MULTIVARIATE STATISTIK

Prof. Dr. Nina Baur
Technische Universität Berlin
Sommersemster 2015
Zusammenfassung von Juliane Pfeiffer

1. Einführung

Einordnung der Multivariaten Statistik

- Mehr als 2 Variablen
- Eine abhängige durch mehrere unabhängige Variablen erklären

Systematik

Deskriptive / Beschreibende Statistik	Induktive Statistik
= Verdichtung der Information & Beschreibung der Stichprobe	= Verallgemeinerung auf Grundgesamtheit
z.B. Kausalanalyse, Mehrebenenanalyse, Dimensionsbildung, SNA, Clusteranalyse, Längsschnittanalyse	z.B. Schätzung eines multiplen linearen Modells, Multiples Testen von Koeffizienten und Parameter

nach N. Baur

Struktur-prüfende Verfahren	Struktur-entdeckende Verfahren
= Unterstützung kausaler Hypothesen über multivariate Zusammenhänge	= Identifizierung von Mustern in Daten
z.B. Regressionsanalyse	z.B. Faktorenanalyse, Clusteranalyse

nach S. Fromm

Grundsätzliches Vorgehen

1. Umformung der Forschungsfrage in mathematisches Modell
2. Auswahl des multivariaten Modells
   → Möglichkeit zur Beantwortung der Forschungsfrage
3. Überprüfung der Anwendungsvoraussetzungen
   (z.B. Vorhandensein angemessener Daten, Mindeststichproben, Zufallsstichproben, Skalenniveau, bestimmte Verteilungsvoraussetzungen, Variablen streuen genug, weitere Modellspezifische Annahmen)
4. Datenaufbereitung
5. Berechnung (z.B. mit SPSS)
6. Dateninterpretation
   → Statistisch
   → Theoretisch
   → Kontextspezifisches Wissen

Umgang mit Anwenungsvoraussetzungen

Dilemma
= In der Praxis Anwendungsvoraussetzungen fast nie alle erfüllt

Was tun?
a) anderes Verfahren
b) Verfahren trotzdem durchführen und vorsichtig interpretieren

→ hängt von der Robustheit des Verfahrens ab

Robustheit
→ Viele multivariate Verfahren liefern trotz vorliegender (geringer) Verstöße gegen die Voraussetzungen richtige Resultate

Variablentypen

Variablentyp	Charakteristika der Ausprägung	Beispiel
Manifeste Variable	Direkte Messung	Bruttoeinkommen in € 0 – x Mio.
Latente Variable	Indirekte Messung	Sozial Schicht
Indikator Variable	Zur Konstruktion latenter Variablen	Bildung, Einkommen, etc

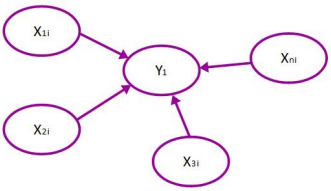
Individual- & Kollektivmerkmale

Individualmerkmale	Kollektivmerkmale
= Beziehen sich auf Individuen (z.B. Personen, Texte, Situationen etc.) Arten: - Absolute Variable: Genuine Eigenschaften des Individuums - Relationale Variable: Beziehung des Individuums zu anderen Individuen - Kontextvariable: Merkmale des Kollektivs, dem das Individuum angehört	= Beziehen sich auf Kollektive, die sich aus anderen Individuen zusammensetzen Arten: - Global: Genuine Eigenschaften des Kollektiv als Individuum - Analytisch: Berechnung aus absoluten Merkmalen niederer Ebenen - Strukturell: Berechnung aus relationalen Merkmalen niederer Ebenen

2. Multiple lineare Regressionsanalyse

Ziele

- multiple: Aufklärung von Kausalbeziehungen zwischen einer Abhängigen (*Kriterium*) und mehreren unabhängigen Variablen (*Prädiktoren*)



- linear: wenn sich die Ausprägung der abhängigen Variablen proportional mit der Veränderung der unabhängigen Variable verändert
- Schätzgleichung zur Beschreibung der durchschnittlichen linearen Abhängigkeit einer Variable von mehreren anderen
  - Stärke und Richtung des Einflusses der einzelnen unabhängigen Variablen auf die abhängige Variable
  - Erklärungskraft allen unabhängigen Variablen zusammen (Modellgüte)

- Schätzung von Ausprägungen der abhängigen Variable bei Merkmalsträgern, bei denen diese nicht bekannt ist.

