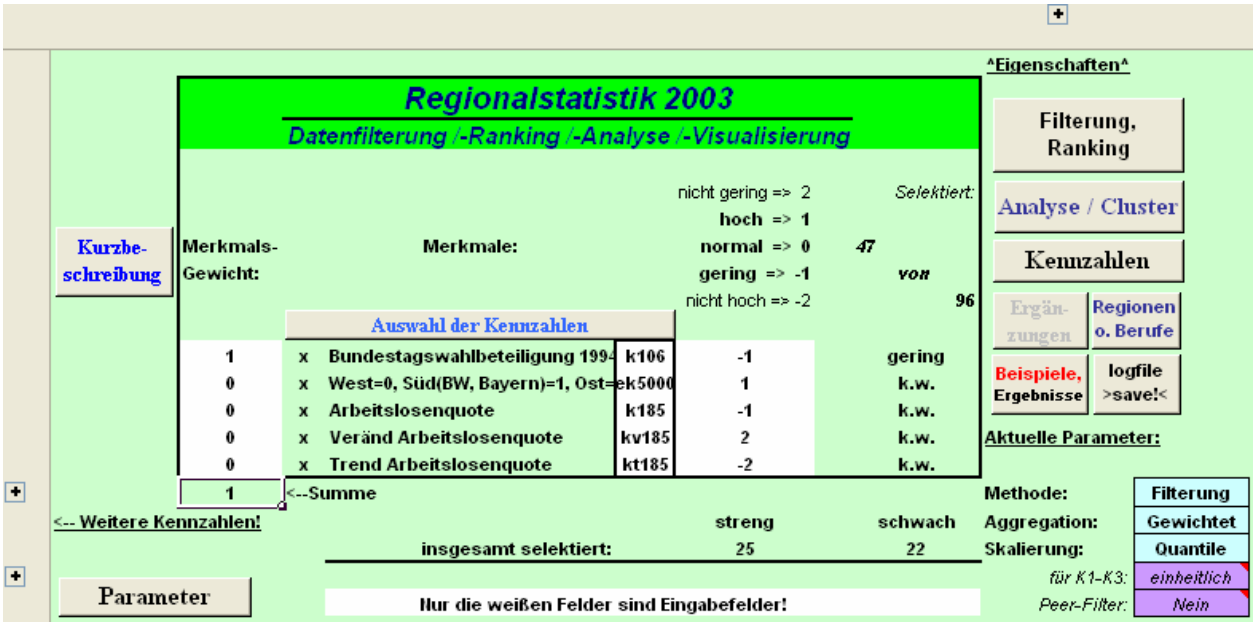


Abbildung 1: Blatt ***Dialog*** (Beispiel 1)

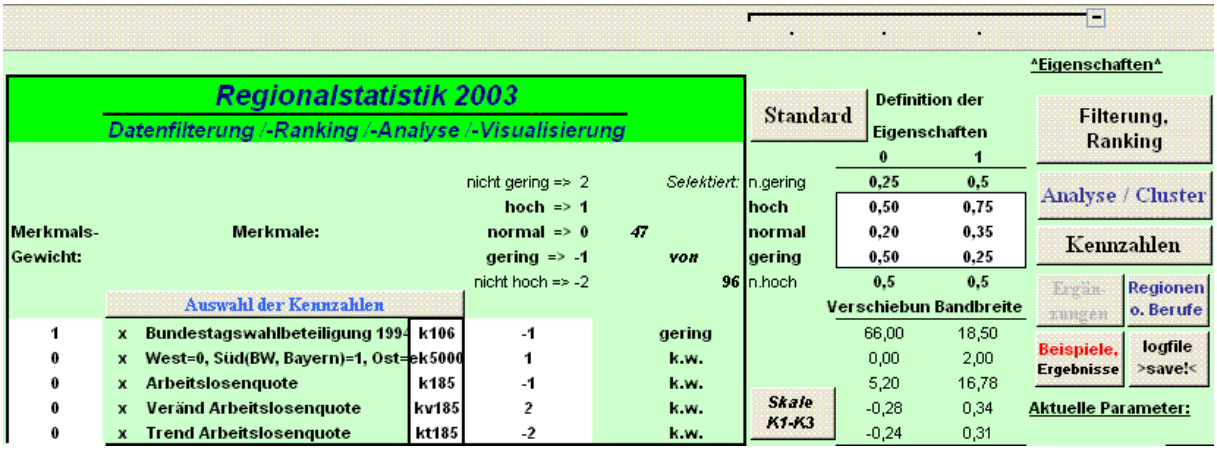
Mit dem Blatt Dialog werden über ***Kennzahlschlüssel,*** in Abbildung 1 zum Beispiel k106, ek5000, k185, kv185 und kt185, die zu betrachtenden Merkmale, im Beispiel die Bundestagswahlbeteiligung des Jahres 2002 (k106), eine Kennzahl die mit den Ausprägungen 0, 1, 2 zwischen dem Westen, dem Süden und dem Osten Deutschlands unterscheidet (ek5000) sowie die Arbeitslosenquote (k185), deren Veränderung (kv185) und deren Trend (kt185) ausgewählt. Es wird hierbei zwischen Kennzahlen ***k***, deren Veränderung ***kv***, deren Trend ***kt*** und eigenen bzw. weiteren Kennzahlen ***ek*** unterschieden (In den Versionen zur Versicherungs- und Bausparwirtschaft sind auch Kennzahl-Durchschnitte ***kd*** betrachtet). Über den Button ***Kennzahlauswahl*** ist es möglich, die Kennzahlen sowie deren Typ über Listboxes auzuwählen. Mit dem Gruppierungssymbol [+] können - neben den

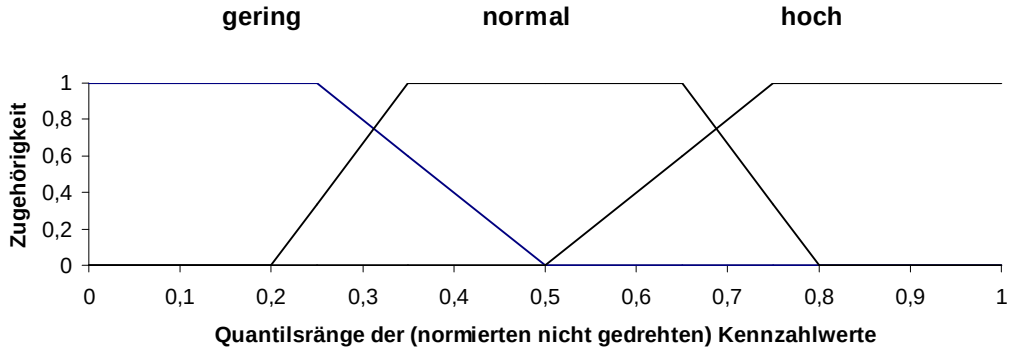
[1] Einen ausführlicheren Einstieg in die Techniken ermöglicht die Datei http://www.rankingweb.de/MANUAL.pdf oder die mit Beispielen versehenen Buchpublikationen zu speziellen Themen, die über die Seite http://www.rankingweb.de/Buch.html zugänglich sind.

aufgeführten - 7 weitere Kennzahlen festgelegt werden. Zur Verfügung stehende Kennzahlen sind über den Button Kennzahlen einsehbar.

Nicht dargestellt ist die in der Buchversion gegebene Möglichkeit über den Button „***Daten Regio2002***“ auch die jeweils aktuelle Kennzahlauswahl der Datei ***Regio2002.xls*** des Vorjahres zu übernehmen sofern vorhanden und geöffnet. So können auch die Beispiele aus Vorjahresdateien aktualisiert betrachtet werden.

Sollen die Merkmale für ein Ranking oder eine Filterung verwendet werden, so ist ihnen ein ***Gewicht*** und eine ***Eigenschaft*** zuzuordnen. Als ausgewählt gelten Kennzahlen, die ein von 0 verschiedenes Gewicht erhalten, zumeist 1. Die Eigenschaften können über das so unterschriebene Gruppierungssymbol [+] verändert werden, wozu hier auch die ***Standardeinstellung*** abrufbar ist.





Strenge Filtereigenschaften

Zugehörigkeit	Region	Bundestagswahlbeteiligung 2002
100%	Zwickau, Kreisfreie Stadt	70,8
100%	Trier, Kreisfreie Stadt	74,8
100%	Thüringen	74,8
100%	Schwerin, Kreisfreie Stadt	72,8
100%	Sachsen-Anhalt	68,8
100%	Sachsen	73,7
100%	Rostock, Kreisfreie Stadt	71,4
100%	Offenbach am Main, Kreisfreie Stadt	74,1
100%	Neumünster, Kreisfreie Stadt	74,7
100%	Mönchengladbach, Kreisfreie Stadt	74,4
100%	Mecklenburg-Vorpommern	70,6
100%	Magdeburg, Kreisfreie Stadt	68,9
100%	Leipzig, Kreisfreie Stadt	73,8
100%	Kaiserslautern, Kreisfreie Stadt	73,1
100%	Halle (Saale), Kreisfreie Stadt	70,2
100%	Gera, Kreisfreie Stadt	73,9
100%	Gelsenkirchen, Kreisfreie Stadt	75,0
100%	Flensburg, Kreisfreie Stadt	74,6
100%	Erfurt, Kreisfreie Stadt	75,0
100%	Dresden, Kreisfreie Stadt	75,0
100%	Dessau, Kreisfreie Stadt	70,1
100%	Cottbus, Kreisfreie Stadt	71,3
100%	Chemnitz, Kreisfreie Stadt	74,5
100%	Brandenburg an der Havel, Kreisfreie Sta	66,0
100%	Brandenburg	73,7
96%	Worms, Kreisfreie Stadt	75,1
96%	Ingolstadt, Kreisfreie Stadt	75,1
84%	Bremerhaven, Kreisfreie Stadt	75,4
68%	Pforzheim, Kreisfreie Stadt	75,8
60%	Mannheim, Universitätsstadt, Kreisfreie S	76,0
48%	Fürth, Kreisfreie Stadt	76,3
44%	Lübeck, Hansestadt, Kreisfreie Stadt	76,4
44%	Hagen, Kreisfreie Stadt	76,4
40%	Ludwigshafen am Rhein, Kreisfreie Stadt	76,5
40%	Kassel, Kreisfreie Stadt	76,5
40%	Duisburg, Kreisfreie Stadt	76,5
24%	Krefeld, Kreisfreie Stadt	76,9
24%	Köln, Kreisfreie Stadt	76,9
20%	Wilhelmshaven, Kreisfreie Stadt	77,0
20%	Wiesbaden, Landeshauptstadt, Kreisfreie	77,0
20%	Frankfurt am Main, Kreisfreie Stadt	77,0
16%	Herne, Kreisfreie Stadt	77,1
12%	Remscheid, Kreisfreie Stadt	77,2
12%	Oberhausen, Kreisfreie Stadt	77,2
12%	Delmenhorst, Kreisfreie Stadt	77,2
12%	Augsburg, Kreisfreie Stadt	77,2
8%	Potsdam, Kreisfreie Stadt	77,3

Segment: Bundestagswahlbeteiligung 2002 "gering"

(Filterergebnis des Beispiels 2)

Das Ergebnis von **Filterungen** wird mit der Anzahl der streng (Zugehörigkeit 1 bzw 100%) bzw. schwach (Zugehörigkeiten größer 0 und kleiner 1) gefilterten Objekte angezeigt. Die Eigenschaften hoch, normal, gering, nicht hoch und nicht gering sind Transformationen der Kennzahlwerte auf Werte von 0 bis 1 und damit auf Kennzahlwerte, die die Eigenschaften vollständig (strenge Zuordnung 1) bzw. schwach erfüllen (schwache Zuordnung größer 0 und kleiner 1).

