TECHNISCHE UNIVERSITÄT MÜNCHEN-WEIHENSTEPHAN
MATHEMATIK UND STATISTIK
INFORMATIONS- UND DOKUMENTATIONSZENTRUM R. KRAFT

Biometrische und Ökonometrische Methoden II SS 00

Clusteranalyse

Clusteranalyse

Fragestellung

Einordnung von Objekten in Gruppen anhand von mehreren Merkmalen

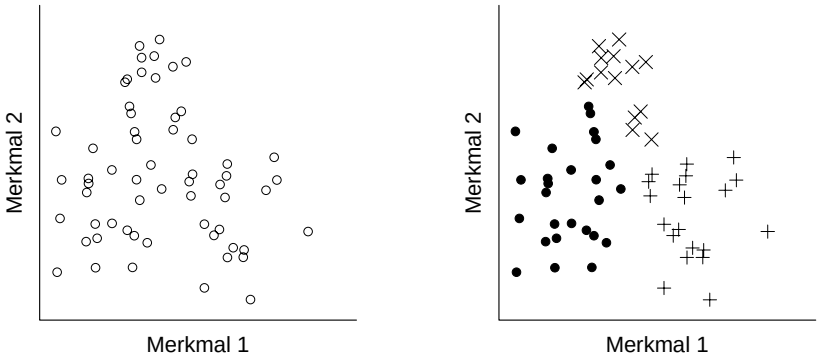
Beispiel:

Untersuchungsobjekte: Landwirtschaftliche Betriebe

Betriebskenngrößen: Landwirtschaftliche Nutzfläche
Großvieheinheiten
Betriebseinkommen
etc.

Einteilung in Betriebe ähnlicher Wirtschaftsweise, evtl. Pflanzenbaubetriebe, Tierhaltungsbetriebe, Gemischtbetriebe

n Objekte o_i
p Merkmalsvariablen x_k
x_{ik} Ausprägung des k-ten Merkmals des i-ten Objekts (i=1,...,n, k=1,...,p)



Aufgaben

Klassifikation

Optimale Aufteilung einer Menge von Objekten in möglichst homogene Gruppen anhand ihrer Merkmale

Beispiel: Einteilung von Pflanzenbeständen in Fruchtarten anhand ihres Rückstreuverhaltens bei verschiedenen Wellenlängen

Datenreduktion

Vereinfachte Darstellung einer Menge von Objekten
Auffinden von Mustergruppen oder Mustertypen

Beispiel: Welches Rückstreuverhalten ist für einen gesunden Zuckerrübenbestand typisch?

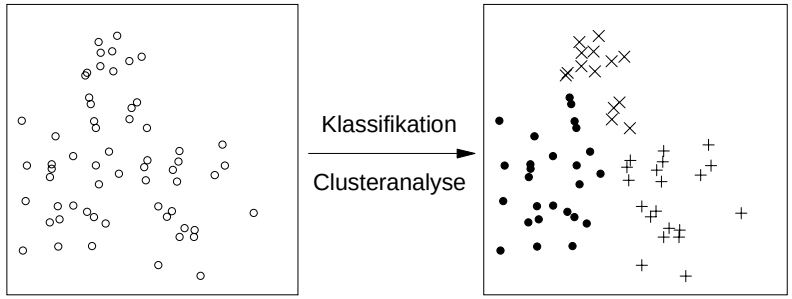
Hypothesenerzeugung

Auffinden von Hypothesen über die Merkmale bzw. die Gruppeneinteilung.

Beispiel: Hat ein gesunder Zuckerrübenbestand ein signifikant anderes Rückstreuverhalten als ein kranker Bestand?

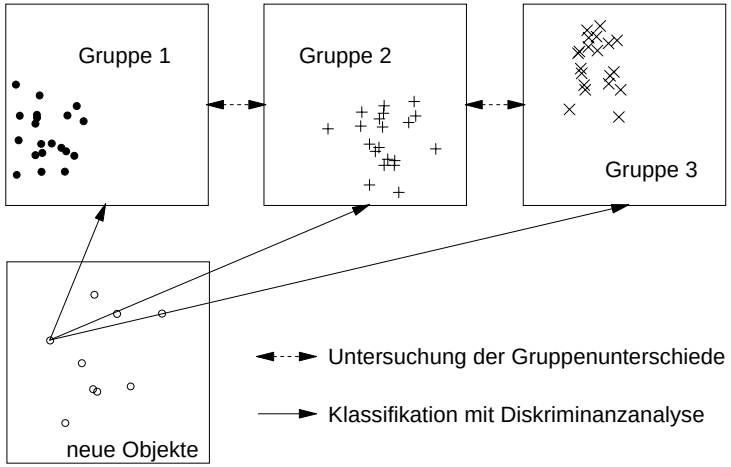
Clusteranalyse

Gruppenzugehörigkeit unbekannt
Einordnung von Objekten in ähnliche Gruppen
unüberwachte Klassifikation (unsupervised classification)



