

Clusteranalyse mit SPSS

Einführung in die Clusteranalyse

- Vorstellung der Verfahren
- Praktische Arbeit mit der VSKT

Referentin: Vera Beitner

1. Allgemeines zur Clusteranalyse
2. Die Verfahren der Clusteranalyse
 - Divisives Verfahren
 - Partitionierendes Verfahren anhand von K-Means
 - Anwendung mit SPSS
 - Hierarchisch agglomeratives Verfahren
 - Anwendung mit SPSS
 - [Two-Step-Clusteranalyse]
3. Quellen

Clusteranalyse – wozu?

- Exploratives Verfahren (↔ konfirmatorisches Verfahren)
- Ziel:
Finden von Gruppen (Clustern) in den Daten
- Anwendungsbereiche:
Marktforschung, Psychologie, Medizin, Soziologie etc.
- Voraussetzung:
 - Cluster in sich möglichst homogen (Intracluster-Homogenität)
 - Cluster unter sich möglichst heterogen (Intercluster-Heterogenität)

Maßzahlen der Clusteranalyse

Proximitätsmaße

- Ähnlichkeitsmaße

Je größer, desto ähnlicher

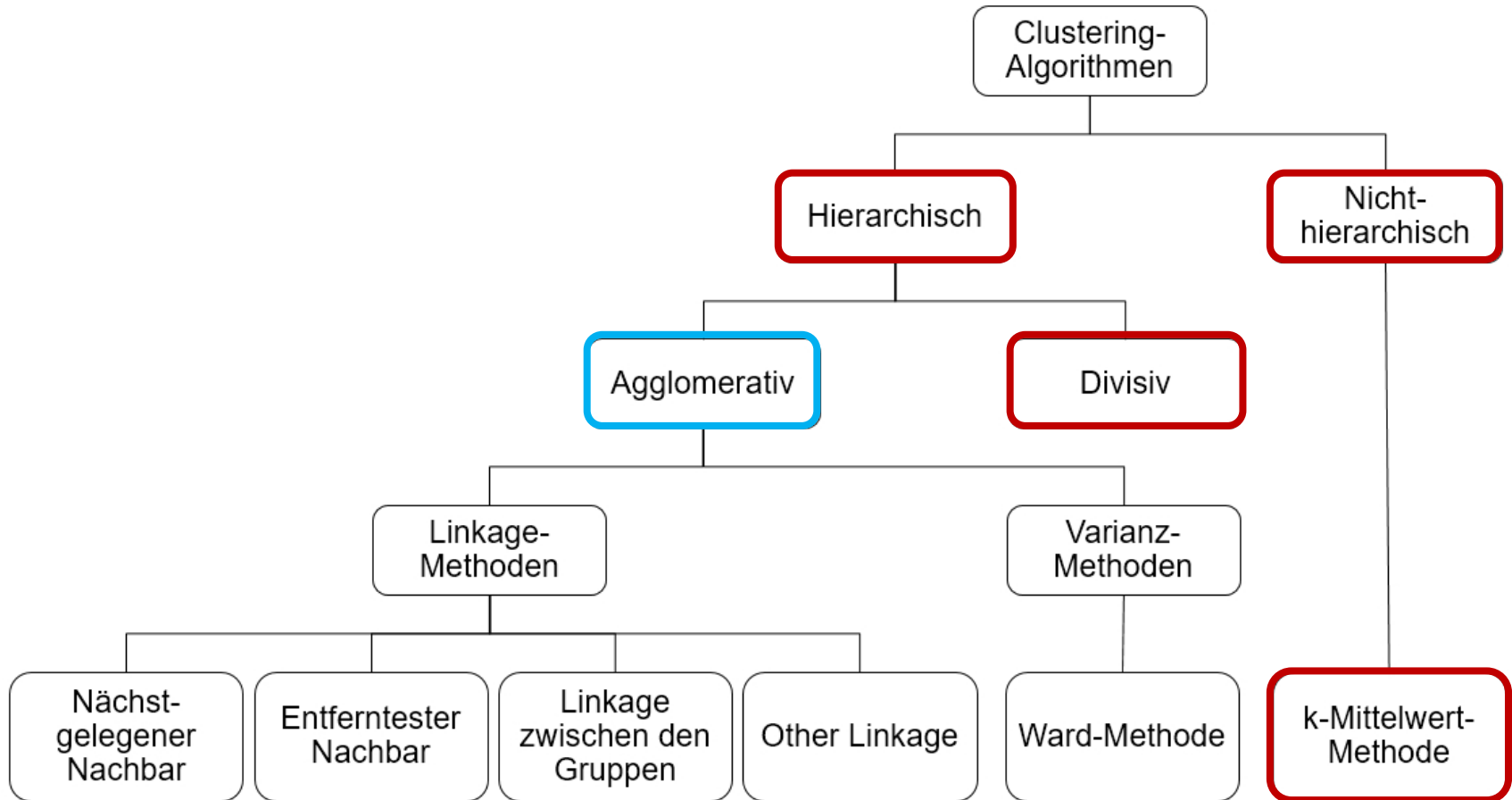
1. Ähnlichkeitskoeffizient $[0,1]$
2. Korrelationskoeffizient $[-1,1]/[0,1]$