Durch betätigen des Buttons „***Filterung/Ranking***“ oben werden die Regionen links, deren Wahlbeteiligung der Bundestagswahl 2002 „gering“ war, gefiltert. Diesem Segment werden 22 Regionen schwach und weitere 25 Regionen streng zugeordnet, was mit der Definition der Eigenschaft „gering“ festgelegt ist und wobei hier die nicht gleichmäßige Aufteilung durch Wiederholungen von Werten der Wahlbeteiligungen hervorgerufen ist. In der strengen Zuordnung sind die Regionen alphabetisch angeordnet.

Mit Wahl des Parameters *Methode = 0* und der Eigenschaft „gering“ werden alle Regionen bezüglich der Wahlbeteiligung absteigend mittels eines ***Rankings*** dargestellt.

Zur Verknüpfung mehrerer Merkmale bzw. deren Eigenschaften sind verschiedene ***Aggregationen*** möglich, die das logische UND, das logische ODER und verschiedene mittelnde Verknüpfungen der Zugehörigkeitswerte realisieren, so auch deren Produkt. Von Bedeutung ist hier im wesentlichen die Verknüpfung „***Gewichtet***“, die eine exclusive gewichtete Durchschnittsbildung darstellt, wobei exclusive bedeutet, dass nur Objekte, die alle gewählten Eigenschaften mit von 0 verschiedenen Zugehörigkeiten erfüllen, gefiltert werden. Als ***Methode*** sollten Sie ***Filterung*** eingestellt lassen sofern Sie nicht ein vollständiges Ranking aller Objekte bezüglich ausgewählter Merkmale anstreben. Nur die Methode Filterung unterstützt die im Blatt Analyse vornehmbare Zufallsbereinigung der Kennzahl- bzw. Merkmalsausprägungen.

Es sei noch erwähnt, dass über den Button „***Regionen o. Berufe***“ zur Berufsgruppenstatistik gewechselt werden kann. Die Berufsgruppenstatistik war Gegenstand des Ergänzungsbandes zum Großstädte-Ranking 2002 und ist hier erweitert um die Arbeitsamtsstatistik der Vollständigkeit halber aktualisiert aufgenommen. In den sonstigen Analyse-Dateien kann hier zur Aufnahme eigener Kennzahlen oder anderen speziellen Kennzahlbereichen gewechselt werden.

- **Analyse**

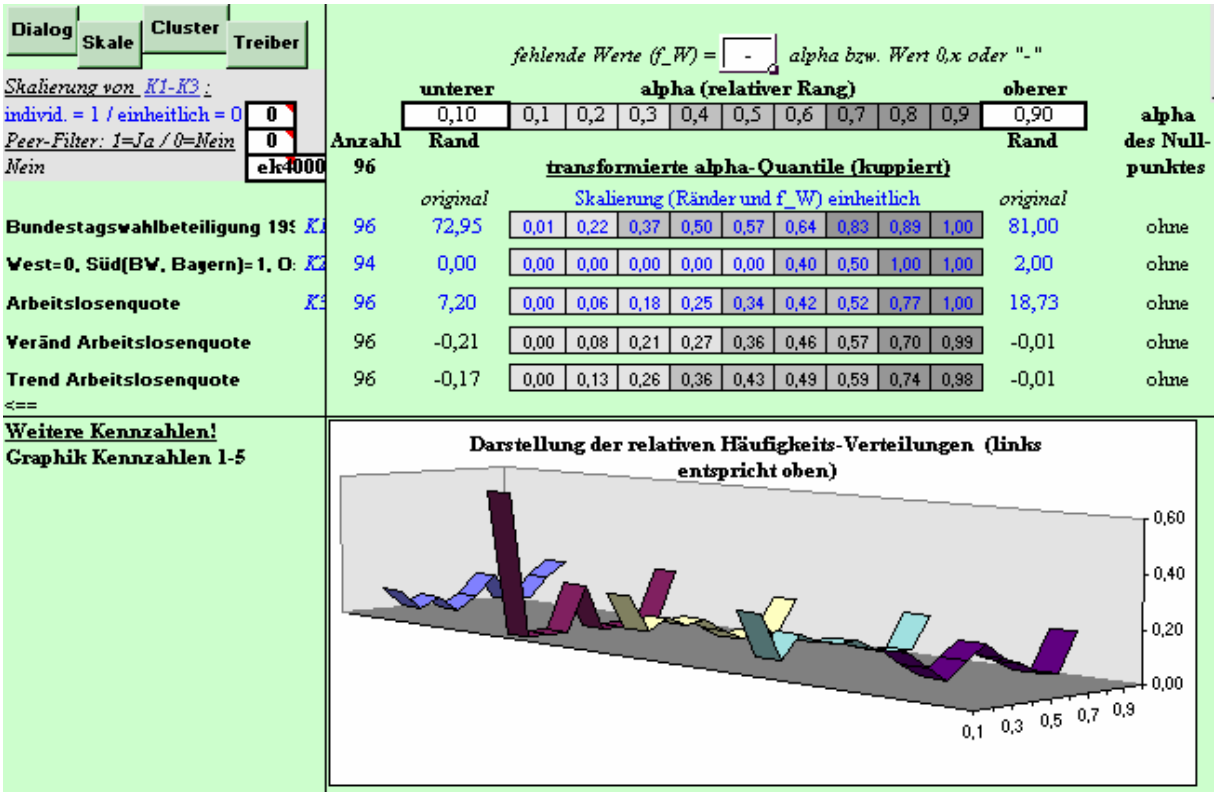


Abbildung 2: Blatt ***Analyse***

Das Blatt Analyse stellt für die ersten fünf Merkmale die Häufigkeitsverteilungen deren normierten Merkmalsausprägungen ***graphisch*** dar sowie alle Merkmale auch mit den normierten Skalenwerten anhand von Skalen[2]. Für eine ***Zufallsbereinigung*** der Kennzahlwerte können die Verteilungen am unteren und oberen Ende gestutzt (kupiert) werden. ***Fehlende Kennzahlwerte*** können in den Aggregationen für Filterungen gänzlich unberücksichtigt bleiben (so die Voreinstellung) oder aber mit transformierten Kennzahlwerten zwischen 0 und 1 beliebig festgelegt werden. Für die Clusterung wird bei der Wahl ``-`` für fehlende Werte 0,5 ersetzt, da hier ein numerischer Wert vorliegen muß. Ist im Blatt Dialog die ***Skalierung*** mit „Quantile“ eingestellt, werden die Kupierungsgrenzen und die Wahl für den Wert fehlender Werte sowie auch die Grenzen der Eigenschaften oben als Quantilsränge interpretiert, was die sinnvollste Wahl für die Erzielung ausreichender Segmentstärken darstellt. Mit

[2] Ist als Skalierung “Quantile” gewählt, sind in den Skalen die Quantilswerte dargestellt, ist die Skalierung „linear“ gewählt, sind die Quantilsränge der äquidistanten Aufteilung der Kennzahlwerte dargestellt.

der Einstellung der Skalierung als „linear“ werden die Werte als die normierten Skalenwerte interpretiert.

Des Weiteren können über den Button ***Skale*** die ersten drei Merkmale im Sinne der Filter-Eigenschaften und der Zufallsbereinigung individuell transformiert werden sofern hierzu auf dem Blatt Analyse die entsprechende Option mit 1 belegt wird.

Über sogenannte **Peer-Filter** kann zusätzlich die Datenbasis eingeschränkt werden. In der Datenbasis des Vorjahres sind die Peer-Filter post, pnrw und pwest zur Einschränkung auf die Regionen der Neuen Bundesländer, Nordrhein-Westfalens und der Alten Bundesländer enthalten, in der aktuellen Datenbasis nur noch der Peer-Filter ek4000, der die enthaltenen Nicht-Städte ausblendet . Sie können eigene Peer-Filter bilden und über „Ergänzungen“ einbinden oder aus der Datei des Vorjahres übernehmen. Peer-Filter sind Kennzahlen die alle Objekte ausblenden, denen in der Peer-Kennzahl ein von 1 verschiedener Wert zugeordnet ist.

Diskriminanz und Segment-Statistik

Dialog | Skale | Cluster | Treiber | Skalen, Graphik | Segment ablegen | Zugehörigkeit | schwach | streng | Segment-Statistik

Summe: 33,4 | 22 | 25

Skalierung von K1-K3 :
individ. = 1 / einheitlich = 0: 0
Peer-Filter: 1=Ja / 0=Nein: 0
Nein: ek4000

	Mittelwert original (kuppiert)	Streuung original (kuppiert)	Segment Gewicht	Eigenschaft	Streuung Träger	Streuung Träger	Mittelwert gewichtet (relative Abweichungen)	Streuung streng	Streuung streng	Anzahl:
Bundestagswahlbeteiligung 19 K1	77,383	2,662	1	gering	-41%	-3%	-4%	-5%	-70%	47
Vest=0, Süd(BW, Bayern)=1, O: K2	0,617	0,814	0	k.w.	16%	48%	89%	133%	10%	47
Arbeitslosenquote K3	11,844	3,843	0	k.w.	2%	18%	28%	39%	-23%	47
Veränd Arbeitslosenquote	-0,127	0,068	0	k.w.	9%	22%	37%	55%	-8%	47
Trend Arbeitslosenquote	-0,098	0,049	0	k.w.	6%	22%	35%	53%	-19%	47

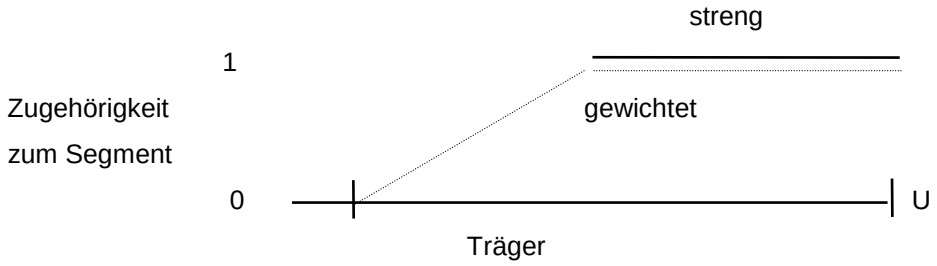
Abbildung 3: *Segment-Statistik*

Über das mit „***Segment-Statistik***“ unterschriebene Gruppierungssymbol [+] kann rechtsseitig der Skalen zur Ansicht der Segment-Statistik gewechselt werden. In der Abbildung oben ist so das Segment der Regionen mit „gering“ (-wertigen) Wahlbeteiligungen von oben dargestellt.

Die vorhandenen Informationen zum gebildeten Segment werden geschichtet betrachtet, um auch innere Aussagen bezüglich des betrachteten Segmentes erhalten zu können. Hierzu wird das jeweils betrachtete Segment über die Segment definierenden Eigenschaften – hier „gering(e)“ Wahlbeteiligung - in Regionen aufgeteilt, die ihm vollständig (streng) angehören und in solche, die die Segmenteigenschaften mehr oder weniger (schwach) erfüllen. So kann mit einfachen Mitteln eine Segment-Statistik gebildet werden, die die an die Korrelation angelehnte Blickrichtung

„***je mehr die Eigenschaften des Segmentes gegeben sind, desto ...***“

für beliebige ein- oder mehrdimensionale Objekteausschnitte ermöglicht.



