SAS im Forschungsdatenzentrum der Statistischen Landesämter

Olaf Schoffer
Forschungsdatenzentrum der Statistischen Landesämter (regionaler Standort Kamenz im Statistischen Landesamt des Freistaates Sachsen)
Macherstraße 63
01917 Kamenz
Olaf.schoffer@statistik.sachsen.de

Zusammenfassung

Das Forschungsdatenzentrum der Statistischen Landesämter ermöglicht wissenschaftlichen Nutzern den komfortablen Zugriff auf faktisch anonymisierte Einzeldaten der amtlichen Statistik. Nachfolgend werden die mit SAS im Forschungsdatenzentrum umgesetzten Arbeiten sowie die Nutzungsmöglichkeiten dieser Serviceeinrichtung beschrieben.

Schlüsselwörter: Forschungsdatenzentrum, amtliche Statistik, Anonymisierung.

1 Einleitung und Hintergrund

Das Forschungsdatenzentrum der Statistischen Landesämter (FDZ) ist eine Serviceeinrichtung der amtlichen Statistik für die Wissenschaft und wird derzeit als Pilotprojekt vom Bundesministerium für Bildung und Forschung (BMBF) gefördert. Mit der Einrichtung von Forschungsdatenzentren steht erstmals ein breites Spektrum tief gegliederter amtlicher Mikrodaten aller 16 Bundesländer für die wissenschaftliche Analyse zur Verfügung. Von besonderem Vorteil ist dabei das so genannte „Wissenschaftsprivileg“: die Einzeldaten müssen zur wissenschaftlichen Nutzung nicht absolut sondern nur faktisch anonymisiert werden (§ 16 Abs. 6 BStatG). Im Rahmen dieses Piliotprojektes werden amtliche Daten zur Wirtschafts-, Sozial-, Gesundheits-, Umwelt-, Rechtspflegestatistik u.v.m. aufbereitet und deren Eigenschaften sowie Analysepotential in begleitenden Metadaten dokumentiert. SAS wird dabei im FDZ neben anderen Programmen insbesondere bei der Aufbereitung und Anonymisierung von Mikrodaten eingesetzt, steht aber ebenso den wissenschaftlichen Nutzern zur Analyse der aufbereiteten Mikrodaten zur Verfügung.

2 Datenaufbereitung

Im FDZ werden die verschiedenen Mikrodaten fachlich zentralisiert aufbereitet. Dabei werden die Datenbestände bestimmter Statistiken aus allen 16 Bundesländern an einem Standort zusammengeführt, aufbereitet und vorgehalten. Genutzt werden können die

aufbereiteten Datenbestände jedoch an allen Standorten des FDZ. Nachfolgend werden typische mit SAS bearbeitete Aufgaben der Datenaufbereitung im FDZ beschrieben.

2.1 Einlesen

Die Datenbestände, welche im Rahmen der „fachlich zentralisierten Datenhaltung“ jeweils aus 16 Bundesländern an einem Standort zusammengeführt und eingelesen werden müssen, weisen mitunter komplexe Datenstrukturen auf.
Im nachstehenden Beispiel gibt die auf eine laufende Nummer folgende so genannte Satzart (fett gedruckt) an, welche Struktur und welchen Inhalt die nachfolgenden Daten aufweisen.

```
0004134623
0004256823963299 6
0004239463946933 9
0004256793383473 3
0005145944
0005256800576597 5
```

Eingelesen werden können solche Strukturen beispielsweise mit der nachstehenden Syntax, welche aus den vorliegenden Rohdaten zwei verschiedene Datensätze erzeugt.

```
DATA Satzart1 Satzart2;
 INFILE FILE='<...>';
 INPUT lfdn 4. sa 1. @;
 IF sa=1 THEN INPUT v1 5. @@;
 IF sa=2 THEN INPUT v2 6. v3 6. @@;
 IF sa=1 THEN OUTPUT Satzart1;
 IF sa=2 THEN OUTPUT Satzart2;
RUN;
```

2.2 Umcodieren im DATA-Step

Die in der amtlichen Statistik oft als Großrechnerdateien im EBCDIC-Format vorliegenden Daten werden im FDZ als PC-lesbare CSV-Files angeboten und müssen entsprechend umcodiert werden.
Im der nachstehend aufgeführten Syntax wird beispielhaft ein im EBCDIC-Format vorliegender Datenbestand mittels DATA-Step zu einer CSV-Datei umgewandelt.

```
DATA _NULL_;
 INFILE "...\EBCDIC.txt" MISSOVER;
 INPUT ef1 $EBCDIC2. ef2 $EBCDIC4. ef2 $EBCDIC1. @@;
 FILE "...\ASCII.csv" DELIMITER=';' DSD DROPOVER;
 IF _N_ = 1 THEN PUT 'ef1' ';' 'ef2' ';' 'ef3';
 PUT ef1 ef2 ef3;
RUN;
```