Gleichung

$$y = b_0 + b_1 x + ... + b_n x + e\,(Stichprobe)$$
$$y = \beta_0 + \beta_1 x + ... + \beta_n x + U\,(Allgemein)$$

① ② ③

① Regressionskonstante
② Regressionskoeffizient
③ Störvariable

Güte

Bestimmtheitsmaß r^2 $r^2 = \frac{erklärte\ Streuung}{Gesamtstreuung}$ Es gilt $0 \leq r^2 \leq 1$

Residuen

beobachtete Abweichungen zwischen dem Messwert y_i und dem Schätzwert (nicht durch Störvariable)
→ sollte möglichst klein sein (Güte)

$$y'_1 = y_1 - e_i$$

Modellvoraussetzung

- Skalenniveau: Intervallskalenniveau mindestens, ggf. dichotomisieren
- Linearität zwischen Kriterium und Prädiktoren: Kann man ggf. herstellen, z.B. durch Logarithmisierung → linearer Zusammenhang darf nicht zu groß sein, da sonst redundant
- Normalverteilung der Residuen $\varepsilon_i \sim N(0, \sigma^2)$
- Keine Heteroskedastizität: Die Streuung der Residuen soll in jedem Wertebereich der abhängigen Variablen gleich sein (und keinem Muster folgen)
- keine Autokorrelation der Residuen: keine systematische Verbindung zwischen den Residun benachbarter Fälle
- keine Nullvarianz der Variablen: Variablen weisen ein Mindestmaß an Streuung auf
- keine Multikollinearität: Unabhängige Variablen dürfen nicht miteinander korrelieren → Additivität

Vorgehen

1. Fragestellung mit theoretischem Modell überlegen
   → Umformung der Forschungsfrage in kausalanalytisches Modell
2. Daten
   → Variablen operationalisieren / Fragebogen erstellen
   → Erhebung durchführen
3. Überprüfung der Anwendungsvoraussetzungen
   → Manche Voraussetzungen können erst nach Berechnung des Modells überprüft werden
4. Datenaufbereitung
5. Regressionsmodell berechnen
   → Ermittlung der relevanten Einflussvariablen (Modellökonomie)
   → Ermittlung der relativen Stärke des Einflusses der Einflussvariablen
6. Güte des Modells bestimmen
7. Dateninterpretation: statistisch (Gleichung), inhaltlich, Verallgemeinerung auf die GG

Interpretation der SPSS-Ausgabe

Korrelationstabelle
- Korrelation nach Pearson sollte deutlich von Null abweichen
- Signifikanz der abhängigen Variable mit unabhängigen Anschauen
- Pearson-Korrelation der unabhängigen Variable anschauen (→ Multikollinearität)

Modellzusammenfassung
- R^2 ist aufgeklärte Varianz im Gesamtmodell ($0 \leq R^2 \leq 1$)
- Autokorrelation über Durbin-Watson-Statistik (≈2 → keine)

Koeffizienten
- für Regressionsgleichung:
  - B bei Konstante = b_0
  - B bei Variable n = b_n

3. Zeitreihenanalyse

Zeitreihe und deren Komponenten
Zeitreihe $\{Z_t = y_{ti}, t = 1, ..., n\}$ → Eine Variable Y_{ti} wir zu verschiedenen Zeitpunkten $t_1,..., t_n$ gemessen

$$x_t = T_t + K_t + S_t + R_t$$

$$Zeitreihe = Trend + Konjunkturkomponente + Saisonkomponente + Residuum$$

Trendkomponente
Langfristige Änderung im Mittelwert der Zeitreihe

Saisonkomponente
Schwankungen, die sich relativ regelmäßig wiederholen (Tag, Woche, Jahr)

