Makro zum Auffinden mehrfach vergebener Schlüsselnummern

Martina Rossi
Medizinisches Institut für Umwelthygiene
an der Heinrich-Heine-Universität
Abt. Epidemiologie
Auf'm Hennekamp 50
40225 Düsseldorf

1. Einleitung

Eines der Hauptforschungsgebiete der Abteilung Epidemiologie sind die Studien zur „Wirkung von Luftfremdstoffen auf Kinder unter den besonderen nach der Wiedervereinigung sich entwickelnden Unweltbedingungen“[1]. Die Daten dieser Studien werden von mir in meiner Funktion als Data Managerin zur Auswertung bereitgestellt und gepflegt. Die Datenhaltung ist objektorientiert unter SAS\VMS, so dass alle Informationen zu einem Kind (=1 Beobachtung) zentral vorliegen. Da das Design der Studien viele Untersuchungen in unterschiedlichen Subkollektiven vorsieht, muss ich also viele unterschiedliche Datendateien zu einer Auswertedatei zusammenfügen. Beispielhaft möchte ich für die Schulanfängerstudie West / Ost 1997 einige Zahlen nennen:

Schulanfängerstudie West /Ost 1997	
Einzeldateien	4128
Kooperationspartner	8
Kontaktpersonen	21
Fragebogenmodule	8
Stoffuntersuchungen	10
SAS-Datei	4133 Obs; 759 Var
Einträge in SAS-Datei	3.136.947

Aus diesen Zahlen kann man leicht erkennen, dass die Gewährleistung der Datenintegrität, d.h. Korrektheit und Vollständigkeit der Daten in der Praxis, ein recht schwieriges Problem ist, zumal bei diesen Studien die Fragebogendateneingabe dezentral erfolgt.

2. Problemstellung

Die Vielzahl der Einzeldateien, die in die Auswertedatei eingehen, bergen das Problem der mehrfachen Schlüsselnummern. Nur über ein korrektes Schlüsselkriterium (Probandennummer) können die Mess- und Untersuchungsergebnisse eines Kindes richtig zugeordnet und ausgewertet werden. Ursachen für mehrfache Nummern können Schreibfehler bei der Beschriftung von Probemedien (z.B. Urintöpfchen, Zusatzfragebögen) sein oder manuelle Übertragungsfehler (z.B. bei der Eingabe der Fragebögen). Durch eine doppelte Dateneingabe kann man zwar die manuellen Übertragungsfehler minimieren, allerdings bleiben bei der Datenzusammenstellung einige Problemfälle übrig.

[1] Forschungsplan 1997

Welche Auswirkung die Nichtbeachtung der mehrfachen Probandennummern bei der Datenzusammenführung und deren späterer Auswertung hätte, zeigt die nachfolgende Abbildung.

Datei bsp.urin

OBS	PROBNR	HGU_M	KREATU_M	ORT	DATUM
26	12165	0.11	64.3	7	11/02/97
27	12165	0.11	64.3	7	11/02/97
28	12165	.	.	.	.
29	12188	0.09	44.7	5	11/02/97
30	12188	0.10	85.6	28	11/05/97

Datei bsp.auswert

OBS	PROBNR	GEBDAT	GEBDATMU
31	12165	25/05/92	19/04/61
54	12188	14/01/92	13/08/61
55	12189	06/05/92	14/04/68
56	12190	01/07/91	23/09/67
57	12191	07/06/92	05/02/65

Datei auswert nach MERGE Statement

OBS	PROBNR	HGU_M	KREATU_M	DATUM	GEBDAT	GEBDATMU	GEBGEW	GEBGROES	FRUEHGEB	SWOCHE
57	12165	0.11	64.3	11/02/97	25/05/92	19/04/61	4530	59	2	40
58	12165	0.11	64.3	11/02/97	25/05/92	19/04/61	4530	59	2	40
59	12165	.	.	.	25/05/92	19/04/61	4530	59		
82	12188	0.09	44.7	11/02/97	14/01/92	13/08/61	3240	50	2	41
83	12188	0.10	85.6	11/05/97	14/01/92	13/08/61	3240	50	2	40