Diskriminanzanalyse

Gruppenzugehörigkeit bekannt
Einordnung neuer Objekte in bekannte Gruppen
überwachte Klassifikation (supervised classification)



Ausgangssituation

n Objekte o_i, $i = 1,2,...,n$
p quantitative, qualitative oder binäre Merkmalsvariablen
Rohdatenmatrix $X_{n\times p} = (x_{ik})$, $k = 1,2,...,p$

	Merkmal 1	...	Merkmal k	...	Merkmal p
Objekt 1	x_{11}	...	x_{1k}	...	x_{1p}
⋮	⋮		⋮		⋮
Objekt i	x_{i1}	...	x_{ik}	...	x_{ip}
⋮	⋮		⋮		⋮
Objekt n	x_{n1}	...	x_{nk}	...	x_{np}

Für jedes Objektpaar (o_i,o_j) ist eine Distanz d_{ij} definiert
Distanzmatrix $D_{n\times n} = (d_{ij})$, $i,j = 1,2,...,n$

	Objekt 1	...	Objekt i	...	Objekt n
Objekt 1	d_{11}	...	d_{1i}	...	d_{1n}
⋮	⋮		⋮		⋮
Objekt i	d_{i1}	...	d_{ii}	...	d_{in}
⋮	⋮		⋮		⋮
Objekt n	d_{n1}	...	d_{ni}	...	d_{nn}

Gesucht: Aufteilung der n Objekte in g Gruppen
Objekte einer Gruppe möglichst ähnlich (kleine Distanz)
Gruppen untereinander möglichst unähnlich (große Distanz)

Fragen: Welches Proximitätsmaß wird gewählt?
Welches Klassifikationsverfahren legt Einteilung fest?
Wieviele Gruppen sind zu unterscheiden?

Proximitätsmaße

Ähnlichkeitsmaße: Quantifizieren die Ähnlichkeit zwischen zwei Objekten. Je größer der Wert, desto ähnlicher sind die beiden Objekte.

Distanzmaße: Quantifizieren die Unterschiedlichkeit zwischen zwei Objekten. Je größer der Wert, desto unähnlicher sind die beiden Objekte.

Merkmalsvariablen

Binär: Merkmal vorhanden (1) oder nicht (0)

Tanimoto-Koeffizient
Russel & Rao (RR)-Koeffizient
Simple Matching (M)-Koeffizient

Nominal und Ordinal: Transformation in binäre Merkmale

Quantitativ: Metrische Merkmale

Euklidische Distanz
Pearsonsche Distanz
Manhattan-Metrik
Gower-Distanz
Mahalanobis-Distanz

Distanzmaße bei quantitativen Merkmalen

Euklidische Distanz

$$d_2(O_i,O_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

$$d_2^2(O_i,O_j) = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \qquad \text{(quadriert)}$$

Pearsonsche Distanz

$$d_P(O_i,O_j) = \sqrt{\sum_{k=1}^{p} \frac{(x_{ik} - x_{jk})^2}{s_k^2}}$$

Manhattan-Metrik

$$d_1(O_i,O_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

Gower-Distanz

$$d_G(O_i,O_j) = \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{r_k} \qquad \text{mit } r_k = \max_k x_{ik} - \min_k x_{ik}$$

Mahalanobis-Distanz

$$d_M^2 = (x_i - x_j)' S^{-1} (x_i - x_j) \qquad \text{mit } s_{rc} = \frac{1}{n}\sum_{i=1}^{n} (x_{ir} - \bar{x}_{.r})(x_{ic} - \bar{x}_{.c})$$

