Deutsche Rentenversicherung Bund

# Clusteranalyse mit SPSS

## Einführung in die Clusteranalyse

- Vorstellung der Verfahren
- Praktische Arbeit mit der VSKT

Referentin: Vera Beitner

# Ablauf

Deutsche Rentenversicherung Bund

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?

Deutsche Rentenversicherung Bund

- 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

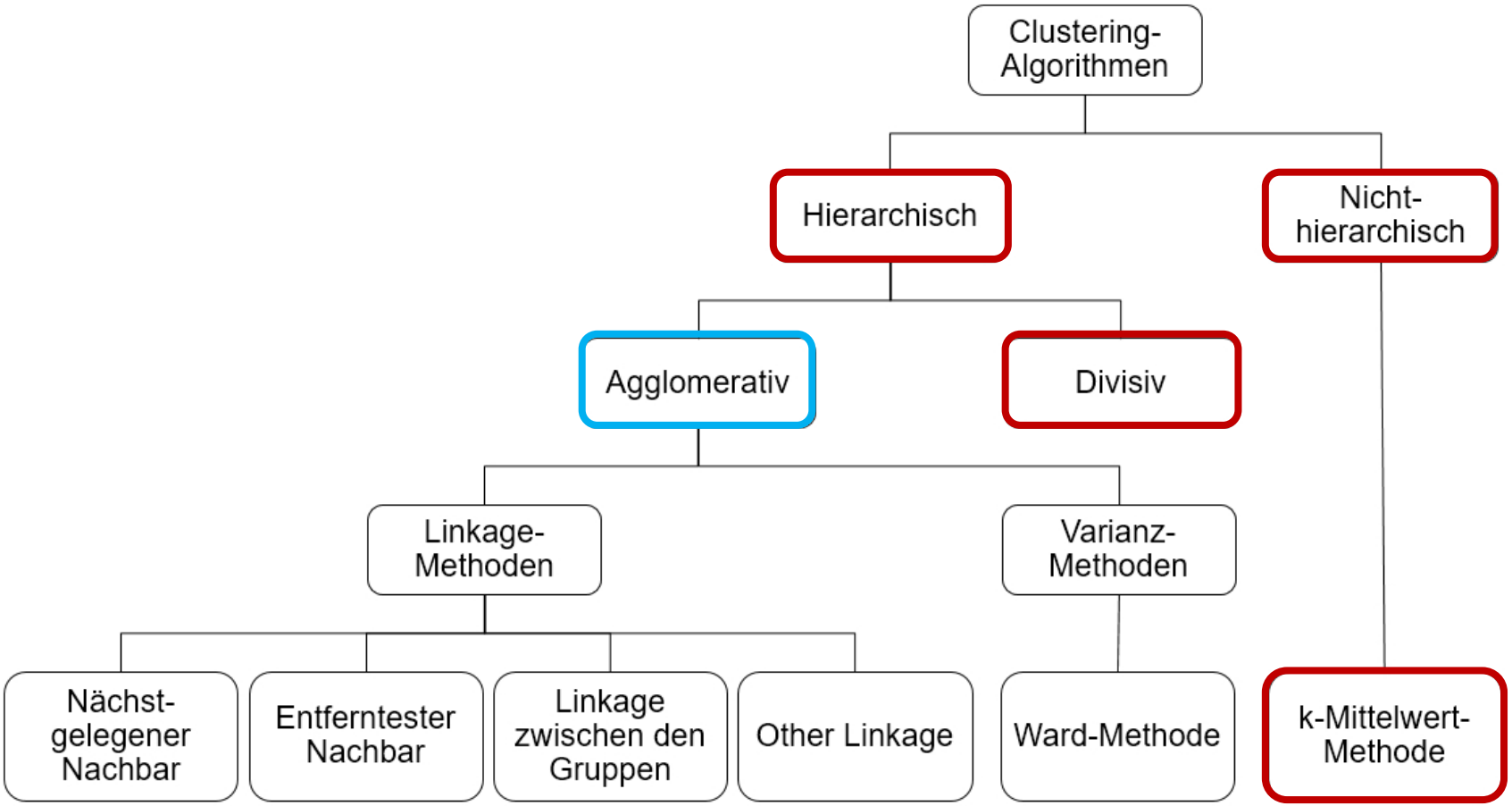
1. Ähnlichkeitskoeffizient [0,1]
2. Korrelationskoeffizient [-1,1]/[0,1]

