Gesis-Projekt: Sozialer und ökonomischer Wandel in (West-)Deutschland

Zwischenbericht

Schätzung nichtvorhandener Bildungsinformationen in den Mikrozensen der 60er Jahre aus sekundären Quellen

Inga Höhne, Peter H. Hartmann
Dezember 2007
Heinrich-Heine-Universität Düsseldorf

Inhaltsverzeichnis

1. Einleitung

In Kooperation mit Gesis werden im Projekt „Sozialer und ökonomischer Wandel in (West-) Deutschland“ Analysen mit den Daten des Mikrozensus von 1962 bis 2004 durchgeführt.[1] Das von uns betreute Teilprojekt „Lebenslagen in Lebensformen“ befasst sich mit dem Wandel– insbesondere wirtschaftlicher – Lebenssituationen in verschiedenen Lebensformen (traditionelle Familien, Alleinerziehende, Singles etc.). Ein wichtiges, historisch durchaus vergleichbares Merkmal, das die Lebenslage strukturiert, ist die Schulbildung. Eine Aufschlüsselung von Haushalten nach der Schulbildung ihrer Mitglieder oder zumindest nach der Schulbildung einer Bezugsperson ist in den Mikrozensen der 60er und frühen 70er Jahre nicht möglich, weil die Schulbildung damals nicht erhoben wurde.

Um das Ziel der Berichterstattung über lange Zeiträume realisieren zu können, wird aber eine Information über die Bildungsabschlüsse von Mikrozensus-Befragten auch in den 60er Jahren benötigt. Im vorliegenden Papier versuchen wir die Schätzung dieser Abschlüsse zumindest bei einer Personengruppe, nämlich bei den Erwerbspersonen.

2. Entwicklung von Strategien zur Synthese von Bildungsvariablen für die Mikrozensen der 60er Jahre

Ziel des hier dokumentierten Projektschritts ist die Schätzung nichtvorhandener Bildungsangaben in den Mikrozensen der 60er Jahre.[2] Um eine solche Schätzung zu ermöglichen, wird zunächst nach Merkmalen gesucht, die in diesen Mikrozensen vorhanden sind und die mit der Schulbildung korreliert sind. Auf Basis dieser Merkmale wird dann die Schulbildung geschätzt, wobei zur Bestimmung der Schätzfunktion zeitlich den 60er Jahren möglichst nahe kommende amtliche Mikrodaten verwendet werden.

Insbesondere bieten sich hierfür der Mikrozensus 1976 und die Volkszählung 1970 an. Der Mikrozensus 1976 ist der erste Mikrozensus, in dem Schulbildung im Rahmen der Regelbefragung erhoben wurde. Auch in der den 60er Jahren noch näher liegenden Volkszählung 1970 wurde die Schulbildung erhoben. Auf den Mikrozensus 1976 kann in einfacher Weise

[1] Lengerer, Schroedter und Hubert (2007) dokumentieren die Aufbereitung der Daten, die die vergleichende Analyse der Mikrozensen ermöglicht.
[2] Auch der Mikrozensus 1973 enthält keine Schulbildung und wird deshalb von uns wie die Mikrozensen der 60er Jahr behandelt.

zugegriffen werden, da er Bestandteil der Datenkumulation des Gesis-Projekts „Sozialer und ökonomischer Wandel in (West-)Deutschland“ ist. Ein Zugriff auf Daten der Volkszählung 1970 ist über die in-house-Nutzung bei ZUMA möglich, dabei wurde die 10%-Stichprobe dieser Volkszählung verwendet.

Zunächst werden relevante Prädiktoren für die Schulbildung im Mikrozensus 1976 und bei der Volkszählung 1970 ermittelt, indem eine dreistufige Bildungsvariable geschätzt wird. Die Bildungsvariable enthält die folgenden Kategorien:

1) Volks- und Hauptschulabschluss
2) Realschulabschluss bis Fachhochschulreife
3) Abitur

Es wird also lediglich die allgemeine Schulbildung und nicht die berufliche Bildung betrachtet, eine Schätzung der beruflichen Bildungsabschlüsse in Verbindung mit allgemeinen Abschlüssen würde sich als zu schwierig erweisen, da die Gruppenzugehörigkeit mit zunehmender Zahl der Kategorien immer schwieriger zu schätzen ist.

Die unabhängigen Variablen, mit denen die Bildung geschätzt wird, sind Geschlecht, Geburtskohorte, Stellung im Beruf und Wirtschaftssektor. Die Vorhersageleistung der einzelnen unabhängigen Variablen und ihrer Interaktionen wird zunächst mittels multinomialer logistischer Regressionsanalysen ermittelt. Die Zuweisung der Bildungskategorien zu einzelnen Befragten geschieht einerseits mittels eines Cold-Deck-Verfahrens[3], bei dem in den Mikrozensen der 60er Jahren jeder Kombination der unabhängigen Variablen eine Bildungsverteilung zugewiesen wird, und dann innerhalb der jeweils gegebenen Kombination jedem Befragten zufällig eine Schulbildungskategorie zugewiesen wird. Bei diesem Verfahren werden, in ähnlicher Weise wie bei einem gesättigten loglinearen Modell, alle Interaktionsterme der unabhängigen Variablen berücksichtigt. Nachteil ist eine größere Zahl nicht oder schwach besetzter Zellen. Ein ungesättigtes loglineares Modell vermeidet diesen Nachteil, indem nur wenige, begründete Interaktionen berücksichtigt werden. Deshalb wird ein solches Modell als Alternative zum Cold-Deck-Verfahren entwickelt und getestet.

Die Validität der mit den verschiedenen Verfahren erreichten Schätzungen wird beim Mikrozensus 1976 wie folgt überprüft: Die Schätzung wird durchgeführt und dann mit der Verteilung der tatsächlichen Schulbildung bei den Befragten verglichen. In einem zweiten Schritt wird der Zusammenhang zwischen dem Einkommen der Befragten und ihrer Schulbil-

[3] Unter einem Cold-Deck-Verfahren wird normalerweise die Ersetzung fehlender Werte bei einer Erhebung durch Werte einer früheren Erhebung verstanden. In unserem Fall ersetzen wir allerdings fehlende Werte älterer Mikrozensen durch Werte neuerer.

dung für die geschätzten Bildungswerte ermittelt und mit dem Zusammenhang für die tatsächlichen Bildungswerte verglichen.

2.1 Ermittlung der relevanten Schätzparameter

Um möglichst vielen Befragten das Merkmal Schulbildung zuweisen zu können, wurden alle Analysen auf Basis der wohnberechtigten Bevölkerung durchgeführt. Weiterhin wurden die Analysen und Vorhersagen auf Erwerbspersonen beschränkt. Nur bei diesen steht eine hinreichende Zahl von Prädiktoren der Schulbildung zur Verfügung.

Um herauszufinden, anhand welcher Variablen die nicht erfragten Schulabschlüsse in den Mikrozensen der 60er Jahre vorhergesagt werden können, werden zunächst multinomiale logistische Regressionen der Schulbildung auf die für die Schätzung der Bildungsinformationen in Frage kommenden unabhängigen Variablen berechnet. Basis für diese Analyse ist der älteste Mikrozensus mit Bildungsinformationen (1976), entsprechende Fälle aus dem Befragungsjahr 1976 wurden aus der kumulierten Gesamtdatei des Projektes ausgewählt. Um eine Vergleichbarkeit der Ergebnisse zu gewährleisten und um die Analysen zu erleichtern, wurde nicht auf die Originaldaten des Mikrozensus 1976 zurückgegriffen, sondern auf die 1976er Daten der kumulierten Datei.

Insgesamt wurde ein Modell mit möglichst wenig Prädiktoren gesucht, da durch jeden weiteren Prädiktor die Tabellendatei für das Schätzverfahren vergrößert wird und mehr Zellen mit kleiner Besetzung entstehen würden. Dadurch steigt der Anteil der Zufallsfehler sowohl beim Cold-Deck-Verfahren als auch bei den loglinearen Modellen. Deshalb musste auf den Einbezug einiger Variablen verzichtet werden, nicht berücksichtigt wurden das Bundesland und die Gemeindegrößenklasse. Weiterhin wurde das (mit der Schulbildung stark korrelierte) persönliche Einkommen der Erwerbspersonen nicht berücksichtigt, weil es weiter unten zu Validierungszwecken als abhängige Variable dient. Darüber hinaus können Einkommensunterschiede in verschiedenen Haushaltsformen Gegenstand späterer inhaltlicher Analysen sein, so dass eine analytische Bindung des Einkommens an die Schulbildung zu diesem Zeitpunkt nicht opportun erscheint.

