AUSTRIAN JOURNAL OF STATISTICS
Volume 34 (2005), Number 4, 327–343

Das Datenmanagement im neuen Mikrozensus – Eine Prozessbeschreibung

Winfried Moser
Statistik Austria, Vienna

Abstract: The Austrian microcencus, by now existing for more than thirty years, has been redesigned completely at the beginning of the year 2004. This article gives an overview about the rearrangement in the area of data-editing and describes the basic structures of relevant subprocesses (data-transformation, plausibility-checks, imputation). The implementation of these processes had to meet input requirements concerning automation, transparency and expandability. This article also deals with the implications of these basic preconditions for the configuration of the data-editing.

Zusammenfassung: Mit Anfang des Jahres 2004 wurde der seit nunmehr über dreißig Jahren bestehende Mikrozensus grundlegend umgestaltet. Der vorliegende Artikel gibt einen Überblick über die Neugestaltung im Bereich des Datenmanagements. Es werden die Grundstrukturen der verschiedenen für die Aufarbeitung notwendigen Teilprozesse (Datentransformation, Plausibilitätsprüfungen, Imputation) beschrieben. Die Umsetzung dieser Prozesse erfolgte unter den Vorgaben Automatisierbarkeit, Transparenz und Ausbaufähigkeit. In dem Artikel wird auch auf die prozesstechnischen Implikationen eingegangen, die diese Grundvoraussetzungen auf die Gestaltung des Datenmanagements haben.

Keywords: Mikrozensus, Imputation, Datenmanagement, SPSS.

1 Einleitung

Die Statistik Austria führt seit nunmehr über dreißig Jahren die größte Stichprobenerhebung Österreichs durch, in der jedes Quartal die Mitglieder von etwa 20.000 Haushalten befragt werden: den Mikrozensus. Mit Anfang des Jahres 2004 wurde diese Erhebung grundlegend umgestaltet. Ein wichtiger Grund dafür liegt in der Vorgabe, die Mikrozensus-Stichprobe gleichmäßig auf die Kalenderwochen eines Jahres zu verteilen (EU-Verordnung Nr. 577/98, Erwerbs- und Wohnungsstatistik-Verordnung – EWStV) – die Organisation des alten Mikrozensus war für die Anforderungen einer kontinuierlichen Erhebung nicht geeignet.

Zu den wichtigsten Neuerungen zählen:

(a) Auswahl für die Stichprobe bildet ab 2004 das Zentrale Melderegister (ZMR) und nicht mehr die Volkszählungsdaten. Dieses Register bietet auf längere Sicht eine bessere Voraussetzung für Stichprobenziehungen. Pro Kalenderwoche kommen rund 1700 Haushalte in die Stichprobe.

(b) Jeder Haushalt bleibt fünf Quartale lang in der Stichprobe (vor 2004: acht Quartale). Die Erstbefragung erfolgt face-to-face, Erhebungsinstrument bleibt weiterhin ein Papierfragebogen, der aber völlig neu konzipiert wurde. Für die Folgebefragung wurde ein

Telefonstudio eingerichtet, als Software kommt dabei das vom Nationalen Statistischen Institut der Niederlande entwickelte Programm BLAISE zum Einsatz.

(c) Erhebung und Aufarbeitung des alten Mikrozensus stammen konzeptuell im Wesentlichen aus den 1970er-Jahren. So wurde die Datenerfassung mittels scannerfähiger Papierbelege bewerkstelligt, die Datenaufarbeitung wurde am Großrechner durchgeführt. Die technische Entwicklung machte hier eine Neukonzeption dringend notwendig. Ab 2006 wird bei den face-to-face-Interviews CAPI eingesetzt werden, für das Datenmanagement wurde eine PC-basierte Lösung geschaffen[1].

Ich möchte hier *einen* der genannten Punkte näher beleuchten – das neue Datenmanagement. Der Prozess war von Grund auf neu aufzubauen und bei der Entwicklung musste – neben der Beschäftigung mit vielen Detailproblemen – eine Reihe von Basislösungen gefunden werden. Ziel des Artikels ist es, diese Problemstellungen und Basislösungen in einer möglichst verallgemeinerten Form vorzustellen.

Unter Datenmanagement wird hier die *strukturelle Grundlage* verstanden, innerhalb der alle notwendigen Prozessschritte durchgeführt werden können, um aus vorliegenden Rohdaten ein auslieferbares Datenfile zu erstellen. Zu den notwendigen Prozessschritten gehören insbesondere Plausibilisierung und Imputation – eine Thematik, auf die in der Statistik Austria schon immer besonderes Augenmerk gelegt wurde. Der Artikel wird sich allerdings nur am Rande mit diesbezüglichen methodischen Aspekten befassen, das Hauptaugenmerk soll auf der Beschreibung der Gesamtprozessstruktur des Datenmanagements liegen, in die Plausibilisierung und Imputation eingebettet sind.

Ein wichtiger Grundsatz bei der Konzeption des Datenmanagements war: Es soll mit dem gesamten System ein solider Standard für den Mikrozensus geschaffen werden, der einen möglichst großen Spielraum für Weiterentwicklung und Weiterausbau offen lässt. Eine besondere Schwierigkeit bei der Entwicklung war der Umstand, dass der Prozess der Datenaufarbeitung für eine große Zahl von Variablen weitestgehend automatisiert ablaufen muss. Das ist schon allein deshalb notwendig, weil EUROSTAT für den Mikrozensus eine Lieferfrist von drei Monaten nach Quartalsende vorschreibt – das sind nur sechs Wochen nach Eingang der letzten Interviews. Für die Variablen des *Grundprogramms* im alten Mikrozensus gab es schon vor 2004 eine solche Prozessautomatisierung. Das betraf damals jedoch nur einen Satz von 24 Variablen, der über die Jahre hinweg weitgehend stabil geblieben ist (im Wesentlichen seit 1967 mit einer einzigen größeren Änderung im Jahr 1994)[2]. Das Frageprogramm im neuen Mikrozensus ist demgegenüber deutlich größer – es umfasst derzeit über 150 Variablen, die durch eine komplexe Fragebogenstruktur miteinander in Verbindung stehen.