Fette - Fettsäuremuster

	[g / 100 g]							
	Buttersäure	Laurinsäure	Myristinsäure	Palmitinsäure	Stearinsäure	Ölsäure	Linolsäure	Linolensäure
Kuhmilch	3.0	2.9	10.6	26.9	13.1	27.8	3.1	1.0
Maisöl	0.0	0.0	0.0	7.1	2.0	33.7	52.3	2.0
Palmkernfett	0.0	46.9	16.9	7.6	2.0	14.1	2.1	0.0
Rindertalg	0.0	0.0	1.1	27.0	13.0	48.4	7.9	1.0
Schweinefett	0.0	0.0	1.9	27.6	26.0	38.1	2.1	0.0
Sojaöl	0.0	0.0	0.0	7.9	3.0	27.6	53.0	8.1

Fette - Euklidische Distanzen

	Kuhmilch	Maisöl	Palmkernfett	Rindertalg	Schweinefett	Sojaöl
Kuhmilch	0					
Maisöl	55.7	0				
Palmkernfett	51.7	73.4	0			
Rindertalg	23.6	52.0	64.5	0		
Schweinefett	19.2	59.5	63.1	17.6	0	
Sojaöl	66.0	8.7	73.0	54.6	60.7	0

Hierarchische Agglomerative Klassifikationsverfahren

Anfang: n Gruppen mit jeweils einem Objekt
Schrittweise Fusion ähnlicher Objekte oder Gruppen
Reduzierte Distanzmatrix
Heterogenitätsmaß der Partition $h(P) = \min d_{\nu\mu}$
Darstellung der Partitionen P durch ein Dendrogramm

Verfahren	$d_{(\nu\mu)\lambda}$
Complete-Linkage	$\max(d_{\nu\lambda}, d_{\mu\lambda})$
Single-Linkage	$\min(d_{\nu\lambda}, d_{\mu\lambda})$
Average-Linkage	$\frac{n_\nu d_{\nu\lambda} + n_\mu d_{\mu\lambda}}{n_\nu + n_\mu}$
Ward	$\frac{(n_\nu + n_\lambda) d_{\nu\lambda} + (n_\mu + n_\lambda) d_{\mu\lambda} - n_\lambda d_{\nu\mu}}{n_\nu + n_\mu + n_\lambda}$
Centroid	$\frac{n_\nu d_{\nu\lambda} + n_\mu d_{\mu\lambda}}{n_\nu + n_\mu} - \frac{n_\nu n_\mu d_{\nu\mu}}{(n_\nu + n_\mu)^2}$
Median	$\frac{d_{\nu\lambda} + d_{\mu\lambda}}{2} - \frac{d_{\nu\mu}}{4}$
McQuitty	$\frac{d_{\nu\lambda} + d_{\mu\lambda}}{2}$

Fette - Complete-Linkage

Euklidische Distanzmatrix D:

	1	2	3	4	5	6
1	*					
2	55.7	*				
3	51.7	73.4	*			
4	23.6	52.0	64.5	*		
5	19.2	59.5	63.1	17.6	*	
6	66.0	8.7	73.0	54.6	60.7	*

Fusion von 2 (Maisöl) und 6 (Sojaöl)

$h_1 = \min d_{v\mu} = d_{26} = 8.7$

Bestimmung der neuen Distanzen:

$d_{(26)1} = \max(d_{21},d_{61}) = d_{61} = 66.0$
$d_{(26)3} = \max(d_{23},d_{63}) = d_{23} = 73.4$
$d_{(26)4} = \max(d_{24},d_{64}) = d_{64} = 54.6$
$d_{(26)5} = \max(d_{25},d_{65}) = d_{65} = 60.7$

Reduzierte Distanzmatrix D':

	1	(26)	3	4	5
1	*				
(26)	66.0	*			
3	51.7	73.4	*		
4	23.6	54.6	64.5	*	
5	19.2	60.7	63.1	17.6	*

Fette - Complete-Linkage

Reduzierte Distanzmatrix D':

	1	(26)	3	4	5
1	*				
(26)	66.0	*			
3	51.7	73.4	*		
4	23.6	54.6	64.5	*	
5	19.2	60.7	63.1	17.6	*

Fusion von 4 (Rindertalg) und 5 (Schweinefett)

$h_2 = \min d'_{v\mu} = d_{45} = 17.6$

Bestimmung der neuen Distanzen:

$d_{(45)1} = \max(d_{41},d_{51}) = d_{41} = 23.6$
$d_{(45)(26)} = \max(d_{4(26)},d_{5(26)}) = d_{5(26)} = 60.7$
$d_{(45)3} = \max(d_{43},d_{53}) = d_{43} = 64.5$

Reduzierte Distanzmatrix D'':