Die betrachteten Objekte u erhalten also bezüglich der Segment definierenden (abhängigen) Merkmale einen Zugehörigkeitswert $\mu(u) \in [0; 1]$ zu dem jeweils betrachteten Segment, wie dies auch oben mit dem Beispiel 2 der Datei *Regio2003.xls* dargestellt ist.

Für die **Segment-Statistik** werden dann sowohl in den Segment definierenden Merkmalen wie auch in den übrigen (erklärenden bzw. unabhängigen) Merkmalen

- mit dem **Träger** alle Objekte betrachtet, die dem Segment mit einer Zugehörigkeit $\mu(u) > 0$ zugeordnet sind und deren (ungewichteter) Mittelwert m_1 sowie deren ebensolche Streuung s_1 ermittelt.

- Weiter wird der bezüglich der Zugehörigkeiten **gewichtete** Mittelwert m_2 betrachtet

- und schließlich der (ungewichtete) Mittelwert m_3 und die ebensolche Streuung s_3 der Objekte, die dem Segment mit Zugehörigkeit $\mu(u) = 1$ **streng** zugeordnet sind.

So läßt sich mittels verringerter Streuungen gegenüber der Gesamtstreuung s_0 der jeweiligen Kennzahlwerte beurteilen ob Beobachtungen **charakteristisch** für das betrachtete Segment sind, während die Mittelwerte m_1 bis m_3 für sich und im Vergleich zum Mittelwert m_0 aller Unternehmen eine eventuell vorhandene **Systematik** bezüglich der vorgegebenen Blickrichtung erkennbar machen. Wir nennen genauer Mittelwerte m_1 bis m_3 **systematisch im Sinne der Blickrichtung**, wenn $m_0 \le m_1 \le m_2 \le m_3$ oder $m_0 \ge m_1 \ge m_2 \ge m_3$ gilt. In den Segmentstatistiken sind die Abweichungen der Mittelwerte bzw. Streuungen vom Gesamtmittelwert bzw. von der Gesamtstreuung relativ in der Form

$$\frac{a_i - a_0}{\mathrm{abs}(a_0)}$$

aufgeführt, womit auch die Richtung der Abweichung direkt sichtbar wird. Als Streuungsmaß ist jeweils die (empirische) Standardabweichung betrachtet.

Im Beispiel oben wird somit deutlich, dass dort, wo die Wahlbeteiligungen geringer sind, diese in der strengen Zuordnung im Mittel um 5% geringer gegenüber dem zufallsbereinigten Gesamtmittel von 77,4% ist und dies mit um 70% reduzierter Streuung. Konstruktionsbedingt sind die Beobachtungen im

Segmentbestimmenden Merkmal auch systematisch und charakteristisch im obigen Sinn. Systematisch aber großteils mit gegenüber der Gesamtstreuung erhöhter Streuung und damit nicht charakteristisch sind dann die übrigen erklärenden Merkmale jeweils deutlich systematisch (im Sinne der Blickrichtung) erhöht. In der Kennzahl ek4000 bedeutet dies, das wir hier ein größerer Anteil der Regionen des Ostens Deutschlands (Wert 2) beobachtet wird, wobei der hohe Anteil der mit beobachteten Werte des Westens (Wert 0) die höhere Streuung verursacht.

		Zugehörigkeit	schwach	streng		
		Summe:	36,0	21	24	
Segment		Streuung		Mittelwert		Streuun
		Träger	Träger	gewichtet	streng	streng
Gewicht	Eigenschaft	relative Abweichungen				
1	hoch	-58%	3%	4%	4%	-85%
0	k.w.	-46%	-59%	-56%	-46%	-42%
0	k.w.	-37%	-17%	-21%	-24%	-50%
0	k.w.	-28%	-20%	-27%	-28%	-34%
0	k.w.	-27%	-20%	-26%	-26%	-29%

Für eine **Diskriminanz-Analyse** kann über den Button **„Segment ablegen“** der aktuelle Segment-Ausschnitt am rechten Seiten-Ende abgelegt werden, so dass bei zusätzlicher Auswahl eines konträren Segmentes eine Diskriminanz-Betrachtung durchgeführt werden kann.

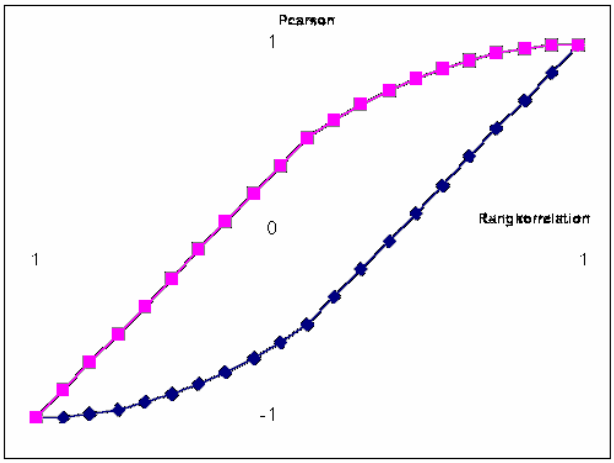
Das dem obigen analoge Segment „hoch“ der Regionen mit hohen Wahlbeteiligungen, weist dann zunächst Konstruktions bedingt systematisch und charakteristisch höhere Wahlbeteiligungen auf und auch in den übrigen Kennzahlen werden alle Beobachtungen sehr charakteristisch also mit deutlich verringerter Streuung gemacht. Die Veränderungen der Arbeitslosigkeit sowie die Arbeitslosigkeit selbst ist darüber hinaus umso geringer je deutlicher die Wahlbeteiligungen der Regionen mit den hohen Wahlbeteiligungen zugeordnet werden können. Nicht systematisch aber ebenfalls deutlich werden hier vor allem Regionen des Westens und des Südens beobachtet.

Korrelationen

Peer-Filter: 1=Ja / 0=Nein	0								
Nein	ek4000	Anzahl:	Rangkorrelationen					Merkmal_1:	4
								Merkmal_2:	5
			[1]	[2]	[3]	[4]	[5]	[6]	
Bundestagswahlbeteiligung 19	K1	45	(1)	-0,5	-0,5	-0,4	-0,4		
West=0, Süd(BW, Bayern)=1, O	K2	43	-0,5	(1)	0,3	0,3	0,2		
Arbeitslosenquote	K3	45	-0,7	0,5	(1)	0,6	0,6		
Veränd Arbeitslosenquote		45	-0,6	0,5	0,8	(1)	0,8		
Trend Arbeitslosenquote		45	-0,6	0,5	0,8	0,9	(1)		

Über das mit „***Korrel***“ unterschriebene Gruppierungssymbol [+] ist es ferner möglich die paarweisen Korrelationen der Merkmale zu betrachten. Die übliche metrische sogenannte Pearsonsche Korrelation ist jeweils aktuell im unteren Dreieck dargestellt. Die Rangkorrelationen des oberen Dreiecks sind unter Angabe des Kennzahl-Paares jeweils einzeln zu berechnen, da dies zeitaufwendiger ist. Folgende Graphik verdeutlicht im inneren der beiden dargestellten Kurven die

möglichen Größenverhältnisse zwischen den metrischen und den Rangkorrelationen. Eine Rangkorrelation von 0 kann mit Pearsonschen Korrelationen von –0,5 bis 0,5 beobachtet werden.



Im Beispiel oben zeigen sich sehr deutliche Korrelationen der Merkmale. Die Rangkorrelation ist absolut immer schwächer als die metrische Korrelation des unteren Dreiecks, was allgemein gilt und wie auch das Schaubild oben zeigt[3].

Die aufgeführten Anzahlen sind die des ausgeblendeten Segmentes und keine Einschränkung der Datenbasis für die Korrelationen, wie sie aber über Peer-Filter vorgenommen werden könnte. Einfluß auf die Korrelationen hat die Kupierung und die Festlegung fehlender Werte, nicht der Ausschnitt des aktuellen Segmentes.

Diskriminanz-Analyse

Sind wie in Abbildung 4 zwei unterschiedliche Segmente gegenübergestellt (diskriminiert), kann aus der Kenntnis, das der über sogenannte Mahanalobis-Distanzen[4] gemessenene Abstand der

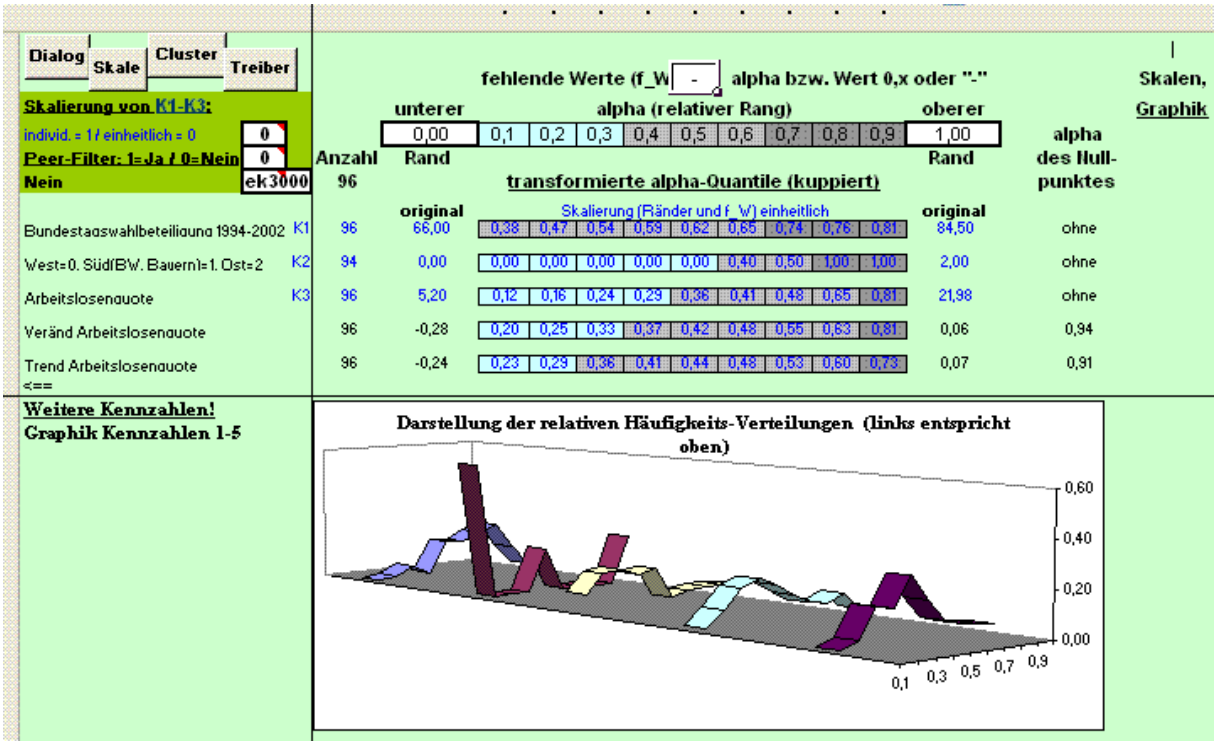
[3] Vgl Nelsen 1999: „An Introduction to Copulas“, S. 144

[4] Vgl J. Bortz: „Statistik für Sozialwissenschaftler“, Springer 5. Aufl. 1999, 551f und J. Van Eeghen, E.K. Greup, J.A. Nijssen: „Rate Making“, Surveys of Actuarial Studies No. 2, 1983, Nationale-Nederlanden N.V., Rotterdam, Niederlande.