Die Datendatei bsp.urin enthält als ausgewählten Problemfall die Probandennummer 12165. Es handelt sich hierbei um eine Dublette und eine Beobachtung mit lediglich der Probandennummer als vorhandenen Variablenwert. Bei der Probandennummer 12188, deren 2 Beobachtungen sich komplett voneinander unterscheiden, könnte es sich um einen Schreibfehler bei der Probenbeschriftung handeln. Fügte man diese Dateien jetzt mit dem MERGE Statement zusammen, erhielte man den unerwünschten Effekt, dass die Urinmesswerte von Kind 12165 sich verdreifachen würden. Bei Probandennummer 12188 existieren unterschiedliche Urinwerte für ein Kind.

Offensichtlich ist, dass mehrfache Schlüsselnummern Daten verfälschen und sie somit zur Gewährleistung der Datenintegrität eliminiert werden müssen. Im Nachfolgenden werde ich mich daher mit 4 Lösungsmöglichkeiten beschäftigen.

3. Lösungsmöglichkeiten

Die SAS Procedure SORT bietet zwei Optionen, die Beobachtungen mit mehrfachen Schlüsselnummern eliminieren.

```
PROC SORT DATA=bsp.urin NODUP;
BY probnr;
```

Durch die Option NODUP werden alle Dubletten eliminiert. In der Datei bsp.urin wäre das eine Beobachtung mit der Probandennummer 12165. Allerdings bliebe sowohl die „leere“ Beobachtung 12165 als auch 12188 mit unterschiedlichen Einträgen erhalten.

Benutzt man stattdessen die Option NODUPKEY, so werden alle mehrfach vorhandenen Schlüsselnummern auf jeweils eine Beobachtung reduziert. Das führt im konkreten Fall zu folgendem Ergebnis:

```
OBS     PROBNR     HGU_M     KREATU_M     ORT        DATUM
26      12165      0.11       64.3        7      11/02/97
27      12188      0.09       44.7        5      11/02/97
```

Die Dublette aus der Datei bsp.urin ist gelöscht worden und die zwei doppelten Beobachtungen der Nummern 12165 und 12188. Vergleicht man die Werte mit denen der Ursprungsdatei, so erkennt man, dass die Eliminierung von der Position in der aufgerufenen Datei abhängig ist. Diese Vorgehensweise ist willkürlich, da die Position in der Datei keinerlei Aussage über der Richtigkeit der Einträge zulässt. Ausserdem kann man als Nutzer dieser Optionen nicht erkennen, welche konkreten Probandennummern elimiert worden sind, da im Log-File lediglich die Anzahl ausgegeben wird.

Da die Gewinnung unserer Messergebnisse sehr zeit- und kostenintensiv ist, sind wir bestrebt alle vorhandenen Ergebnisse zu nutzen und versuchen deshalb die Daten, trotz der mehrfachen Nummer, korrekt zuzuordnen. Man braucht die Einzelwerte um konkrete Fragen an die Ansprechpartner stellen zu können, die dann (hoffentlich) eine richtige Zuordnung ermöglichen.

Daher habe ich, als praktisches Arbeitsinstrument, das Makro (dopkey) geschrieben, dass Dubletten eliminiert und die Einzelwerte doppelt vorhandener Schlüsselnummern ausgibt.

```
Die Datei bsp.urin enthaelt doppelte Schluesselnummern
mit den folgenden Einzelwerten:
                                        09:38 Wednesday, February 17, 1999
OBS     PROBNR     HGU_M     KREATU_M     ORT        DATUM
1       12165      0.11       64.3         7      11/02/97
2       12165       .           .           .            .
3       12188      0.09       44.7         5      11/02/97
4       12188      0.10       85.6        28      11/05/97
```

Abbildung 1: OUTPUT nach Makroaufruf

Die „leere“ Beobachtung mit der Nummer 12165 würde ich ohne Rückfrage löschen. Für das Problem bei Probandennummer 12188 kann aufgrund der unterschiedlichen Einträge bei den Variablen ‘ort’ und ‘datum’ bei den Kollegen aus der Chemie Rücksprache gehalten werden und somit sind die Ergebnisse ggfs. der korrekten Probandennummer zuordenbar.