Restkomponente
Zusammenfassung aller sonstigen Einflüsse und Störungen

Ziel
→ Analyse langfristigen sozialen Wandels
- Beschreibung der Kurvenform (Maßzahl, Diagramme)
  - Besseres Verständnis längerfristigen sozialen Wandels
  - Verknüpfung mit anderen Verfahren (z.B. qualitative Fallstudie)
- Modellierung der Kurvenform und der Modellparameter
  - Trendkomponente: Regressionsanalyse, Gleitmittelverfahren
- Suche nach Ursachen bzw. Wirkungen für diese Verlaufsform
  - Regression mehrerer Zeitreihen
  - Gefahr der Scheinkausalität

Einordnung
- baut mathematisch auf der linearen Regression auf
- andere Verfahren: Kohortenanalyse, Ereignisanalyse, Squenzanalyse

Voraussetzungen

- Längsschnittdaten
- möglichst viele Messzeitpunkte
- Messzeitpunkte (Zeit) als eigene Variable im Datensatz

Herstellung von Zeitreihendaten auf der Kollektivebene

Erhebungsdesign für Kollektivmerkmale im Längsschnitt	
Direkte Erhebung von globalen Merkmalen - Laufende Protokollierung - Nachträgliche Konstruktion von Datensätzen	Aggregation von Individualmerkmalen - Replikationsstudien & Trenddesign - Paneldesign - Retroperspektive Befragung

Bestimmung der Komponenten

Trendkomponente

1. Variante: Methode der kleinsten Quadrate (Regressionsanalyse)
   unabhängige Variable: Zeit, abhängige Variable Y
   Problem: Zeitreihen meist nicht linear

2. Variante: Methode der Reihenhälfte (Gleitmittelverfahren)
   Trendwert zu einem Zeitpunkt: Mittelwert des Stützbereichs → Ergebnis geglättete Kurve
   → Je größer der Stützbereich desto stärker die Glättung

Periodenkomponente

Zeitreihe bereinigt vom Trend

Restkomponente

bereinigt von Trend und Periodenkomponente

4. Mehrebenenanalyse

Ziel

- Gegenstände verschiedener Ordnungen werden in einer Analyse simultan verrechnet
- Bestimmung der Wirkung und des relativen Einflusses folgender unabhängiger Variablen auf die abhängige Variable
  - Individualvariable
  - Kontext- bzw. Aggregationvariable
  - Interaktionseffekt → Zusammenwirken der Individual- und Kontextvariablen
- Mögliche Wirkungsrichtung
  - Makrodeterminismus (abhängige: Individualvariable, unabhängige: Kontextvariable)
  - Mikrodeterminismus (abhängige : Kontextvariable, unabhängige: Individualvariable)

Einordnung

- baut auf Regressionsanalyse auf
- andere Mehrebenenverfahren: ML/3, VARCL, MLA, Mixed Models

Gleichung

$$\hat{y}_{ij} = \hat{\beta}_{0j} + \hat{\beta}_{1j} x_{ij} + r_i$$

J → im Kontext j

Vorgehen

1. Fragestellung mit theoretischen Modell überlegen
   → Umformung der Forschungsfrage in ein statistisch bearbeitbares Modell
2. Daten
   → Variablen operationalisieren / Fragebogenkonstruktion
   → Erhebung durchführen
3. Überprüfung der Anwendungsvoraussetzung
4. Datenaufbereitung
5. Statistisches Modell berechnen. Berechnung...
   … des Nullmodells: Ist eine Mehrebenenanalyse überhaupt sinnvoll?
   … der Stärke der Individualeffekte
   … der Stärke der Kontexteffekte (Konditioniertes Modell)
   … der Stärke der Interaktionseffekte (Unkonditioniertes Modell)
   … des Gesamtmodells
6. Güte des Modells bestimmen
7. Dateninterpretation

Berechnung des statistischen Modells

1. Nullmodell

- Lohnt sich eine Mehrebenenanalyse überhaupt? → Vergleich Regressionsanalyse & MLH
- Zusätzliche Varianzkomponente u_j → wenn 0, dann lohnt es sich nicht
  a) Stärke des Einflusses des Kontextes auf die abhängige Variable bestimmen
  b) Modellgüte (R^2) muss hoch sein

2. Bestimmung der Individualeffekte

Erklärungskraft der Individualvariable für sich genommen

- Erklärung eines möglichst hohen Anteils der Varianz auf Ebene 1
- Minimierung der Residuen ε_i