- 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

Deutsche Rentenversicherung Bund

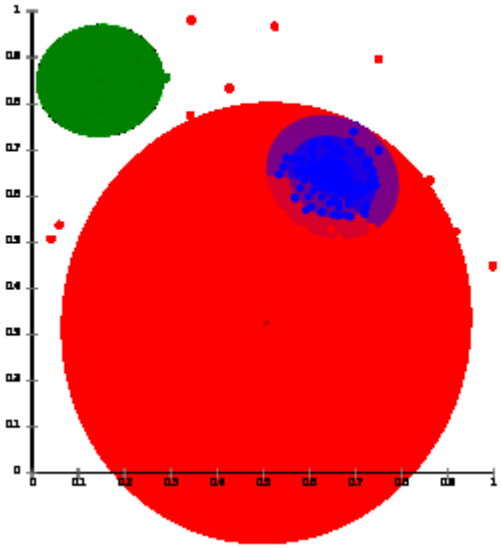
- Divisives Verfahren
  - Begonnen mit einem großen Cluster
  - Wird immer weiter geteilt
  - Kaum praktische Relevanz
- Partitionierendes Verfahren
  - Für metrische Variablen
  - Vorgegebene Gruppen
  - Kontinuierliche Umsortierung bis zur idealen Gruppierung
  - Ausschlaggebend sind die Distanzen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

# Partitionierende Clusteranalyse

Deutsche Rentenversicherung Bund

Clusterzentrenanalyse
(SPSS: K-Means)

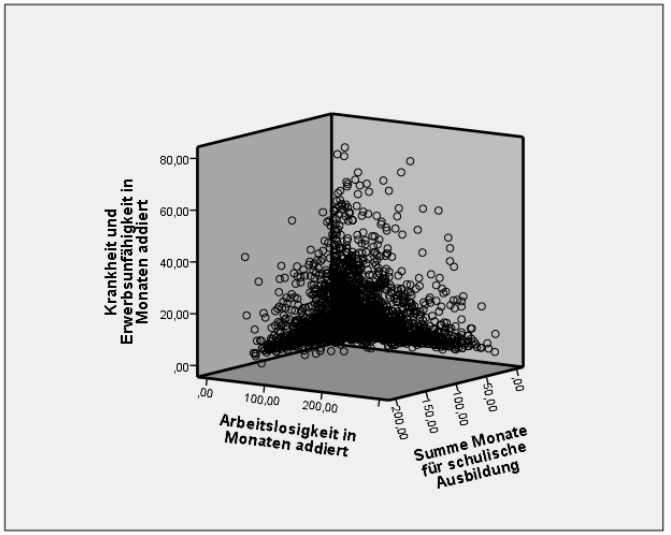
- Anwendung bei großen Fallzahlen
- Variablen müssen quantitativ sein
- Reihenfolge der Variablen wirkt sich auf Ergebnis aus!