Als unabhängige Variablen wurden zunächst Geschlecht (gesch), Geburtskohorten (ermittelt aus dem Geburtsjahr)[4], Wirtschaftssektor (sek) und Stellung im Beruf (stib_1) aus-

[4] Insgesamt wurden 13 Geburtskohorten gebildet, die die berufstätigen Jahrgänge zu beiden Erhebungszeitpunkten abdecken. Die jüngste Kohorte wird beim Mikrozensus 1976 ausgeschlossen, weil diese Personen noch nicht arbeiten konnten. Aus dem gleichen Grund wird die älteste der ermittelten Kohorten bei der Analyse mit der

gewählt. Dabei sollte geprüft werden, ob sich die genannten Variablen für die Schätzung als relevant erweisen.

Da sich die Bildungsverteilung von Frauen und Männern historisch sehr unterschiedlich entwickelt hat – Frauen haben, ausgehend von einer sehr schlechten Position, von den Bildungsreformen stärker profitiert als Männer - wurden die Interaktionen der Kohorten mit dem Geschlecht berücksichtigt. Die Gemeindegrößenklasse dagegen fand keine Berücksichtigung. Statt dessen wurde die Interaktion der Stellung im Beruf mit dem Wirtschaftssektor in das Modell aufgenommen, was regionale Unterschiede der Wirtschaftsstruktur berücksichtigt. Modelle für einzelne Bundesländer wurden nicht geschätzt –auch hier wurde auf eine weitere Differenzierung zugunsten höherer Fallzahlen verzichtet.

Tabelle 1 gibt die Ergebnisse der multinomialen logistischen Regression mit Schulbildung als abhängiger Variable wieder. Referenzkategorie bei den abhängigen Variablen ist der Volksschulabschluss. Die Ergebnisse werden sowohl für den Mikrozensus 1976 als auch für die Volkszählung 1970 berichtet. Höhere Bildungsabschlüsse finden sich in besonderem Maße bei Selbständigen, Beamten, und in Angestelltenberufen. Im Primärsektor sind die Bildungsabschlüsse am niedrigsten, im Sekundärsektor sind sie etwas höher, am höchsten sind sie im tertiären Sektor. Zunächst überraschend sind die relativ höheren Bildungsabschlüsse bei den älteren Kohorten, sie sind vermutlich durch ein selektiv häufigeres Ausscheiden niedrig Gebildeter aus dem Erwerbsleben erklärbar.

Volkszählung 1970 nicht berücksichtigt – hier ist die älteste Kohorte bereits nicht mehr im Erwerbsleben. Die Referenzkategorie ist aber für beide Jahrgänge gleich.

Tabelle 1: Regressionskoeffizienten Exp(B) in multinominalem logistischem Regressionsmodell der Schulbildung, Teil 1 (ohne Interaktionen)

	MZ 1976		VZ 1970	
Volksschule (Referenz)	Realschule	Abitur	Realschule	Abitur
Geschlecht				
männlich (Referenz)				
Weiblich	0,93	0,77**	1,03*	0,74**
Stellung im Beruf				
(Heim-)Arbeiter (Referenz)				
Kaufm./ techn. Auszubildender	16,84**	11,88**	22,46**	83,09**
Gewerblicher Auszubildender	4,56**	4,74**	7,98**	35,31**
Selbstständiger, […]	9,22**	7,56**	9,55**	30,99**
Mithelfender Familienangehöriger	7,66**	3,32**	6,97**	8,71**
Beamter, Richter, Polizist, […]	13,23**	22,71**	11,35**	80,48**
Angestellter, Zivildienstleistender	11,89**	5,55**	12,66**	22,31**
Wirtschaftssektor				
Tertiärsektor (Dienstleistung) (Referenz)				
Primärsektor (Landwirtschaft, Bergbau)	0,51**	0,29**	0,57**	0,32**
Sekundärsektor (Industrie)	0,69**	0,40**	0,72**	0,49**
Gebutskohorten				
1926 bis 1930 (Referenz)				
1956 oder später	1,33**	0,19**		
1951 bis 1955	1,97**	0,49**	0,75**	0,02**
1946 bis 1950	1,66**	1,12*	1,31**	0,26**
1941 bis 1945	1,51**	1,32**	1,67**	0,97*
1936 bis 1940	1,08*	0,98	1,17**	1,00
1931 bis 1935	0,98	0,98	1,01	0,92**
1921 bis 1925	0,93	1,05	0,89**	1,06**
1916 bis 1920	1,09	1,12*	1,10**	1,04**
1911 bis 1915	1,28**	1,52**	1,12**	0,96**
1906 bis 1910	1,31**	2,08**	1,10**	1,03
1901 bis 1905	1,16	2,09**	1,23**	1,40**
1900 und früher			1,12**	1,65**
[Interaktionen, siehe Tabelle 2]				
Pseudo-R^2				
Cox und Snell	0,25		0,20	
Nagelkerke	0,32		0,30	
McFadden	0,19		0,19	
N	167.464		2.736.294	

* p<0,05 ** p<0,01

Tabelle 2 berichtet weiterhin die berücksichtigen Interaktionseffekte. Dabei zeigen sich besonders deutlich die geringe allgemeine Schulbildung bei den älteren Jahrgängen der Frauen und die vergleichweise besseren Chancen bei den jüngeren Frauen. In der Kombination von Wirtschaftssektor und Stellung im Beruf zeigen sich (relativ) höhere Bildungsabschlüsse bei Angestellten und teilweise bei Beamten im Primärsektor, ebenso bei Angestellten im sekundären Sektor.

Tabelle 2: Regressionskoeffizienten Exp(B) in multinominalem logistischem Regressionsmodell der Schulbildung, Teil 2 (nur Interaktionen)

	MZ 1976		VZ 1970	
Volksschule (Referenz)	Realschule	Abitur	Realschule	Abitur
[Fortsetzung, nur Interaktionen]				
Mann, 1926 bis 1930 (Referenz)				
Frau * 1956 oder später	1,15*	1,01		
Frau * 1951 bis 1955	0,92	1,51**	0,95**	0,55**
Frau * 1946 bis 1950	0,92	1,19*	1,08**	2,02**
Frau * 1941 bis 1945	1,07	1,04	1,03	1,49**
Frau * 1936 bis 1940	1,08	0,95	1,06**	1,06*
Frau * 1931 bis 1935	1,05	0,90	1,00	0,90**
Frau * 1921 bis 1925	1,17*	0,78**	1,12**	0,87**
Frau * 1916 bis 1920	1,35**	0,75**	1,25**	0,81**
Frau * 1911 bis 1915	1,20*	0,67**	1,19**	0,97
Frau * 1906 bis 1910	0,97	0,39**	1,05*	0,70**
Frau * 1901 bis 1905	0,93	0,25**	0,80**	0,37**
Frau * 1900 und früher			0,60**	0,24**
Tertiärsektor, Arbeiter (Referenz)				
Primärsektor * Kaufm./ techn. Auszub.	1,48	1,98	1,46**	1,35
Primärsektor * Gewerblicher Auszub.	1,15	2,01	1,46**	3,04**
Primärsektor * Selbstständiger […]	0,30**	0,13**	0,29**	0,13**
Primärsektor * Mithelfender Fam.	0,23**	0,20**	0,23**	0,21**
Primärsektor * Beamter, […]	8,02**	4,69**	3,09**	1,21
Primärsektor * Angestellter, […]	1,95**	2,96**	1,43**	2,15**
Sekundärsektor * Kaufm./ techn. Auszub.	1,11	1,48	1,44**	1,66**
Sekundärsektor * Gewerblicher Auszub.	0,75**	0,97	0,92**	0,86
Sekundärsektor * Selbstständiger […]	0,93	0,58**	0,75**	0,42**
Sekundärsektor * Mithelfender Fam.	0,86	0,79	1,00	0,79**
Sekundärsektor * Beamter,	1,78*	0,49	0,96	1,01
Sekundärsektor * Angestellter, […]	1,38**	1,73**	1,29**	1,52**
Pseudo-R^2				
Cox und Snell	0,25		0,20	
Nagelkerke	0,32		0,30	
McFadden	0,19		0,19	
N	167.464		2.736.294	

* p<0,05 ** p<0,01

2.2 Entwicklung einer Schätzmethode durch Cold-Deck-Verfahren

Ziel dieses Arbeitsschritts ist es, jedem Individuum in den Mikrozensen der 60er Jahre einen geschätzten Wert für den Bildungsabschluss zuzuordnen. Beim Cold-Deck-Verfahren wird dabei auf Daten einer anderen Datei, in unserem Fall des Mikrozensus 1976 bzw. der Volkszählung 1970, zurückgegriffen. Beim Mikrozensus 1976 wird der scientific use file verwendet, bei der VZ 1970 die 10%-Stichprobe.

