Informationen zur Nutzung des SAS-Setups für das Mikrozensus Scientific Use File 2003

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) der GESIS ein Setup (setup03.sas) bereitgestellt.

Ausführliche Informationen über den Mikrozensus und die Mikrozensus Scientific Use Files finden Sie auf den Websites der GESIS unter:
http://www.gesis.org/dienstleistungen/daten/amtliche-mikrodaten/ mikrozensus/ grundfile/

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 2003 *(setup03.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 Statistischen Bundesamtes enthält alle Variablen des Mikrozensus 2003. 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 2003 (ohne Gewichtung, ohne Selektion) = 502.873

Des Weiteren können die Verteilungen ausgewählter Variablen des Mikrozensus SUF 2003 zur Prüfung des fehlerfreien Ablaufs des Setups herangezogen werden (vgl.: http://www.gesis.org/dienstleistungen/daten/amtliche-mikrodaten/mikrozensus/grundfile/mz2003/randverteilungen/).

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 2003 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 2003 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 2003 vgl.: http://www.gesis.org/dienstleistungen/daten/amtliche-mikrodaten/mikrozensus/grundfile/mz2003/abgrenzungen-hochrechnung/).

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).

Weiterhin wurden verschiedene generierte Merkmale (sog. Bandsatzergänzungen) des Mikrozensus 2003 auf ihre Plausibilität hin überprüft. Dabei zeigten sich keine Inkonsistenzen. Es wurde lediglich festgestellt, dass einige Haushalts- und Familienbezugspersonen unter 15 Jahre alt sind, was gemäß den Vorgaben des Statistischen Bundesamtes nicht der Fall sein dürfte.

Im Zuge der Aufbereitung des Mikrozensus Scientific Use Files 2003 wurden die Variablenlabels nach einer neuen Systematik erstellt. Die Labels enthalten nun die Nummer der entsprechende Frage im Selbstausfüllerbogen, den grundsätzlichen inhaltlichen Bezug der Variable und eventuelle Spezifizierungen, außerdem sind Freiwilligkeit und Zugehörigkeit zur Unterstichprobe gekennzeichnet. Ein im Vergleich mit früheren Files abweichendes Label bedeutet also i.d.R. nicht, dass sich am Inhalt der entsprechenden Variable etwas geändert hat.

GESIS - Leibniz-Institut für Sozialwissenschaften
German Microdata Lab
B2,1
68159 Mannheim
Tel: 0621-1246-265
Fax: 0621-1246-100

http://www.gesis.org/das-institut/wissenschaftliche-arbeitsbereiche/dauerbeobachtung-der-gesellschaft/german-microdata-lab/

Kontakt: Andrea Lengerer, Julia H. Schroedter, Hossein Shahla (GML, Mikrozensus Grundfiles)
Email: gml@gesis.org