- Unähnlichkeits- / Distanzmaße

Je größer, desto unähnlicher

1. Euklidische Distanz
2. City-Block-Abstand
3. Gewichtete Euklidische Distanz

Die Verfahren der Clusteranalyse



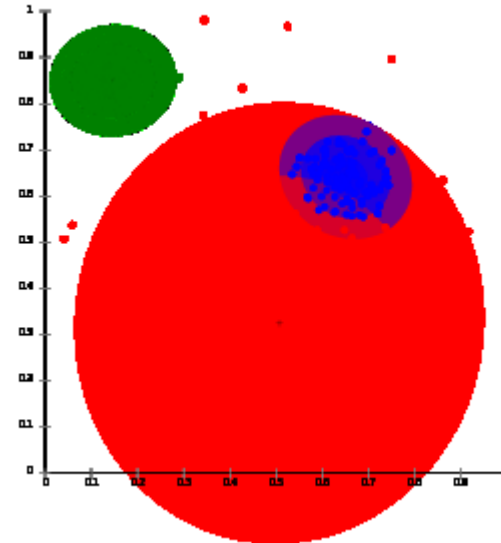
Quelle: Universität Zürich Methodenberatung

Die Verfahren der Clusteranalyse

- Divisives Verfahren
 - Begonnen mit einem großen Cluster
 - Wird immer weiter geteilt
 - Kaum praktische Relevanz
- Partitionierendes Verfahren
 - Für metrische Variablen
 - Vorgegebene Gruppen
 - Kontinuierliche Umsortierung bis zur idealen Gruppierung
 - Ausschlaggebend sind die Distanzen zum Clusterzentrum

Clusterzentrenanalyse (SPSS: K-Means)