Darüber hinaus unterliegt der neue Variablensatz von Quartal zu Quartal stärkeren Änderungen als das einstige Grundprogramm. Diese änderungen müssen *vor* einer Prozessautomatisierung abgefangen werden. Wie dabei vorgegangen wurde, wird im Folgenden beschrieben.

[1] Eine detaillierte Beschreibung der strukturellen Änderungen liefern Kytir and Stadler (2004, S. 511 ff.)

[2] Seit 1994 wurde die Arbeitskräfteerhebung, die seit 2004 in das Grundprogramm des Mikrozensus eingebunden ist, einmal jährlich (im Frühjahr) durchgeführt. Vgl. dazu Bartunek (1994, S. 905 ff.)

2 Beschreibung der Programm- und Prozessstruktur

Die Programmierung des Datenmanagements erfolgte in der Syntaxsprache des Statistikpaketes SPSS. Es wurde damit – aufgrund der besseren Integration von Datenaufarbeitung und Analyse – eine PC-basierte Lösung geschaffen, die die Aufarbeitung am Großrechner ersetzt.

Die wichtigsten Elemente des Gesamtprozesses sind in Abbildung 1 dargestellt. Die Beschreibung der Programm- und Prozessstruktur wird sich weitestgehend an der dort dargestellten Struktur orientieren. Die Elemente sind (1) die Erstellung eines internen Standardrecode-Files („Extraktion"), (2) der Plausibilisierungsblock, (3) der Imputationsblock (mit nach Variablen verschiedenen Imputationsmethoden) und (4) die Erstellung der verschiedenen User-Files, in dem zum Einen standardmäßig zusätzlich generierte Variablen wie Hochrechnung, sozialgeographische Informationen und sonstige abgeleitete Variablen eingebaut werden und zum Anderen bestimmte Informationen wieder verschlüsselt oder ausgeblendet werden (für externe Nutzer).

Es sollte anhand der Darstellung deutlich werden, dass die gesamte Datenmanagementprogrammierung in einer Art „Containerbauweise" erstellt ist, sodass bestimmte Änderungen oder Erweiterungen, die im Wesentlichen die Prozessschritte Extraktion, Plausibilisierung und Imputation betreffen, gut möglich sind.

2.1 Generelle Namens- und Programmierkonventionen

Innerhalb der dargestellten Strukturen sind die derzeit etwa 450 Syntaxdateien des Systems angeordnet. Um den Überblick zu bewahren, wurden sowohl auf Syntax- als auch auf Variablenebene spezielle Namenskonventionen erarbeitet. Auf der *obersten Strukturierungsebene* gibt es die fünf unterschiedlichen Präfixe meta., proz., var., pla., und imp., die auf die Position rückschließen lassen, die die jeweilige Syntaxdatei in der dargestellten Struktur einnimmt. meta.- und proz.-Dateien gehören zur Prozesssteuerung, var.-, pla.- und imp.-Dateien referenzieren auf Variablenextraktion, Plausibilisierung und Imputation einzelner Variablen.

Tabelle 1: Systematisierung der Syntax durch Syntaxpräfixe

meta.	Globale Prozeduren und Variablen, Strukturelle Syntax
proz.	Produktionssyntax
var.	Variablenselektion
pla.	Plausibilisierung
imp.	Imputation

2.1.1 Die Datengenerationen

Um später nachvollziehen zu können, was in welchem Schritt passiert, werden Zwischenergebnisse des Prozesses in unterschiedlichen Files mit einer generalisierten Benennung