Gemäß den Ergebnissen der logistischen Regressionsanalyse wird eine Tabelle mit allen möglichen Merkmalskombinationen erstellt. Im Einzelnen handelt es sich im die Kombi-

nation der folgenden Merkmale (in Klammern werden die Variablennamen aus der kumulierten Gesamtdatei des Mikrozensus und die Merkmalsausprägungen angegeben):

1) Geschlecht (gesch, 2 Kategorien: männlich, weiblich)

2) Geburtskohorten, ermittelt aus dem Geburtsjahr, 11 Kategorien[5]: für Berechnungen mit dem Mikrozensus 1976: 1956 bis 1960, 1951 bis 1955, 1946 bis 1950, 1941 bis 1945, 1936 bis 1940, 1931 bis 1935, 1926 bis 1930, 1921 bis 1925, 1916 bis 1920, 1911 bis 1915 und 1906 bis 1910. Für Berechnungen mit der Volkszählung 1970: 1951 bis 1955, 1946 bis 1950, 1941 bis 1945, 1936 bis 1940, 1931 bis 1935, 1926 bis 1930, 1921 bis 1925, 1916 bis 1920, 1911 bis 1915, 1906 bis 1910 und 1901 bis 1905. Die in den Modellen berücksichtigten Kohorten umfassen so jeweils die Kohorten, die im Erhebungsjahr berufstätig sind, und zwar im Mikrozensus 1976 im Alter zwischen 16 und 70 Jahren und in der Volkszählung 1970 im Alter zwischen 15 und 69 Jahren.

3) Wirtschaftssektor (sek, 3 Kategorien: Primärsektor, Sekundärsektor, Tertiärsektor),

4) Stellung im Beruf (stib_1, 7 Kategorien: (Heim-)Arbeiter; Kaufm./ techn. Auszubildender; Gewerblicher Auszubildender; Selbstständiger; Mithelfender Familienangehöriger; Beamter, Richter, Polizist, …; Angestellter, Zivildienstleistender)

5) abhängige Variable: Schulbildung (3 Kategorien: Volksschulabschluss, Realschulabschluss/Fachhochschulreife, Abitur)

Die Tabelle der Kombination dieser Merkmale enthält 2*11*3*7*3=1386 Zellen. Beschränkt man sich auf die Erwerbspersonen mit gültiger Angabe bei allen 5 Merkmalen, erhalten wir die in Tabelle 3 dargestellten Kombinationen aus Fallzahl und Zellenzahl.

Tabelle 3: Gültige Fälle, Zahl besetzter Zellen und Zahl der Nullzellen

	Zahl gültiger Fälle	Zahl besetzter Zellen	Zahl der Nullzellen
Volkszählung 1970	2.704.386	1161	225
Mikrozensus 1976	165.964	957	429

Fasst man alle drei Bildungskategorien zusammen, erhält man eine Tabelle von 1386/3, also von 462 Zellen. Diese Zellenzahl steht uns bei den Mikrozensen der 60er Jahre ohne Bildungsangabe in gleicher Weise wie beim Mikrozensus 1976 oder der Volkszählung 1970 zur Verfügung. Für jede dieser 462 Zellen berechnen wir die Anteile der drei Bildungsgruppen bei Volkszählung 1970 bzw. Mikrozensus 1976.

Beim Cold-Deck-Verfahren verteilen wir nun diese Anteile auf die Befragen der Mikrozensen der 60er Jahre. Dies geschieht innerhalb der 462 Zellen so, dass die Randvertei-

[5] Abweichend von der logistischen Regression, bei für die Volkszählung 1970 und den Mikrozensus 1976 jeweils 12 Kategorien verwendet wurden.

lung innerhalb jeder Zelle in den Mikrozensen der 60er Jahre der Randverteilung derselben Zelle in der Volkszählung 1970 bzw. im Mikrozensus 1976 exakt angepasst wird. Die Zuordnung zu den einzelnen Befragten der Mikrozensen der 60er Jahre dagegen geschieht innerhalb der 462 Zellen zufällig, aber so, dass die Randverteilung von Volkszählung 1970 bzw. Mikrozensus 1976 erhalten bleibt. So entsteht auch in den Mikrozensen der 60er Jahre eine Matrix von 462 Zellen ohne Bildungsangabe * 3 Zellen mit Bildungsangabe =1386 Zellen.

Das folgende Beispiel bezieht sich auf einen Validitätstest, bei dem die Zuweisung von Bildungsmerkmalen aus dem Mikrozensus 1976 an der gleichen Datei, aus der zuvor das Bildungsmerkmal entfernt wurde, durchgeführt wurde. Im Folgenden wird die Datei, die zu Validierungszwecken gebildet wurde und bei der späteren Zuweisung durch die Mikrozensen der 60er Jahre ersetzt wird, durch den Bestandteil „_val“ im Dateinamen gekennzeichnet. Die Analysen wurden mit dem Programmpaket SPSS durchgeführt, die Syntaxbefehle werden für jeden Arbeitsschritt protokolliert.

Der Logik von Cold-Deck-Verfahren zufolge sind zwei Dateien die Basis der Analysen, und zwar einer mit dem zu übertragenden Merkmal und einer ohne. In diesem Fall ist die Datei mit dem Merkmal „Schulabschluss“ (im Folgenden „Geberdatei“ genannt) beispielsweise der Mikrozensus 1976, die Datei ohne das entsprechende Merkmal (im Folgenden „Empfänger“-Datei genannt) der Mikrozensus 1976_val (bei dem zu Validierungszwecken die Variable Schulbildung entfernt wurde).

1) Zunächst wird für jeden Fall in der Empfängerdatei eine Fallnummer und eine reproduzierbare Zufallszahl ermittelt, nach der die Dateien sortiert werden. So können Effekte der Sortierung innerhalb von Zellen ausgeschlossen werden und die Fälle zu einem späteren Zeitpunkt wieder in die ursprüngliche Reihenfolge gebracht werden.

```
*(für die jeweilige Empfängerdatei).
set rng= mt mtindex=2000000.
compute fallnr = $casenum.
compute zufall=uniform(1).
sort cases by zufall.
```

2) Für die beiden jetzt nach einer Zufallszahl sortierten Dateien wird jeweils eine Arbeitsdatei aggregiert, die alle Erwerbspersonen mit gültigen Angaben für die Stellung im Beruf und den Wirtschaftssektor in Gruppen zusammenfasst. Die Gruppen werden definiert nach Geschlecht, Geburtskohorte, Stellung im Beruf und Wirtschaftssektor, in der „Geber“-Datei (MZ 1976) zusätzlich nach Schulabschluss.

```
get file="PFAD\MZ1976_val.sav".
select if erw_1 ne 3.
```

```
select if range(stib_1,1,7) and range(sek,1,3).
AGGREGATE
  /OUTFILE='PFAD\Empfaengerdatei.sav'
  /BREAK=gesch kohort stib_1 sek
  /N_BREAK76_val=N.

get file="PFAD\MZ1976.sav".
select if erw_1 ne 3.
select if range(stib_1,1,7) and range(sek,1,3).
select if range(schul3,1,3).
AGGREGATE
  /OUTFILE='PFAD\Geberdatei.sav'
  /BREAK=gesch kohort stib_1 sek schul3
  /N_BREAK76=N.
```