2.3 Weitere Aufgaben

Neben dem Einlesen, Zusammenführen und Umwandeln der vorliegenden Datenbestände werden die vorliegenden Datenbestände u.a. für Analysen mit SAS nutzerfreundlich aufbereitet. Dabei werden beispielsweise spezielle Merkmalsausprägungen mittels vor- und selbstdefinierter Formate bzw. Informate (`PROC FORMAT`) geeignet gelabeled. Um identisch strukturierte Daten mehrerer Erhebungsjahre und Bundesländer flexibel bearbeiten zu können, werden darüber hinaus Makros eingesetzt.

3 Anonymisierung

In der amtlichen Statistik existieren drei Anonymitätsbegriffe:
Absolut anonym sind Daten, mit denen eine Identifizierung der Auskunftgebenden unmöglich ist. Diese strengste Form der Anonymität führt in der Regel zu stark aggregierten Datensätzen und findet Anwendung bei Publikationen und Auskünften der Amtlichen Statistik.
Faktisch anonym sind Daten, mit denen eine Identifizierung nur mit einem unverhältnismäßig hohen Aufwand an Zeit, Kosten und Arbeitskraft möglich ist (§ 16 Abs. 6 BStatG). Diese mildere Form der Anonymität erlaubt meist die Beibehaltung von Einzeldaten und findet Anwendung bei der Erstellung von Mikrodaten für die wissenschaftliche Forschung über das FDZ.
Formal anonym sind Daten, in denen nur eindeutige identifizierende Personen- und Hilfsmerkmale entfernt sind. In dieser Form wird ein Großteil der Originaldaten in der Amtlichen Statistik gespeichert, welche der Öffentlichkeit jedoch so nicht zur Verfügung gestellt werden können..

Zur Erreichung der faktischen Anonymität im FDZ werden verschiedene Anonymisierungsverfahren eingesetzt, von denen nachfolgend ausgewählte vorgestellt werden.

3.1 Systemfreie Sortierung

Damit eine Identifikation nicht bereits durch die Position der Merkmalsträger in den Ausgangsdaten möglich ist, wird das Datenmaterial oft systemfrei sortiert. D.h. der Datensatz wird gemäß einer stetigen Pseudo-Zufallsvariable (`RANUNI(<seed>)`) und gegebenenfalls anderen Kriterien sortiert. Da in der amtlichen Statistik oft große Datensätze bearbeitet werden, spielt dabei der sinnvolle Einsatz der Optionen `<NO>EQUALS` und `TAGSORT` in `PROC SORT` eine wichtige Rolle.

3.2 Vergröberung

Um zu vermeiden, dass Auskunftgebende anhand seltener Ausprägungen einzelner Merkmale zu identifizieren sind, kann mittels Vergröberung oder Mikroaggregation sichergestellt werden, dass mindestens drei Merkmalsträger pro Ausprägung im Datenbestand vorhanden sind.

Eine Vergröberung von Merkmalsausprägungen wird erreicht, indem metrische Merkmale gruppiert, Ausprägungen ordinaler oder kardinaler Merkmale zusammengefasst oder bereits gruppiert vorliegende Merkmale gröber gruppiert werden. Die Umsetzung in SAS erfolgt mittels `IF`-Anweisungen oder mathematischer Funktionen (z.B. `ROUND(<var,unit>)`).

3.3 Mikroaggregation

Bei der Mikroaggregation (genauer: unabhängige eindimensionale Mikroaggregation) werden stetige Merkmale der Größe nach sortiert, anschließend jeweils drei benachbarte Werte durch Ihren gemeinsamen Durchschnitt ersetzt und zuletzt die ursprüngliche Reihenfolge wieder hergestellt. Damit wird ebenfalls erreicht, dass mindestens drei Merkmalsträger pro Ausprägung im Datenbestand vorhanden sind.
Der nachstehender Programmausschnitt zeigt gerade das Ersetzen der Merkmalswerte durch das arithmetische Mittel von jeweils drei benachbarten Merkmalswerten in einem bereits sortierten Datensatz.

```
PROC EXPAND DATA=sortiert OUT=sortiert(DROP=time);
 CONVERT v=lead2_v / TRANSFORM=(LEAD 2);
 CONVERT v=lead1_v / TRANSFORM=(LEAD);
 CONVERT v=lag1_v / TRANSFORM=(LAG);
 CONVERT v=lag2_v / TRANSFORM=(LAG 2);
RUN;

DATA mikroagg;
 SET sortiert;
 IF MOD(_N_,3)=1 THEN v_ma=MEAN(lead2_v,lead1_v,v);
 IF MOD(_N_,3)=2 THEN v_ma=MEAN(lead1_v,v,lag1_v);
 IF MOD(_N_,3)=0 THEN v_ma=MEAN(v,lag1_v,lag2_v);
 <...>
RUN;
```

Die Behandlung der „Ränder“ bei Datensätzen mit einer nicht durch drei teilbaren Zahl von Beobachtungen ist hier der Übersichtlichkeit halber nicht aufgezeigt.