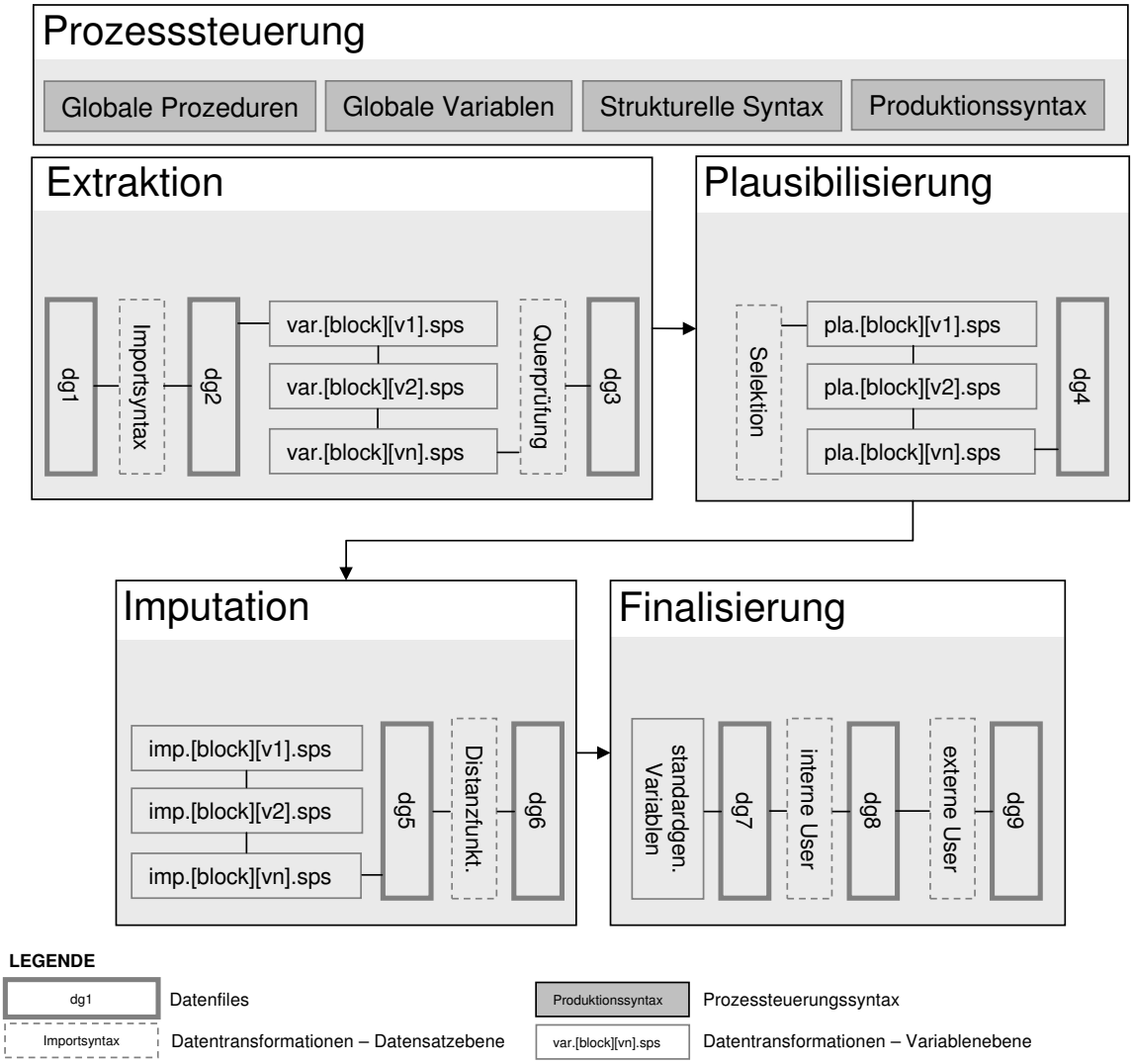

Abbildung 1: Darstellung der Prozessstruktur des Datenmanagements für den neuen Mikrozensus

abgelegt: Ein Präfix (dg1 bis dg9) kennzeichnet die Position im Prozess, an der das jeweilige File erstellt wurde, daran anschließend wird im Dateinamen Jahr und Quartal der jeweiligen Erhebung indiziert. Das Kürzel „dg“ steht für „Datengeneration“. Die angesprochenen Namenskonventionen haben sich mittlerweile etabliert und ersetzen umständliche Benennungen der unterschiedlichen Datenbestände. Als Beispiel: Die geplausten und nicht imputierten Daten des zweiten Quartals des Jahres 2004 finden sich in der Datei *dg4.mz2004q2.sav.* Die verschiedenen Datengenerationen sind folgendermaßen definiert:

Datengeneration 1 enthält das rohe Flatfile, das von der Exportschnittstelle der CATI-Software (Blaise) erstellt wird. Es entspricht in seiner Struktur nicht mehr der ursprünglichen Blaise-Datenbank, in der die Informationen auf Haushaltsebene verspeichert sind. Antworten auf *eine* bestimmte Frage werden damit zum Beispiel im Fall eines Drei-Personen-Haushalt in der Zeile des gegebenen Haushalts drei mal getrennt verspeichert. Mittels einer Prozedur, die im Blaise eigens erstellt werden muss, werden diese Informa-

tionen in der dg1 auf Personenebene gebracht.

Datengeneration 2: Mit „Camäleon“, einem Blaise-Modul, wird automatisch eine SPSS-Einlesesyntax generiert, mit der die dg1 in die SPSS-lesbare und gelabelte dg2 überführt wird. Dieses File umfasst etwa 700 Variablen, der größte Teil davon sind Systeminformationen. Die Datengenerationen eins und zwei werden automatisiert einmal wöchentlich erzeugt. Sie beinhalten neben den bei der Befragung erhobenen Daten auch Grundinformationen über Haushalte, die im Zuge der Befragung nicht erreicht wurden (Alter und Geschlecht der Haushaltsmitglieder aus dem zentralen Melderegister). Die in dem vorliegenden Artikel beschriebenen Prozesse setzen auf die dg2 auf.

In **Datengeneration 3** erfolgt der erste starke Eingriff in die Daten: Die Variablenextraktion. Im Wesentlichen werden hier Variablenstrukturierung und Missingsystematisierung durchgeführt. Eine genauere Beschreibung der entsprechenden Vorgänge finden sich im Abschnitt 2.3 „Die Basis des Systems: Die Transformation der Rohdaten“.

Datengeneration 4 wird erstellt, nachdem alle Plausibilisierungsvorgänge abgeschlossen sind. Am Beginn des entsprechenden Prozessschrittes erfolgt auch die Auswahl der relevanten Datensätze. Ausgeschlossen werden Datenzeilen von (1) nicht befragten Haushalten, (2) von nicht mehr vorhandenen Personen, die noch in der Stichprobe sind und (3) von Personen, deren Daten bestimmte Qualitätskriterien nicht erfüllen.

Diese Datengeneration stellt das wichtigste Zwischenergebnis im Aufarbeitungsprozess dar. Zum Einen enthält es die plausiblen transformierten Rohdaten und zum Anderen sind hier auch sämtliche fehlenden Werte klassifiziert (siehe dazu Kapitel 2.3.2 „Missingsystematisierung: Die Standardisierung von Missing-Codes“). Dieses File ist für jene interessant, die mit den Daten Längsschnittanalysen durchführen wollen oder aus anderen Gründen unimputierte Daten brauchen. Aus diesem Grund wird es auch gemeinsam mit dem endgültigen Datenbestand *(Datengeneration 8)* veröffentlicht. Vor der Veröffentlichung werden sensible Informationen (z.B. genaues Geburtsdatum, Wohngemeinde) gelöscht bzw. vergröbert.

Datengenerationen 5 und 6 sind die ersten „vollständigen“ Files, in denen fehlende Werte mittels verschiedener Imputationsprozeduren durch plausible Werte ersetzt sind (logische Imputation, Hot-Deck und nachgelagert die Distanzfunktion).

In **Datengeneration 7** ist das fertige *interne* Datenfile. Hier werden verschiedene systemexterne Variablen eingebunden (z.B. geographische Informationen, Hochrechnung, etc.). Die Files bis zu dieser Datengeneration befinden sich im Datenmanagementordner und sind aus Sicherheitsgründen nur einem sehr kleinen Nutzerkreis zugänglich (derzeit sieben Personen).

In **Datengeneration 8** schließlich werden aus datenschutzrechtlichen Gründen bestimmte Informationen wieder entfernt bzw. vergröbert. Dieses File wird auf einem externen Ordner einem größeren Nutzerkreis zugänglich gemacht.