Mit $D_k^2 = (\bar{x}_1 - \bar{x}_2)^T S_k^{-1} (\bar{x}_1 - \bar{x}_2)$, k die Zahl der berücksichtigten Merkmale

$$s_{pq} = \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^{2} \sum_{t=1}^{n_i} (x_{ipt} - \bar{x}_{ip})(x_{iqt} - \bar{x}_{iq}), \quad S_J = \{s_{p,q} : p, q = 1,..,J\}$$

Mittelwerte beteiligter Merkmale von diskriminierten Segmenten in der Veränderung bei Hinzunahme weiterer Merkmale einer Fischer-Verteilung genügt, die Signifikanz des Erklärungsgehaltes weiterer Merkmale für die Unterschiedlichkeit der Segmente geprüft werden, sofern die Grundvoraussetzung multivariat Normalverteilter Merkmalskombinationen gegeben ist.



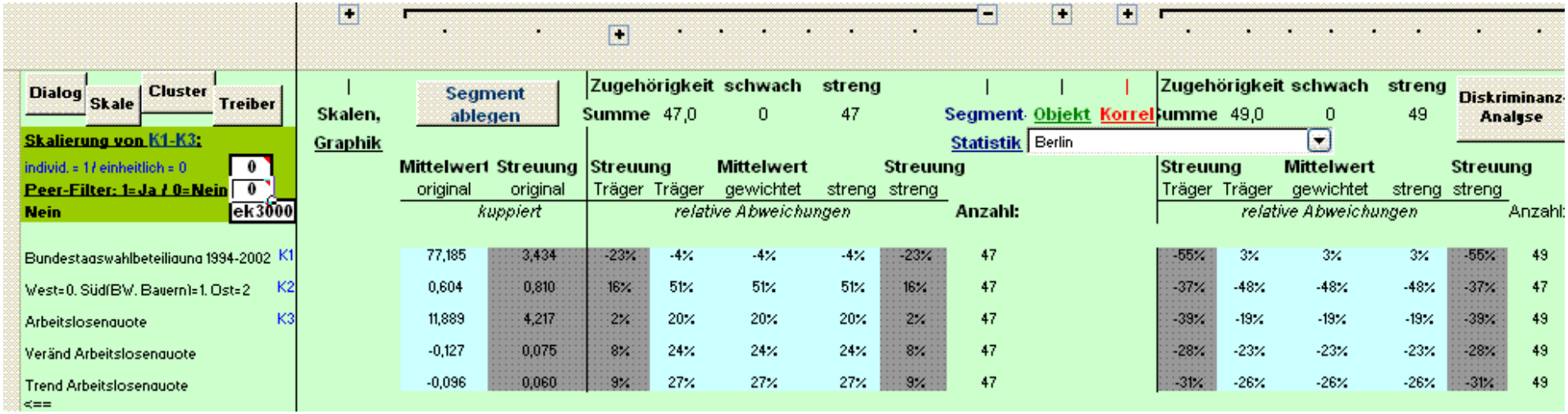
Blatt Analyse mit den Einstellungen des Beispiels 8

---

und der Kenntnis von $$F = \frac{(n_1 n_2)(n_1 + n_2 - k - 2)}{(n_1 + n_2)(n_1 + n_2 - 2)} \frac{\left[D^2_{k+1} - D^2_k\right]}{(1 + \frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 2)} D^2_k)}$$ als Fischer-Verteilung

mit (1, n_1+n_2-k-2) Freiheitsgraden zum Niveau α wird dann sukzessive über $F > F_\alpha(1, n_1+n_2-k-2)$ geprüft ob das Merkmal k+1 signifikant den Abstand der Segmente erhöht.

Intuitiv wird ersichtlich, dass je größer D^2_k ist, desto größer auch der diskriminierende Charakter der gewählten Gruppen im Sinne des Abstandes der Merkmale ist, wobei über S Abhängigkeiten der Merkmale mittels der (paarweisen) Korrelation der Merkmale in die Abstandsmessung mit einfliessen. Hierbei können für die Ermittlung der Matrix S_k^{-1}, k<12, Abhängigkeiten zu den übrigen jeweils betrachteten Merkmalen als abgegrenzte Betrachtung vernachlässigt werden oder aber über die mögliche Ermittlung aus der Matrix für k=12 berücksichtigt bleiben, was zu unterschiedlichen Ergebnissen führen kann.



Dialog | Skale | Cluster | Treiber

Skalierung von K1-K3:
individ. = 1 / einheitlich = 0 | 0
Peer-Filter: 1=Ja / 0=Nein | 0
Nein | ek3000

Skalen, Graphik | Segment ablegen

Zugehörigkeit schwach streng
Summe 47,0 0 47

Segment Objekt Korrel
Statistik | Berlin

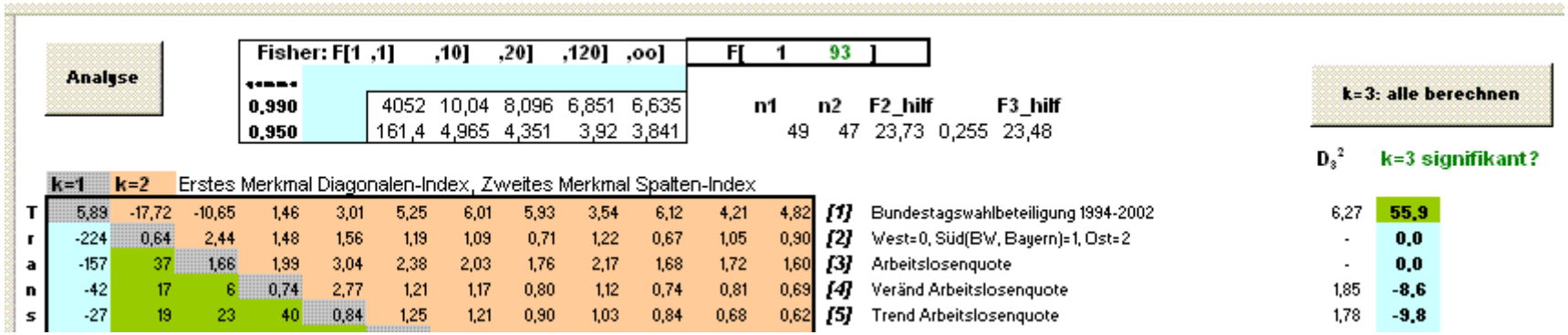
Zugehörigkeit schwach streng
Summe 49,0 0 49

Diskriminanz-Analyse

	Mittelwert original	Streuung original	Streuung Träger	Träger	Mittelwert gewichtet	streng	Streuung streng	Anzahl:	Streuung Träger	Träger	Mittelwert gewichtet	streng	Streuung streng	Anzahl:
	kuppiert		relative Abweichungen						relative Abweichungen					
Bundestagswahlbeteiligung 1994-2002 K1	77,185	3,434	-23%	-4%	-4%	-4%	-23%	47	-55%	3%	3%	3%	-55%	49
West=0, Süd(BW, Bayern)=1, Ost=2 K2	0,604	0,810	16%	51%	51%	51%	16%	47	-37%	-48%	-48%	-48%	-37%	47
Arbeitslosenquote K3	11,889	4,217	2%	20%	20%	20%	2%	47	-39%	-19%	-19%	-19%	-39%	49
Veränd Arbeitslosenquote	-0,127	0,075	8%	24%	24%	24%	8%	47	-28%	-23%	-23%	-23%	-28%	49
Trend Arbeitslosenquote	-0,096	0,060	9%	27%	27%	27%	9%	47	-31%	-26%	-26%	-26%	-31%	49

<==

Abbildung 4: Diskriminierung (Kennzahlen und Eigenschaften des Beispiels 8)



Analyse

Fisher: F[1 ,1] ,10] ,20] ,120] ,oo] | F[1 93]

0,990	4052	10,04	8,096	6,851	6,635
0,950	161,4	4,965	4,351	3,92	3,841

n1	n2	F2_hilf		F3_hilf
49	47	23,73	0,255	23,48

k=3: alle berechnen

	k=1	k=2	Erstes Merkmal Diagonalen-Index, Zweites Merkmal Spalten-Index												D_s^2	k=3 signifikant?
T	5,89	-17,72	-10,65	1,46	3,01	5,25	6,01	5,93	3,54	6,12	4,21	4,82	[1]	Bundestagswahlbeteiligung 1994-2002	6,27	55,9
r	-224	0,64	2,44	1,48	1,56	1,19	1,09	0,71	1,22	0,67	1,05	0,90	[2]	West=0, Süd(BW, Bayern)=1, Ost=2	-	0,0
a	-157	37	1,66	1,99	3,04	2,38	2,03	1,76	2,17	1,68	1,72	1,60	[3]	Arbeitslosenquote	-	0,0
n	-42	17	6	0,74	2,77	1,21	1,17	0,80	1,12	0,74	0,81	0,69	[4]	Veränd Arbeitslosenquote	1,85	-8,6
s	-27	19	23	40	0,84	1,25	1,21	0,90	1,03	0,84	0,68	0,62	[5]	Trend Arbeitslosenquote	1,78	-9,8

Abbildung 5: Diskriminanz-Analyse / Blatt Diskriminanz

Zur Erzielung größerer Übereinstimmung mit der Annahme multivariat normalverteilter Merkmalsausprägungen beteiligter Merkmale, sind mit den Einstellungen des Beispiels 8 wie oben dargestellt keine Kupierungen der Kennzahlwerte vorgenommen und die Eigenschaften hoch und gering des Merkmals „Bundestagswahlbeteiligung 2002“ für eine strenge Aufteilung in 49 und 47 Regionen gewählt.

Die Abbildung 5 oben stellt für k=1 in der Diagonalen grau unterlegt und für k=2 im oberen Dreieck orange unterlegt jeweils die Abstände aller möglichen Merkmalskombinationen der abgegrenzten Betrachtung dar, ein zusätzliches Merkmal (k=3) kann durch Berechnung hinzugefügt werden.

Das a-priori in zwei Gruppen aufteilende Merkmal [1] der „Bundestagswahlbeteiligung 2002“ erzeugt so einen Abstand von 5,89 der hohen und geringen Wahlbeteiligungen. Mit dem unteren Dreieck hellblau und grün dargestellt, sind die Werte der Teststatistik für die relative Veränderung der Abstände bei Hinzunahme eines Merkmals kleiner und größer oder gleich 5 unterschieden. Werte größer 5 deuten auf Signifikanz zum Niveau 0,05. Negative Werte zeigen hingegen die Fehlannahme multivariat normalverteilter Merkmalskombinationen auf.

Im Beispiel deutet die Merkmalskombination der Merkmale [3] und [5] mit dem Abstand von 3,04 die beste Erklärung der a-priori Aufteilung an, dies mit einem Wert der Teststatistik von 23 signifikant zum gewählten Niveau und darüber hinaus. Ähnlich gut erklärt mit einem Abstand von 2,77 die Merkmalskombination [4] und [5] und mit einem Abstand von 2,44 die Merkmalskombination [2] und [3] ebenfalls signifikant im Sinne der Teststatistik die a-priori Aufteilung.

Nur das Merkmal [1] selbst lässt sich dann signifikant Abstands erhöhend den genannten Merkmalskombinationen hinzufügen, dies in der Kombination der Merkmale [2], [3] und [1] mit dem höchsten generierten Abstand von 6,27 gegenüber 5,89 auf Basis nur des Merkmals [1]. Offensichtlich kann dann hier aber von der gegebenen Grundannahme normalverteilter Merkmalskombinationen aufgrund der schon paarweise in Kombination mit dem Merkmal [1] nicht gegebenen Annahme nicht ausgegangen werden.

