

Informationen zur Nutzung des SAS-Setups für das Mikrozensus Scientific Use File 1995

Der vorliegende Text soll Ihnen dabei helfen, den Rohdatensatz des Statistischen Bundesamtes, den Sie im ASCII-Format erhalten haben, korrekt in SAS einzulesen. Dazu wird vom German Microdata Lab (GML) des Zentrums für Umfragen, Methoden und Analysen (ZUMA) ein Setup (*setup95.sas*) bereitgestellt.

Ausführliche Informationen über den Mikrozensus und die Mikrozensus Scientific Use Files finden Sie auf den Websites des Zentrums für Umfragen, Methoden und Analysen (ZUMA) unter:

<http://www.gesis.org/Dauerbeobachtung/GML/index.htm>

Weitere Informationen sind auf den Seiten des Statistischen Bundesamtes abrufbar:

http://www.destatis.de/themen/d/thm_mikrozen.php

Das vom GML bereitgestellte SAS-Setup für das Mikrozensus Scientific Use File 1995 (*setup95.sas*) dient zum Einlesen des Rohdatenmaterials und zum Erstellen eines SAS-Systemfiles. Es beinhaltet Programmanweisungen zum Ersetzen fehlender Werte, zur Umwandlung alphanumerischer Variablen in numerische sowie zum Versehen der Variablen und ihrer Ausprägungen mit entsprechenden Labels.

Das Setup gliedert sich in verschiedene Bereiche, wobei Folgendes zu beachten ist:

Zunächst werden Formate definiert, um diese später den Ausprägungen der Variablen zuzuweisen. Sollen diese Formate dauerhaft verwendet werden, müssen sie in einer separaten Datei gespeichert werden. Das entsprechende Verzeichnis ist dann mit dem speziell dafür vorgesehenen Bibliotheksnamen LIBRARY zu referenzieren. Um in späteren SAS-Sitzungen auf die Formate zugreifen zu können, muss dem Verzeichnis, in dem die Formate abgespeichert sind, wiederum der Name LIBRARY zugewiesen werden.

Beim Einlesen des Rohdatenfiles ist in der Option LRECL der INFILE-Anweisung angegeben, wie viele Stellen eine Zeile (d.h. eine Beobachtung) im Rohdatenfile umfasst. Alphanumerische Variablen sind beim Einlesen mit einem \$-Zeichen versehen und werden in einem nachfolgenden Schritt in numerische Variablen umgewandelt. Fehlenden Werte, d.h. Leerstellen im Rohdatenfile, werden über IF-Anweisungen durch gültige Werte ersetzt und später über die Format-Anweisung mit Labels versehen. Weil SAS keine benutzerdefinierten Missings kennt, müssen die Werte für spätere Auswertungen gegebenenfalls manuell auf Missing zurückgesetzt werden. Optional können die IF-Anweisungen im Setup auch auskommentiert werden. Dann werden die fehlenden Werte allerdings auch nicht gelabelt.

Das Rohdatenfile des Statistischen Bundesamtes enthält alle Variablen des Mikrozensus 1995. Damit das SAS-Setup für jede mögliche Variablen-Auswahl verwendbar ist, wurden alle Variablen erfasst. Variablen, die Sie nicht bestellt haben, sind in Ihrem Rohdatenfile

auf 0 oder Leerzeichen gesetzt. Um ein Systemfile zu erzeugen, das nur die von Ihnen bestellten Variablen enthält, fügen Sie im Setup am Ende des DATA-step die Anweisung KEEP ein und nennen dort Ihre bestellten Variablen.

Um eine reibungsfreie Aufbereitung der Daten zu gewährleisten, empfiehlt es sich, die in dem Setup vorgesehenen Voreinstellungen zu berücksichtigen. Insbesondere die Anweisung COMPRESS=YES im DATA-step bewirkt, dass die erzeugte Datei deutlich reduzierten Speicherplatz benötigt.

Sind die Daten eingelesen und ist das entsprechende Systemfile erstellt, kann die folgende Fallzahl (ohne Gewichtung, ohne Selektion) zur Kontrolle, ob der Rohdatensatz fehlerfrei eingelesen wurde, mit der des eingelesenen Datensatzes verglichen werden. Unterscheiden sich die Fallzahlen, weist dies auf einen Fehler beim Einlesen hin.

Fallzahl Mikrozensus 1995 (ohne Gewichtung, ohne Selektion) = 512.509

Des Weiteren können die Verteilungen ausgewählter Variablen des Mikrozensus SUF 1995 zur Prüfung des fehlerfreien Ablaufs des Setups herangezogen werden (vgl.: http://www.gesis.org/Dauerbeobachtung/GML/Daten/MZ/mz_1995/randverteilungen_mz95.htm).

Als Referenz zur Plausibilitätsprüfung der Mikrozensus Scientific Use Files dienen die in den Fachserien des Statistischen Bundesamtes veröffentlichten Ergebnisse des jeweiligen Mikrozensus. Für den Mikrozensus 1995 sind dies die Reihe 3 (Haushalte und Familien), die Reihe 4.1.1 (Stand und Entwicklung der Erwerbstätigkeit) und die Reihe 4.1.2 (Beruf, Ausbildung und Arbeitsbedingungen der Erwerbstätigen) der Fachserie 1.