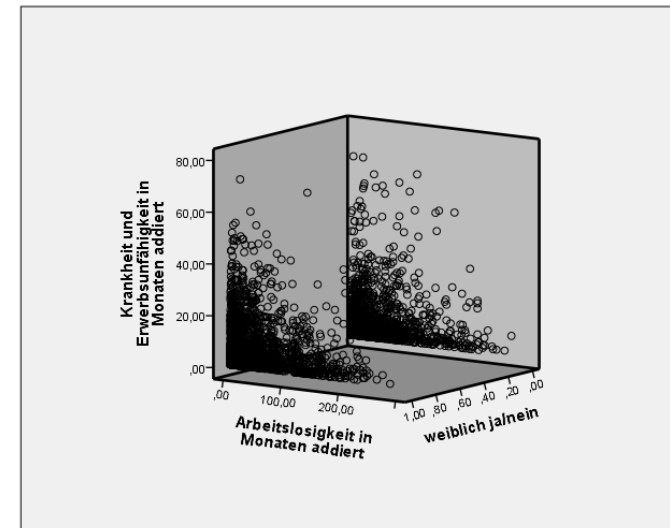
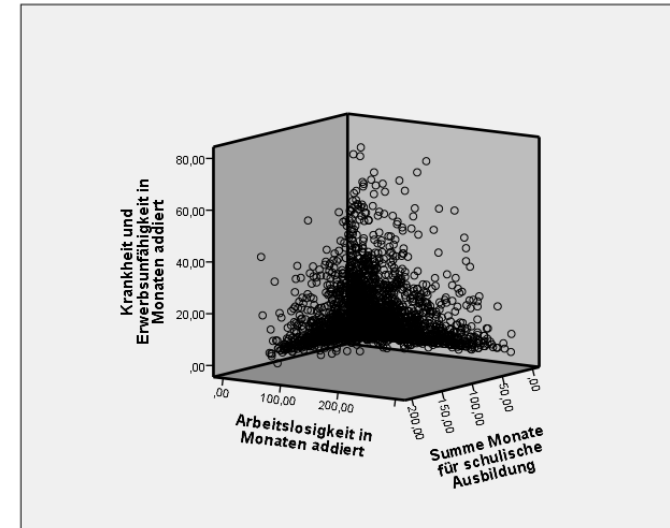
- Anwendung bei großen Fallzahlen
- Variablen müssen quantitativ sein
- Reihenfolge der Variablen wirkt sich auf Ergebnis aus!



Quelle: Wikipedia

Daten für die Analyse - VSKT

- SUFs der VSKT 2017
- Fix-Daten, Krankheit, Arbeitslosigkeit, soziale Erwerbssituation
- GEH → Frau (0-1 codiert)
- Altersgruppe 40-45 (GBJA 1973-1977)
- Fallzahl: 3587



Merkmale des SUFs

- Alle relevanten aus fix-Datensatz
- Addierte Arbeitslosigkeitsmonate (ARBEITSL_ALL)
- Addierte Krankheitsmonate (KRANK_ALL)
- Ausprägungen des SES dummycodiert
 - sumschule
 - sumausbild
 - sumpflege
 - sumkind
 - sumkrank
 - sumalos
 - sumALG1
 - etc.

Anwendung mit SPSS

K-Means I

SUF VSKT FDZ.sav [DataSet1] - IBM SPSS Statistics Dateneditor

File Bearbeiten Ansicht Daten Transformieren **Analysieren** Grafik Extras Erweiterungen Fenster Hilfe

Name	Typ	Breite
1 SK	Numerisch	2
2 JA	Numerisch	4
3 CASE	Numerisch	5
4 GBJA	Numerisch	4
5 PSGR	Numerisch	2
6 TLRT	Numerisch	1
7 ZTPTRTBEJJ	Numerisch	4
8 VSAT	Numerisch	2
9 VSKN	Numerisch	1
10 HRF	Numerisch	14
11 WHOT_BLAND	Numerisch	3
12 GDEGPTDX	Numerisch	6
13 VGEGPTDX	Numerisch	6
14 RTZTMO	Numerisch	3
15 BUEZT	Numerisch	3
16 BUEZTEGPT	Numerisch	8
17 BUEZTPE	Numerisch	3
18 BUEZTPEEGPT	Numerisch	8
19 VSKTF1	Numerisch	1
20 VSKHRF1	Numerisch	16
21 VSKTF2	Numerisch	1
22 KRANK_DII	Numerisch	8

Analysieren > Klassifizieren > **K-Means-Cluster...**

K-Means-Clusteranalyse

Variablen:

- SK
- JA
- CASE
- GBJA
- PSGR
- TLRT
- ZTPTRTBEJJ
- VSAT
- VSKN

Fallbeschriftung:

Anzahl der Cluster: 2

Methode: Iterieren und klassifizieren Nur klassifizieren

Clusterzentren:

Anfangswerte einlesen:

- Geöffnetes Dataset
- Externe Datendatei

Endwerte schreiben in:

- Neues Dataset
- Datendatei

Buttons: Iterieren..., Speichern..., Optionen..., OK, Einfügen, Zurücksetzen, Abbrechen, Hilfe

Anwendung mit SPSS

K-Means II

The image shows the SPSS K-Means Cluster Analysis dialog box and its sub-dialogs. The main dialog box has the following settings:

- Variables: (empty)
- Fallbeschriftung: (empty)
- Anzahl der Cluster: 2
- Methode: Iterieren und klassifizieren Nur klassifizieren
- Clusterzentren: Anfangswerte einlesen: Geöffnetes Dataset Externe Daten/datei (Datei...)
- Endwerte schreiben in: Neues Dataset Datendatei (Datei...)