3. Bestimmung der Kontexteffekte

Erklärungskraft der Kontextvariable für sich genommen

- Erklärung der Regressionskonstante β_{0j}
- Erklärung eines möglichst hohen Anteils der Varianz auf Ebene 2
- Minimierung der Residuen

4. Bestimmung der Interaktionseffekte

Interaktion zwischen Individual- und Kontextebene

- Erklärung des Regressionskoeffizienten β_{kj} (Slopes)
- Erklärung eines möglichst hohen Anteils der Varianz auf Ebene 2
- Minimierung der Residuen

Interaktionseffekt: Die Individualvariablen wirken in unterschiedlichen Kontexten unterschiedlichen → unterschiedliche Steigerung der Regressionsgeraden je nach Kontext

5. Bestimmung der Güte

Im Unkoordinierten Modell muss für jeden Regressionskoeffizienten eigene Regressionsanalyse durchgeführt werden

- Verwendung von Likelihood-Verfahren
- Berechnung der Devianz für das Nullmodell und das Alternativmodell – 2log Likelihood
  → Je näher de Devianz an 0, desto besser prognostiziert das Modell die empirischen Daten

5. Faktorenanalyse

Fragestellung

- Verallgemeinerung der Dimensionsanalyse nach dem Modell der Likert-Skala
- untersucht, ob mehrere latente Dimensionen vorliegen und wie diese inhaltlich bestimmt werden
  → Items einer Dimension korrelieren miteinander, andere nicht
  → inhaltlich und statistische Konstruktion theoretischer Hintergrundvariablen

Vorgehen

1. Datenaufbereitung (v.a. Umpolung + Skalenniveau)
2. Korrelationsmatrix
3. Extraktion und Rotation der Faktoren
   → Wie viele und welche Dimensionen
4. Berechnung der Dimensionsausprägung bei den Merkmalsträgern
   → Welche Items gehen ein?

Dispositionen

Konsistente situationsübergreifende Reaktionstendenzen in Hinsicht auf

- Einstellungen
- Fähigkeiten
- Verhalten → graduell

Ziele

= ausgehend von den Reaktionen der Befragten auf die Items, eine oder mehrere Dimensionsvariblen zu konstruieren

→ sind dimensionle Strukturen überhaupt erkennbar und welche bzw. wie viele Dispositionen sich sinnvoll unterscheiden lassen

→ Konstruktion geeigneter Skalen zur Messung von Dispositionen

→ Über Traceline (Item-Charakteristik) Zusammenhand Dimensionsausprägung und Bejaungswahrscheinlichkeit

Schritte

1. Dimensionsbestimmung
2. Itemselektion
3. Aufstellung einer Messvorschrift

Likert-Skalierung
→ Methode der summierten Ratings

- mindestens Intervallskalierung
- Traceline monoton bzw. annähernd linear
  → die Bejahungswahrscheinlichkeit eines Items ist umso größer, je stärker die Dimension ausgeprägt ist
- keine Rangfolge der Items vorausgesetzt

Cronbachs Alpha
- für die Homogenität der gesamten Skala
  → bei welchem Item steigt es → Item muss rein
  → bei welchem Item sinkt es → Item raus

$$\alpha = \frac{i * \bar{r}}{1 + \bar{r}(i-1)}$$

$0 \leq \alpha \leq 1$ → Ideal $\alpha \leq 0{,}8$

Trennschärfekoeffizient
- Maß, wie gut ein Item die Dimension erfasst
- Korrelation des betrachteten Items mit dem Gesamtpunktwert aller anderen Items

Berechnung und Inspektion der Korrelationsmatrix
- Alle Korrelation gegen Null → keine dimensionale Struktur
- sehr hohe Korrelation und hochsignifikant

Barlett-Test
Ist signifikant, wenn in der Grundgesamtheit nicht alle Korrelationskoeffizienten den Wert haben

KMO-Test
ist groß, wenn die partielle Korrelation klein sind, also die gemeinsame Streuung von Items durch ein Faktor bestimmt wird (0-1)

Extraktion & Rotation der Faktoren
→ Bestimmen der Rotation

Verfahren
= Hauptkomponentenmethode (HKM)
→ Annahme: Gesamte Varianz wird im Modell erklärt
→ Messwerte lassen sich als Linearkombination darstellen

1. Extraktion der Hauptkomponente, den den größten Teil der Gesamtvarianz aller Item im statistischen Sinn erklärt
   → Konstruktion neuer Achsen
2. nächste Hauptkomponente extrahieren usw.