Clusterung

Clusterverfahren bilden anhand gegebener Merkmale betrachteter Objekte **Gruppen ähnlicher Objekte**, den sogenannten Clustern. Die Technik hier setzt durch metrische Merkmale charakterisierte Objekte voraus, womit eine Ähnlichkeit der Objekte direkt mittels üblicher Abstandsmaße beurteilt werden kann. Es können aber auch beliebige kardinal differenzierte Merkmalsausprägungen verwendet werden sofern mindestens zwei verschiedene Ausprägungen gegeben sind, wie dies beispielsweise mit der Kennzahl ek5000 zur Differenzierung in Regionen des Westens, des Südens und des Ostens Deutschlands gegeben ist.

Die Identifikation von Clustern wird bei der FCM-Clusterung durch die Clustermittelpunkte vorgenommen[5]. Für die Ermittlung von „optimalen“ Clusterzentren ist die **Clusteranzahl** fest vorzugeben, hierzu kann auf dem Blatt ***Cluster*** (vgl. Abbildung 6) eine Clusteranzahl zwischen 2 und maximal 12 ausgewählt werden. Mit einem Iterationsverfahren werden dann Clusterzentren gebildet, die eine Struktur festlegen, bei der die Objekte jeweils dem ihnen ähnlichsten Clusterzentrum zugeordnet werden.

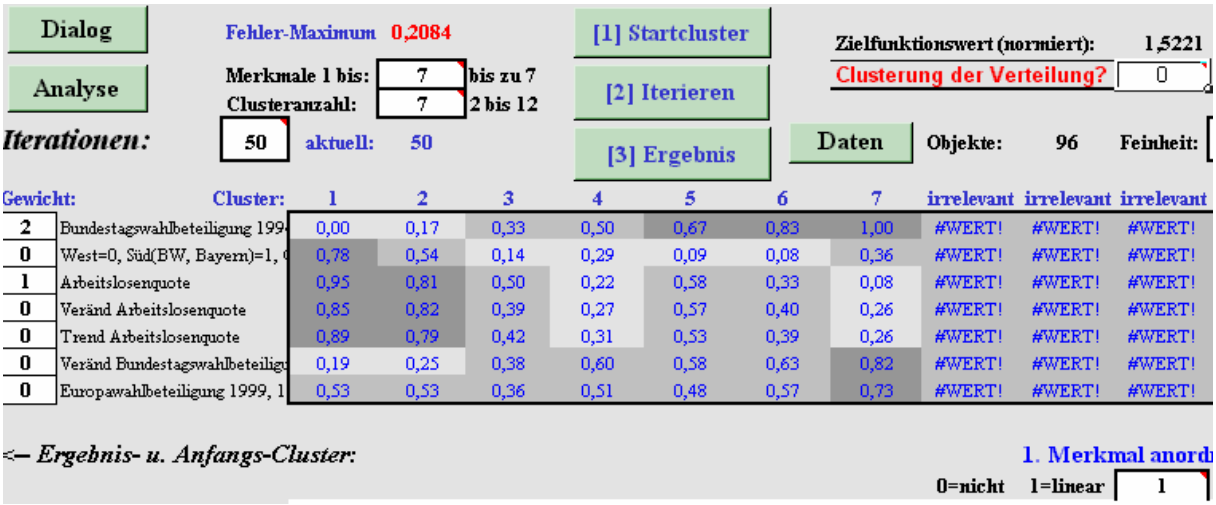


Abbildung 6: Blatt ***Cluster***

Für die Clusterung stehen jeweils die 7 zuerst genannten Merkmale im Blatt Dialog zur Verfügung. Im Blatt Cluster ist die Zahl der hieraus ausgewählten wiederum zuerst genannten **Merkmale** festzulegen. Die Clusterung kann dann nur bezüglich des ersten oder weiterer zuerst genannter Merkmale bis hin zu den 7 in Frage kommenden Merkmalen vorgenommen werden. Auch ist es möglich die Merkmale über die **Merkmalsgewichte** auf dem Blatt Cluster mit dem Gewicht 0

[5] Vgl zur FCM-Clusterung auch R.Holz: „Fuzzy Sets in der Tarifierung“, Shaker 1996 oder besser J.C. Bezdek: „Pattern recognition with fuzzy objective function algorithms“, Plenum Press, New York 1981.

Mit $\tilde{z}(u,\tilde{v}_i)=\dfrac{1/d_2^2(u,\tilde{v}_i)}{\sum_{j=1}^{c} 1/d_2^2(u,\tilde{v}_j)}$, die Zugehörigkeitswerte i = 1,..,c, c die Clusteranzahl $u\in U\subset IR^p$,

$d_2(u_r,v_j):=\sqrt{\sum_{i=1}^{p} g_i(u_{r(i)}-v_{j(i)})^2}$, $g_i\in IR$, die Abstände, $\tilde{v}_j=\begin{pmatrix}\frac{\sum_{r=1}^{n}\tilde{z}^2(u_r,\tilde{v}_j)u_{r(1)}}{\sum_{r=1}^{n}\tilde{z}^2(u_r,\tilde{v}_j)}\\ \vdots \\ \frac{\sum_{r=1}^{n}\tilde{z}^2(u_r,\tilde{v}_j)u_{r(p)}}{\sum_{r=1}^{n}\tilde{z}^2(u_r,\tilde{v}_j)}\end{pmatrix}$, j = 1,..,c die

Clustermittelpunkte, können über g_i auch die Mahanalobisdistanzen oben abgebildet werden.

auszublenden oder ihnen besonderes Gewicht zu geben. Die Merkmale mit Gewicht 0 haben keinen Einfluß auf die Clusterbildung. Die Gewichte des Blattes Cluster sind unabhängig von den Gewichten des Blattes Dialog. Die Anzahl der ***Iterationen*** für eine Clusterung kann variabel gewählt werden, in der Abbildung 6 sind 50 Iterationsschritte eingestellt.

Die Clusterung selbst ist in die Schritte

[1] Startcluster, festlegen der Ausgangsclusterzentren;
[2] Iterieren, optimieren der Struktur bis ein kleines „Fehler-Maximum“ erreicht ist;
[3] Ergebnis, Darstellung des Clusterergebnisses

unterteilt.

Sollten sich bei der Iteration **Fehler** ergeben, so liegt dies fast immer an der Auswahl nicht existenter Objekte für die Startcluster oder aber am nicht einhalten der Reihenfolge *[1], [2], [3]*. Die Ausgangsclustermatrix muß mit Werten aus [0; 1] belegt sein.

Da das numerische Ergebnis der normierten Cluster-Ergebnis-Matrix nur mühselig interpretierbar ist, werden über den Button **[3] Ergebnis** symbolische Vergröberungen der Merkmalsausprägungen vorgenommen, um das Clusterergebnis anschaulicher zu gestalten. Es können hier verschiedene Feinheiten der Betrachtung gewählt werden.

Das Clusterergebnis zur Wahlbeteiligung der Bundestagswahlen 2002 entspricht dem Beispiel 5 der Begleit-Software. Die **Feinheit der Struktur** ist mit 7 gewählt, womit die transformierten Merkmalsausprägungen äquidistant von 0 bis 1/7 mit ``---`` (sehr niedrige Werte) bis zu 6/7 bis 1 mit ``+++`` (sehr hohe Werte) symbolisiert werden. Ebenfalls möglich sind die Feinheiten 3, 5, 9 und 11.

"Min"	1	2	3	4	5	6	7	"Max"		
72,950	---	--	-	**o**	+	++	+++	*81,000*	*2,0*	Bundestagswahlbeteiligung 2002
0,000	+++	+	---	--	---	---	-	*2,000*	*0,0*	West=0, Süd(BW, Bayern)=1, Ost=2
7,200	+++	++	-	---	**o**	--	---	*18,725*	*1,0*	Arbeitslosenquote
-0,214	+++	++	-	--	**o**	--	--	*-0,006*	*0,0*	Veränd Arbeitslosenquote
-0,168	+++	++	-	--	**o**	-	--	*-0,014*	*0,0*	Trend Arbeitslosenquote
-0,087	--	--	**o**	+	**o**	+	++	*-0,005*	*0,0*	Veränd Bundestagswahlbeteiligung
35,750	**o**	+	-	-	-	**o**	+	*53,250*	*0,0*	Europawahlbeteiligung 1999, 1994
	14	**7**	**13**	**12**	**14**	**19**	**17**		96	**Anzahl:**
	0,82	*0,69*	*0,52*	*0,72*	*0,59*	*0,66*	*0,81*		*0,69*	***avg.-max.-memb.***

Clusterung *(Beispiel 5) Bundstagswahlbeteiligung*

Da dem Merkmal Wahlbeteiligung mit 2 ein vergleichsweise hohes Gewicht gegeben ist, ist das Ergebnis bei dem die Ausprägungen der Wahlbeteiligungen in den Iterationen festgehalten wurden an der Höhe der Wahlbeteiligungen ausgerichtet. Die zufallsbereinigte geringste Wahlbeteiligung ist mit 72,95% festgelegt, die zufallsbereinigte höchste Wahlbeteiligung liegt bei 81,0%. Auf dem Blatt Analyse wird ersichtlich, dass dies die 10%- und 90%-Quantile sind. Die Zuordnung einzelner

Regionen kann über den Button ***Daten*** eingesehen werden, da hier die (normierten) Daten inclusive Clusterzuordnung und Zugehörigkeitsgrad dargestellt sind. Darüber hinaus enthält die Kennzahl ek1000 jeweils die aktuelle Clusterzuordnung des letzten Clusterergebnisses. Über den Parameter „***Clusterung der Verteilung***“ der nicht in die log-files aufgenommen ist, kann zusätzlich ausgewählt werden, ob wie üblich die Skalenwerte der Merkmale oder deren Quantilsränge geclustert werden sollen.

Da beispielsweise die Kennzahl Europawahlbeteiligung mit dem **Gewicht 0** in die Clusterung eingeht, hat das Merkmal keinen Einfluß auf die Strukturfindung und es ist lediglich die mittlere Ausprägung der Europawahlbeteiligungen zu den gefundenen Clustern in den Clusterzentren mit dargestellt. Offensichtlich zeigte die Europawahl 1999 ein weitgehend anderes Wählerverhalten als die Bundestagswahlbeteiligung 2002, was allein schon an den mit 35,75% bis 53,25% sehr viel geringeren Wahlbeteiligungen ablesbar ist.

Die Zuordnung zu den aufsteigend angeordneten Bundestagswahlbeteiligungen ist durch das ***Festhalten des 1. Merkmals*** in linearer Anordnung in den Iterationen erzwungen, wozu der in Abbildung 6 nicht vollständig abgebildete Parameter „1. Merkmal anordnen“ mit dem Parameter-Wert 1 versehen wurde. Mit dem Parameter-Wert 2 kann auch die **Wertevorgabe** in „**1:**“ im ersten Merkmal beibehalten werden. Darüber hinaus kann mit dem Parameter 0 auch das erste Merkmal variabel in die Clusterung eingehen.

Eine Eigenart der verwendeten Clustertechnik ist, dass der Algorithmus eine Ordnung der Zugehörigkeit der Objekte zu den Clusterzentren herstellt, ähnlich wie dies auch für die Zugehörigkeit in der Segmentstatistik oben gegeben ist und über den Button Daten unter **[+] membs** eingesehen werden kann. Clusterungen können insofern auch als einerseits allgemeinere Segmentstatistik angesehen werden, da auf weiteren Segmenten beruhend. Andererseits ist die Information zu den Segmenten aber reduziert. Im jeweils dargestellten Clusterergebnis sind die Objekte dem Cluster mit höchster Zugehörigkeit des Objektes streng zugeordnet, woraus die aufgeführten Objektanzahlen resultieren. Außerdem sind die durchschnittlichen maximalen Zugehörigkeitsgrade (**avg-max-memb**) der Cluster angegeben, die bezogen auf die einzelnen Objekte jeweils auch mit der Kennzahl ek2000 zur Verfügung stehen.