The sub-dialogs are:

- K-Means-Clusteranalyse: Iterieren**: Maximalzahl der Iterationen: 10 (circled in red), Konvergenzkriterium: 0, Gleitende Mittelwerte verwenden. Buttons: Weiter, Abbrechen, Hilfe.
- K-Means-Clusteranalyse: Neue Va...**: Clusterzugehörigkeit, Distanz vom Clusterzentrum. Buttons: Weiter, Abbrechen, Hilfe.
- K-Means-Clusteranalyse: Optionen**: Statistik: Anfängliche Clusterzentren, ANOVA-Tabelle, Clusterinformationen für jeden Fall. Fehlende Werte: Listenweiser Fallausschluss, Paarweiser Fallausschluss. Buttons: Weiter, Abbrechen, Hilfe.

Annotations:

- A red arrow points from the text "Bei K-Means muss Clusteranzahl vorgegeben werden!" to the "Anzahl der Cluster" field in the main dialog.
- A red circle highlights the value "10" in the "Maximalzahl der Iterationen" field, with a red arrow pointing to the text "Stufen der Zusammenfassung".
- Black arrows point from the "Iterieren...", "Speichern...", and "Optionen..." buttons in the main dialog to their respective sub-dialogs.

Bei K-Means muss Clusteranzahl vorgegeben werden!

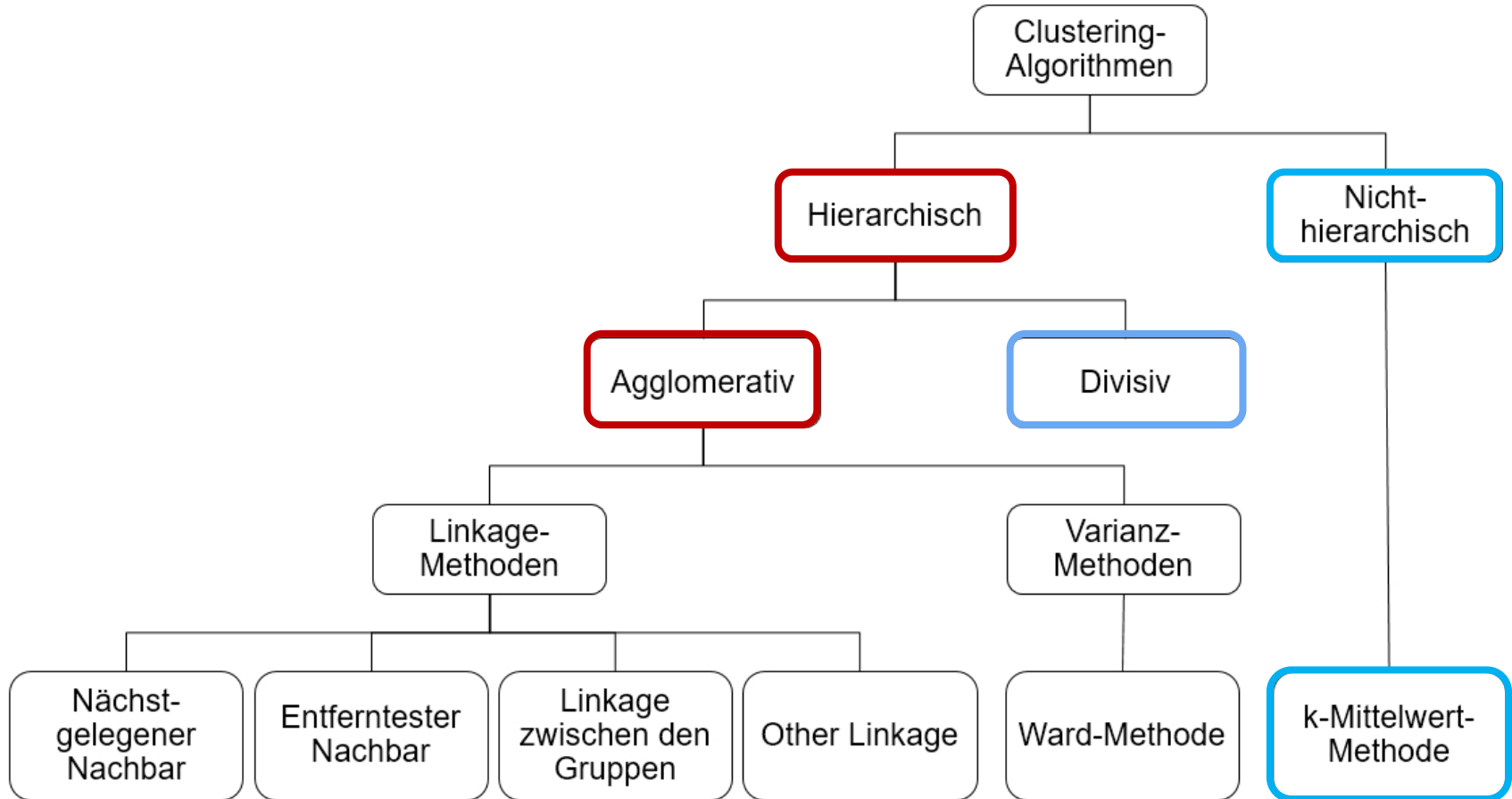
Vorüberlegungen:

→ Welche Merkmale kommen in Frage

- KRANK_ALL (in Monaten)
- ARBEITSL_ALL (in Monaten)
- Alle sum~ Merkmale (in Monaten)
- RTZTMO (in Monaten)

→ Standardisierung notwendig?

Die Verfahren der Clusteranalyse



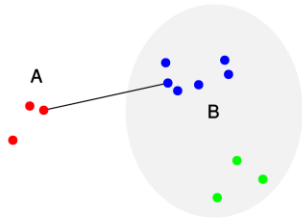
Quelle: Universität Zürich Methodenberatung

Hierarchisch-agglomerative Clusteranalyse

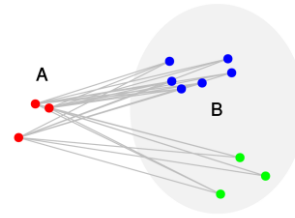
- Beginnend mit n Clustern, wobei n =Anzahl der Objekte
- Schrittweise Zusammenführung der Objekte bzw. Aggregate
- Verschiedene Skalenniveaus können nicht gemischt werden
- Verschiedene Messpunkte ausschlaggebend
→ Verschiedene Verfahren

Hierarchisch-agglomerative Clusteranalyse

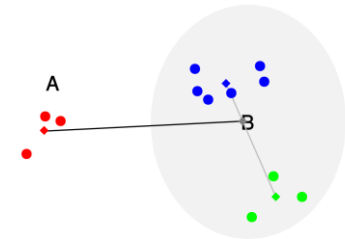
Linkage-Methoden



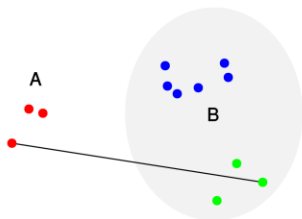
Single Linkage
Minimaler Abstand



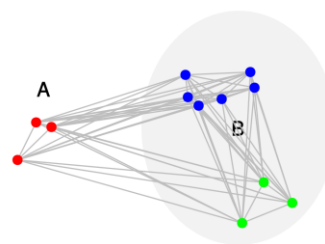
Average Linkage
Durchschnittlicher Abstand