Bestimmung der Komponenten
- sehr hohe Korrelation zwischen allen Items → 1 Achse
- alle Korrelationen gegen Null → Achsenzahl = Itemzahl
- dazwischen → 1 bis Itemzahl → statistisch + inhaltlich

Komponentenmatrix
- Einfachstruktur: jedes Item lädt ideal nur auf einen Faktor hoch

Kommunalität
- Extraktion möglichst hoch

Gesamtvarianz
- Anfänglicher Eigenwert deutlich größer als 1
- Rotierte Summer der quadrierten Ladung möglichst hoch

Bestimmung der Items & Berechnung der Dimensionsausprägung
- statistisch (nach Komponentenmatrix) & inhaltlich
- Messer der Ausprägung der Faktoren bei den einzelnen Mermalsträgern
  → Factorscore, regressionsanalytische Schätzung
  → Compute

6. Clusteranalyse

Typus
= Kombination aus mehreren (mindestens 2) nicht linear Zusammenhängender Variablen

→ überzufällige häufige Merkmalskombination im theoretisch möglichen Ereignisraum
→ Wiederholung in Zeit und Raum

Identifikation
→ Abstraktion gemeinsamer Nenner am Inhalt der Typisierung machen
→ Sinnbezug: Nenner = Beziehung zu den Sinnwelten der Handelnden

Varianten der Typenbildung
- Bildung übergeordneter Typen
- Aufspaltung einer Dimension in unterschiedliche Typenbildung
- Zusammenführen von Typen, die häufig gemeinsam auftreten

Ziele
Gruppieren einer Menge von Fällen (=Objekten) in Klassen (=Gruppen, Cluster)
→ Fälle werden durch eine Menge von Merkmalsträgern beschrieben
→ hinsichtlich dieser Merkmale:
- Homogenität innerhalb der Klasse
- Heterogenität zwischen den Klassen

Ergebnis
- Informationsreduktion
  → Äußerlich verschiedene Phänomene als eine Sinnstruktur gedeutet
  → leichtere Kmmunizierbarkeit
- Informationserweiterung
  → Interpretation geben über die Einzelaspekte der Variablen hinaus

Einordnung

- Andere Verfahren des Typenbildung: Korrespondenzanalyse
- Keine zwingende Kausalität (≠ Regressionsanalyse)
- nicht (linearer) Zusammenhang (≠ Fakorenanalyse)

Vorgehen

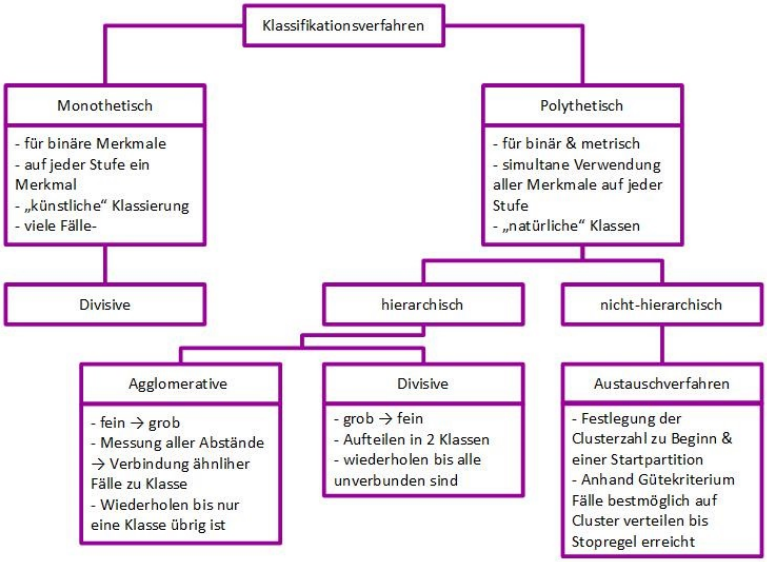
1. Auswahl von Fällen

- Die zu klassifizierende Objekte (=Fälle) werden durch das Forschungsziel bestimmt
  → Die Fälle müssen klar abgrenzbar sein
- Gesamtereignis wird beeinflusst durch
  - Zahl der Fälle
  - Ausreißer