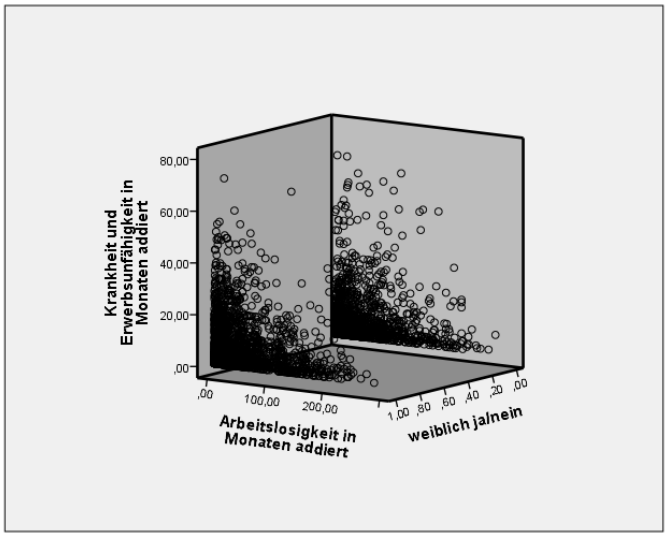


Quelle: Wikipedia

# 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



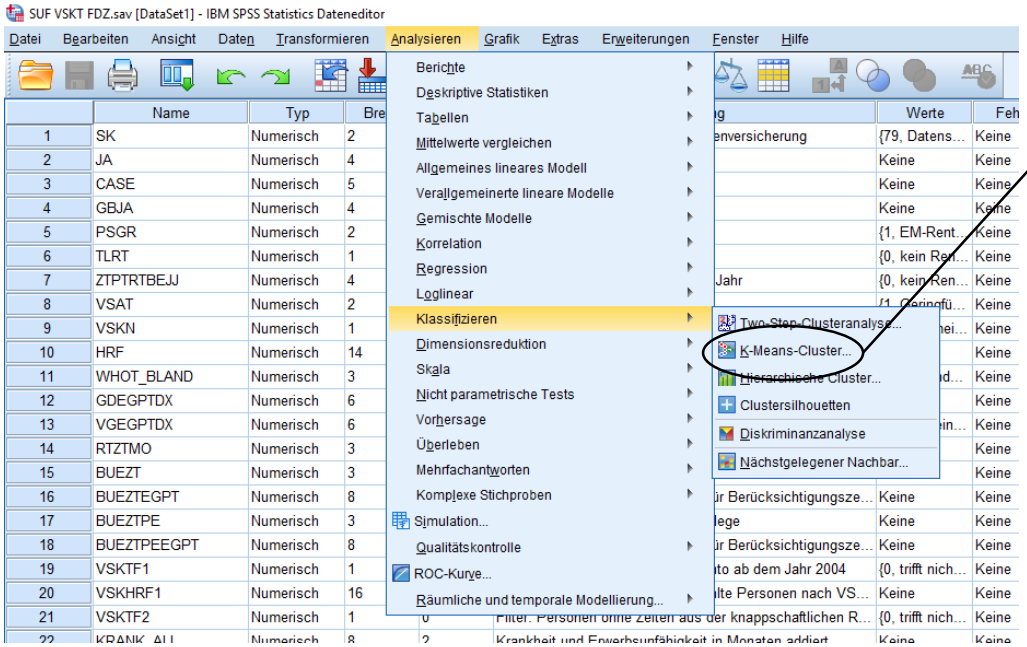
# Merkmale des SUFs

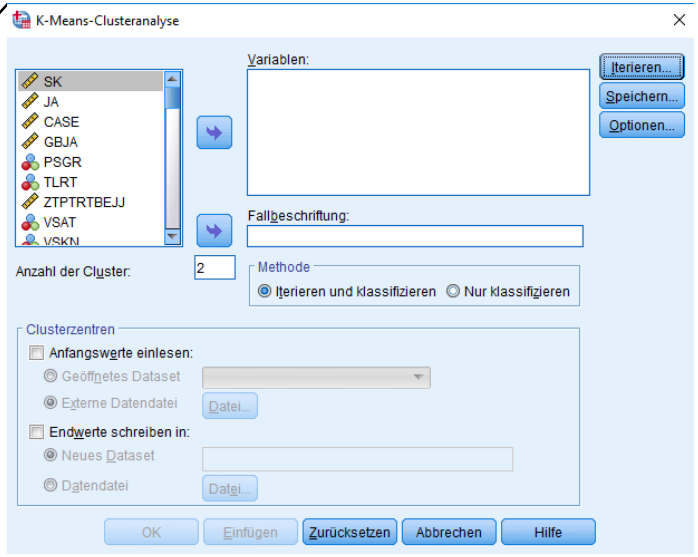
Deutsche Rentenversicherung Bund