4 Wissenschaftliche Analysen

Wissenschaftler können auf die im FDZ aufbereiteten Daten über verschiedene Nutzungswege zugreifen.
Als besondere Form der absolut anonymisierten Public-Use-Files stehen über die Internetseite www.forschungsdatenzentrum.de **CAMPUS-Files** für Lehrzwecke an Hochschulen unentgeltlich zum Download zur Verfügung.
Ausgewählte Statistiken sind bereits als **Scientific-Use-Files**, d.h. standardisierte faktisch anonymisierte Mikrodatenfiles zur Nutzung außerhalb der Statistischen Ämter, aufbereitet und stehen somit schnell für Analysen zur Verfügung. Sie verfügen über ein deutlich höheres Informationspotential als Public-Use-Files.

Alle im FDZ bereits aufbereiteten Statistiken können nach einer projektbezogenen faktischen Anonymisierung an **Gastwissenschaftler-Arbeitsplätzen**, d.h. PC-Analysearbeitsplätzen mit SPSS, SAS und/oder STATA in den geschützten Räumen der amtlichen Statistik, genutzt werden. Aufgrund der durch diese Regulierung des Zugangs erhöhten Schutzwirkung dürfen die Daten noch wesentlich detailliertere Informationen enthalten als die per Datenträger übermittelten Scientific-Use-Files.
Formal anonymisierte Originaldaten ohne eine weitere Anonymisierung stehen mittels **Kontrollierter Datenfernverarbeitung** für Analysen zur Verfügung. Statt des Zugriffs auf die Einzeldaten erhalten die Datennutzer zunächst Strukturdatensätze (Dummy-Dateien), die in Aufbau und Merkmalsausprägungen dem Originalmaterial gleichen und so die Erstellung einer SAS-, SPSS- oder STATA-Auswertungssyntax erlauben. Mit diesen Programmen werden die Originaldaten in den Statistischen Ämtern ausgewertet und die Ergebnisse dieser Auswertung nach einer Geheimhaltungsprüfung an den Datennutzer zurückgeliefert.
Die verschiedenen Nutzungswege können auch kombiniert werden. Für die zur wissenschaftlichen Nutzung aufbereiteten Datenbestände wird derzeit pro Material (d.h. ein Erhebungsjahr einer Statistik) und Nutzungsweg eine Schutzgebühr von 65 EUR erhoben.
Etwa 20 % der zur Verfügung gestellten Materialien werden dabei mit SAS analysiert. Weiterhin werden beispielsweise SPSS und Stata eingesetzt. An den FDZ-Arbeitsplätzen stehen zur Analyse u.a. die SAS-Module BASE, SAS/STAT, SAS/GRAPH, SAS/ETS, SAS/IML zur Verfügung.
Am häufigsten werden dabei deskriptive Analysen (Kennzahlen, Häufigkeitstabellen, Histogramme, Kartogramme), jedoch auch induktive Verfahren (Regression, Verteilungsanalysen) sowie spezielle Verfahren der Ökonometrie und Biometrie angewendet.
Im nachfolgenden Beispiel wird die amtliche Krankenhausstatistik - Teil II Diagnosen von stationär behandelten Krankenhauspatienten betrachtet. Es werden die Häufigkeiten bestimmter im Jahr 2001 in Krankenhäusern gestellter Diagnosen nach dem Wohnort des Patienten bezogen auf die Einwohnerzahl berechnet und kartografisch dargestellt. Das Kartogramm in Abbildung 1 stellt diese relativen Häufigkeiten in den einzelnen Bundesländern für Endokrine, Ernährungs- und Stoffwechselkrankheiten dar. Die Klassengrenzen wurden dabei so gewählt, dass die Bandbreite der empirischen Häufigkeiten bestmöglich wiedergegeben wird.