2. Auswahl der Variablen (Klassifikationsmerkmale)

- Auswahl der Variablen beeinflusst Klassifikationsergebnis
- Sachliche Gesichtspunkte: Alle Variablen sollen annähernd gleichbedeutend für das Erkenntnisinteresse sein
- Ähnliche Wertebereiche (da sonst verschiedene Gewichtung)
  → oder Bewusst zur Gewichtung
- Einheitliches Skalenniveau

3. Auswahl eines Klassifikationsverfahrens



4. Auswahl eines Verfahrens der Abstandsmessung zwischen den Fällen

Grad der Ähnlichkeit (geringe Streuung & Distanz → Homogenität in den Klassen) und Unähnlichkeit (hohe Streuung & Distanz → Heterogenität zwischen den Klassen) muss gemessen werden

(Un-)Ähnlichkeitsmaße:

- Berechnung: merkmalsweiser Vergleich der Ausprägungen für Paare von Fällen
- Grad der (Un-)Ähnlichkeit
  → Ähnlichkeitsmaß
  → Distanzmaß
- bei binären Merkmalen: Über 4-Felder-Tafeln (Simple Matching Coeffizient, Jaccrads Coefficient)
- bei metrischen Merkmalen: über Distanzmaß (Euklidische Distanz, Quadrierte Euklidische Distanz, City-Block-Metric)

5. Auswahl eines Verfahrens der Abstandsmessung zwischen Gruppen (Fusionsalgorithmus)

Problem: Wie misst man den Abstand zwischen Gruppe?
→ Fusionsalgorithmus: Es muss Regel festgelegt werden, wie ähnlich Gruppen zusammengefügt werden

Verschiedene Verfahren:

- Single Linkage → Identifikation von Ausreißern
- Complete Linkage → zum Vergleich mit Single Linkage und Gefühl für die Struktur der Daten
- Average Link → Durchschnitt der Objekte einer Klassen ODER Ward-Verfahren → niedrigster Grad an Heterogenität und niedrigster Zuwachs

→ Dendogramm: Je früher sich Cluster vereinigen, desto Kompakter sind sie

6. ggf. zweiter Klassifikationsprozess um das Ereignis des erstens zu verbessern

7. Bestimmung der Klassenzahl & 8. Bestimmung der Güte (Gütemaß, Klassendiagnose)

Gütefunktion: Varianz, Determinantenfunktion, Kriterium der adaptiven Distanz → Das Invarianzproblem wird auf die „richtige" Modellbildung verlagert

- Knick im Struktogram → wo Kurve abknickt maximale Clusterzahl
- Sprung im Dendogramm → nicht zu große Heterogenität
- Sprung in der Fusionstabelle → wo bei Koeffizienten großer Sprung: Schritt + 1 - Sprungstufe

Klassendiagnose → Entscheidung über beste Clusterlösung + Interpretation

- Mittelwertvergleiche zwischen Clustern u. Gesamtdatensatz
- Streuung in, zwischen und zum Gesamtdatensatz
- t-Werte:

$$t = \frac{Mittelwert\ Variable\ J\ \text{im}\ Cluster\ G - Mittelwert\ Variable\ J\ \text{in}\ Gesamtdatensatz}{Standardabweichung\ Variable\ J\ \text{im}\ Gesamtdatensatz}$$

  → negativ: Im betreffenden Cluster wurden bezüglich der Variablen im Schnitt niedrigere Werte angegeben als im Gesamtdatensatz → Unterrepräsentation
  → positiv: Im betreffenden Cluster wurden bezüglich der Variablen im Schnitt höhere Werte angegeben als im Gesamtdatensatz → Überrepräsentation
- F-Wert:

$$F = \frac{Varianz\ der\ Variable\ J\ \text{im}\ Cluster\ G}{Varianz\ der\ Variable\ J\ \text{im}\ Gesamtdatensatz}$$