- 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

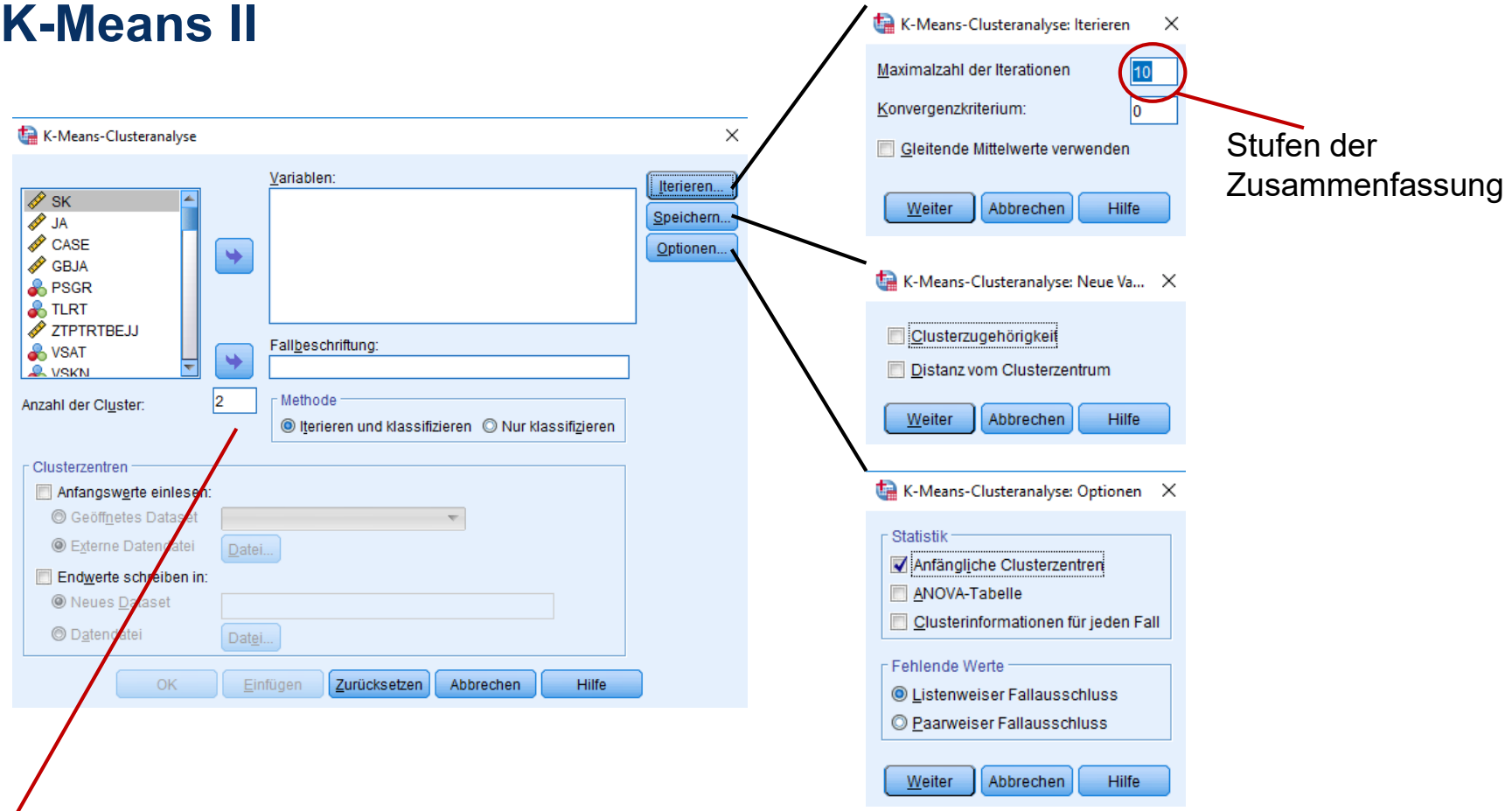
## K-Means I



# Anwendung mit SPSS

## K-Means II




Stufen der Zusammenfassung

Bei K-Means muss Clusteranzahl vorgegebenen werden!

