

Egozentrierte Netzwerke in Massenumfragen 3: Datenorganisation in einer SIR-Datenbank

Mohler, Peter Ph.; Pfenning, Uwe

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:

Mohler, P. P., & Pfenning, U. (1987). Egozentrierte Netzwerke in Massenumfragen 3: Datenorganisation in einer SIR-Datenbank. *ZUMA Nachrichten*, 11(20), 51-56. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-210278>

Nutzungsbedingungen:

Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:

This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Egozentrierte Netzwerke in Massenumfragen 3: Datenorganisation in einer SIR-Datenbank

1. Anforderungen an das Datenmanagement

Im folgenden Beitrag werden die besonderen Anforderungen der zuvor skizzierten Netzwerkstudie an die Datenaufbereitung und Datenverwaltung (Datenmanagement) erörtert. Für jeden Befragten der Netzwerkstudie wurden wie o.a. zwei Typen von Fragebögen erstellt:

- a. ein Basisfragebogen, der die soziodemographischen und bestimmte Einstellungsmerkmale des Befragten und Partner/-in beinhaltet, und
- b. einer von drei möglichen Netzwerkgeneratoren (GSS/Fischer/Global).

Dieses Instrumentarium wurde in einem Kreuzdesign in zwei Wellen den Befragten vorgelegt (vgl. Abschnitt 3 des Beitrags "Zum Design des Methodenforschungsprojektes" in diesem Heft) und würde nur mäßige Anforderungen an das Datenmanagement stellen, wenn für jede Kombination von Basisfragebogen und Netzwerkgenerator die Variablenzahl und -struktur ähnlich blieben. In diesem Fall könnte nämlich in konventioneller Weise je ein Datensatz pro Welle gebildet werden, der im Prinzip wie ein komplexeres "Split Half" (in diesem Fall ein "Split Third") aussehen würde.

Dies ist aber nicht der Fall. Die drei Netzwerkgeneratoren unterscheiden sich grundlegend in der Variablenzahl und -struktur. Dies ergibt sich aus der äußerst unterschiedlichen Art der Namensgeneratoren des GSS, Fischers und des globalen Generators (vgl. Abschnitt 2 des Beitrags "Zum Design des Methodenforschungsprojektes" in diesem Heft). Damit verliert sich von vornherein die Möglichkeit, auf gleichen Lochkartenpositionen pro Fragebogenversion unterschiedliche Informationen abzuliegen.

Es müßte also pro Variable jeder Fragebogenversion je ein Datenfeld (Lochkartenposition) vorgesehen werden. Dies würde bedeuten, daß für jede befragte Person Leerfelder für die beiden nicht abgefragten Fragebogenversionen "mitgeschleppt" werden müßten. Schematisch ergäbe sich für jeden Befragten eine der folgenden Kombinationen:

ZUMA

Basisfragebogen	Netzwerkgeneratoren		
	GSS	Fischer	Global
Informationen	Informationen	leer	leer
Informationen	leer	Informationen	leer
Informationen	leer	leer	Informationen

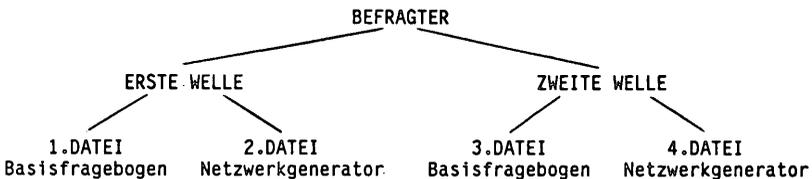
Diese Vorgehensweise würde offenkundig nicht nur Speicherplatzverschwendung, sondern auch erheblichen Zeitverlust bei Analysen bedeuten, weil z.B. SPSS vor jeder Datenverarbeitung den gesamten Datensatz auf eine Arbeitsdatei überträgt. Je größer die Analysedatei ist, um so länger dauert das Einlesen. Hinzu kommen die Beschränkungen bei der Variablenanzahl und Probleme bei der Verarbeitung von Textdaten aus offenen Fragestellungen.

Deshalb wurde ein Datenverarbeitungsprogramm favorisiert, das diese Probleme durch Bereitstellung folgender Möglichkeiten vermeidet:

- Verarbeitung variabler Dateiformate und Vermeidung von Leerpositionen,
- Verarbeitung unterschiedlicher Variablentypen (numerische und Textdaten),
- keine Variablenanzahlbeschränkung und stets mögliche Aufnahme weiterer Fälle unter der gleichzeitigen Bedingung der Vergabe identischer Fallidentifikationsvariablen (Panelansatz, 2. Welle),
- Interface zu Statistikprogrammpaketen.