Die Clusterergebnisse können aus einer **Optimierungstheorie** heraus interpretiert werden, stellen aber unabhängig davon jeweils Strukturierungen der Regionen dar, denen jedes Objekt (fast immer[6]) eindeutig zugeordnet ist. Ohne mathematische Vorkenntnisse kann die Aussagekraft der gefundenen Strukturen anhand der Ähnlichkeit der gebildeten Cluster sowie an den hiervon nicht unabhängigen durchschnittlichen maximalen Zugehörigkeitsgraden beurteilt werden.

Ein wesentlicher Vorteil des verwendeten Verfahrens ist abgesehen von der Praktikabilität, die größere Robustheit im Vergleich zu üblichen Clusterverfahren. So sind nicht zentrale

[6] Bei Mehrdeutigkeit wird dem erstgenannten Cluster zugeordnet.

Merkmalsverteilungen und insbesondere statistische Ausreißer in fast allen Verfahren von besonderer Bedeutung, was durch die Zuordnung zu allen Clustern bei der FCM-Clusterung deutlich abgemildert ist, da die Berechnung der Clusterzentren auf den auch „schwachen“ Zuordnungen basiert.

Ist das Clusterergebnis oben eines beim dem die Werte der Merkmale für die Abstandsbestimmungen verwendet werden, so kann durch Wahl des Parameters „Clusterung der Verteilung“ =1 auch die Verteilung der Merkmalswerte für die Abstandsbestimmung verwendet werden. Das Ergebnis unten mit der Parameter Festlegung der übrigen Parameter wie im Beispiels 8 bringt dann die a-priori Aufteilung der Regionen des Beispiels zur Diskriminanz-Analyse oben hervor und wobei der Einblick in die Streuung der Komponenten verloren geht.

"Min"	1	2	"Max"		
66,000	---	+++	84,500	2,0	Bundestagswahlbeteiligung 2002
0,000	o	--	2,000	1,0	West=0, Süd(BW, Bayern)=1, Ost=2
5,200	+	-	21,975	1,0	Arbeitslosenquote
-0,281	+	-	0,057	0,0	Veränd Arbeitslosenquote
-0,240	+	-	0,072	0,0	Trend Arbeitslosenquote
-0,118	-	+	0,101	0,0	Veränd Bundestagswahlbeteiligung
22,500	o	+	63,800	0,0	Europawahlbeteiligung 1999, 1994
	47	49		96	**Anzahl:**
	0,78	0,81		0,80	***avg.-max.-memb.***

(Beispiel 8) *Clusterung der Verteilung*

Teiber- bzw. Ursachen-Analyse

Das über das Blatt Analyse zugängliche Blatt Treiber erlaubt die Prüfung des ***Erklärungsgehaltes von bis zu 11 Merkmalen für das zu erklärende (erste) Merkmal [0]***, das selbst auch aus der Aggregation mehrerer Merkmale hervorgegangen sein kann, wenn es dann als neues Merkmal aufgenommen wurde.

Zunächst ist es jedoch notwendig die Merkmals-Ausprägungen sowohl des abhängigen zu erklärenden wie auch aller potentiellen Treiber-Merkmale zu Klassen zu vergröbern. Sofern in der Zeile Ausprägungsanzahl eine jeweils angemessene Anzahl eingegeben ist (im Beispiel der Abbildung 7 ist auch für das Merkmal Region die Feinheit 7 gewählt), werden über die Betätigung des Button ***"[1] Merkmale setzen"*** automatisch die Vergröberungen der Merkmale vorgenommen, dies ***äquidistant***, bei Wahl der Skalierung linear und ***gleichverteilt***, bei Wahl der Skalierung Quantile[7].

[7] Bei Merkmalen mit weniger Ausprägungen als der gewählten Feinheit ergeben sich Sondersituationen.
Für das angewendete Verfahren zur Bestimmung des Treiber-Fits ist es zur Herstellung stochastischer Verhältnisse notwendig, dass das Merkmal [0] gleichverteilte Ausprägungsanzahlen aufweist, weshalb für das Merkmal [0] bei Wahl der Skalierungsart „linear“ erfragt wird ob dies sichergestellt werden soll. Das Verfahren nach Dempster-Shafer bringt die Überlegenheit gegenüber

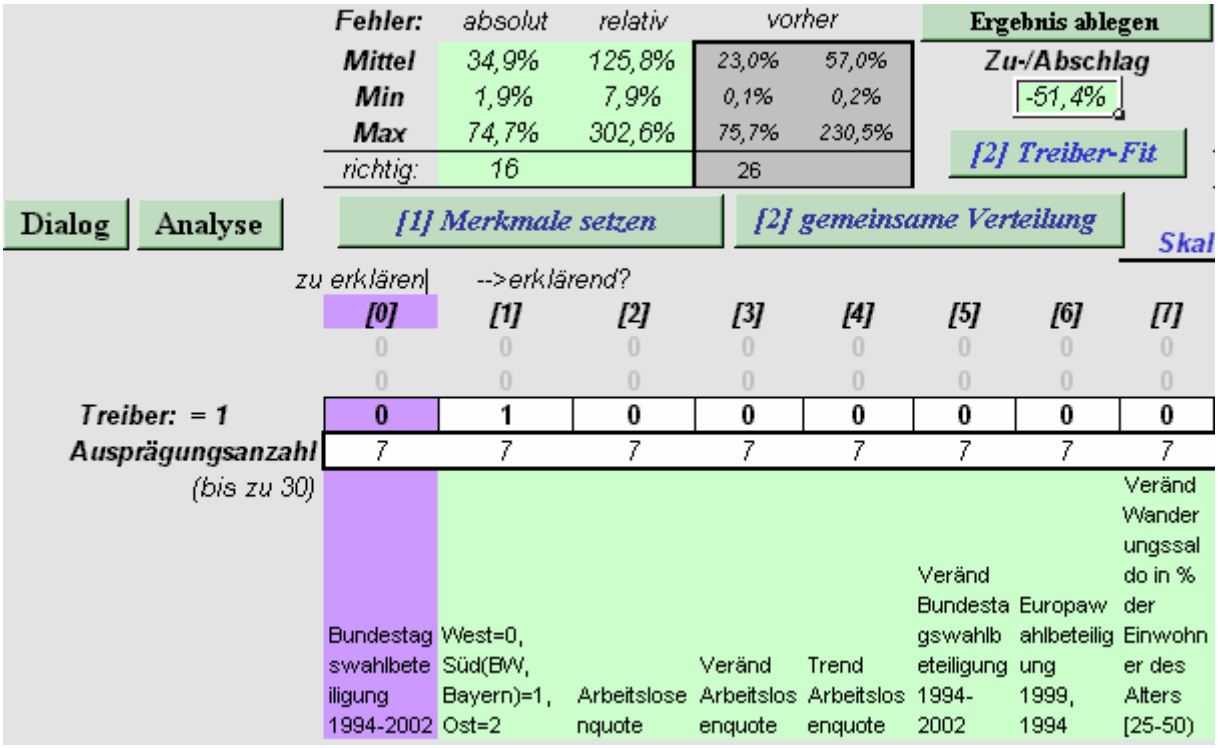


Abbildung 7: Blatt ***Treiber***

<u>Die Fehlermaße zur Bestimmung des Fits sind:</u>

a) Der mittlere absolute Fehler sowie der mittlere relative Fehler der absoluten Abweichungen.

b) Der minimale absolute und der minimale relative Fehler der absoluten Abweichungen.

c) Der maximale absolute und der maximale relative Fehler der absoluten Abweichungen.

d) Die Anzahl der korrekt zugeordneten Objekte (Fehler 0 bezüglich der vergröberten Ausprägungen)

Der Fit kann sowohl bezüglich der aggregierten ermittelten A-posteriori-Verteilung ermittelt werden, wie auch ohne weitere Regularitätsannahmen mit einer "***abhängigen Betrachtung",*** die jedes Objekt für sich betrachtet und bei verschiedenen Objekten mit gleichen erklärenden Merkmals-Ausprägungen die geschätzte zu erklärende Merkmals-Ausprägung mittelt. Durch Gegenüberstellung der Verfahren wird insbesondere die Wirkung der Annahme der stochastischen Unabhängigkeit - als wesentliche Regularitätsannahme für die Berechnung der gemeinsamen A-posteriori-Verteilung der ausgewählten Treiber - hinterfragbar. Über den Button **„Ergebnis ablegen“** können Sie die Fehlermaße der aktuellen Berechnung festhalten.

---

dem Bayes’schen Verfahren hingegen besonders bei nicht gleichverteilten Merkmalsausprägungen zu Tage. Andererseits können dann aber auch empirisch nicht haltbare Ergebnisse resultieren, was nicht Gegenstand der Betrachtung hier sein soll. Eine ausführlichere Darstellung ist mit der Datei http://www.rankingweb.de/MANUAL.pdf gegeben.