2.1.2 Programmierung auf Variablenebene

Ein zentraler Bestandteil der Prozessprogrammierung ist auch, dass es für jede einzelne der Variablen des internen Standardrecodefiles getrennt nach den Prozessschritte Extraktion, Plausibilisierung und Imputation eigene Files mit *wohldefinierten Namenskonventionen* gibt (Prozesspräfix-Variablenname). Diese Files werden dann durch eine übergeord-

nete Syntax sequenziell abgearbeitet. Dieses Vorgehen ist aus zwei Gründen nützlich: Einerseits, wenn man nachvollziehen will, was bei einer bestimmten Variable passiert ist und andererseits, wenn man Änderungen oder Weiterentwicklungen einbauen will.

Diese Nachvollziehbarkeit ist auch historisch gegeben. Der Mikrozensus ist ein lebendes System, das sich jedes Quartal ein wenig verändert, manchmal stärker, manchmal weniger stark. Die Syntax wird in jedem Quartal gemeinsam mit den Datenfiles abgelegt, und es kann später nachvollzogen werden, was in welchem Quartal mit den Daten passiert ist.

Dieses System der Syntax-Organisation hat übrigens auch einen Nachteil: Es muss deutlich extensiver programmiert werden. Prozessschritte, die sonst vereinfacht über mehrere Variablen hinweg programmiert werden können, müssen hier Variable für Variable extra erstellt werden. Der bisherigen Erfahrung nach überwiegen die genannten Vorteile jedoch diesen Nachteil deutlich. Letztlich soll es ja möglich sein, das ganze System im Falle eines Personalwechsels in andere Hände weiterzugeben, und dafür ist eine größtmögliche Transparenz der Programmierung nötig.

2.2 Der Kopf des Systems: Die Prozesssteuerung

Über allem steht die Prozesssteuerung, in der einerseits eine Reihe von Prozeduren und Programmvariablen definiert sind, auf die im allgemeinen Ablauf immer wieder zurückgegriffen wird (globale Prozeduren und Parameter) und in der andererseits der Ablauf der Prozedur zentral geregelt wird (strukturelle Syntax). Darüber hinaus befinden sich im Rahmen der Prozesssteuerung noch die verschiedenen Produktionssyntaxdateien, mittels derer Spezialauswertungen produziert werden.

Um diese doch sehr theoretische Beschreibung mit Inhalt zu füllen, einige Beispiele:

(1) *Globale Prozeduren:* Im Zuge des Aufarbeitungsprozesses ist es sehr oft notwendig, bestimmte Subgruppen im Datensatz zu aggregieren und das Ergebnis dieser Aggregierung wieder den Individuen der jeweiligen Subgruppen zuzuordnen (etwa: jedem Mitglied eines Haushalts das Durchschnittsalter dieses Haushalts zuordnen). Eine derartige Prozedur gibt es im SPSS bislang nicht. Deshalb wurde für diesen speziellen Zweck eine Prozedur erstellt, die in der Datei meta.deklarationen.sps abgelegt wird (eine Sammeldatei für zahlreiche solcher Prozeduren). Bevor die Aufarbeitung gestartet wird, muss diese Datei geladen werden, und im Bedarfsfall wird mit einem deutlich vereinfachten Befehl auf die jeweilige Prozedur zurückgegriffen.

(2) *Globale Parameter:* Bei jedem Quartalswechsel wird ein neuer Ordner erstellt, der sich auf das jeweilige Quartal bezieht. Damit ändern sich jedes Quartal auch die Pfade, unter denen bestimmte Dateien abzulegen sind. Aus diesem Grund werden in der Datei meta.pfade.sps alle vom Programm benötigten Pfade zentral definiert. In der Prozessprogrammierung selbst greifen generalisierte Pfadvariablen dann auf die jeweils aktuellen Pfade in der genannten Datei zurück. Das selbe gilt für die Dateinamen.

(3) *Struktursyntax:* Aus Gründen, auf die ich später noch eingehen werde, sind Prozeduren, die Extraktion, Plausibilisierung oder Imputation von Variablen betreffen, „auf Variablenebene“ programmiert. Das heißt, es gibt für jede einzelne Variable und für jeden der genannten drei Prozessschritte eigene Syntaxfiles. Strukturelle Syntaxdateien fassen diese Einzeldateien wieder zusammen und definieren den sequenziellen Ablauf der Ab-

arbeitung dieser Dateien. Eine Datei dieser Art (meta.steuerungszentrum.sps) steuert den Gesamtablauf der Prozeduren.

(4) *Produktionssyntax:* Es gibt eine Vielzahl denkbarer Spezialauswertungen, die in einem Aufarbeitungsprozess notwendig sind. Eine solche Syntax-Datei etwa erstellt automatisch eine Liste von Personen, bei denen das Geschlecht offensichtlich falsch eingetragen war, auf die Erhebungsseitig zugegriffen werden kann. Eine andere produziert eine Spezialauswertung der Daten, die zur Berechnung des Verbraucherpreisindex benötigt wird. Eine dritte erstellt durch eine spezielle Aggregierung der Daten einen Überblick die verschiedenen Arten von fehlenden Werten. Die jeweiligen Dateien werden von der Struktursyntax an den geeigneten Positionen im Datenproduktionsprozess aufgerufen.

2.3 Die Basis des Systems: Die Transformation der Rohdaten (Extraktion)

2.3.1 Standardisierung der Variablennamen und -inhalte

Die wichtigste strukturelle Vorbedingung für ein automatisiertes Datenmanagement sind standardisierte und stabile Variablennamen und Codierungen. Es ist ein Hauptzweck der Extraktion der Rohdaten, diese Vorbedingung zu erfüllen. Es war eine sehr grundsätzliche Entscheidung, nicht mit den Rohdatenvariablen direkt zu arbeiten, sondern als ersten Schritt ein *internes Standardrecodefile* zu erstellen. Bei der Erstellung dieser Datei werden *standardisierte Variablennamen* mit bestimmten Namenskonventionen vergeben und gleichzeitig *Variableninhalte standardisiert*.