3) Im nächsten Schritt werden beide Dateien anhand der Schlüsselvariablen Geschlecht (gesch), Geburtskohorte (kohort), Stellung im Beruf (stib_1) und Wirtschaftssektor (sek) zusammengeführt. Es ergibt sich in Kombination von Geschlecht, Geburtskohorten, Stellung im Beruf, Wirtschaftssektor und Schulbildung (aus der Geberdatei) eine Matrix mit theoretisch möglichen 1386 Zellen. So enthält die Datei sowohl die Gruppengrößen mit – als auch ohne die Information des Schulabschlusses. Tabelle 4 zeigt einen Ausschnitt der zusammengefügten aggregierten Datei, die Grundlage weiterer Berechnungen ist.

```
match files
 /file='PFAD\Empfaengerdatei.sav'
 /file='PFAD\Geberdatei.sav'
 /by gesch kohort stib_1 sek.
execute.
save outfile='Empfänger_Geberdatei.sav'.
```

Tabelle 4: Schematische Darstellung zum Cold-Deck-Verfahren, Ergebnis von Arbeitsschritt 3

Geschlecht (gesch)	Kohorte (kohort)	Stellung im Beruf (stib_1)	Wirtschafts-sektor (sek)	Fallzahl 1976 (ohne Bildung) (n_break76_val)	Bildung (schul3)	Fallzahl 1976 (mit Bildung) (n_break76)
Mann	1956-1960	Arbeiter	Primärsektor	160	Volksschule	139
Mann	1956-1960	Arbeiter	Primärsektor	.	Realschule	9
Mann	1956-1960	Arbeiter	Primärsektor	.	Abitur	3
Mann	1956-1960	Arbeiter	Sekundärsektor	2084	Volksschule	1862
Mann	1956-1960	Arbeiter	Sekundärsektor	.	Realschule	124
Mann	1956-1960	Arbeiter	Sekundärsektor	.	Abitur	19

4) Da geprüft werden soll, ob eine geringe Fallzahl der Gruppen (ohne Bildung) Einfluss auf die Schätzleistung bzw. die Ergebnisse der Validitätstests hat, wird durch Aggregation eine Variable gebildet, die die Fallzahl der gültigen Fälle jeder Gruppe ohne Schulbildung auf Basis der Zellen mit Information über Schulbildung darstellt. Ein Ausschnitt aus der Arbeitsdatei wird in Tabelle 5 dargestellt. Es werden einzelne Modelle mit Cold-Deck-Verfahren berechnet, und zwar jeweils ohne Fallzahlbegrenzung, mit Fallzahlbegrenzung für Fallzahl pro Zelle

kleiner/gleich 50, 100 und 150. Die Fallzahlbegrenzung geschieht durch die Selektion von Fällen nach der Häufigkeit der Variable „N_BREAK76_OB“. Berichtet wird ein Beispiel ohne Fallzahlbegrenzung.

```
WEIGHT BY N_BREAK76.
AGGREGATE
  /OUTFILE=*
  MODE=ADDVARIABLES
  /BREAK=gesch kohort sek stib_1
  /N_BREAK76_OB=N.
weight off.
select if N_BREAK76_OB>0.
```

Tabelle 5: Schematische Darstellung zum Cold-Deck-Verfahren, Ergebnis von Arbeitsschritt 4

Geschlecht (gesch)	Kohorte (kohort)	Stellung im Beruf (stib_1)	Wirtschaftssektor (sek)	Fallzahl 1976 (ohne Bildung) (n_break76_val)	Bildung (schul3)	Fallzahl 1976 (mit Bildung) (n_break76)	Fallzahl für Beschränkung (n_break76_ob)
Mann	1956-1960	Arbeiter	Primärsek.	160	Volksschule	139	151
Mann	1956-1960	Arbeiter	Primärsek.	.	Realschule	9	151
Mann	1956-1960	Arbeiter	Primärsek.	.	Abitur	3	151
Mann	1956-1960	Arbeiter	Sekundärsek.	2084	Volksschule	1862	2005
Mann	1956-1960	Arbeiter	Sekundärsek.	.	Realschule	124	2005
Mann	1956-1960	Arbeiter	Sekundärsek.	.	Abitur	19	2005

5) Weiter werden Anteile der Schulabschlüsse innerhalb der einzelnen Gruppen ermittelt. Es wird durch Aggregation eine neue Datei erzeugt, die für die Empfängerdatei in jeder Zelle Anteile der Schulabschlüsse enthält, die exakt mit der Randverteilung der Schulabschlüsse innerhalb einzelner Zellen der Geberdatei übereinstimmen.

Das Ergebnis der Aggregation wird in Tabelle 6 berichtet. Nach der Aggregation werden die Variablen mit den absoluten Fallzahlen der Schulabschlüsse aus den Anteilen berechnet, diese Variablen stehen ganz rechts in Tabelle 6.

```
weight by N_BREAK76.
AGGREGATE
  /OUTFILE=*
  MODE=replace
  /BREAK=gesch kohort stib_1 sek
  /N_BREAK76_val=first(N_BREAK76_val)
  /schul3_1 = FIN(schul3 0.9 1.1)
  /schul3_2 = FIN(schul3 1.9 2.1)
  /schul3_3 = FIN(schul3 2.9 3.1).
compute n_schul3_1=n_break76_val*schul3_1.
compute n_schul3_2=n_break76_val*schul3_2.
compute n_schul3_3=n_break76_val*schul3_3.
compute n_schul3_1=rnd(n_schul3_1).
compute n_schul3_2=rnd(n_schul3_2).
compute n_schul3_3=rnd(n_schul3_3).
```

Tabelle 6: Schematische Darstellung zum Cold-Deck-Verfahren, Ergebnis von Arbeitsschritt 5

Geschlecht (gesch)	Kohorte (kohort)	Stellung im Beruf (stib_1)	Wirtschaftssektor (sek)	Fallzahl 1976 (ohne Bildung) (n_break 76_val)	Hauptschule (schul 3_1)	Realschule (schul 3_2)	Abitur (schul 3_3)	N Hauptschule (n_schul3_1)	N Realschule (n_schul3_2)	N Abitur (n_schul 3_3)
Mann	1956-60	Arbeiter	Primärsek.	160	0,921	0,060	0,020	147	10	3
Mann	1956-60	Arbeiter	Sekundärsek	2084	0,929	0,062	0,009	1935	129	20

6) Auf Basis der ermittelten Fallzahlen werden drei Hilfsdateien erzeugt, eine für jede Art des Schulabschlusses. Jede dieser drei Hilfsdateien enthält die geschätzte Fallzahl des jeweiligen Schulabschlusses in jeder der definierten Zellen, und zwar unter dem Variablennamen „gewicht“.

```
do if n_schul3_1 ne 0.
compute val_schul3=1.
compute gewicht=n_schul3_1.
xsave  outfile="PFAD\Bildung1.sav"
 /keep=gesch kohort stib_1 sek val_schul3 gewicht.
end if.
do if n_schul3_2 ne 0.
compute val_schul3=2.
compute gewicht=n_schul3_2.
xsave  outfile="PFAD\Bildung2.sav"
 /keep=gesch kohort stib_1 sek val_schul3 gewicht.
end if.
do if n_schul3_3 ne 0.
compute val_schul3=3.
compute gewicht=n_schul3_3.
xsave  outfile="PFAD\Bildung3.sav"
 /keep=gesch kohort stib_1 sek val_schul3 gewicht.
end if.
execute.
```

7) Im nächsten Schritt werden diese Hilfsdateien vertikal untereinander geschrieben. Es entsteht eine Matrix mit der gleichen Zellenanzahl wie in der ursprünglichen aggregierten Geberdatei (1386 Zellen). Die Fälle werden nicht nach Bildung, sondern wie in der ursprünglichen Datei nach Geschlecht, Kohorte, Stellung im Beruf und Wirtschaftssektor sortiert. Ein Ausschnitt dieser Datei zeigt Tabelle 7.

```
add files
  file="PFAD\Bildung1.sav"
 /file="PFAD\Bildung2.sav"
 /file="PFAD\Bildung3.sav"
execute.
sort cases by gesch kohort stib_1 sek val_schul3.
```

Tabelle 7: Schematische Darstellung zum Cold-Deck-Verfahren, Arbeitsschritt 7

Geschlecht (gesch)	Kohorte (kohort)	Stellung im Beruf (stib_1)	Wirtschafts-sektor (sek)	Bildung (val_schul3)	Fallzahl (gewicht)
Mann	1956-1960	Arbeiter	Primärsek.	Volksschule	147
Mann	1956-1960	Arbeiter	Primärsek.	Realschule	10
Mann	1956-1960	Arbeiter	Primärsek.	Abitur	3
Mann	1956-1960	Arbeiter	Sekundärsek.	Volksschule	1935
Mann	1956-1960	Arbeiter	Sekundärsek.	Realschule	129
Mann	1956-1960	Arbeiter	Sekundärsek.	Abitur	20