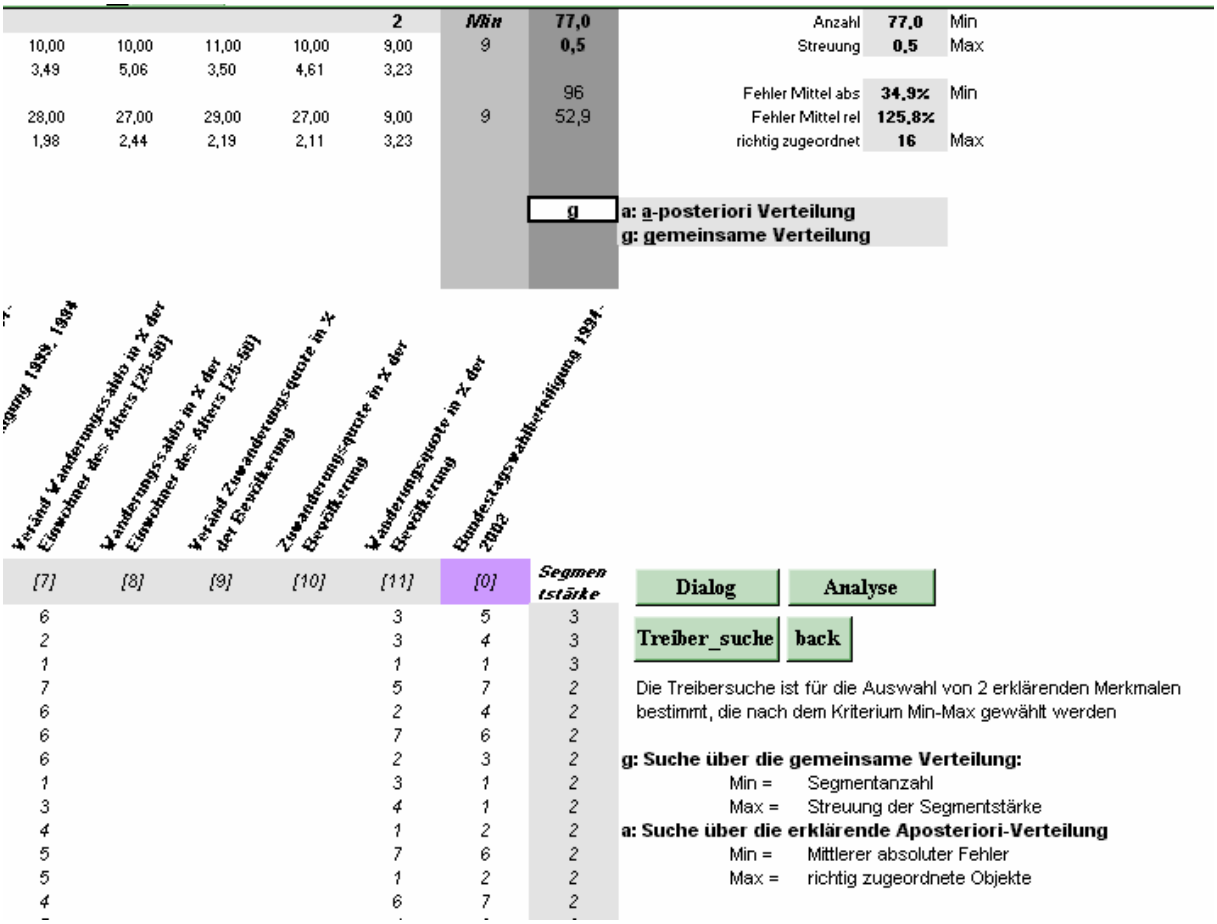


Abbildung 8: Blatt ***Verteilung***

Intuitiver besteht über den alternativen Button ***[2] gemeinsame Verteilung*** auch die Möglichkeit die gemeinsame Verteilung des zu erklärenden Merkmals [0] mit ausgewählten (möglichen) Treibern bezüglich der vergröberten Ausprägungen zu ermitteln. Es erscheint das Blatt ***Verteilung*** auf dem die ermittelten Segmente nach der Segmentstärke angeordnet dargestellt sind.

Über den Button ***Treiber_suche*** ist es dann möglich zwei Merkmale im Sinne zweier Min-Max Regeln automatisch auszuwählen. Dies über die gemeinsame Verteilung oder die erklärende empirische A-posteriori Verteilung. Für beide Verfahren sind ausser der vorgenommenen Vergröberung der Merkmale keine weiteren Modellannahmen notwendig.

Wendet man die Treiber_suche mit der Parametrisierung des Beispiels 9 an also mit einer Differenzierung des abhängigen Merkmals „Bundestagswahlbeteiligung 2002“ - hier Merkmal [0] - in zwei Gruppen und der Feinheit 3 für die Vergröberung der übrigen Merkmale, so finden sich wieder die Merkmale Region und Arbeitslosenquote - hier Merkmal [1] und [2] - der Diskriminanz-Analyse oben als die Merkmale mit dem höchsten Erklärungsgehalt.

Fortsetzung des Beispiels zur Wahlbeteiligung

Der bereits mit der hohen negativen metrischen Korrelation von -0,7 oben angedeutete Zusammenhang zwischen der Bundestagswahlbeteiligung und der Arbeitslosigkeit soll nun weiter konkretisiert werden:

1. Mehrdimensionale Segmentstatistik

Das Beispiel 10 der Datei *Regio2003.xls* wählt die Merkmale „Bundestagswahlbeteiligung 2002“ und „Arbeitslosigkeit“ als Segment definierende Merkmale aus und legt die Eigenschaften „hoch“ und „gering“ für eine strenge Aufteilung der Regionen im Median der Kennzahlen fest, wobei folgende Segmentstärken resultieren:

	Bundestags-wahlbeteiligung	**Arbeitslosig-keit**	**Anzahl der Regionen**	**Veränd Arbeitslosig-keit**	**Trend Arbeitslosig-keit**
I	gering	gering	13	-41%	-40%
II	gering	hoch	33	+51%	+54%
III	hoch	gering	34	-39%	-41%
IV	hoch	hoch	15	+13%	+8%

In den betrachteten Segmenten zeigen sich dann besonders auch die Veränderungen der Arbeitslosenquoten jeweils mit verringerter Streuung und damit charakteristisch gegenüber dem Mittel aller Regionen deutlich einheitlich erhöht beziehungsweise verringert.

2. Clusterung

Mit der Clusterergebnis des Beispiels 10 das sich mittels der Clusterung der Quantilsränge (der Verteilung) anstelle der üblichen Clusterung der Werte ebenfalls deutlicher an der Häufung der Regionen orientiert ist im ersten Merkmal der zweimal der Wert 0,25 (gering) und zweimal der Wert 0,75 (hoch) festgehalten. Im Clusterergebnis, das nur die beiden Segment definierenden Merkmale von oben mit Gewicht berücksichtigt, findet sich dann im Cluster 1 mit ähnlicher Anzahl an Regionen das Segment II mit den gegenüber dem Gesamtmittel am deutlichsten erhöhten Veränderungen der Arbeitslosenquoten wieder. Ähnlich das Segment I im Cluster 2, das Segment III im Cluster 4 und das Segment IV im Cluster 3.

"Min"	1	2	3	4	"Max"		
66,000	--	--	++	++	84,500	1,0	Bundestagswahlbeteiligung 2002
0,000	o	--	---	-	2,000	0,0	West=0, Süd(BW, Bayern)=1, Ost=2
5,200	++	-	o	---	21,975	1,0	Arbeitslosenquote
-0,281	++	-	o	--	0,057	0,0	Veränd Arbeitslosenquote
-0,240	++	-	o	--	0,072	0,0	Trend Arbeitslosenquote
-0,118	--	-	+	++	0,101	0,0	Veränd Bundestagswahlbeteiligung
22,500	o	-	o	+	63,800	0,0	Europawahlbeteiligung 1999, 1994
	30	19	25	22		96	**Anzahl:**
	0,78	0,66	0,67	0,73		0,72	***avg.-max.-memb.***

(Beispiel 10) *Clusterung der Verteilung mit Anordnung*

"Min"	1	2	3	4	"Max"		
66,000	-	+	+	++	84,500	1,0	Bundestagswahlbeteiligung 2002
0,000	+++	--	--	--	2,000	0,0	West=0, Süd(BW, Bayern)=1, Ost=2
5,200	++	o	--	---	21,975	1,0	Arbeitslosenquote
-0,281	++	o	-	--	0,057	0,0	Veränd Arbeitslosenquote
-0,240	++	o	-	-	0,072	0,0	Trend Arbeitslosenquote
-0,118	--	-	-	o	0,101	0,0	Veränd Bundestagswahlbeteiligung
22,500	o	o	o	+	63,800	0,0	Europawahlbeteiligung 1999, 1994
	19	30	26	21		96	**Anzahl:**
	0,78	0,68	0,66	0,78		0,72	***avg.-max.-memb.***

(Beispiel 12) *Clusterung der Werte ohne Anordnung*

"Min"	1	2	3	4	"Max"		
66,000	+	---	-	+++	84,500	1,0	Bundestagswahlbeteiligung 2002
0,000	---	+	--	--	2,000	0,0	West=0, Süd(BW, Bayern)=1, Ost=2
5,200	+	+++	--	--	21,975	1,0	Arbeitslosenquote
-0,281	+	++	-	--	0,057	0,0	Veränd Arbeitslosenquote
-0,240	o	++	-	--	0,072	0,0	Trend Arbeitslosenquote
-0,118	o	--	o	++	0,101	0,0	Veränd Bundestagswahlbeteiligung
22,500	o	o	o	+	63,800	0,0	Europawahlbeteiligung 1999, 1994
	27	26	19	24		96	**Anzahl:**
	0,68	0,85	0,72	0,76		0,72	***avg.-max.-memb.***

(Beispiel 12) *Clusterung der Verteilung ohne Anordnung*

Mit den beiden weiteren Clusterergebnissen des Beispiels 12 ist die Clusterung des Beispiels 10 ohne festhalten eines der Merkmale sowohl für die Quantilsränge wie auch für die Werte durchgeführt, dies um aufzuzeigen, dass hier jeweils die Übereinstimmung in den Veränderungen und den Trends der Arbeitslosenquoten der Segmente oben deutlich wiederzufinden ist.

3. Gemeinsame Verteilung

Das Beispiel 11 ermöglicht die Ermittlung der gemeinsamen Verteilung der Segment definierenden Merkmale oben mit gleicher Feinheit, was wie folgt deutlicher die Segmentstärken von oben wiedergibt:

	Bundestags-wahlbeteiligung	Arbeitslosig-keit	Anzahl der Regionen	Segmentstärke bei exakter Aufteilung im Median
I	gering	gering	17	**17**
II	gering	hoch	34	**34**
III	hoch	gering	31	**34**
IV	hoch	hoch	14	**15**

Die Unterschiede resultieren hier aus einer unterschiedlichen Ermittlung der Quantile in Häufungspunkten, weshalb zur Erzielung disjunkter Segmente oben auch nicht exakt im Median aufgeteilt wurde und was auch den Verzicht auf eine Region oben erklärt.

4. Diskriminanz-Analyse

Die betrachteten 4 Segmente bieten offensichtlich zwei Möglichkeiten einer Gegenüberstellung im Sinne negativer oder positiver Korrelationen, die folgend für eine a-priori Aufteilung der Regionen verwendet werden, um hierauf eine Diskriminanz-Analyse auzubauen.

Diskriminierung der Segmente I und IV des Beispiels 10:

Oberes Dreieck Distanzwerte; Unteres Dreieck Werte der Teststatistik

k=1	k=2	Erstes Merkmal Diagonalen-Index, Zweites Merkmal Spalten-Index						D_3^2	k=3 signifikant?
5,99	-0,14	25,65	18,00	7,39	*[1]*	Bundestagswahlbeteiligung 1994-2002		-	**0,0**
-16	0,24	6,52	3,38	2,25	*[2]*	West=0, Süd(BW, Bayern)=1, Ost=2		21,06	**-3,9**
51	39	6,20	-0,04	0,56	*[3]*	Arbeitslosenquote		-	**0,0**
31	20	-16	3,13	2,47	*[4]*	Veränd Arbeitslosenquote		20,48	**-4,4**
4	13	-14	-2	1,47	*[5]*	Trend Arbeitslosenquote		26,36	**0,6**

Dem hohen Diskriminanzwert von 25,65 im Vergleich zur Diskriminanz allein des Merkmals [1] von 5,99 und 6,20 des Merkmals [3] kann dann kein Merkmal signifikant Distanz erhöhend hinzugefügt werden. Die Hinzunahme des Merkmals [5] des Trends der Arbeitslosenquote bringt hingegen eine nicht signifikant höhere Distanz der Merkmalebezüglich der a-priori Aufteilung von 26,36 mit sich.

Diskriminierung der Segmente II und III des Beispiels 10:

Die Segmente II und III bilden die gemessene negative Korrelation nach, die auch die hier größere Zahl der Regionen erklärt.

Oberes Dreieck Distanzwerte; Unteres Dreieck Werte der Teststatistik

k=1	k=2	Erstes Merkmal Diagonalen-Index, Zweites Merkmal Spalten-Index						D_3^2	k=3 signifikant?
7,21	-25,2	1,56	5,28	6,30	*[1]*	Bundestagswahlbeteiligung 1994-2002		11,71	**-9,8**
-187	0,95	14,54	5,34	5,38	*[2]*	West=0, Süd(BW, Bayern)=1, Ost=2		-	**0,0**
-33	180	10,68	5,27	9,90	*[3]*	Arbeitslosenquote		-	**0,0**
-11	58	-24	3,56	10,38	*[4]*	Veränd Arbeitslosenquote		11,13	**-11,8**
-5	59	-3	59	3,59	*[5]*	Trend Arbeitslosenquote		11,13	**-11,8**

Offensichtlich verhindert hier schon die nicht gegebene Annahme der multivariat normalverteilten Merkmalskombination der Merkmale [1] und [3] für die a-priori Aufteilung die reguläre Anwendung der Teststatistik, die auf die positiven reellen Zahlen konzentriert sein müsste.