	1	(26)	3	(45)
1	*			
(26)	66.0	*		
3	51.7	73.4	*	
(45)	23.6	60.7	64.5	*

Fette - Complete-Linkage

Reduzierte Distanzmatrix $D´´$:

	1	(26)	3	(45)
1	*			
(26)	66.0	*		
3	51.7	73.4	*	
(45)	23.6	60.7	64.5	*

Fusion von (45) und 1 (Kuhmilch)

$h_3 = \min d´´_{v\mu} = d_{(45)1} = 23.6$

Bestimmung der neuen Distanzen:

$d_{(1(45))(26)} = d_{(145)(26)} = \max(d_{1(26)},d_{(45)(26)}) = d_{1(26)} = 66.0$
$d_{(1(45))3} = d_{(145)3} = \max(d_{13},d_{(45)3}) = d_{(45)3} = 64.5$

Reduzierte Distanzmatrix $D´´´$:

	(145)	(26)	3
(145)	*		
(26)	66.0	*	
3	64.5	73.4	*

Fette - Complete-Linkage

Reduzierte Distanzmatrix $D´´´$:

	(145)	(26)	3
(145)	*		
(26)	66.0	*	
3	64.5	73.4	*

Fusion von (145) und 3 (Palmkernfett)

$h_4 = \min d´´´_{v\mu} = d_{(145)3} = 64.5$

Bestimmung der neuen Distanzen:

$d_{(3(145))(26)} = d_{(1345)(26)} = \max(d_{3(26)},d_{(145)(26)}) = d_{3(26)} = 73.4$

Reduzierte Distanzmatrix $D´´´´$:

	(1345)	(26)
(1345)	*	
(26)	73.4	*

Fusion von (1345) und (26) zu (123456)

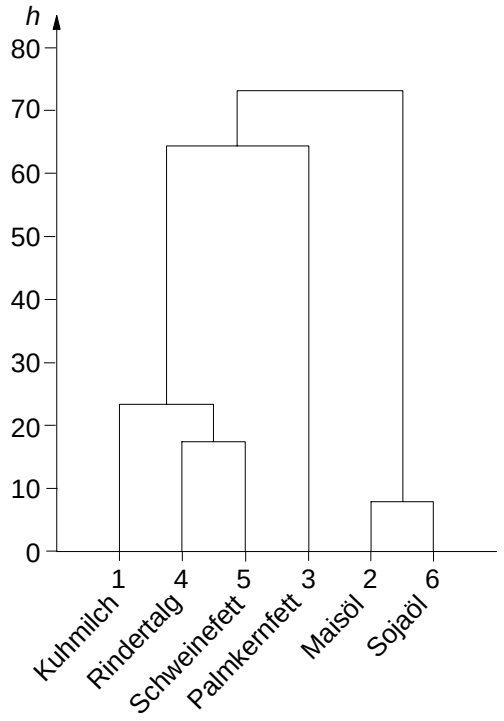
$h_5 = \min d´´´´_{v\mu} = d_{(1345)(26)} = 73.4$

Fette - Complete-Linkage

Partitionen

Partition	Index	Gruppen
P_0	0.0	(1), (2), (3), (4), (5), (6)
P_1	8.7	(1), (26), (3), (4), (5)
P_2	17.6	(1), (26), (3), (45)
P_3	23.6	(145), (26), (3)
P_4	64.5	(1345), (26)
P_5	73.4	(123456)

Dendrogramm



Iterative Klassifikationsverfahren

Iterative Verbesserung einer gegebenen Anfangspartition

***k*-Mittelwerte-Verfahren** (*k*-Means-Procedure)

Anfangspartition aufgrund hierarchischem Verfahren
zufällig
willkürlich

Verschiebung von Objekten in verschiedene Gruppen

Kriterien: Abstandsquadratsummenkriterium
Varianzkriterium
Determinantenkriterium
Spurkriterium

Lokales (jedoch i.a. kein globales) Minimum

Betriebe

```
MTB > Retrieve "TPG_BETR.MTW".
Retrieving worksheet from file: TPG_BETR.MTW

MTB > Print 'LN' 'GV/ha' 'AK'.
```