8) Um aus der aggregierten Datei mit 1386 Zellen wieder eine Individualdatei zu erzeugen, werden für jede Kombination aus Geschlecht, Kohorte, Stellung im Beruf, Wirtschaftssektor und dem geschätzten Schulbildungsmerkmal so viele Fälle herausgeschrieben, wie in der ermittelten und als Fallzahl (gewicht) bezeichneten Variable für jeden Schulabschluss für jede Zelle angegeben. So entsteht eine synthetische Individualdatei mit den Variablen Geschlecht, Kohorte, Stellung im Beruf, Wirtschaftssektor und dem geschätzten Schulbildungsmerkmal. Die Fallzahl entspricht den Fällen mit gültigen Angaben in der ursprünglichen Empfängerdatei, dem Mikrozensus 1976 ohne das Merkmal Schulbildung.

```
loop x=1 to gewicht.[6]
xsave outfile="PFAD\synthetische_Individualdatei_mz76_val.sav".
end loop.
```

9) Ziel des letzten Schritts ist das Anhängen des geschätzten Bildungsmerkmals an die ursprüngliche Empfängerdatei, in diesem Beispiel den Mikrozensus 1976 (ohne das Merkmal Schulbildung). Dafür sind zunächst zwei Zwischenschritte nötig, um beide Dateien für das Zusammenfügen vorzubereiten. In der synthetischen Individualdatei wird eine neue Variable erzeugt, die für jeden Fall die Ausprägung 1 hat. Diese steht für vollständige Angaben bei den Variablen Geschlecht, Kohorte, Stellung im Beruf, Wirtschaftssektor und dem geschätzten Schulbildungsmerkmal. Weiter wird die synthetische Datei nach den Merkmalen Geschlecht, Kohorte, Stellung im Beruf, Wirtschaftssektor und dem geschätzten Schulbildungsmerkmal sortiert.

Auch in der Empfänger-Individualdatei wird die Variable erzeugt, die für vollständige Fälle steht, die mit denen in der synthetischen Datei übereinstimmen können. Fälle, die gültige Werte bei den Variablen Geschlecht, Kohorte, Stellung im Beruf, Wirtschaftssektor und Schulbildung haben, bekommen bei dieser Variable die Ausprägung 1. Die Datei wird nach der Variable der Vollständigkeit und dann nach Geschlecht, Kohorte, Stellung im Beruf und

[6] Bei SPSS 15 wird beim xsave-Befehl die irreführende Warnung ausgegeben, dass loop keinen Effekt auf diesen Befehl habe. Es ist aber das Gegenteil der Fall, loop hat den gewünschten Effekt. SPSS 14 gibt an dieser Stelle keine Warnung aus.

Wirtschaftssektor sortiert. Dann können beide Dateien anhand der Schlüsselvariablen Vollständigkeit, Geschlecht, Kohorte, Stellung im Beruf, Wirtschaftssektor und Schulbildung zusammengefügt werden.

```
get file="PFAD\synthetische_Individualdatei_mz76_val.sav".
compute vollst=1.
sort cases by vollst gesch kohort stib_1 sek schul3.
save outfile="PFAD\synthetische_Individualdatei_mz76_val.sav".

get file="PFAD\MZ1976_val.sav".
compute vollst=0.
if (erw_1 ne 3) and range(stib_1,1,7) and range(sek,1,3) vollst=1.
sort cases by vollst gesch kohort stib_1 sek.
save outfile="PFAD\MZ1976_val.sav".

match files
 /file="PFAD\synthetische_Individualdatei_mz63.sav".
 /file="PFAD\MZ1976_val.sav".
 /by vollst gesch kohort stib_1 sek.
```

10) Wegen Rundungen bei der Ermittlung konkreter Fallzahlen aus Anteilen im Schritt 5) sind in Ausnahmefällen (18 Fälle) wenige Fälle mehr synthetisch erzeugt worden, als tatsächlich Fälle in den jeweiligen Zellen der Originaldatei enthalten sind. Diese Bildungsinformationen können dementsprechend keinem Originalfall zugeordnet werden. Um diese durch Rundungsdifferenzen erzeugten Fälle aus der Datei zu entfernen, werden nur diejenigen Fälle ausgewählt, die in allen Originaldaten einen gültigen Wert aufweisen. Da die Variable Familienstand bei allen Fällen der Originaldatei, aber nur bei diesen, gültig ist, werden lediglich Fälle mit gültiger Angabe des Familienstands ausgewählt.

Für jeden Fall mit gültigen Werten in den oben genannten Schlüssel- bzw. Gruppenvariablen hat das neue Merkmal „Schulbildung“ jetzt einen gültigen Wert. Die Randverteilung entspricht der anhand der in der Geberdatei zurückgerechneten Randverteilung, die im Mikrozensus 1976 ermittelt worden wäre, wenn Schulbildung abgefragt worden wäre, und stimmt so im Validierungsbeispiel überein.

```
select if not sysmis(famst).
save outfile="PFAD\synthese1976val.sav"
 /drop gewicht eins x vollst.
```

Ein Problem bei der Schätzung besteht in der durch die Geburtsjahrgänge bedingten Selektivität. Während in den jüngsten betrachteten Jahrgängen vor allem die Personen mit dem Volksschulabschluss bereits das Bildungssystem verlassen haben und erwerbstätig sind, sind dies in den älteren Jahrgängen besonders die höher gebildeten Personen in den gehobenen Berufspo-

sitionen. Dies kann, wenn die Daten auf ältere Mikrozensen übertragen werden, zu einer Unterschätzung der Bildung in den jüngeren und zu einer Überschätzung in den älteren Jahrgängen führen.

2.3 Entwicklung einer Schätzmethode durch loglineare Modelle

Alternativ zu der Schätzung durch Cold-Deck-Verfahren können auch loglineare Modelle zur Schätzung von Randverteilungen herangezogen werden (Little & Rubin 2002: Abschnitt 13.4). Für das saturierte loglineare Modell wird die Matrix aus der Kombination der unabhängigen Variablen in gleicher Weise wie beim Cold-Deck-Verfahren erzeugt, die Bildungsverteilung wird analog geschätzt.

Eine modellbasierte Alternative zum saturierten Modell ist die Anwendung eines nichtsaturierten Verfahrens. Dabei werden die zu berücksichtigenden Parameter in gleicher Weise wie bei der logistischen Regressionsanalyse spezifiziert. Nach der Anpassung der Schulbildung an die Randverteilungen von Geschlecht (gesch), Geburtskohorte (ermittelt aus dem Geburtsjahr gebj), Wirtschaftssektor (sek) und Stellung im Beruf (stib_1) werden weiterhin die Interaktionen dieser Variablen mit der Schulbildung betrachtet, darüber hinaus die (Dreiweg-) Interaktion von Schulbildung, Geschlecht und Kohortenzugehörigkeit sowie die Interaktion von Schulbildung, Wirtschaftssektor und Stellung im Beruf.[7]

Auch dieses Verfahren besteht aus mehreren Schritten und gleicht in den Grundzügen der oben beschriebenen Schätzmethode durch das Cold-Deck-Verfahren, bei der die aus dem Mikrozensus 1976 ermittelten Bildungsinformationen auf Individualdaten der Mikrozensen der 60er Jahre übertragen werden können.

1) Zunächst werden die Individualdaten der Geber- und der Empfängerdatei zu zwei Arbeitsdateien aggregiert, und zwar ebenso wie bei der Schätzmethode nach Cold-Deck-Verfahren nach den Merkmalen Geschlecht, Kohorte, Stellung im Beruf und Wirtschaftssektor, in der Geberdatei zusätzlich nach Schulabschluss. Dies entspricht den Schritten 1) und 2) in der Beschreibung der Cold-Deck-Verfahren.