Dazu ein konkretes Beispiel: Die Mikrozensusfrage nach dem Wohnort vor einem Jahr besteht aus fragebogentechnischen Gründen aus vier verschiedenen Fragen: (1) „Haben Sie vor einem Jahr an der gleichen Adresse wie jetzt gewohnt?“, (2) wenn nein: „Haben Sie damals im gleichen Bundesland gewohnt?“, (3) wenn nein: „In welchem Bundesland haben Sie gewohnt?“ und (4) wenn kein Bundesland: „In welchem Staat haben Sie gewohnt?“. Diese Art der Fragestellung soll erhebungsseitig die Respondentenbelastung vermindern und auch die Interviewtätigkeit erleichtern – komplexe Fragen werden nur gestellt, wenn es unbedingt notwendig ist.

Es werden hier zwar vier Fragen gestellt, deren Antworten in den Rohdaten auch in vier verschiedenen Variablen verspeichert werden, es geht jedoch im Grunde nur um eine einzige Information: den Wohnort vor einem Jahr. Bei der Erstellung des internen Standardrecodefiles werden also die vier aus diesen Fragen resultierenden Variablen zu einer einzigen zusammengeführt. Im derzeitigen Mikrozensusfragebogen gibt es zahlreiche solcher Fälle; es geht hier darum, redundante Information zu minimieren. Der Schritt ist aus verschiedenen Gründen notwendig beziehungsweise vorteilhaft:

(1) Durch Änderungen der Fragebogenstruktur kann es – bedingt durch Programmkonventionen des Blaise (das Programm für die Erstellung des Cati-Fragebogens) – zu Verschiebungen der Variablennamen in den Rohdaten kommen. Wenn etwa an einer bestimmten Stelle im Fragebogen eine neue Frage eingefügt wird, verschieben sich die Variablennamen aller darauf folgenden Variablen des jeweiligen Fragenblocks. Auf dieser Grundlage ist keine standardisierte Plausibilitätsprüfung und Imputation möglich.

(2) Bestimmte Filterführungen im Fragebogen, bei denen es ausschließlich um die Re-

spondentenentlastung geht – wie etwa bei der vorhin vorgestellten Frage nach dem Wohnort vor einem Jahr – werden an dieser Stelle bereits abgefangen und müssen so bei Plausibilitätsprüfung und Imputation nicht mehr berücksichtigt werden. Die Durchführung dieser beiden Prozesse wird damit über weite Strecken erleichtert.

(3) Nachdem die Standardisierung der Variablen sich an Inhalten orientiert und damit Fragebogenkomplexitäten ausblendet, ist der Datensatz für jene die ihn analysieren, einfacher handhabbar. Es bleiben nach Abschluss der Prozedur auch deutlich weniger Variablen übrig.

Im alten Mikrozensus waren die Variablen nummeriert, was zwar programmiertechnisch praktikabel ist, aber auch einen großen Nachteil mit sich bringt: Neu hinzukommende Variablen bringen das Nummerierungssystem durcheinander, und es muss dann erst wieder zu alphanumerischen Bezeichnungen übergegangen werden. Eine inhaltliche Sortierung der Variablen ist auch schwer möglich. Deshalb wurde zu einem System mit folgender Namenskonvention übergegangen: [Blockbuchstabe][Variablenbezeichner]. Der Blockbuchstabe besteht aus einem Zeichen, der Variablenbezeichner sollte ein sprechendes Kürzel sein. Als Beispiel: bwov1j steht für die Variable „**W**ohn**o**rt **v**or **1** **J**ahr“ aus dem **B**-Block. Die derzeit bestehenden Blocks sind in Tabelle 2 dargestellt.

Tabelle 2: Blocks im neuen Mikrozensus

(a)	Stichprobeninformation
(w)	Wohnungserhebung
(b)	demographische Informationen
(c)	Bestimmung der Erwerbstätigkeit
(d)	derzeitige berufliche Tätigkeit
(e)	Zweittätigkeit
(h)	Arbeitssuche
(k)	Ausbildung
(j)	frühere Tätigkeit
(l)	Situation vor einem Jahr
(x)	wichtige Systemvariablen, nicht zuordenbare abgeleitete Variablen
(z)	Zusatzerhebung

2.3.2 Missingsystematisierung: Die Standardisierung von Missing-Codes

Als zweiter wichtiger Schritt werden in der Extraktion die Missings standardisiert und systematisiert. Die Missingsystematisierung ist die wichtigste Grundvoraussetzung für Plausibilitätsprüfung und Imputation. Letztlich ist dieser Schritt auch so etwas wie eine „Grundplausibilisierung“. Bezogen auf eine bestimmte Zelle im Datensatz geht es hier um die Fragen: „Warum steht hier kein gültiger Wert?“ bzw. „Warum steht hier ein Wert, obwohl per definitionem keiner hier stehen soll?“.

Die Missingsystematisierung gibt darauf eine Antwort. Grundsätzlich werden für fehlende Werte negative Codes vergeben. Nachdem negative Werte im Datensatz nicht vorkommen, wird damit auf einfache Weise verhindert, dass eine bestehende Codierung gleichzeitig als Missingcode verwendet wird.

Konkret geschehen an dieser Stelle folgende Prozesse: Zum einen werden die Blaise-Missingcodes standardisiert – im Blaise werden diese automatisiert vergeben. Sie können sich somit von Variable zu Variable unterscheiden, was für ein automatisiertes Datenmanagement wenig brauchbar ist.

Für die restlichen Zellen wird die Filterführung systematisch überprüft. Die Systematik unterscheidet zwischen befüllten und leeren Zellen und überprüft, ob der jeweilige Zustand (befüllt/leer) im Einklang mit den Routingregeln ist oder nicht. Auf Ebene der Routingregeln wird noch unterschieden, ob aufgrund eines regulären Werts aus der Befragung geroutet wurde oder aufgrund eines fehlenden Werts. Die Systematik ist übersichtlich in Tabelle 3 dargestellt, die Missingcodes, die sich daraus ergeben, sehen Sie in Tabelle 4.