Data Display

Row	LN	GV/ha	AK
1	33	1.9	2.0
2	41	1.1	1.5
3	40	0.7	1.5
4	43	0.1	0.5
5	38	1.1	1.0
6	28	1.9	2.5
7	32	1.9	2.0
8	51	0.3	1.0
9	34	1.9	1.5
10	58	0.4	1.5
11	41	0.8	1.5
12	42	1.2	1.5
13	60	0.2	1.0
14	50	0.7	1.0
15	33	1.4	1.5
16	23	2.3	3.0
17	43	0.9	1.5
18	37	1.0	2.0
19	44	0.5	1.0
20	57	0.4	1.0

Betriebe - Klassifikation

```
MTB > Cluo 'LN' 'GV/ha' 'AK';
SUBC> Complete;
SUBC> Euclidean;
SUBC> Dendrogram.
```

Cluster Analysis of Observations: LN; GV/ha; AK

Euclidean Distance, Complete Linkage

Amalgamation Steps

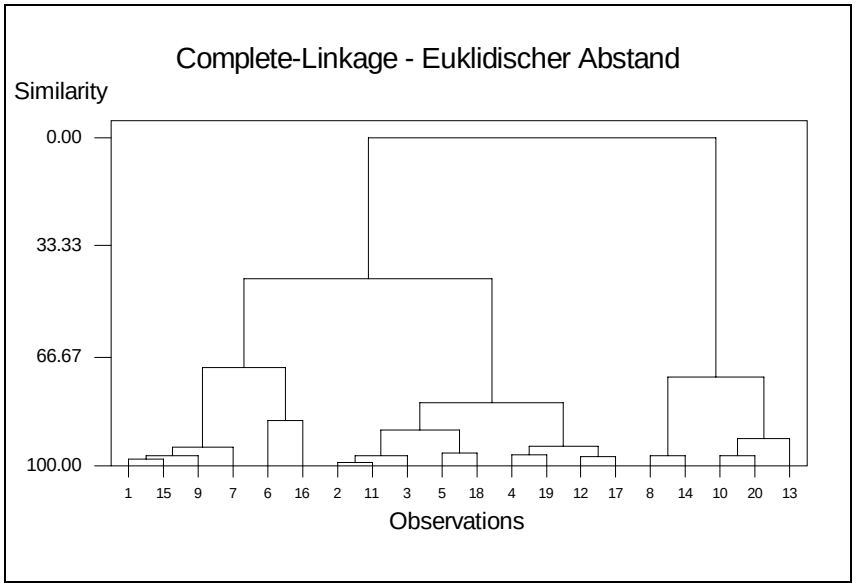
Step	No. of clusters	Simil. level	Dist. level	Clusters joined		New cluster	No. of obs. in new cluster
1	19	99.19	0.300	2	11	2	2
2	18	98.09	0.707	1	15	1	2
3	17	97.19	1.044	12	17	12	2
4	16	97.10	1.077	8	14	8	2
5	15	97.10	1.077	2	3	2	3
6	14	96.99	1.118	10	20	10	2
7	13	96.99	1.118	1	9	1	3
8	12	96.80	1.187	4	19	4	2
9	11	96.18	1.418	5	18	5	2
10	10	94.45	2.062	1	7	1	4
11	9	94.13	2.177	4	12	4	4
12	8	91.90	3.007	10	13	10	3
13	7	89.13	4.036	2	5	2	5
14	6	86.42	5.041	6	16	6	2
15	5	80.90	7.089	2	4	2	9
16	4	73.02	10.012	8	10	8	5
17	3	70.07	11.109	1	6	1	6
18	2	42.95	21.172	1	2	1	15
19	1	0.00	37.113	1	8	1	20