# Vorüberlegungen:

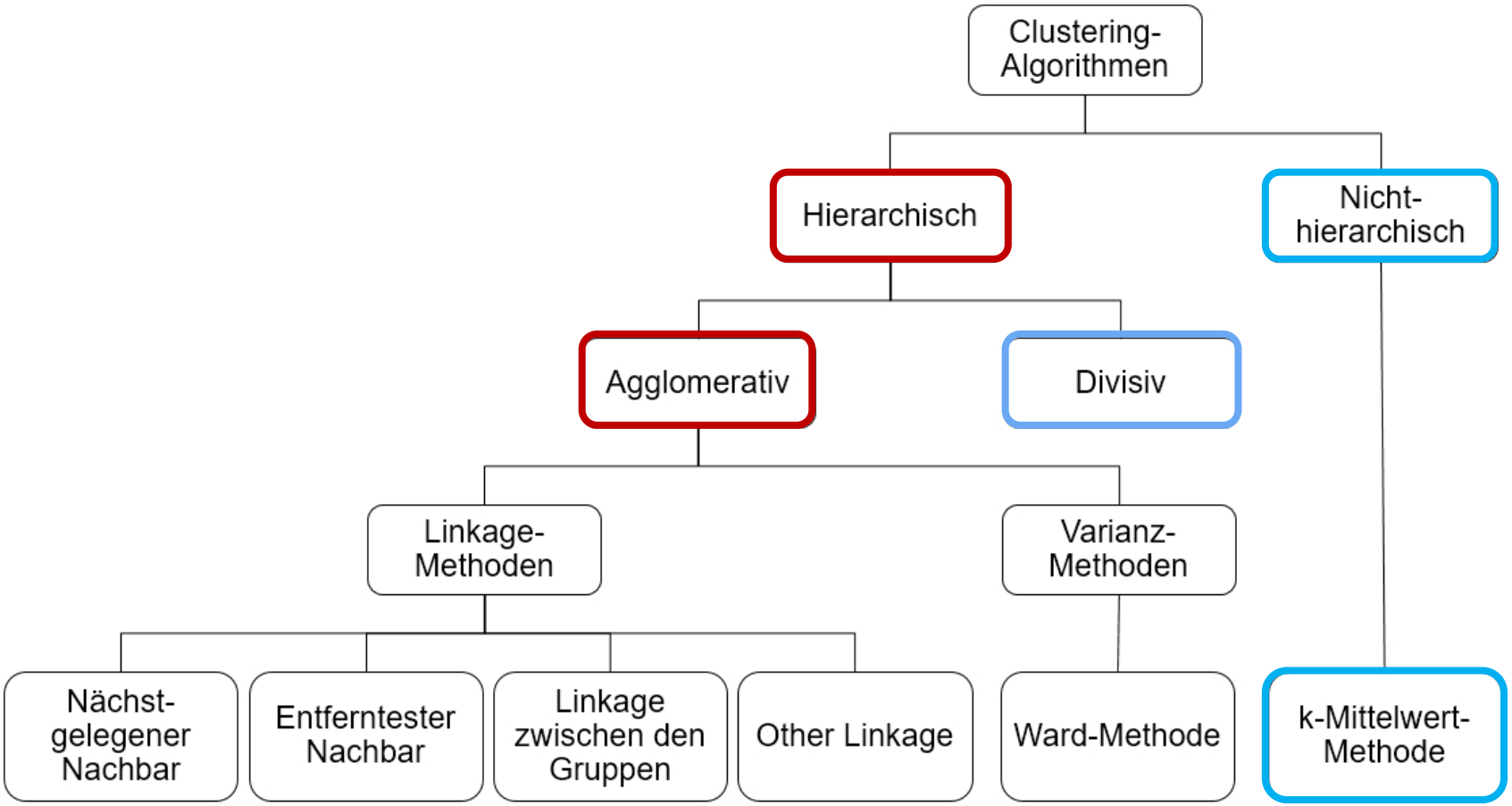
Deutsche Rentenversicherung Bund

→ 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

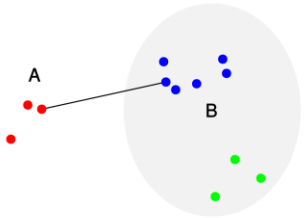
# Hierarchisch-agglomerative Clusteranalyse

Deutsche Rentenversicherung Bund

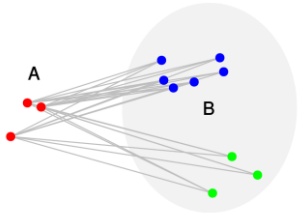
- Beginnend mit n Clustern, wobei n=Anzahl der Objekte
- Schrittweise Zusammenführung der Objekte bzw. Aggregate
- Verschiedene Skalenniveaus können nicht gemischt werden
- Verschiedene Messpunkte ausschlaggebend
  → Verschiedene Verfahren