Tabelle 3: Systematik der Missingkategorien

	In der Zeile steht ...	
Mit Routingregeln ...	ein Wert	ein Missing
übereinstimmend (normales Routing)	gültiger Wert	-3
nicht übereinstimmend (normales Routing)	-4	-5
übereinstimmend (Missingrouting)	gültiger Wert	-6
nicht übereinstimmend (Missingrouting)	-10	-5

Tabelle 4: Systematisierung der fehlenden Werte

Code	Label	Inhalt
-1	Verweigert	aus CATI oder Papierfragebogen
-2	Weiß nicht	aus CATI oder Papierfragebogen
-3	Filter	fehlende, die per definitionem fehlen müssen
-4	Filter gelöscht	Werte, die per definitionem Filter sein müssen
-5	Unbekannt (System)	nicht zuordenbare fehlende Werte
-6	Filter Missingrouting	gefiltert aufgrund eines fehlenden Werts
-7	Teilmissing	aus einer zusammengeführten Variable fehlt ein Wert
-9	Unplausibel	wird bei der Plausibilitätsprüfung zugeordnet
-10	Filter gelöscht Missingrouting	Werte, die aufgrund eines Missingroutings Filter sein müssen

Die Kategorisierungen mögen teilweise komplex erscheinen, sie sind jedoch nur ein Zwischenergebnis auf dem Weg zum authentischen Datenbestand, das die Analyse der Daten und das Zusammenspiel der verschiedenen beteiligten Programmierer erleichtert. Vor allem der letztere Punkt ist wichtig: Blaise- und SPSS-Programmierung erfolgen unabhängig voneinander, müssen jedoch möglichst gut zusammenspielen.

2.4 Damit alles seine Richtigkeit hat: Die Plausibilitäts-Prozeduren

Die Prozeduren im Einzelnen zu beschreiben würde wohl den Rahmen dieses Artikels sprengen, darum wird die Beschreibung hier eher kursorisch bleiben. Bei Plausibilitätsprüfungen ist ein axiomatisches Grundproblem zu berücksichtigen: *Ist das Ergebnis einer*

Plausibilitätsprüfung, in die zwei oder mehr Variablen einbezogen sind, negativ, so ist daraus nicht ableitbar, welche der Variablen den falschen Wert beinhaltet.

Dieses Problem könnte man – relativ aufwändig – lösen, indem man jede der geprüften Variablen mit einem Satz weiterer Variablen plausibilisiert. Jene Variable, bei der eine größere Anzahl von Prüfungen negativ ausgeht, wird dann mit größerer Wahrscheinlichkeit den falschen Wert beinhalten. Das würde bedeuten, dass von Fall zu Fall entschieden wird, welcher Wert zu löschen ist. Man könnte in den entsprechenden Fällen auch nachtelefonieren, wie das bei manchen Surveys auch passiert. Weil es beim neuen Mikrozensus jedoch – neben einer möglichst weitgehenden Richtigkeit der Einzelinformationen – auch darum geht, dass der Datensatz in sehr kurzer Zeit auslieferbar ist, haben wir uns für eine einfachere Lösung entschieden, und zwar aus mehreren Gründen:

(1) Die oben genannten Möglichkeiten wären für den neuen Mikrozensus in der Phase des Systemaufbaus technisch schwer umsetzbar gewesen.

(2) Grundlegend lösbar ist diese Problemstellung nur im Zuge der Erhebung: Baut man die Plausanweisungen direkt in die Erhebungssoftware ein, ist es möglich, durch Nachfragen in der Befragungssituation den „richtigen falschen Wert“ zu bestimmen. Ist es der zeitlich früher erhobene Wert, muss der Fragebogen von der entsprechenden Stelle an ein weiteres Mal durchlaufen werden – die Befragten werden in so einem Fall also zurückgeroutet. Es gibt bereits sehr viele Plausibilitäts-Checks im Blaise. Kommen im Zuge der Aufarbeitung größere Ungereimtheiten in der Datensatzstruktur zum Vorschein, werden die Blaise-Checks entsprechend erweitert.

(3) Solange der Anteil der unplausiblen Werte relativ niedrig ist, ist es statistisch irrelevant, welcher Wert gelöscht wird. Ist der Anteil aber hoch, deutet das meist auf grundlegendere Probleme bei der Erhebung hin, etwa auf eine schlecht umgesetzte Fragestellung. Auch eine noch so durchdachte Plausibilitätsprüfung kann hier nur wenig helfen.

Ein wichtiger Teil der Plausprüfungen – jener, der mit der Filterführung zu tun hat – wird schon bei der Variablenextraktion durchgeführt (siehe Kapitel 2.3.2 „Missingsystematisierung“). Bei diesem Teil wird hierarchisch vorgegangen. Das heißt, wichtige Variablen, aufgrund derer gefiltert wird (Alter, Präsenzdienst, Erwerbstätigkeit u. ä.), werden mit Ausnahme ein paar kleinerer Checks grundsätzlich als richtig angenommen. Das ist anders als im alten Mikrozensus, wo zum Beispiel bei Altersplausprüfungen aufgrund bestimmter Regeln manchmal auch die zweite Information als richtig angenommen und das Alter gelöscht wurde. Ein so komplexes Vorgehen ist beim neuen Mikrozensus nicht notwendig. Da durch die CATI-Struktur die Daten sequenziell verspeichert werden, kann – zumindest auf Ebene der Filterführung – auch die Überprüfung sequenziell geschehen.

Der zweite Plausibilisierungsschritt – die Prüfung auf konsistentes Antwortverhalten der Respondenten hinsichtlich unterschiedlichster Aspekte bzw. automatische Umcodierungen bestimmter Variablen – erfolgt erst im Anschluss an den vorhin beschriebenen Schritt. Hier werden gemeinsam mit den Fachleuten für die jeweiligen Themen laufend Plausanweisungen entwickelt und in den Prozess eingebaut. Die Plausibilitätsprüfungen werden so sukzessive genauer. Durch die Programmierung auf Variablenebene ist auch der nachträgliche Ein- und Weiterausbau von Plausprozeduren leicht möglich und durch das Abspeichern der Syntax mit den Datenbeständen des jeweiligen Quartals können Änderungen und Weiterentwicklungen relativ einfach nachvollzogen werden.