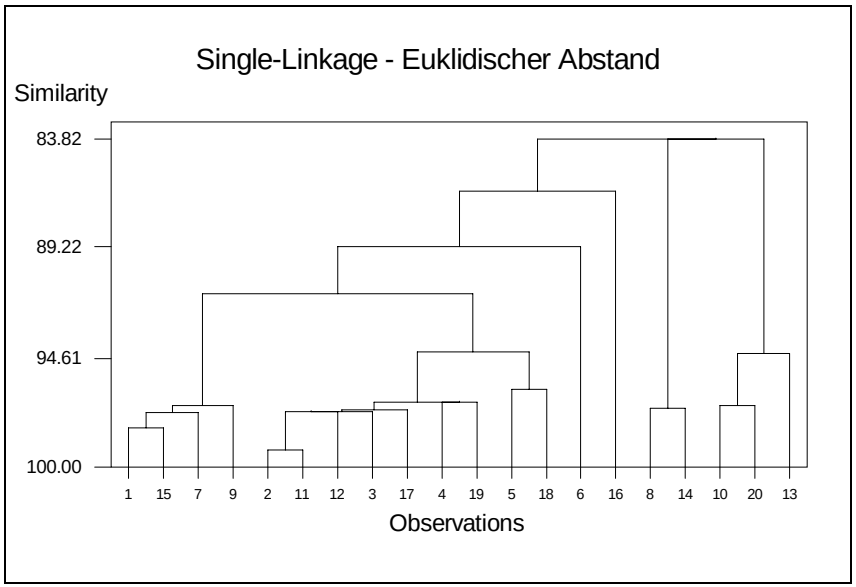
Final Partition

Number of clusters: 1

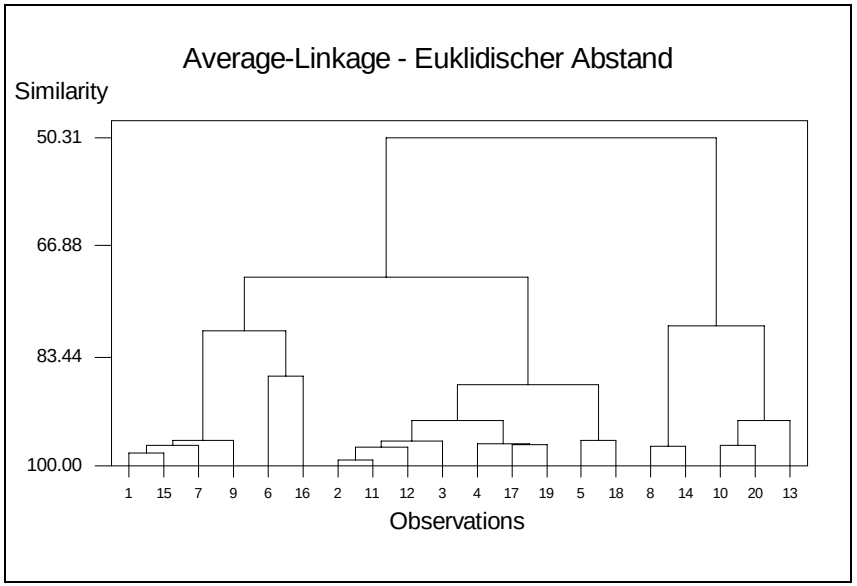
	No. of obs.	Within cluster sum of squares	Aver. dist. from centroid	Max. dist. from centroid
Cluster1	20	1893.566	7.603	18.625