# Hierarchisch-agglomerative Clusteranalyse

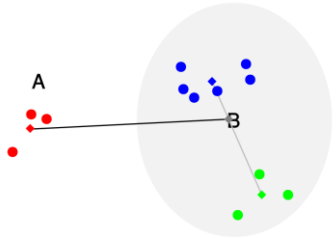
## Linkage-Methoden



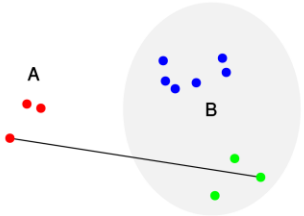

*Single Linkage*
Minimaler Abstand



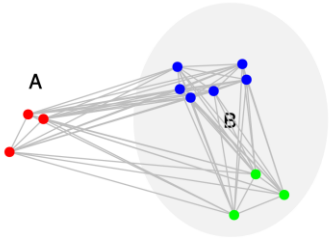

*Average Linkage*
Durchschnittlicher Abstand



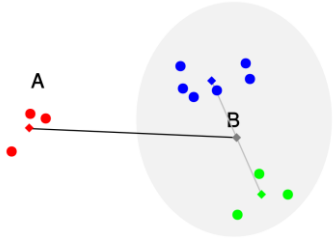

*Zentroid Clustering*
Abstand der Zentren




*Complete Linkage*
Maximaler Abstand




*Average Group Linkage*
Durchschnittlicher Abstand aus
der Vereinigung A und B




*Median Clustering*
Abstand der Zentren

Quelle: Wikipedia