  → F < 1: Cluster G ist bezüglich Variable J homogener als Gesamtdatensatz
  → F > 1: Cluster G ist bezüglich Variable J heterogener als Gesamtdatensatz

9. Interpretation der Ergebnisse

7. Strukturelle Netzwerkanalyse

Perspektiven

Egozentriertes Netzwerkanalyse
= Einzelner Akteur zu Zentralität / Prestige
→ Degree (in/out), Closenedd, Betweeness, Strukturelle Löcher, Cutpoints

Gesamtnetzwerk
= Struktur des Gesamtnetzwerks zu Kohäsion (Grad und Qualität der Vernetzung)
→ Dichte, durchschnittliche Distanz, Maße der Zentralisierung, Hierarchisierung

Sub-/Teilgruppen
= Gruppenanalyse
→ Innendichte, Außendichte, n-Cliquen, k-plexe, k-cores

Verhältnis Individuum und Netzwerk

Akteur
→ Mikroperspektive, egozentriertes Netzwerk
- Das Verhältnis des Akteurs zum Gesamtnetzwerk kann beschrieben werden
- Voraussetzung: Daten über das Gesamtnetzwerk

Gesamtnetzwerk
→ Makroperspektive
- kann durch die Beziehung der Akteure untereinander beschrieben werden

2 Perspektiven der Charakterisierung von Mikro-Makro-Beziehung
- Kohäsion / Zusammenhalt des Netzwerkes
  → Mögliche Ressourcen für die Kooperation und kollektives Handeln
- Zentralität und Prestige von Akteuren
  → Mögliche Ressource für Einflussausübung, Machtkämpfe und Konkurrenz

Maße in Egozentrierten Netzwerken

1. Univariate Verteilung ermitteln und interpretieren
2. Homophilie → Eine überzufällige Ähnlichkeit in den Merkmalen und Einstellungen der Personen im egozentierten Netzwerk anhand von soziodemographischen Merkmalen und Einstellungsfragen über bivariate Zusammenhänge
3. Multiplexität von Beziehungen → Beziehung, die in unterschiedlichen (min. 2) Kontexten zugleich von Bedeutung ist
4. Multiplexität von Rollen → Zusammenfassung nach Rollen bei Personen (Verdichtung der Informationen)

   → Ohne Berücksichtigung des Gesamtnetzwerkes

Maße im Gesamtnetzwerk

Dichte
= Verhältnis zwischen realisierten und möglichen Beziehungen → Stärke der Verflechtung zwischen allen Akteuren im Netzwerk: Berechnung (N: Akteursanzahl)

- ungerichtete Graphen $\frac{realisierte\ Beziehung}{(N^2-N)\div 2}$ [0;1]
- gerichtete Graphen $\frac{realisierte\ Beziehung}{N\cdot(N-1)}$

Degree
= Anzahl der direkten Verbindungen eines Punktes (Zeilen oder Spalten auszählen)

In-/Outdegree
→ Bei gerichteten Beziehungen
= ein- und ausgehende Beziehung
- Indegree: Spaltensumme
- Outdegree: Zeilensumme

Degreebasierte Zentralisierung
Extremformen → Stern, Kreis, Doppelstern / Pyramide, Kette

Closeness
= Entfernung zu anderen Punkten
→ über Pfaddistanz

Wege (Walks)
= eine direkte Linie zwischen zwei Punkten

Pfad (Paths)
= eine indirekte Verbindung zwischen zwei Punkten durch Wege (über andere Punkte)
- Path = keine Wiederholung von Knoten
- Trail = keine Wiederholung von Kanten
- Walk = unbeschränkt

Zyklus
= Geschlossener Pfad zum Ausgangspunkt

Länge (Length)
eines Pfades = Zahl der Kanten von A nach B

Pfaddistanz
= Kürzeste Verbindung (max. N-1 bei Kette) → Pfaddistanzmatrix

Durchmesser (Diameter)
= Länge des längsten Pfades
→ Kompaktheit des Netzwerkes

Durchschnittliche Distanz
= Kürzeste im Schnitt

Betweenness
→ „Gatekeeper“

Erreichbarkeit
→ Erreichbarkeitsmatrix
→ Welche Akteure können sich (direkt oder indirekt) erreichen? Über wie viele Alternative Wege?