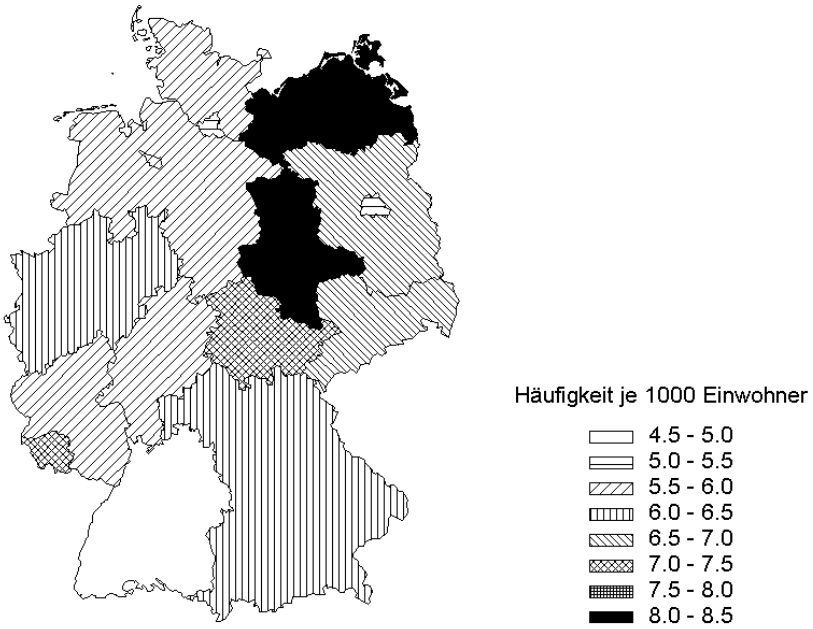

Abbildung 1: Kartogramm der relativen Fall-Häufigkeiten in Deutschland im Jahr 2001 (Endokrine, Ernährungs- und Stoffwechselkrankheiten) - Quelle: eigene Forschungsergebnisse (Daten: Forschungsdatenzentrum der Statistischen Landesämter – Krankenhausstatistik, Teil II Diagnosen 2001).

Eine weitere hier vorgestellte Analyse befasst sich mit der Altersstruktur der behandelten Krankenhauspatienten. Dabei ist jedoch die Fallbezogenheit der Krankenhaus-Diagnosestatistik zu berücksichtigen: Patienten, welche im Jahr 2001 mehrfach in einem oder mehreren Krankenhäusern behandelt wurden, werden mehrfach gezählt (d.h. die Alterspyramide in der Abbildung 2 ist bezogen auf die Fälle, nicht aber auf die Patienten).

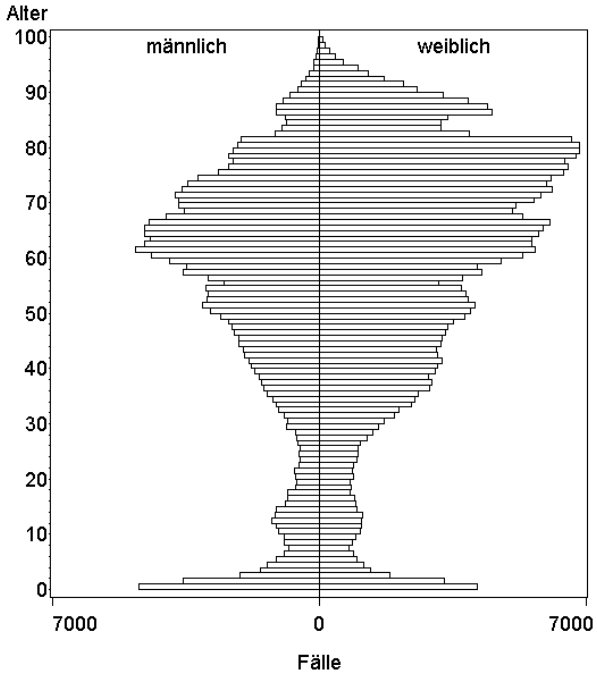

Abbildung 2: Fallbezogene Altersstruktur von im Jahr 2001 in Krankenhäusern behandelten Patienten in Deutschland (Endokrine, Ernährungs- und Stoffwechselkrankheiten) - Quelle: eigene Forschungsergebnisse (Daten: Forschungsdatenzentrum der Statistischen Landesämter – Krankenhausstatistik, Teil II Diagnosen 2001).

5 Ausblick

Im FDZ werden typische Aufgaben wie Datenaufbereitung und Anonymisierung u.a. mit SAS umgesetzt, um die Einzeldaten der amtlichen Statistik zur wissenschaftlichen Nutzung verfügbar zu machen. Diese Nutzung kann ebenfalls mit SAS erfolgen. Eine ausführliche Beschreibung des bisher aufbereiteten Datenangebotes inklusive der erwähnten Metadaten sowie weitere Angaben zu den möglichen Nutzungswegen ist unter www.forschungsdatenzentrum.de zu finden.