Augenscheinlich wird aber, dass die Merkmale [4] und [5] für sich signifikant Diskriminanz erhöhend eine hohe Trennung im Sinne der a-priori Aufteilung ergeben, die vergleichbar derer allein auf Basis des Merkmals [3] ist.

5. Lineare Skalierung

Das Beispiel 15 hält die Parameter für eine Betrachtung bereit, die sich an der linearen Skalierung orientiert:

	Bundestags-wahlbeteiligung	Arbeitslosig-keit	Anzahl der Regionen	Veränd Arbeitslosig-keit	Trend Arbeitslosig-keit
I'	gering	gering	8	-18%	-22%
II'	gering	hoch	19	+84%	+88%
III'	hoch	gering	63	-26%	-27%
IV'	hoch	hoch	6	+34%	+30%

Die gemeinsame Verteilung mit der entsprechenden Vergröberung der Merkmale anhand der linearen Aufteilung in der Mitte der Skalen gibt dann exakt die Segmente I' bis IV' wieder.

Diskriminierung der Segmente I' und IV' des Beispiels 15:

Oberes Dreieck Distanzwerte; Unteres Dreieck Werte der Teststatistik

k=1	k=2	Erstes Merkmal Diagonalen-Index, Zweites Merkmal Spalten-Index						D_3^2	k=3 signifikant?
7,74	5,44	46,7	-10,5	-2,73	*[1]*	Bundestagswahlbeteiligung 1994-2002		-	**0,0**
-2	0,01	7,22	1,55	1,31	*[2]*	West=0, Süd(BW, Bayern)=1, Ost=2		22,00	**-5,4**
38	23	6,81	1,14	3,58	*[3]*	Arbeitslosenquote		-	**0,0**
-18	5	-6	1,44	2,67	*[4]*	Veränd Arbeitslosenquote		31,67	**-3,3**
-10	4	-3	3	1,29	*[5]*	Trend Arbeitslosenquote		52,53	**1,3**

Mittels einer deutlich verringerten Datenbasis von nur noch 14 gegenüber oben 28 betrachteten Regionen wiederholt sich das Ergebnis mit deutlicherer Diskriminanz, in deren Berechnung die metrischen Korrelationen wesentlich mit eingehen.

Diskriminierung der Segmente II' und III' des Beispiels 15:

Oberes Dreieck Distanzwerte; Unteres Dreieck Werte der Teststatistik

k=1	k=2	Erstes Merkmal Diagonalen-Index, Zweites Merkmal Spalten-Index						D_3^2	k=3 signifikant?
10,99	6,8	14,87	18,69	15,51	*[1]*	Bundestagswahlbeteiligung 1994-2002		-	**0,0**
-20	8,64	28,64	16,84	16,49	*[2]*	West=0, Süd(BW, Bayern)=1, Ost=2		33,07	**70,6**
19	112	17,45	12,87	16,53	*[3]*	Arbeitslosenquote		-	**0,0**
37	46	-16	8,01	15,43	*[4]*	Veränd Arbeitslosenquote		21,54	**25,9**
22	44	-3	43	6,93	*[5]*	Trend Arbeitslosenquote		20,48	**21,8**

Mit der der negativen Korrelation angelehnten a-priori Aufteilung zeigt sich ebenfalls wieder das Merkmalspaar [4] und [5] als die a-priori Aufteilung weitgehend erklärend. Dem Merkmalspaar [1] und [3] können nun alle Merkmale signifikant zusätzlich trennend hinzugefügt werden.

6. Lineare Skalierung und Segmentbildung über die Merkmalsmittelwerte

Das Beispiel 17 hält die Parameter für eine Betrachtung bereit, die sich an der linearen Skalierung orientiert und wobei die Aufteilung in geringe und hohe Werte anhand der hier nicht zentralen Mittelwerte der Merkmalsausprägungen mittels individueller Skalierung vorgenommen ist:

	Bundestags-wahlbeteiligung	Arbeitslosig-keit	Anzahl der Regionen	Veränd Arbeitslosig-keit	Trend Arbeitslosig-keit
I*	gering	gering	16	-40%	-37%
II*	gering	hoch	30	+56%	+57%
III*	hoch	gering	39	-34%	-35%
IV*	hoch	hoch	11	+28%	+23%

Das Merkmal Region verursacht im Folgenden jeweils eine Undefiniertheit in der Bestimmung der Inversen-Matrix und die Ergebnisse von oben wiederholen sich weitgehend:

Diskriminierung der Segmente I* und IV* des Beispiels 17:

Oberes Dreieck Distanzwerte; Unteres Dreieck Werte der Teststatistik

k=1	k=2	Erstes Merkmal Diagonalen-Index, Zweites Merkmal Spalten-Index					D_3^2	k=3 signifikant?
4,32	#####	15,6	17,2	8,41	*[1]*	Bundestagswahlbeteiligung 1994-2002	-	**0,0**
#####	#####	#####	#####	#####	*[2]*	West=0, Süd(BW, Bayern)=1, Ost=2	#######	**#VALUE!**
33	#####	5,59	3,56	0,32	*[3]*	Arbeitslosenquote	-	**0,0**
38	#####	-5	4,56	-1,85	*[4]*	Veränd Arbeitslosenquote	15,46	**-0,1**
12	#####	-13	-18	1,55	*[5]*	Trend Arbeitslosenquote	12,83	**-3,4**

Diskriminierung der Segmente II* und III* des Beispiels 15:

Oberes Dreieck Distanzwerte; Unteres Dreieck Werte der Teststatistik

k=1	k=2	Erstes Merkmal Diagonalen-Index, Zweites Merkmal Spalten-Index					D_3^2	k=3 signifikant?
8,76	#####	4,17	6,90	7,37	*[1]*	Bundestagswahlbeteiligung 1994-2002	-	**0,0**
#####	#####	#####	#####	#####	*[2]*	West=0, Süd(BW, Bayern)=1, Ost=2	#######	**#VALUE!**
-24	#####	11,97	6,06	11,50	*[3]*	Arbeitslosenquote	-	**0,0**
-10	#####	-24	4,03	10,85	*[4]*	Veränd Arbeitslosenquote	12,92	**71,0**
-7	#####	-2	56	3,83	*[5]*	Trend Arbeitslosenquote	13,06	**72,2**

7. Copula und Resümee

Die Kennzahl ek31 „Copula der Wahlbeteiligung, der Arbeitslosigkeit und deren Veränderungen“ ist aus dem Produkt der Rankings mit Richtung „hoch“ der genannten Merkmale entstanden und im Ergebnis als neue Kennzahl aufgenommen.

Die Clusterung der Verteilung der Copula mit den Randverteilungen genannter Merkmale ermöglicht dann einen weiteren Einblick in die Abhängigkeiten der Merkmale:

"Min"	1	2	3	4	5	6	7	"Max"		
0,000	---	--	-	**o**	+	++	+++	0,348	5,0	Produkt-Copula: Wahlbeteiligung, Arbeitslos
66,000	++	**o**	-	+	+	--	--	84,500	1,0	Bundestagswahlbeteiligung 1994-2002
5,200	---	--	-	-	+	++	+++	21,975	1,0	Arbeitslosenquote
-0,281	---	--	-	**o**	+	++	+++	0,057	1,0	Veränd Arbeitslosenquote
-0,240	---	--	-	**o**	+	++	+++	0,072	1,0	Trend Arbeitslosenquote
	8	**17**	**15**	**16**	**15**	**14**	**11**		96	**Anzahl:**
	0,76	0,55	0,54	0,54	0,55	0,62	0,69		0,59	**avg.-max.-memb.**

Beispiel 20 Clusterung der Verteilung

Deutlich zeigt sich eine gegenläufige Tendenz der Höhe der Wahlbeteiligungen und der Arbeitslosenquoten und deren Veränderungen, deutlich sind auch die hohen Korrelationen der Kennzahlen zur Arbeitslosigkeit wiederfindbar.

"Min"	1	2	3	4	5	6	7	"Max"		
0,000	---	--	-	**o**	+	++	+++	0,348	5,0	Produkt-Copula: Wahlbeteiligung, Arbeitslos
66,000	+	+	**o**	-	**o**	**o**	**o**	84,500	1,0	Bundestagswahlbeteiligung 1994-2002
5,200	--	-	+	++	++	++	++	21,975	1,0	Arbeitslosenquote
-0,281	--	**o**	+	++	++	++	+++	0,057	1,0	Veränd Arbeitslosenquote
-0,240	--	**o**	+	++	++	++	+++	0,072	1,0	Trend Arbeitslosenquote
	35	**33**	**15**	**6**	**5**	**0**	**2**		96	**Anzahl:**
	0,76	0,69	0,54	0,63	0,59	-	0,77		0,69	**avg.-max.-memb.**

Beispiel 20 Clusterung der Werte

Das analoge Clusterergebnis der Werte macht zusätzlich deutlich, dass die Ergebnisse insbesondere durch wenige Regionen mit hohen Arbeitslosigkeitsquoten und höheren weiteren Zunahmen beeinflusst sind, in denen die Wahlbeteiligung eher nicht typisch weiter verringert ist.

Insgesamt kann jedoch aufgrund der durch die verschiedenen Verfahren gewonnenen Einblicke in die Datenlagen resümiert werden, dass ein Zusammenhang zwischen der Wahlbeteiligung und dem Niveau der Arbeitslosigkeit beziehungsweise dessen Veränderung vorliegt und was deutlicher auch folgende Segmentstatistik vor Augen führt, die aufzeigt:

Segmentstatistiken des Beispiels 21

			gering						hoch					
	Mittelwert	**Streuung**	**Streuung** Träger	Träger	**Mittelwert** gewichtet	streng	**Streuung** streng	***Anzahl:***	**Streuung** Träger	Träger	**Mittelwert** gewichtet	streng	**Streuung** streng	***Anzahl:***
Bundestagswahlbeteiligung 1994-	77,185	3,434	-29%	2%	2%	2%	-35%	42	10%	-2%	-3%	-6%	-9%	42
Veränd Bundestagswahlbeteiligur	-0,044	0,033	-20%	31%	38%	40%	-28%	42	15%	-28%	-33%	-49%	51%	42
Arbeitslosenquote	11,889	4,217	-51%	-26%	-29%	-31%	-50%	42	-8%	28%	35%	52%	-41%	42
Veränd Arbeitslosenquote	-0,127	0,075	-57%	-51%	-59%	-72%	-71%	42	-30%	54%	64%	90%	-56%	42
Trend Arbeitslosenquote	-0,096	0,060	-50%	-52%	-62%	-74%	-64%	42	-25%	55%	66%	96%	-40%	42

- dass dort wo sich die Arbeitslosenquoten weniger zurückentwickelt haben (Segment hoch), diese charakteristisch und systematisch auch deutlich höher vorzufinden ist und die Wahlbeteiligung systematisch verringert ist sowie nicht charakteristisch aber systematisch der Rückgang der Wahlbeteiligungen deutlich höher ist,
- dem entgegen zeigt sich auch charakteristisch jeweils systematisch mit den deutlicheren Rückgängen der Arbeitslosigkeit diese verringert sowie auch die Veränderung der Wahlbeteiligung weniger rückläufig bei insgesamt etwas erhöhter Wahlbeteiligung im Ergebnis.