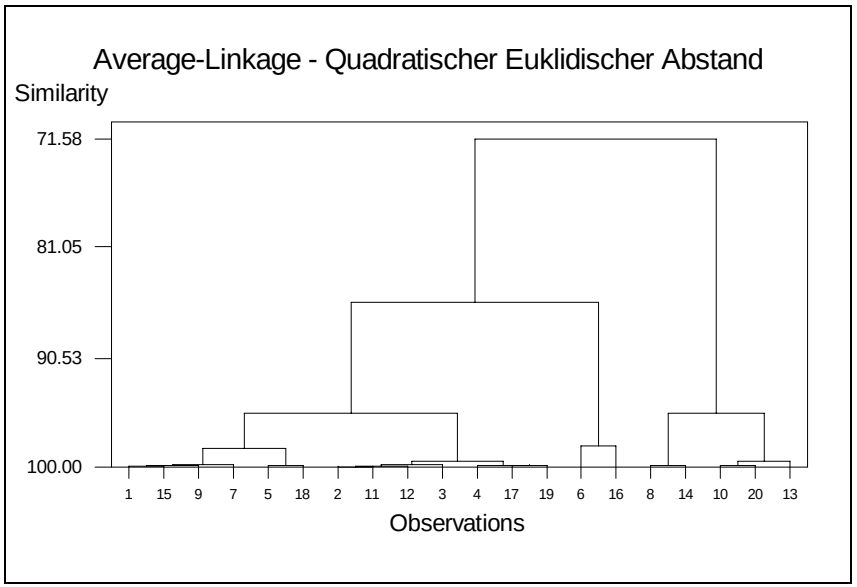
Betriebe - Dendrogramme



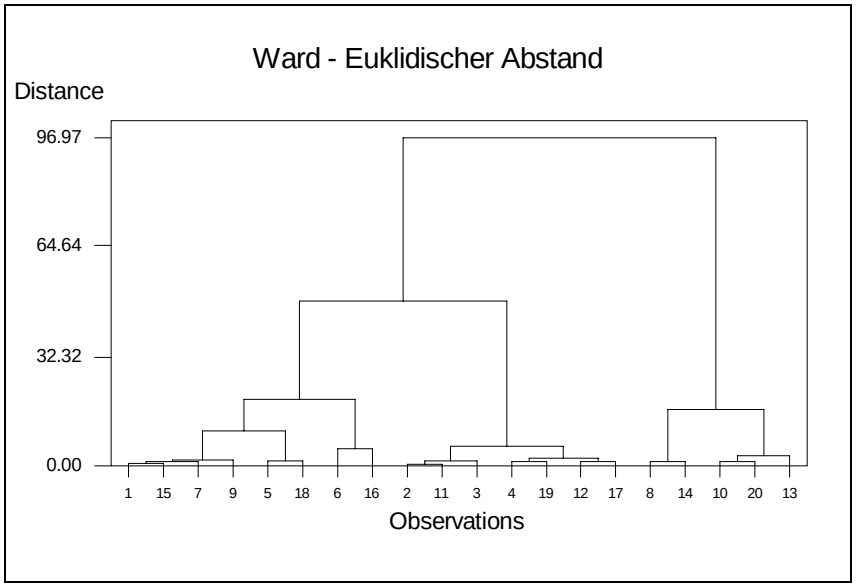


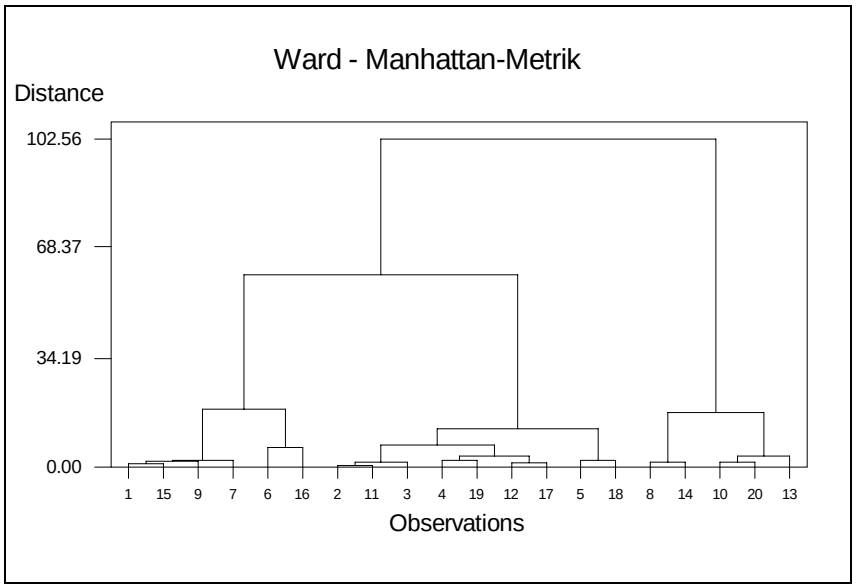
Betriebe - Dendrogramme





Betriebe - Dendrogramme





Betriebe - Ward-Klassifikation

```
MTB > Name c4 'Ward'
MTB > Cluo 'LN' 'GV/ha' 'AK';
SUBC>   Ward;
SUBC>   Pearson;
SUBC>   Standardize;
SUBC>   Number 3;
SUBC>   Dendrogram;
SUBC>   Title "Ward - Pearsonscher Abstand";
SUBC>   Type 2 3 4;
SUBC>   Member 'Ward'.
```

Cluster Analysis of Observations: LN; GV/ha; AK

```
Standardized Variables, Pearson Distance, Ward Linkage

Amalgamation Steps

Step  No. of  Simil. Dist. Clusters   New    No. of obs.
     clusters level  level  joined  cluster in new cluster
  1     19    98.32  0.101   1    7    1         2
  :      :      :     :      :    :    :         :
  :      :      :     :      :    :    :         :
 19      1  -133.81 13.965   1    2    1        20

Final Partition

Number of clusters:   3

          No. of Within cluster  Aver. dist.    Max. dist.
           obs.  sum of squares from centroid from centroid
Cluster1    6       6.837          0.904          1.850
Cluster2    7       2.186          0.485          0.918
Cluster3    7       4.780          0.746          1.301

Cluster Centroids

Variable  Cluster1  Cluster2  Cluster3  Grand centrd
LN         -1.0961   -0.1121    1.0516      0.0000
GV/ha       1.2862   -0.0964   -1.0061      0.0000
AK          0.9973    0.0000   -0.8549     -0.0000

Distances Between Cluster Centroids

          Cluster1  Cluster2  Cluster3
Cluster1    0.0000    1.9684    3.6466
Cluster2    1.9684    0.0000    1.7066
Cluster3    3.6466    1.7066    0.0000
```