2) Im nächsten Schritt unterscheiden sich die loglinearen von den Cold-Deck-Modellen. Zunächst wird wiederum eine Matrix mit theoretisch möglichen 1386 Zellen aufgebaut, die sich aus der Kombination von Geschlecht, Geburtskohorten, Stellung im Beruf, Wirtschaftssektor und Schulbildung (aus der Geberdatei) ergeben. Die Informationen aus der

[7] Die Berechnungen wurden mit der Prozedur GENLOG im SPSS durchgeführt.

Empfängerdatei mit theoretisch möglichen 462 Zellen ohne Schulbildung werden aber hier drei mal als jeweils eine Variable für jeden der drei möglichen Bildungsabschlüsse hinzugefügt. Die noch zu schätzenden Zellenhäufigkeiten für Schulabschlüsse in der Empfängerdatei bleiben zunächst offen.

```
add files
  files=PFAD\Geberdatei.sav' /in=quelle
 /files='PFAD\Empfaengerdatei.sav' /in=bildung1
 /files='PFAD\Empfaengerdatei.sav' /in=bildung2
 /files='PFAD\Empfaengerdatei.sav' /in=bildung3.
if bildung1=1 valschul3=1.
if bildung2=1 valschul3=2.
if bildung3=1 valschul3=3.
```

Tabelle 8: Schematische Darstellung zum Schätzverfahren mit loglinearen Modellen, Ergebnis von Arbeitsschritt 2

Geschlecht (gesch)	Kohorte (kohort)	Stellung im Beruf (stib_1)	Wirtschaftssektor (sek)	Bildung 1976 / Geberdatei (valschul3)	Fallzahl 1976 (mit Bildung) (n_break76)	Fallzahl 1976 (ohne Bildung) (n_break 76_val)	Hilfs-Var. (quelle)	Hilfsvar. (bildung1)	Hilfsvar. (bildung2)	Hilfsvar. (bildung3)
Mann	1956-60	Arbeiter	Prim.sek.	1	139	.	1	0	0	0
Mann	1956-60	Arbeiter	Prim.sek	2	9	.	1	0	0	0
Mann	1956-60	Arbeiter	Prim.sek	3	3	.	1	0	0	0
Mann	1956-60	Arbeiter	Sek..sek	1	1862	.	1	0	0	0
Mann	1956-60	Arbeiter	Sek..sek	2	124	.	1	0	0	0
Mann	1956-60	Arbeiter	Sek..sek	3	19	.	1	0	0	0
[…]										
Mann	1956-60	Arbeiter	Prim.sek	1	.	160	0	1	0	0
Mann	1956-60	Arbeiter	Sek..sek	1	.	2084	0	1	0	0
[…]										
Mann	1956-60	Arbeiter	Prim.sek	2	.	160	0	0	1	0
Mann	1956-60	Arbeiter	Sek..sek	2	.	2084	0	0	1	0
[…]										
Mann	1956-60	Arbeiter	Prim.sek	3	.	160	0	0	0	1
Mann	1956-60	Arbeiter	Sek..sek	3	.	2084	0	0	0	1

3) Im nächsten Schritt wird ein loglineares Modell berechnet, das auf Basis der Bildungsverteilung des 76er Mikrozensus (Geberdatei) die Randverteilung der Bildungsvariable in der Empfängerdatei schätzt, und zwar unter Berücksichtigung von Kohorte, Stellung im Beruf und Wirtschaftssektor. Neben den Haupteffekten werden auch Wechselwirkungen berücksichtigt, und zwar zwischen der Bildungsvariablen und allen Merkmalen, zu dem aber auch zwischen Geschlecht und Kohorte, Stellung im Beruf und Wirtschaftssektor, Geschlecht und Kohorte und Bildung sowie zwischen Stellung im Beruf, Wirtschaftssektor und Bildung. Das loglineare Modell nimmt die Fälle mit einer Gewichtungsvariable>0 als Berechnungsbasis für

die Schätzung. Die Schätzung selber wird auch für die Fälle der Empfängerdatei durchgeführt, diese können daran erkannt werden, dass die Gewichtungsvariable („gewicht“, berechnet aus n_break76) zunächst auf 0 gesetzt wurde.

Die Berechnung wird für zwei verschiedene Versionen durchgeführt: In der ersten Version (a) wird ein nichtsaturiertes Modell berechnet, das die Haupteffekte und Interaktionen in gleicher Weise wie bei der logistischen Regressionsanalyse (vgl. Abschnitt 2.1) spezifiziert. In der zweiten Version (b) wird ein saturiertes Modell berechnet.[8]

```
compute gewicht=n_break76.
recode gewicht(sysmis=0).
weight by gewicht.

*Version a) modellbasiert, nichtsaturiertes Modell.
GENLOG
  mann rec_kohort rec_stib_1 sek valschul3
  /MODEL = MULTINOMIAL
  /PRINT = FREQ RESID
  /PLOT = none
  /CRITERIA = CIN(95) ITERATE(100) CONVERGE(.001) DELTA(.5)
  /DESIGN valschul3 mann rec_kohort sek rec_stib_1 valschul3*mann val-
schul3*rec_kohort valschul3*sek
          valschul3*rec_stib_1 mann*rec_kohort sek*rec_stib_1 val-
schul3*mann*rec_kohort valschul3*sek*rec_stib_1
  /SAVE=PRED.

*Version b) saturiertes Modell.
GENLOG
  valschul3 mann rec_kohort rec_stib_1 sek
  /MODEL = POISSON
  /PRINT = FREQ RESID
  /Plot = NONE
  /CRITERIA = CIN(95) ITERATE(1000000) CONVERGE(.01) DELTA(.5)
  /DESIGN
  /SAVE=PRED.
```

5) Da in den loglinearen Modellen die Zellenhäufigkeit in der Tabellendatei geschätzt wird, stimmt die geschätzte Bildungsverteilung auf Aggregatebene perfekt mit der tatsächlichen Bildungsverteilung überein. Um aus der Tabellendatei wieder eine Individualdatei zu erstellen, müssen die geschätzten Zellenhäufigkeiten allerdings gerundet werden. In diesem Schritt treten Probleme der modellbasierten Version (a) auf: Da 8,6% der geschätzten Zellengewichte unter 0,5 liegen und somit auf 0 gerundet werden, wird in diesem Schritt die Fallzahl extrem minimiert. Da kleine Zellenhäufigkeiten vor allem bei den seltenen hohen Bildungsabschlüssen geschätzt werden, werden vor allem bei diesen die Zellenhäufigkeiten von kleiner 0,5 auf

[8] Beim saturierten Modell musste – aus offensichtlich numerischen Gründen, die in der Logik des GENLOG-Programms von SPSS liegen dürften – ein weniger scharfes Konvergenzkriterium gewählt werden, um Konvergenz zu erreichen. Uns ist unklar, aus welchen numerischen Gründen eine Konvergenz beim voreingestellten Konvergenzkriterium von 0,001 nicht erreicht wird.

0 abgerundet, was in diesem Modell zu einer Unterschätzung der höheren Bildungsabschlüsse führt, da diese oft Inhalt kleiner Zellen sind.

```
compute rndpre_1=rnd(PRE_1).
select if rndpre_1>0 and gewicht=0.
```

6) Analog zu der oben beschriebenen Schätzung durch das Cold-Deck-Verfahren wird aus der Tabellendatei im nächsten Schritt wieder eine Individualdatei erzeugt. Anhand der Schlüsselvariablen Geschlecht, Geburtskohorte, Stellung im Beruf und Wirtschaftssektor werden die Fälle mit Bildungsinformation anschließend mit dem zuvor randomisierten Ursprungsdaten zusammengefügt. Für jeden Fall mit gültigen Werten in den oben genannten Schlüssel- bzw. Gruppenvariablen hat das neue Merkmal „Schulbildung“ jetzt einen gültigen Wert. Dieser Arbeitsschritt entspricht den Schritten 8) bis 10) des Cold-Deck-Verfahrens. Die Randverteilung des saturierten Modells entspricht im vorliegenden Test der tatsächlichen Randverteilung der Schulbildung im Mikrozensus 1976.

```
loop x=1 to rndpre_1.
xsave outfile='PFAD\synthetische_Individualdatei_mz76_val.sav'.
end loop.

get file='PFAD\synthetische_Individualdatei_mz76_val.sav'.
compute vollst=1.
sort cases by vollst mann kohort stib_1 sek valschul3.
save outfile='PFAD\synthetische_Individualdatei_mz76_val.sav'.

get file='PFAD\MZ76_val.sav'.
compute vollst=0.
if (erw_1 ne 3) and range(stib_1,1,7) and range(sek,1,3) vollst=1.
sort cases by vollst mann kohort stib_1 sek.
save outfile='PFAD\MZ76_val.sav'.

match files
 /file='PFAD\synthetische_Individualdatei_mz76_val.sav'
 /file='PFAD\MZ76_val.sav'
 /by vollst mann kohort stib_1 sek.

select if not sysmis(famst).
```

3. Ergebnis der Validitätsprüfungen mit dem Mikrozensus 1976

Zur Prüfung der Validität der Zuweisung wurden die Verteilungen der geschätzten Bildungsvariablen mit denen der tatsächlichen Bildungsvariablen verglichen. Rässler (2002:30f) gibt eine Reihe von Qualtitätskriterien für synthetische Variablen an. Zunächst kann die Reproduktion der ursprünglichen Randverteilungen untersucht werden. Tabelle 9 gibt die Randverteilungen der geschätzten und der tatsächlichen Bildungsverteilung für jedes der Modelle an, und zwar die des Cold-Deck-Verfahrens ohne Fallzahlbegrenzung und mit Fallzahlbegrenzung von 50, 100 und 150 sowie diejenige des loglinearen nichtsaturierten und des saturierten loglinearen Modells.