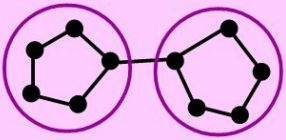
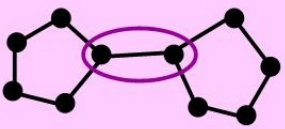
Verbundenheit
= Verbundenheit zwischen Punkten

Fragmentierung
→ Komplementär zu Verbundenheit
= Anteil derer, die sich nicht erreichen können

Flow
→ Informationsfluss
= je mehr und je kürzer die Wege, desto mehr Flow

Eigenvector
= gut vernetzt mit den zentralen Akteuren

Typen von Teilgruppen

Relationale Perspektive: Cliquen / Komponenten	Positionale Perspektive Blöcke
→ Fokus auf Beziehung zwischen Netzwerkmitgliedern	→ Fokus auf Rolle, Position der Netzwerkmitglieder
→ Blick auf erhöhte Kommunikations- und Kooperationschancen im Netzwerk zwischen Akteuren	→ Blick auf Konkurrenz zwischen Akteuren
→ Subgruppe = Akteure, die besonders stark miteinander verbunden sind und wenig nach außen verbunden sind	→ Subgruppen = Akteure die äquivalente Postionen innerhalb des Netzwerkes einnehmen und deshalb wechselseitig ersetzbar sind
	

Blöcke
- Ziel:
  - durch die Gruppierung der Netzwerkelemente anhand einer Äquivalenzdefintion soll die grundlegende Netzwerkstruktur sichtbar gemacht werden
  - Bei der Netzwerkreduzierung sollen möglichst viele Information erhalten bleiben

- Strukturelle Äquivalenz: Zwei Elemente sind strukturell äquivalent, wenn sie dieselbe ein- und ausgehende Beziehung im Netzwerk aufweisen
- Automorphe Äquivalenz: Zwei Elemente sind automorph äquivalent, wenn sie die selbe Zahl und dieselbe Art von ein- und ausgehenden Beziehung zu (strukturell oder funktional) äquivalenten Elementen im Netzwerk aufweisen
- Reguläre Äquivalenz: Zwei Elemente sind regulär äquivalent, wenn sie dieselbe Arten Beziehung zu (strukturell oder funktional) äquivalenten Elementen haben

Cliquen (Bottom-Up-Ansatz)

→ mindestens 3 direkt verbundene Akteure

- Kriterium der Gruppenbildung: Blick auf die Verbundenheit, d.h. Wer steht in direkter oder indirekter Verbindung zu wem

- N-Cliquen: Jedes Cliquenmitglied muss jedes andere in N-Schritten erreichen können
- N-Clans: wie N-Cliquen + nur über Clanmitglieder
- K-Plexe: Teilgraph mit n Akteuren, in dem jeder mindestens (n-k) Akteure erreichen können
- F-Gruppen: Heranziehen bewerteter Beziehungen zur Teilgruppenbildung, Betrachtung transitiver Triaden zur Entdeckung schwacher und starker Transivität

Komponenten (Top-Down-Ansatz)

- Kriterium der Gruppenbildung: Blick auf schwache Stellen im Netzwerk → wo würde das Netzwerk zerfallen, wenn ein Akteur (Cutpoint) oder eine Brücke (Bridge) wegfiel?
  → Isolierte Akteure sind eigene Komponenten

- Bi-Komponenten: Suche nach maximalen Teilgraphen, die nicht durch das Entfernen eines Akteurs zerfallen würden
  → Cutpoints: Bei der Enfernung eines Akteurs zerfällt das Netzwerk in unverbundene Teile
- Lamda-Sets: Algorithmus zur Bewertung von Beziehung zwischen zwei Akteuren anhand der Nutzung durch andere
  → Brücke: hohe Bewertung → Bei Kappung Schwächung des Netzwerkes
- Factions: Suche nach Idealbild (=vollständig verbundene Teilgruppen)