Derzeit erfolgt die Plausibilisierung nur Datensatzintern. Möglichkeiten für eine Wei-

terentwicklung der Plausprüfungen bieten vor allem die Einbindung von Administrativdaten und quartalsübergreifende Prüfungen.

2.5 Der Imputationsblock: Automatisierung und Kohärenz

2.5.1 Einleitung

Die Imputation von Werten kann von zwei Seiten beleuchtet werden – von einer wissenschaftlich-statistischen und von einer praktischen. Erstere soll hier jedoch soweit wie möglich ausgeblendet bleiben, da es Zweck des Artikels ist, die *praktischen* Erfordernisse des Datenmanagements eines großen Surveys zu beschreiben. Nicht zuletzt ergeben sich aus diesen praktischen Erfordernissen oft auch Restriktionen für die Umsetzung wissenschaftlicher Ansprüche.

Es gibt zwei wichtige Gründe für eine Imputation. Aus wissenschaftlicher Sicht geht es um die Verbesserung der Schätzung von Messgrößen – durch Imputation kann, wenn sie gut gemacht ist, Bias ausgeglichen werden. Aus Sicht des Datenhandlings ist es für den Nutzer angenehm, wenn er bei seinen Berechnungen mit stabilen Eckzahlen operieren kann – das ist ohne Imputation nicht möglich.

Man muss sich natürlich auch dessen bewusst sein, dass durch die Imputation Zusammenhänge zwischen Variablen verfälscht werden können. Dieses Problem ist aber nicht gravierend, da der Anteil der imputierten Variablen kaum jemals über 5 Prozent hinausgeht. Interessiert man sich explizit für Zusammenhänge zwischen Variablen, ist auch die dg4 – also der Datensatz mit den plausibilisierten und nicht-imputierten Werten – eine gute Option.

Aus praktischer Sicht gibt es für eine Imputationsprozedur zwei wichtige Kriterien: Zum Einen muss ihre Durchführung *in einen automatisierten Prozessablauf* eingebunden werden können und zum Anderen muss dabei *die logische Kohärenz der Datensätze* gewahrt bleiben.

Die *Automatisierung* der Imputation ist aus Sicht des Mikrozensus unabdingbar, weil hier regelmäßig etwa 150 verschiedene Variablen zu imputieren sind. Das Postulat der *logischen Kohärenz der Datensätze* bedeutet, dass bei den Imputationsprozeduren wie bei den Plausprüfungen eine relativ komplexe Fragebogenstruktur sowie wichtige Interdependenzen zwischen Variablen zu berücksichtigen sind. (1) Das bedeutet – im einfachsten Fall – dass beispielsweise einem Präsenzdiener im Zuge der Imputation nur Werte zugewiesen werden dürfen, die er bei der Befragung aufgrund der Fragebogenroutings *theoretisch* auch bekommen könnte[3]. (2) Und es heißt auch, dass einer Person nur *ein*

[3]Es gibt deutlich kompliziertere Fälle, wie etwa die Imputation des Jahres, in dem die jeweils höchste Bildungsstufe abgeschlossen wurde (kjahr). Kohärenz bedeutet hier, dass das Jahr, in dem diese Ausbildung abgeschlossen wurde, in einem sinnvollen Zusammenhang mit dem Geburtsjahr und der Art der Ausbildung der betreffenden Person stehen muss. Ein Hotdecking, in das Geburtsjahr und Art der Ausbildung als Stratumsvariablen einfließen ist hier nicht möglich, da sich zu viele Stratumgsgruppen ergeben (ca. 80 Geburtsjahrgänge $\times$ 7 Ausbildungsarten = ca. 560 Strata). Aus diesem Grund wird aus den bestehenden Daten erst das Alter bei Abschluss der Ausbildung berechnet, dieses Alter wird danach mit Hotdeck imputiert, mit der Ausbildungsart als Stratumsvariable. Danach wird aus dem imputierten Alter wieder der Jahrgang berechnet.

bestimmter Wert zugewiesen werden darf, obwohl unter Umständen mehrere Werte plausibel und wahrscheinlich sind.

Ginge es nur um Parameterschätzung, wäre diese Vorgabe nicht unbedingt notwendig. Sie könnte in manchen Fällen – im Sinne der Schätzgenauigkeit – sogar kontraproduktiv sein. Für eine Parameterschätzung ist eine genaue Schätzung auf Ebene des Individualdatensatzes weniger wichtig als eine genaue Schätzung auf Aggregatebene. Darüber hinaus liefern bestimmte Schätzmethoden – wie etwa die multiple Imputation[4] – auf individueller Ebene gar keine eindeutigen Werte, dafür aber auf Aggregatebene stabilere Schätzungen[5].

Es wird jedoch trotzdem nicht auf dieses Postulat verzichtet, vor allem weil neben der Veröffentlichung von Schätzungen aus dem Mikrozensus auch anonymisierte Individualdaten weitergegeben werden – sowohl an EUROSTAT als auch an andere private und öffentliche Datennutzer. Die Nutzer sollen die veröffentlichten Schätzungen möglichst leicht nachvollziehen können. Eine Reproduktion von Schätzergebnissen unter Einbeziehung multipler Imputationstechniken wäre aber dem Nutzer kaum zumutbar. Ohne Reproduktionsmöglichkeit ist es aber deutlich schwieriger, Vertrauen in die Daten aufzubauen – ein Aspekt, der nicht unterschätzt werden sollte.