Als grundlegende Anforderung kommt hinzu, daß die Datensätze in beliebiger Auswahl untereinander verknüpfbar sein sollten und aus vorhandenen Datensätzen in ebenso beliebiger Auswahl neue Variablentypen generiert und einem Datensatz angefügt werden können.

Die Datenverarbeitung in einer Datenbank erfüllt diese Voraussetzungen durch Verwendung einer fallorientierten Datenhierarchie, die sich schematisch wie folgt darstellen läßt:

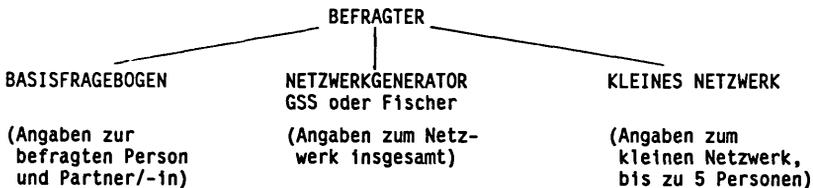


ZUMA

Jedem Befragten sind also vier Dateien zugewiesen; die Identifikationsvariable der Befragten ist die Verknüpfungsvariable zwischen den Dateien. Entsprechend der möglichen Kombinationen zwischen Basisfragebogen und Netzwerkgenerator können acht Dateien zur Datenerfassung erstellt werden (jeweils Basisfragebogen und drei Netzwerkgeneratoren je Welle). Damit ist das Problem der optimalen Speicherung der einzelnen Datensätze gelöst. Technische Details dazu sind in Abschnitt 2 "Technische Umsetzung" erläutert.

Ein zweites und für die spätere Analyse dringlich zu lösendes Problem besteht in der Zusammenführung der Personennennungen des sogenannten "kleinen Netzwerks" mit allen es beschreibenden Merkmalen (für die Netzwerkgeneratoren GSS und Fischer). Ein Beispiel mag dies verdeutlichen: Der Namensgenerator Fischer erhebt pro Befragten acht Situationen möglicher sozialer Kontakte, und bei jeder Situation können bis zu neun verschiedene Personen genannt werden. Hierbei sind Wiederholungsnennungen von Situation zu Situation möglich. Anschließend wurden die Befragten gebeten, die jeweils wichtigsten fünf Personen zu benennen. Hier muß zuerst eine äußerst sorgfältige Vercodung sicherstellen, daß jeder genannte Name eindeutig mit einer Codeziffer belegt wird. So abstrakt dargestellt scheint dies trivial zu sein. Die hohen Anforderungen an die Verocoder lassen sich erst dann angemessen einschätzen, wenn man selbst einmal eine derartige Vercodung durchzuführen versucht.

Die Angaben zu den Personen des "kleinen Netzwerkes" sind wie gesagt ein zentrales Konstrukt des Projekts. Deshalb wurde angestrebt, die Netzwerkpersonencodes mit ihren soziodemographischen Merkmalen (Alter, Bildung, Geschlecht und Beziehung zum Befragten) pro Befragten für die Netzwerkgeneratoren GSS und Fischer gemeinsam auf einer weiteren Datei abzulegen und darüber hinaus auch Variablen zu erstellen, die die Nennungsverteilung dieser fünf Netzwerkpersonen in allen 8 Situationen darstellen. Schematisch ergibt sich daraus für die Datenstruktur untenstehendes Bild:



ZUMA

Für die Analysen können dann kleine, übersichtliche Subdateien gebildet werden, die aus den drei Dateien die jeweils ge- und erwünschten Variablen enthalten. Die Variablen können also nach Analyseaspekten ausgewählt werden. Zur Verdeutlichung wiederum ein Beispiel: Ziel sei der Vergleich der soziodemographischen Merkmale von Befragten und ihren Netzwerkpersonen. Die Angaben zum Befragten und Partner/-in sind in der Datei BASISFRAGEBOGEN vorhanden, die Angaben zur Sozialstruktur der Netzwerkpersonen sind in der zuvor aus der Datei NETZWERKGENERATOR erstellten Datei KLEINES NETZWERK enthalten, zusätzlich können Einstellungsvariablen zugespielt werden. Man erhält pro Befragten einen Record (logischen Satz), der dann in beliebigen Analyseprogrammen verarbeitet werden kann.

Mit einer solchen Datenstruktur gewinnt der Forscher große Freiheiten in der Datenanalyse, da er die Analyserichtungen und -ebenen oft einfacher als bei gewöhnlichen Rechteckdateien ändern kann (z.B. Analysen auf Aggregatsebene der Netzwerkstruktur in Verbindung mit Analysen auf der Individualebene der Befragten).