Tabelle 9: Ergebnisse der Validitätsprüfung, MZ 1976

	MZ 76, original	ohne Fallzahl-begr.	Fallzahl-begr. =50	Fallzahl-begr. =100	Fallzahl-begr. =150	Log-lineares Modell	Log-lineares Modell, saturiert
Bildungsverteilung							
Fallzahl	168.606	171.497	169.378	166.769	164.979	129.123	166.602
Abi	9,5	9,4	9,4	9,4	9,5	6,8	9,5
Realschule, Fachhochs.	18,7	18,7	18,7	18,7	18,7	15,3	18,8
Volks-/Hauptschule	71,8	71,8	71,9	71,8	71,9	77,9	71,7
Dissimilaritätsindex		0,1	0,1	0,1	0,1	6,1	0,1
Validitätstests							
Fallzahl		165.929	163.889	161.377	159.666	124.919	161.210
Gamma		.450	.449	.448	.449	.416	.451
Kappa		.191	.190	.190	.190	.164	.193
Cox/Snell		.060	.060	.060	.060	.043	.061
Nagelkerke		.076	.076	.076	.076	.055	.077
McFadden		.040	.040	.040	.040	.028	.041

Die Randverteilung ist beim Cold-Deck-Verfahren und beim saturierten loglinaren Modell per definitionem gewährleistet. Beim nichtsaturierten loglinearen Modell ergeben sich Abweichungen aufgrund der großen Zahl geschätzter Zellen mit Besetzungen unterhalb von 0,5 (vergleiche Abschnitt 2.3). Diese mit niedriger Besetzung geschätzten Zellen konzentrieren sich auf die selteneren höheren Bildungsabschlüsse. Daher ist bei Verwendung dieses Verfahrens einerseits die Fallzahl geringer als bei allen anderen Verfahren, andererseits wird die Bildung zulasten der selteneren (höheren) Bildungsabschlüsse verzerrt.

Auch bei den Cold-Deck-Verfahren nimmt die Fallzahl mit zunehmender Fallzahlbegrenzung ab, allerdings nicht in dem gleichen Maße wie bei dem nichtsaturierten loglinearen

Modell. Durch die variable Fallzahlbegrenzung wurde bei den Cold-Deck-Verfahren geprüft, ob durch extrem kleine Zellen Zufallsfehler produziert werden. Die Vorhersagegüte konnte allerdings durch die unterschiedlichen Fallzahlbegrenzungen nicht verbessert werden. Da zudem für die Schätzung der Bildungsinformationen in den Mikrozensen der 60er Jahre nicht der Mikrozensus 1976, sondern die 10%-Stichprobe der Volkszählung 1970 als Quelldatei herangezogen wird, spielt durch die viel höheren Fallzahlen eine Fallzahlbegrenzung keine Rolle mehr.

Während die Reproduktion der Randverteilungen problemlos gelingt, erweist sich die Reproduktion der individuellen Bildungsabschlüsse als schwieriger. In den Tabellen 10 bis 12 werden die vorhergesagten Werte der Schulbildung gegen die tatsächliche Schulbildung der Befragten tabuliert. Die jeweils unten aufgeführten Zeilenprozente stehen für die Randverteilung der tatsächlichen Schulbildung. Die in der letzten Spalte aufgeführten Spaltenprozente stehen für die Randverteilung der vorhergesagten Schulbildung. Die Basis dieser Spaltenprozentsätze sind abweichend von Tabelle 9 (oberer Teil) nur diejenigen Befragten mit gültiger Angabe der Schulbildung, während der Tabelle 9 auch Befragte zu Grunde liegen, die zwar keine gültige Angabe bei der Schulbildung machten, bei denen aber die Schulbildung geschätzt werden konnte.

Tabelle 10: Vorhergesagte Schulbildung nach tatsächlicher Schulbildung, MZ 76, Cold-Deck-Verfahren, Fallzahlbegrenzung=0, Spaltenprozente

	tatsächliche Schulbildung			
vorhergesagte Schulbildung	Volksschule	Realschule	Abitur	Gesamt
Volksschule	**78,6**	56,3	50,3	71,8
Realschule	14,8	**30,3**	26,3	18,8
Abitur	6,6	13,4	**23,4**	9,5
Gesamt (Zeilenprozente)	71,7	18,8	9,5	100

n= 165.929

Tabelle 10 gibt zunächst die Ergebnisse für das Cold-Deck-Modell ohne Fallzahlbegrenzung wieder. Für 71,8% der Befragten wurde der Volksschulabschluss vorhergesagt, für 18,8% der Realschulabschluss und für 9,5% das Abitur. Bei Personen, die tatsächlich nur einen Volksschulabschluss hatten, wurden dagegen 78,6% richtige Vorhersagen gemacht. Der Anteil korrekter Vorhersagen bei den (selteneren) höheren Bildungskategorien liegt mit 30,3% bei der Realschule und 23,4% beim Abitur zwar deutlich niedriger, aber immer noch weit über den oben berichteten unbedingten Randverteilungswerten von 18,8 bzw. 9,5 Prozent. Damit wird durch unser Schätzverfahren eine erhebliche Verbesserung bei der Schätzung der Schulbil-

dung erreicht, Tabelle 9 (unterer Teil) gibt eine üblicher Koeffizienten für die Vorhersagegüte an. Besonders die Kappa-Werte sind niedrig, wobei Kappa als Maß der Übereinstimmung für diesen Zweck wegen der Randverteilungsabhängigkeit schlecht geeignet ist. Auch die durch logistische Regression der geschätzten auf die tatsächlichen Bildungsmerkmale ermittelten Pseudo-R-Quadrat-Werte fallen aufgrund der schiefen Randverteilung niedrig aus. Von uns wird die Güte der Anpassung daher auch mit Gamma, einem Zusammenhangsmaß für ordinale Merkmale, beurteilt.

Tabelle 11: Vorhergesagte Schulbildung nach tatsächlicher Schulbildung, MZ 76, Cold-Deck-Verfahren, Fallzahlbegrenzung=100, Spaltenprozente

	tatsächliche Schulbildung			
vorhergesagte Schulbildung	Volksschule	Realschule	Abitur	Gesamt
Volksschule	**78,6**	56,3	50,5	71,8
Realschule	14,8	**30,2**	26,3	18,8
Abitur	56,6	13,5	**23,2**	9,5
Gesamt (Zeilenprozente)	71,8	18,8	9,5	100

n= 161.377

Die Tabellen 11 und 12 geben die Ergebnisse alternativer Verfahren: nämlich eines Cold-Deck-Verfahrens mit Fallzahlrestriktion (Tabelle 11) und eines nichtsaturieren loglinearen Modells (Tabelle 12) wieder. Wie Spalten 4 und 6 in Tabelle 9 zeigen, ergeben sich hierbei keine nennenswerten Verbesserungen der Qualität der Vorhersage.