Die Plausibilitätsprüfung des Mikrozensus SUF 1995 erfolgte durch den Vergleich des aufbereiteten Datensatzes mit den vom Statistischen Bundesamt veröffentlichten Tabellen in den entsprechenden Reihen der Fachserie 1. Dabei wurde nach den vom Statistischen Bundesamt verwendeten Bevölkerungs- und Erwerbskonzepten gewichtet und selektiert. (zur Abgrenzung und Hochrechnung der Bevölkerungsbegriffe im Mikrozensus 1995 vgl.: http://www.gesis.org/Dauerbeobachtung/GML/Daten/MZ/mz_1995/hochrechnungsfaktoren_mz95.htm).

Bedingt durch die Ziehung der 70%-Substichprobe ergeben sich Abweichungen zwischen den Häufigkeiten der Variablen des Mikrozensus Scientific Use Files und den in den Fachserien des Statistischen Bundesamtes veröffentlichten Zahlen (beruhend auf den Original-Mikrozensusdaten). Die meisten Variablen weichen nur in geringem Maße (meist 0% bis 1%, max. 5%) von den veröffentlichten Daten ab. Abweichungen über 5% wurden bei Variablen festgestellt, deren Merkmalsausprägungen mit sehr geringen Fallzahlen besetzt sind (z.B. Staatsangehörigkeit und Wirtschaftsbereiche).

Auf Recodierungen wurde im Setup des Mikrozensus 1995 größtenteils verzichtet. Recodierungen wurden nur vorgenommen, um den im Rohdatensatz nicht gekennzeichneten

Missings einen numerischen Wert zuzuordnen. Im Folgenden wird die Systematik dieser Recodierungen kurz vorgestellt:

- ▶ Soweit die 9 bzw. 99 oder 999 in den einzelnen Variablen nicht vergeben ist, wird die Missing-Kategorie mit diesen Zahlen bezeichnet.
- ▶ Ist die 9 bzw. 99 oder 999 besetzt, so wird die Missing-Kategorie mit der 0 gekennzeichnet.
- ▶ Ist auch die 0 in der Variable schon vergeben, so wird die 8 bzw. die 98 oder 998 gewählt.

Lediglich in den Variablen EF221 und EF231 (Arbeitszeittyp) wurden weitere Recodierungen vorgenommen. Hier wird die Lücke innerhalb der Zahlenfolge geschlossen (1,2,3,4,5,7,8,9,0 und "leer" wurde zu 1,2,3,4,5,6,7,8,0 und 9 als Missing-Kategorie).

Das Setup für den Mikrozensus 1995 weist folgende Besonderheiten auf:

- ▶ Im Gegensatz zu Scientific Use Files (SUF) aus anderen Jahrgängen sind im SUF 1995 oftmals die Kategorien "ohne Angabe" und "entfällt" in eine Missing-Kategorie zusammengefasst.
- ▶ Es finden sich im Datensatz des SUF 1995 bei einigen Variablen einzelne Fälle, die nicht einer der vorhandenen Kategorien zugeordnet sind. Diese werden, soweit möglich, im Setup direkt mit der Missing-Kategorie "Entfällt" zusammengefasst oder aber extra als Missing ausgewiesen.
- ▶ In der Bandsatzerweiterung EF228 "überwiegender Lebensunterhalt, Ehefrau der Familienbezugsperson" befinden sich im Rohdatenfile die zusätzlichen Ausprägungen "8" und "9", die laut Schlüsselverzeichnis mit der Kategorie "1 = Erwerbstätigkeit, Berufstätigkeit" zusammengefasst sein sollten. In den Kategorien "8" und "9" befinden sich entsprechend auch nur erwerbstätige / berufstätige Personen. Für die Analyse empfiehlt es sich daher die drei Kategorien "1", "8" und "9" zusammenzufassen. In der Bandsatzerweiterung EF226 "Wirtschaftszweig der Ehefrau der Familienbezugsperson" gibt es geringe Abweichungen zur Variable EF92 "Wirtschaftszweig". Einige Personen, die in der Variablen EF92 eine gültige Angabe haben, sind in der EF226 der Kategorie "entfällt" zugeordnet. Zur Behebung dieses Problems vergleiche: Schimpl-Neimanns, B./Schmidt, K., 1997: Konstruktion von Bandsatz-Erweiterungen zum Haushalt und zur Familie im Mikrozensus mittels satzübergreifenden Operationen in SPSS. ZUMA-Technischer Bericht Nr. 97/08.

Zentrum für Umfragen, Methoden und Analysen (ZUMA)
German Microdata Lab
B2,1

68159 Mannheim
Tel: 0621-1246-265
Fax: 0621-1246-100

<http://www.gesis.org/ZUMA/index.htm>

Kontakt: Andrea Lengerer, Julia H. Schroedter, Hossein Shahla (GML, Mikrozensus Grundfiles)
Email: lengerer@zuma-mannheim.de; schroedter@zuma-mannheim.de; shahla@zuma-mannheim.de