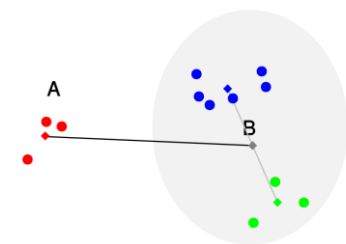
Zentroid Clustering
Abstand der Zentren



Complete Linkage
Maximaler Abstand



Average Group Linkage
Durchschnittlicher Abstand aus
der Vereinigung A und B



Median Clustering
Abstand der Zentren

Anwendung in SPSS

Hierarchisch-agglomerative Clusteranalyse

- Für metrische, nominalskalierte oder gemischt skalierte Variablen
- Es kann Bereich theoretischer Clusteranzahlen festgelegt werden
- Nicht für hohe Fallzahl geeignet (ca. bis 250)
- Standardisierung/Transformierung in SPSS möglich
- Beim Zusammenfügen wird immer die Zahl des niedrigeren Clusters weiterverwendet (Cluster 5 + Cluster 10 → Cluster 5)

- Praktisch größte Relevanz
- Zusammenfügen über minimalen Anstieg der Intraclustervarianz
- Quadrierte euklidische Distanzen einzelner Objekte zum Cluster-Zentroid
 - Quadrierte euklidische Distanzen werden aufsummiert
 - Fusionierung jener zwei Cluster, durch die geringste Erhöhung der Gesamtsumme der Distanzen bewirkt wird

Anwendung in SPSS

Hierarchische Clusteranalyse

Variable(n):

Fallbeschriftung:

Cluster: Fälle Variablen

Anzeige: Statistik Diagramme

Buttons: OK, Einfügen, Zurücksetzen, Abbrechen, Hilfe

Buttons: Statistiken..., Diagramme..., Methode..., Speichern

Hierarchische Clusteranalyse: Statistik

Zuordnungsübersicht
 Ähnlichkeitsmatrix

Clusterzugehörigkeit

Keine
 Einzelne Lösung
 Bereich von Lösungen

Anzahl der Cluster:

Mindestanzahl der Cluster:

Höchstanzahl der Cluster:

Buttons: Weiter, Abbrechen, Hilfe

Hierarchische Clusteranalyse: Spei...

Clusterzugehörigkeit

Keine
 Einzelne Lösung
 Bereich von Lösungen

Anzahl der Cluster:

Mindestanzahl der Cluster:

Höchstanzahl der Cluster:

Buttons: Weiter, Abbrechen, Hilfe

Hierarchische Clusteranalyse: Dia...

Dendrogramm

Eiszapfen

Alle Cluster
 Angegebener Clusterbereich

Start-Cluster:

Stopp-Cluster:

Schritt:

Keine

Ausrichtung

Vertikal
 Horizontal

Buttons: Weiter, Abbrechen, Hilfe

Hierarchische Clusteranalyse: Methode

Clustermethode: Ward-Methode

Maß

Intervall: Quadrierte euklidische Distanz
Exponent: Wurzel:

Häufigkeiten: Chi-Quadrat-Maß

Binär: Quadrierte euklidische Distanz
Vorhanden: Nicht vorhanden:

Werte transformieren

Standardisieren: Keine
 Nach Variablen
 Nach Fällen:

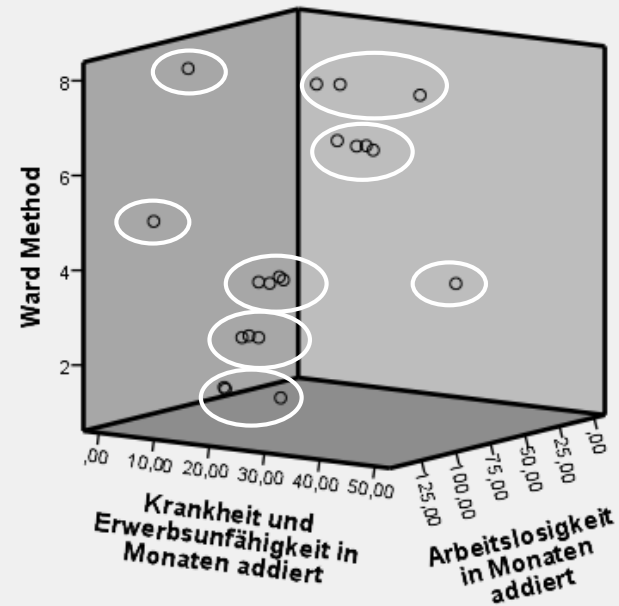
Maße transformieren

Absolutwerte
 Vorzeichen ändern
 Auf Bereich 0-1 skalieren

Buttons: Weiter, Abbrechen, Hilfe

Ergebnis in SPSS

		Häufigkeit
Gültig	1	3
	2	3
	3	4
	4	1
	5	1
	6	4
	7	3
	8	1
Gesamt		20



Two-Step Clusteranalyse

- Soll die Lücken beider anderen Verfahren schließen
- Für hohe Fallzahl geeignet und gemischte Skalenniveaus
- Clusteranzahl muss nicht vorgegeben werden
- Standardisierung in SPSS

Quellen

- Bacher, Johann/Pöge, Andreas/Wenzig, Knut*³ 2010: Clusteranalyse: Anwendungsorientierte Einführung in Klassifikationsverfahren. München: Oldenbourg.
- Backhaus, Klaus et al.*¹³ 2011: Clusteranalyse, in: Multivariate Analysemethoden: eine anwendungsorientierte Einführung. Berlin: Springer, S. 435-496.
- Cleff, Thomas*³ 2015: Clusteranalyse, in: Deskriptive Statistik und Explorative Datenanalyse. Wiesbaden: Gabler Verlag, S. 189-215.
- Eckstein, Peter*⁸ 2016: Angewandte Statistik mit SPSS: Praktische Einführung für Wirtschaftswissenschaftler. Wiesbaden: Springer Fachmedien.
- Fromm, Sabine* 2012: Clusteranalyse, in: Datenanalyse mit SPSS für Fortgeschrittene 2: Multivariate Verfahren für Querschnittsdaten. Wiesbaden: Springer Fachmedien, S. 191-222.
- Handl, Andreas*² 2010: Multivariate Analysemethoden. Berlin/Heidelberg: Springer Verlag.
- Janssen, Jürgen/Laatz, Wilfried*⁷ 2010: Clusteranalyse, in: Statistische Datenanalyse mit SPSS: eine anwendungsorientierte Einführung in das Basissystem und das Modul Exakte Tests. Berlin: Springer, S. 489-519.
- Kuß, Alfred*⁴ 2012: Clusteranalyse, in: Marktforschung: Grundlagen der Datenerhebung und Datenanalyse. Wiesbaden: Springer Gabler, S. 281-284.
- Müller, Wolfgang* 2015: Marketing Analytics: Clusteranalyse, in: Reihe Studienmanuskript, Band 10. Dortmund: Institut für Angewandtes Markt-Management.
- Rudolf, Matthias/Müller, Johannes*² 2012: Clusteranalyse, in: Multivariate Verfahren: eine praxisorientierte mit Anwendungsbeispielen in SPSS. Göttingen: Hogrefe, S. 279-305.
- Schendera, Christian F. G.* 2010: Clusteranalyse mit SPSS: mit Faktorenanalyse. München: Oldenbourg.
- Stein, Petra/Vollnhals, Sven* 2011: Grundlagen clusteranalytischer Verfahren. Universität Duisburg-Essen: Institut für Soziologie.
- Tarnai, Christian* 2010: Clusteranalyse, in: Holling, Heinz(Hrsg.): Handbuch Statistik, Methoden und Evaluation. Göttingen: Hogrefe, S. 548-555.
- Universität Zürich* 2018: Clusteranalyse. Zürich: UZH.
https://www.methodenberatung.uzh.ch/de/datenanalyse_spss/interdependenz/gruppierung/cluster.html, zuletzt geprüft am 24.09.2019.
- Wentura, Dirk/Pospeschill, Markus* 2015: Clusteranalyse, in: Kriz, Jürgen (Hrsg.): Multivariate Datenanalyse: Eine kompakte Einführung. Wiesbaden: Springer Fachmedien, S. 165-179.

Clusteranalyse mit SPSS

Vielen Dank für Ihre Aufmerksamkeit!