Bei häufiger Anwendung dieses Makros, nicht nur auf eingehende Dateien, bietet sich eine Speicherung in compilierter Form an. So dass in jeder Bearbeitungsstufe eine schnelle Prüfung auf mehrfache Schlüsselnummern möglich ist.

```
LIBNAME macro 'c:\1user\rossi\ksfe\compi';

OPTIONS MSTORED SASMSTORE=macro;

%dopkey(bsp.urin,probnr);
```

Abbildung 2: Macroaufruf

4. Quellkode:

```
/*********************************************************************
 Makroname:         dopkey1.sas
 Progammiert von:   M.Rossi [rossim@uni-duesseldorf.de]
 1. Version am:     05.11.96
 Aenderung:         12.01.99 Layout der Ausgabe verbessert
 Aenderung:         19.01.99 mehrzeilige Kommentare hinzugefuegt
 Zweck:             Suche nach doppelten Schluesselnummern
                    in einer Datendatei und ggfs. deren Ausgabe
                    compilierte Speicherung
*********************************************************************/

/*
 Erlaeuterungen zum Makro dopkey
 Das Makro DOPKEY hat zwei Positionsparameter;
  1. DATEN      = Filename der zu pruefenden Datei
  2. KEY        = Variablenname der Schluesselnummer

 Syntax zum Aufruf des Makros DOPKEY:
 %DOPKEY(filename,variablename des schluessels);

 Beispiel:
 %DOPKEY(einl.urin,probnr);
*/

LIBNAME bsp 'c:\1user\rossi\ksfe';
LIBNAME einl ‘c:\1user\rossi\sawo97\sasdata’;
LIBNAME macro 'c:\1user\rossi\ksfe\compi';

*Zur Speicherung des Compilates in permanenten Katalog;
OPTIONS MSTORED SASMSTORE=macro;


%MACRO dopkey(daten,key)
 /STORE DES='Suche nach doppelten Schluesselkriterien';

PROC SORT DATA=&daten NODUP;
BY &key;


DATA key(KEEP= &key);
 SET &daten END=final;
 n+1;
 IF final THEN CALL SYMPUT('anzahl',TRIM(LEFT(n)));

PROC TRANSPOSE DATA=key out=dreh;
%LET anz=%EVAL(&anzahl-2);

DATA double(KEEP=dop1-dop&anzahl);
 SET  dreh;
 ARRAY col col1-col&anzahl;
 ARRAY dop dop1-dop&anzahl;
DO i=0 TO &anz;
```

```
   IF col{i+1}=col{i+2} THEN dop{i+1}=col{i+1};
 END;

PROC TRANSPOSE DATA=double out=doppel;


PROC SQL;
 CREATE TABLE ausgabe AS
 SELECT * FROM &daten
  WHERE &key IN
  (SELECT col1 FROM doppel);

PROC PRINT DATA=ausgabe;
TITLE1 "Die Datei &daten enthaelt doppelte Schluesselnummern";
TITLE2 'mit den folgenden Einzelwerten:';
TILTLE;
RUN;
%MEND dopkey;
RUN;
```

Die Überprüfungen der realen Datei einl.urin durch das Makro dopkey führt zu folgendem Ergebnis.

```
Die Datei einl.urin enthaelt doppelte Schluesselnummern
mit den folgenden Einzelwerten:                        * 14:27 Tuesday, January 19, 1999   2
                                                                                           *
OBS    PROBNR    UR_VOL_M    UR_DIC_M    HGU_M    CDU_M    KREATU_M    ORT               *
                                                                                           *
 1      19125       879        1.018      0.09       .        44.7       5               *
 2      19125       397        1.024      0.10       .        85.6       5               *
```

Literatur

1) SAS Language Reference Version 6
2) SAS Procedures Guide Version 6
3) Thomas Bregenzer, Kursunterlagen zur Einführung in die SAS Macro Progammierung
4) SAS Macro Language