Die o.g. Anforderungen an das Datenmanagement wurden also durch eine Zerlegung der Gesamtinformationen in sinnvolle Teildateien gelöst, die über die Identifikationsvariable des Befragten miteinander verbunden sind. Eine sorgfältige Vercodung der Namensnennungen in den Generatorfragen (Situationen) stellt sicher, daß die Netzwerkpersonen eindeutig identifizierbar sind. Diese Datenstruktur erlaubt unterschiedlichste Analysen der hoch komplexen Daten mittels einfacher Auswahlangaben.

2. Technische Umsetzung

Die Umsetzung der vorgenannten Anforderungen an das Datenmanagement erfolgte durch die Verwendung des SIR-DBMS (Scientific Information Retrieval Database Management System). SIR ist ein fallorientiertes Datenbanksystem, das die Aufnahme variabler Dateiformate und variierender Variablentypen in faktisch unbegrenzter Anzahl und beliebiger Verknüpfung zuläßt. Die Verbindungsvariable ist die sogenannte CASE ID.

Im vorliegenden Fall ist die CASE ID (Fallnummer und oberste Hierarchiestufe) die Variable "Befragtennummer", die verschiedenen Teildatensätze werden

ZUMA

RECORD TYPES (Rectypes) genannt. Mit Hilfe von sogenannten SORT IDs (Fallnummern) können niedrigere Aggregatstufen beschrieben werden (z.B. CASE ID=Befragter, SORT ID1=Partner, SORT ID2=erste Netzwerkperson etc.)

Für die Datenverwaltung der Netzwerkstudie wurde angestrebt, die Fälle des Basisfragebogens in einem Rectype zusammenzufassen. Deshalb wurden die Datensätze in einen für alle Fälle gemeinsamen Rectype BASISFRAGEBOGEN und drei Rectypes NETZWERKGENERATOR aufgeteilt. Der Rectype BASISFRAGEBOGEN enthält die Angaben zur befragten Person und Partner/-in, die drei Rectypes NETZWERKGENERATOR enthalten jeweils die Angaben über die Netzwerkpersonen des Befragten.

Für die wichtigsten Konstrukte der Netzwerkstudie wurden zwei gesonderte Rectypes angelegt, in denen die Angaben der genannten Netzwerkpersonen zu Geschlecht, Alter, Bildung und Beziehung zum Befragten je Fall aufgelistet sind. Schematisch ergibt sich folgende Struktur der Datenbank:

CASE ID=BEFRAGTE=FÄLLE
RECTYPE BASISFRAGEBOGEN, 1. Welle (N=720)
 Rectype1
 Angaben zur befragten Person und deren Partner/-in

RECTYPE NETZWERK-
GENERATOR GSS
(N=240)
Rectype2

RECTYPE NETZWERK-
GENERATOR FISCHER
(N=240)
Rectype3

RECTYPE NETZWERK-
GENERATOR GLOBAL
(N=240)
Rectype4

RECTYPES GROSSES
NETZWERK
Rectype5
Soziodemographische
Variablen

RECTYPES GROSSES
NETZWERK
Rectype9
Soziodemographische
Variablen

Rectype6
Frage- und Rangver-
teilung

Rectype10
Frage- und Rangver-
teilung

RECTYPES KLEINES
NETZWERK
Rectype7
Soziodemographische
Variablen

RECTYPES KLEINES
NETZWERK
Rectype11
Soziodemographische
Variablen

Rectype8
Frage- und Rang-
verteilung

Rectype12
Frage- und Rang-
verteilung

ZUMA

Entsprechend wird beim Aufbau der Teildateien für die zweite Welle verfahren. Für die Generierung der Rectypes für das kleine und große Netzwerk einer Person waren komplexe Datenretrievals notwendig. Diese Datenretrievals greifen auf die Daten der Rectypes zu und können neue Variablen und neue Rectypes generieren. Hinsichtlich der Erstellung der o.g. Netzwerkdateien (Rectypes 5, 6, 7, 8, 9, 10, 11 und 12) wurden durch Suchläufe die jeweils von einer Person angegebenen Netzwerkpersonen im Datensatz identifiziert und die entsprechenden Angaben zum Geschlecht, Alter, Bildung und Beziehung zum Befragten sowie die neuen Variablen Frage und Rangplatz der Nennung zugeordnet (Rectypes 5, 7, 9 und 11). Da als Vercodungskriterium für jede Person nur einmal die soziodemographischen Merkmale angegeben wurden, ist für Wiederholungsnennungen nur noch das Frageitem und gegebenenfalls der Rangplatz der Nennung von Interesse. Diese Angaben befinden sich in den Rectypes für die Frage- und Rangverteilung (Rectypes 6, 8, 10 und 12). Als Sortiervariablen (SORT IDs) wurden bei den Netzwerkrectypes die fünf Personen des kleinen Netzwerks definiert.

