Informationen zur Nutzung des SAS-Setups für das Mikrozensus Scientific Use File 2001

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 (setup01.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 2001 (setup01.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. Die Variable EF643 ist alphanumerisch und wird daher mit einem $-Zeichen versehen. Sie wird in einem nachfolgenden Schritt in eine numerische Variable umgewandelt.
Die fehlenden Werte, d.h. die 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 StBa enthält alle Variablen des Mikrozensus 2001. 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 2001 (ohne Gewichtung, ohne Selektion) = 503.961

Des Weiteren können die Verteilungen ausgewählter Variablen des Mikrozensus SUF 2001 zur Prüfung des fehlerfreien Ablaufs des Setups herangezogen werden (vgl.: http://www.gesis.org/Dauerbeobachtung/GML/Daten/MZ/mz_2001/randverteilungen_mz01.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 2001 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 2001 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 2001 vgl.: http://www.gesis.org/Dauerbeobachtung/GML/Daten/MZ/mz_2001/hochrechnungsfaktoren_mz01.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).

Das Mikrozensus Scientific Use Files 2001 weist einige Besonderheiten auf: In EF512 wird eine falsche Fallzahl für die Missing-Kategorie "Person in Gemeinschafts-/Anstaltsunterkunft" angegeben. Es werden drei Fälle zuviel ausgewiesen (vgl. EF506). Für die

Bandsatzerweiterungen Haushaltsbezugsperson, Bezugsperson in der Familie, Ehefrau der Bezugsperson in der Familie und Lebenspartner der Bezugsperson im Haushalt gibt es jeweils zwei Variablen zum höchsten beruflichen Ausbildungs- oder Hochschul-/ Fachhochschulabschluss, die in den Values variieren. Die Ausprägungen der Variablen EF568, EF603, EF619 und EF667 stimmen mit denen der Mikrozensus Scientific Use Files bis zum Jahr 1998 überein. Die Values der Variablen EF570, EF606, EF622 und EF670 entsprechen den Antwortmöglichkeiten im Erhebungsbogen seit dem Jahr 1999. Bei Zeitvergleichen der angesprochenen Variablen mit Mikrozensen bis zum Erhebungszeitpunkt 1998 bieten sich die Variablen EF568, EF603, EF619 und EF667 an, bei Vergleichen mit Mikrozensen seit dem Jahr 1999 dagegen die Variablen EF570, EF606, EF622 und EF670. Die letztgenannten Variablen können zudem so recodiert werden, dass sie mit den entsprechenden Variablen der Mikrozensen bis zum Jahr 1998 korrespondieren.

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/Dauerbeobachtung/GML/index.htm

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