Betriebe - *k*-Means-Klassifikation

```
MTB > Name c5 'k-Means'
MTB > KMean 'LN' 'GV/ha' 'AK';
SUBC>   Init 'Ward';
SUBC>   Standardize;
SUBC>   Member 'k-Means'.
```

K-means Cluster Analysis: LN; GV/ha; AK

Standardized Variables

Final Partition

Number of clusters: 3

	No. of obs.	Within cluster sum of squares	Aver. dist. from centroid	Max. dist. from centroid
Cluster1	5	4.923	0.866	1.613
Cluster2	8	3.025	0.560	0.888
Cluster3	7	4.780	0.746	1.301

Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Grand centrd
LN	-1.1464	-0.2036	1.0516	0.0000
GV/ha	1.4328	-0.0152	-1.0061	0.0000
AK	1.1968	0.0000	-0.8549	-0.0000

Distances Between Cluster Centroids

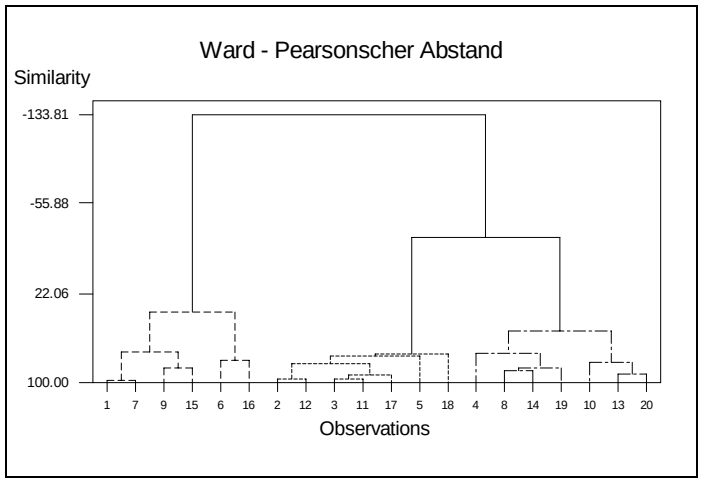
	Cluster1	Cluster2	Cluster3
Cluster1	0.0000	2.1018	3.8715
Cluster2	2.1018	0.0000	1.8134
Cluster3	3.8715	1.8134	0.0000

Betriebe - Vergleich Ward - *k*-Means

```
MTB > Print 'LN' 'GV/ha' 'AK' 'Ward' 'k-Means'.
```

Data Display

Row	LN	GV/ha	AK	Ward	k-Means
1	33	1.9	2.0	1	1
2	41	1.1	1.5	2	2
3	40	0.7	1.5	2	2
4	43	0.1	0.5	3	3
5	38	1.1	1.0	2	2
6	28	1.9	2.5	1	1
7	32	1.9	2.0	1	1
8	51	0.3	1.0	3	3
9	34	1.9	1.5	1	1
10	58	0.4	1.5	3	3
11	41	0.8	1.5	2	2
12	42	1.2	1.5	2	2
13	60	0.2	1.0	3	3
14	50	0.7	1.0	3	3
15	33	1.4	1.5	1	2
16	23	2.3	3.0	1	1
17	43	0.9	1.5	2	2
18	37	1.0	2.0	2	2
19	44	0.5	1.0	3	3
20	57	0.4	1.0	3	3



Eigenschaften der einzelnen Klassifikationsverfahren

Complete-Linkage:	dilatierend kleine homogene Gruppen betont Gruppenunterschiede
Single-Linkage:	kontrahierend große Gruppen betont Zusammenhang in den Gruppen
Average-Linkage:	zwischen Complete- und Single-Linkage
Ward-Verfahren:	leistungsfähig sehr homogene Gruppen
Centroid-Verfahren:	weniger empfehlenswert
Median-Verfahren:	weniger empfehlenswert
McQuitty-Verfahren:	weniger empfehlenswert
k-Means-Verfahren:	Verbesserung einer hierarchischen Anfangspartition