Diese Datenbankstruktur ermöglicht eine analyseorientierte Auswahl von Variablen und gegebenenfalls eine entsprechende Aufbereitung der Variablen (Recodierung, Aggregation und Sortierung), bevor sie über das SIR-interne Interface zu SPSS-, BMDP- oder SAS-Dateien verarbeitet werden. Dieses Interface ermöglicht die Durchführung von komplexen multivariaten Analysen, die innerhalb der Datenbank selbstverständlich nicht möglich sind.

Zur Handhabung der Datenretrievals bleibt anzumerken, daß die vielfältigen Möglichkeiten der Dateiverbindung(en) für den SIR-Anfänger zu Beginn schwer überschaubar sind. Eine maximale Ausschöpfung der Verarbeitungsmöglichkeiten ergibt sich erst mit zunehmender Erfahrung. Dennoch besteht die Möglichkeit zum unkomplizierten Umgang mit komplexen hierarchischen und vernetzten Datenretrievals, wenn von einem erfahrenen Systemverwalter Standardretrievals zur Verfügung gestellt werden. Diese können von den Nutzern aufgerufen und in begrenztem Umfang durch die Angabe von Parametern manipuliert werden. Diese Option erlaubt jedem Benutzer den SIR-gerechten Umgang zur Erstellung von Dateien oder Ausdrucken.

Der vorstehende Beitrag wurde von Peter Ph. Mohler und Uwe Pfennig verfaßt.

ZUMA

Geplante ZUMA-Publikationen

Für das Jahr 1987 ist die Herausgabe von drei ZUMA-Publikationen geplant.

Im Sommer 1987 wird im Campus-Verlag eine weitere ZUMA-Monographie unter dem Titel "Soziale Empfindungen. Ein interkultureller Skalenvergleich bei Deutschen und Amerikanern" herauskommen. Autoren sind Dagmar Krebs (ZUMA) und Karl Schuessler (Indiana University, Bloomington). Gegenstand der Monographie ist die Darstellung der Entwicklung von "Social Life Feeling"-Skalen im interkulturellen Vergleich zwischen der Bundesrepublik und den Vereinigten Staaten. Behandelt werden sowohl methodische Themen wie soziologisch bedeutsame Unterschiede innerhalb und zwischen den verglichenen Ländern. Bei der Diskussion der Methoden wird u.a. auch auf den Einfluß von Antwortstilen auf die erhaltenen Resultate und auf deren Erklärung eingegangen.

Im Laufe dieses Jahres ist auch die Herausgabe des zweiten und des dritten Bandes der ZUMA-Methodentexte geplant. Unter der Autorenschaft von Jürgen Friedrichs (Universität Hamburg) wird ein Band "Telefonische Befragung" vorbereitet. Darin sollen Voraussetzungen, Anwendungen, Probleme der Stichprobenbildung bzw. -gewinnung, das Vorgehen, die Erhebungssituation, Interviewerschulung und -überwachung, einige Varianten (insbesondere CATI-Systeme) und schließlich Fehlerquellen behandelt werden. Es ist geplant, die Einzelheiten mit Erfahrungen aus konkreten empirischen Forschungsprojekten zu illustrieren bzw. zu verbinden.

Gerhard Armingier (Universität-GHS Wuppertal) wird einen Band "Modelle zur Analyse von Panel-Daten" herausgeben. Weitere Autoren werden Ulrich Küsters und Franz Müller (beide Universität-GHS Wuppertal) sein. In diesem Buch werden primär Modelle für die Analyse von metrischen Panel-Daten behandelt. Diese Modelle umfassen die Behandlung von Autokorrelation, Dynamik und un beobachteter Heterogenität. Zur Schätzung wird als einheitliches Verfahren die Maximum-Likelihood-Schätzung mit LISREL angewandt. Insbesondere wird auch auf Probleme fehlender Daten und auf die Behandlung der Anfangswerte in der ersten Panelwelle eingegangen. Alle Modelle werden auch für latente Variablen formuliert. In einem speziellen Kapitel werden die Modelle auf dichotome und ordinale beobachtete Variablen erweitert.