Tabelle 12: Vorhergesagte Schulbildung nach tatsächlicher Schulbildung, MZ 76, loglineares Modell, Spaltenprozente

	tatsächliche Schulbildung			
vorhergesagte Schulbildung	Volksschule	Realschule	Abitur	Gesamt
Volksschule	**83,4**	64,5	62,6	77,8
Realschule	11,7	**25,8**	22,7	15,3
Abitur	5,0	9,7	**14,7**	6,8
Gesamt (Zeilenprozente)	71,8	18,1	10,2	100

n=124.919

Ein weiteres Qualitätskriterium ist die Korrelation der geschätzten Werte der Schulbildung mit externen Kriteriumsvariablen. Dabei bietet sich das persönliche Einkommen der Befragten an, da eine Korrelation von Schulbildung und Einkommen erstens theoretisch anzunehmen ist und zweitens empirisch gut belegt ist.[9] Die Korrelation zwischen geschätzer

[9] Basis für diese Berechnungen ist das persönliche Einkommen (einkp_1, das zuvor im Rahmen der Kumulation harmonisiert wurde).

Schulbildung und und Einkommen kann mit der Korrelation der ursprünglichen Werte (der Schulbildung) mit dem Einkommen verglichen werden.

Tabelle 13: Validitätstests, Einkommen Erwerbstätiger nach Schulbildung

	MZ 76, original	ohne Fallzahl-begr.	Fallzahl-begr. =50	Fallzahl-begr. =100	Fallzahl-begr. =150	Log-lineares Modell	Log-lineares Modell, saturiert
Einkommen Erwerbstätiger nach Schulbildung (in DM)							
Fallzahl	168.606	171.497	169.378	166.769	164.979	129.123	166.602
Abi	1102	868	870	872	876	915	870
Realschule, Fachhochs.	735	694	697	698	704	733	691
Volksschule	548	591	594	599	603	642	592
Summe quadrierter Abweichung		58.286	57.384	56.870	55.062	43.809	57.696
Summe absoluter Abweichungsbeträge		318	316	318	312	283	320

Tabelle 13 gibt die mittleren Werte des persönlichen Nettoeinkommens für Befragte unterschiedlicher Schulbildung an. Bei den Personen mit tatsächlich vorliegender Bildungsangabe verfügen diejenigen mit Abitur im arithmetischen Mittel 1976 über DM 1102, diejenigen mit mittlerer Bildung über DM 735 und diejenigen mit Volksschulabschluss über DM 548. Berechnet man die mittleren Einkommen getrennt nach den geschätzten Bildungsabschlüssen, so erhält man einen statistisch durch Zufallsfehler bei der Schätzung der Bildungsabschlüsse geminderten (attenuierten) Zusammenhang. Nun werden – beim Cold-Deck-Verfahren ohne Fallzahlbegrenzung - für diejenigen mit Abitur lediglich DM 868 geschätzt, für diejenigen mit mittleren Abschlüssen DM 694 und für diejenigen mit lediglich einem Volksschulabschluss DM 591.

In einem weiteren Schritt betrachten wir die Abweichung der hinsichtlich der geschätzten Schulbildung bedingten Einkommensverteilung von der hinsichtlich der tatsächlichen Schulbildung bedingten Einkommensverteilung. Dabei wird die Distanz zwischen den jeweiligen Spalten in Tabelle 13 als Summe der quadrierten Abweichungen und alternativ als Summe der absoluten Abweichungsbeträge ausgedrückt.

Auf Basis der Validitätstests erscheint uns die Nutzung des Cold-Deck-Verfahrens ohne Fallzahlbegrenzung vertretbar.

4. Anwendung des Verfahrens auf die Mikrozensen 1992 bis 1969 und 1973

Das Cold-Deck-Verfahren ohne Fallzahlbegrenzung wurde im Anschluss an die Validitätsprüfungen zum eigentlichen Ziel, der Schätzung von Bildungsinformationen in den Mikrozensen 1962 bis 1973, eingesetzt. Tabelle 14 gibt als Ergebnis der Schätzverfahren die Zahl der Erwerbspersonen in den jeweiligen Mikrozensen an, für die ein Schulabschluss geschätzt wurde, sowie die geschätzte Randverteilung der Schulabschlüsse in den jeweiligen Jahren. Zum Vergleich dazu zeigt Tabelle 15 die tatsächlichen Schulabschlüsse bei Erwerbspersonen mit gültigen Angaben bei den zur Schätzung herangezogenen Variablen Geschlecht, Kohorte, Stellung im Beruf und Wirtschaftssektor sowie Schulbildung.

Tabelle 14: Ergebnisse des Schätzverfahrens, Schulabschlüsse in den Mikrozensen 1962-1973

Jahr	Erwerbspersonen mit geschätztem Schulabschluss	Volks-/Hauptschulabschluss	Realschulabschluss/ Fachhochschulreife	Abitur
1962	244.982	83,2	11,2	5,6
1963	249.838	82,6	11,8	5,6
1964	258.808	81,7	12,5	5,8
1965	262.190	81,4	12,9	5,8
1966	263.506	81,1	13,1	5,8
1967	256.746	81,2	13,1	5,7
1968	258.209	81,6	12,8	5,5
1969	260.702	82,0	12,6	5,3
1973	245.416	81,0	13,6	5,4

Tabelle 15: Tatsächliche Schulabschlüsse in der Volkszählung 1970 und im Mikrozensus 1976

Jahr	Erwerbspersonen mit Angabe zum Schulabschluss	Volks-/Hauptschulabschluss	Realschulabschluss/ Fachhochschulreife	Abitur
VZ70 (10%)	2.665.543	81,4	13,3	5,4
MZ 76	168.606	71,8	18,7	9,5

Die geschätzten Randverteilungen erscheinen für die 60er Jahre plausibel. Vermutlich ergibt sich für 1973 in der Tendenz eine Unterschätzung höherer Abschlüsse. Jedoch ist zu beachten, dass sich die berichten Anteile jeweils auf die gesamte Zahl der Erwerbspersonen beziehen und sich damit sich in Folge der Bildungsexpansion nur sehr langsam ändern.

5. Ausblick und Perspektiven

Ziel der berichteten Analysen war die Vervollständigung der Mikrozensus-Daten der 60er Jahre hinsichtlich der allgemeinen Schulbildungsabschlüsse. Cold-Deck-Verfahren und loglineare Schätzungen lieferten auf Basis des Mikrozensus 1976 und der Volkszählung 1970 ähnliche Ergebnisse, aufgrund der größeren Zeitnähe und der höheren Fallzahl wurde schließlich die 10%-Stichprobe der Volkszählung 1970 als Basis der Zuweisung ausgewählt. Weiterhin ergaben sich die plausibelsten und umfassendsten Resultate beim herkömmlichen Cold-Deck-Verfahren, das deshalb für die Zuweisung angewandt wurde.

Die Validität des Verfahrens wurde anhand des Kriteriums der tatsächlichen Bildungsangaben im Mikrozensus 1976 überprüft, dabei ergab sich eine mäßige Vorhersageleistung. Weiterhin erwiesen sich die geschätzten Bildungsabschlüsse als konsistent mit den zu erwartenden Einkommensstrukturen.

Für Nichterwerbspersonen wurde keine Schätzung der Bildungsabschlüsse vorgenommen. Aufgrund der Haushaltsstrukturen im Deutschland der 60er Jahre erscheint es plausibel, die im nächsten Schritt anstehenden Analysen auf Haushaltsebene auf Basis des „Hauhaltsvorstands“ zu differenzieren, und so auf dessen Bildung zu rekurrieren. Dieser dürfte aber per definitionem bei Nichtrentnerhaushalten in den 60er Jahren in der Regel erwerbstätig gewesen sein. Damit steht auf Haushaltsebene, wenn auch nicht auf Personenebene, zumindest ein Bildungsmerkmal zur Verfügung.

Weitere Analysen zu Vervollständigung historisch fehlender, da nicht erhobener Daten wären in der Zukunft möglich. Dabei könnten neben loglinearen Schätzverfahren und Cold-Deck-Modellen auch Datenfusionsverfahren angewandt werden.

Literatur

Rässler, Susanne (2002): Statistical Matching – A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches. New York: Springer.

Little, Roderick J.A. und Donald B. Rubin (2002): Statistical Analysis with Missing Data. New Jersey: Wiley.

Lengerer, Andrea, Julia H. Schroedter und Tobias Hubert (2007): Harmonisierung der Mikrozensen 1962 bis 2004. ZUMA- Projektbericht Nr. 2007/01. Mannheim: ZUMA.