# 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)

# Ward-Methode

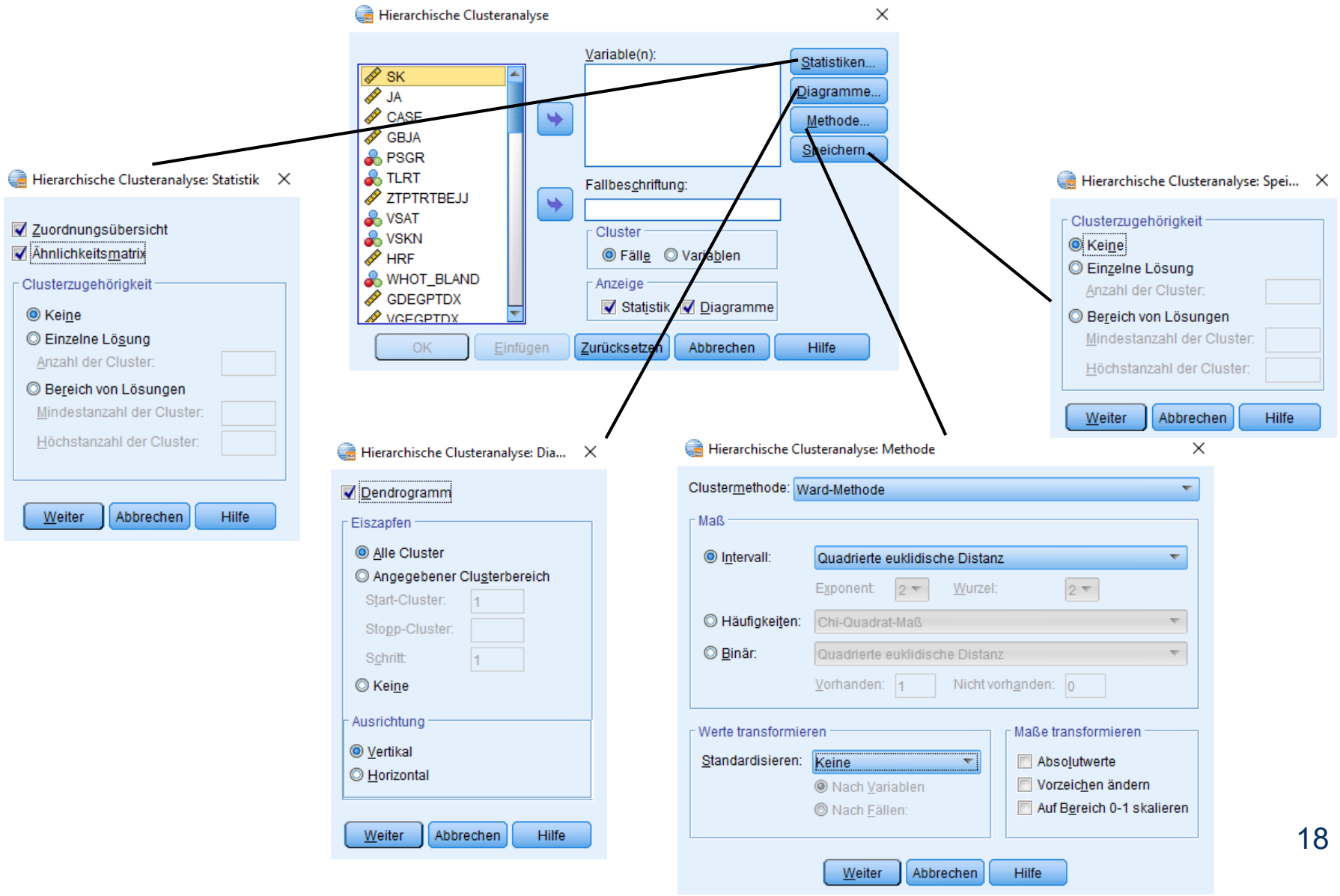
Deutsche Rentenversicherung Bund

- 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

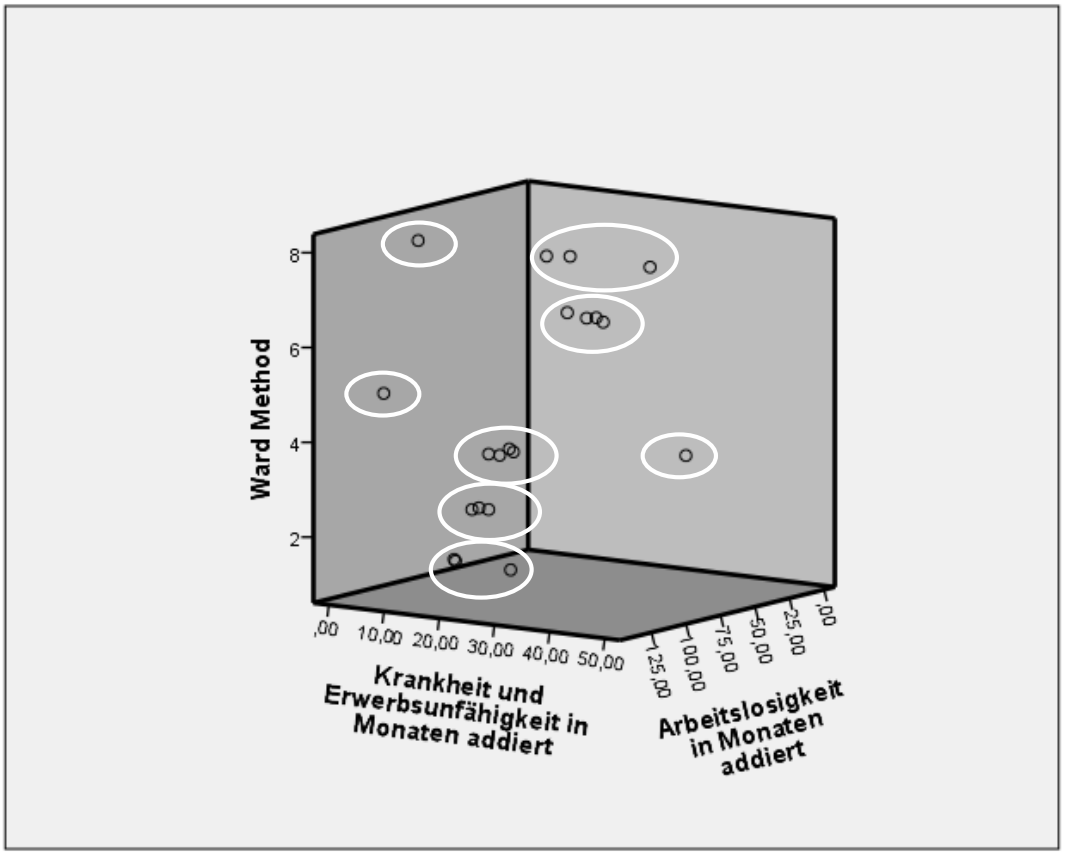


# Hierarchisch-agglomerative Clusteranalyse

Deutsche Rentenversicherung Bund

## 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

Deutsche Rentenversicherung Bund

- 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-Managment.

*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.

Deutsche Rentenversicherung Bund

# Clusteranalyse mit SPSS

## Vielen Dank für Ihre Aufmerksamkeit!