2.5.2 Grundsätzliches Vorgehen

Die wichtigste Entscheidung hinsichtlich der Imputation war, die Unit-Non-Response nicht durch Imputation auszugleichen, sondern durch Gewichtung. Die verbleibende Item-Non-Response wird zum größten Teil mit Hot-Deck[6] (siehe Kapitel 2.5.4) imputiert. Im Folgenden wird die Vorgangsweise kurz beschrieben:

(a) Als erstes verschaffen wir uns einen Überblick über die Non-Response-Situation. Dieser Schritt erfolgt, nachdem die Missingsystematisierung abgeschlossen ist, mittels einer eigens produzierten Überblickstabelle. In den Zeilen dieser Tabelle sind alle Mikrozensus-Variablen verzeichnet und in den Spalten die vorhin beschriebenen verschiedenen Typen von Missings. Anhand der Tabelle kann man dann die Häufigkeiten der verschiedenen Missingkategorien für einzelne Variablen ablesen und damit auch den Anteil der zu imputierenden Werte abschätzen. Genau bestimmen kann man diesen Anteil erst im Nachhinein, da vor der Imputation nicht bekannt ist, wie viele Werte gefiltert werden und damit nicht direkt ins Hot-Deck fallen.

(b) Danach wird imputiert: Durch ein sequenzielles Vorgehen wird sichergestellt, dass nur Werte imputiert werden, die im Einklang mit der Fragebogenstruktur (Filterführung) stehen. Es wird dabei sequenziell vorgegangen, weil *vor* der eigentlichen Imputation entschieden werden muss, ob ein fehlender Wert einen Filter bekommt oder nicht. Dafür müssen alle vorangehenden Informationen bereits vorhanden sein.

[4]Barnard et al. (1998, S. 2772 ff.)

[5]Das selbe gilt übrigens auch für die Hochrechnung: Aus arbeitstechnischen Gründen wird im Mikrozensus nur mit einem einzigen Gewicht gerechnet, obwohl es sich durchaus als sinnvoll erweisen könnte, für einzelne, wichtige Variablen eigene Gewichte zu berechnen.

[6]Lessler and Kalsbeck (1992, S. 213 ff.)

2.5.3 Abbildung der Filterführung in der Imputationsprozedur

Die Abbildung der Filterführung ist direkt in die Imputationsprozedur eingebunden. Letztlich ist sie ja selbst schon eine Art Imputation und eine nötige Grundvoraussetzung für das Hot-Decking, das mittels einer SPSS-Prozedur mit einem sehr einfachen Benutzerinterface durchgeführt wird (Abbildung 2). Ich werde dessen wichtigste Elemente kurz erklären:

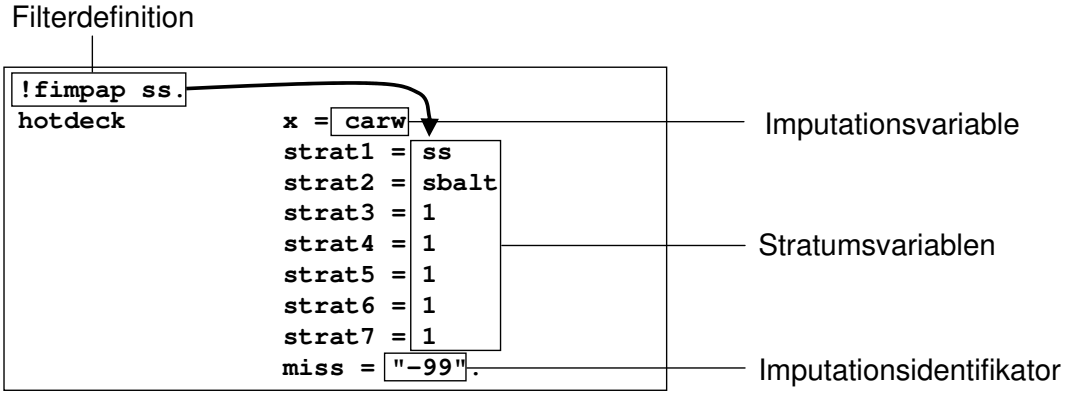

Abbildung 2: Das User-Interface für die SPSS-Hot-Deck-Prozedur

„x“ kennzeichnet die Variable, die man imputieren möchte. Unter „strat1“ bis „strat7“ legt man potenziell sieben Stratumsvariablen an und „miss“ definiert einen Identifikator für die Fälle, die man imputieren möchte. Dieser muss natürlich vorher in den Daten definiert werden. Die Filterführung wird direkt in diesem Prozeduraufruf in die Imputation eingebaut. In dem konkreten Fall werden mit dem Befehl „!fimpap“ unter 15-Jährige und Präsenzdiener angesprochen, die hier einen Filter bekommen. Diese Personen kommen in die eine Subgruppe, alle anderen in die zweite, und die resultierende Variable wird selbst als Stratumsvariable verwendet.

Danach folgt die Imputationsprozedur. Nachdem alle unter 15-Jährigen mit *gültigen* Werten einen Filter haben, bekommen auch die unter 15-Jährigen mit *fehlenden* Werten automatisch einen Filter. Auf der anderen Seite wird durch dieses Vorgehen verhindert, dass Non-Response-Fälle, die gültige Werte bekommen sollten, einen Filter imputiert bekommen. Man schlägt also programmtechnisch zwei Fliegen mit einer Klappe. Dieses Vorgehen hat den Vorteil, dass die „Plaus-Imputations-Schleife“[7] vermieden werden kann.

Die Filterführung ist übrigens für die wichtigsten Gruppen zentral definiert. Das heißt, sollte sie einmal geändert werden, muss man in der Syntax nicht Variable für Variable umprogrammieren, sondern kann das an zentraler Stelle tun.

[7] Alternierendes Durchführen von Plausibilitätsprozedur und Imputationsprozedur, um Fehlzuordnungen sequenziell zu verringern.

2.5.4 Die verschiedenen Imputationsprozeduren

Es wird im Moment mittels dreier verschiedener Verfahren imputiert:

Hot-Deck

Der größte Teil der Variablen wird mittels **Hot-Deck** imputiert. Ein kritischer Punkt beim Hot-Deck ist die Bestimmung der Stratumsvariablen. Im neuen Mikrozensus wurde mit der Deklaration von Sortiervariablen äußerst sparsam umgegangen. Der Grund dafür ist folgender: Grundvoraussetzung für eine sinnvolle Stratifizierung ist, dass zwischen der Sortiervariable und der zu imputierenden Variable ein Zusammenhang besteht. Es wurde also jede einzelne Variable des Datensatzes auf Zusammenhänge mit einem bestimmten Set von Sortiervariablen überprüft. Für eine Imputation wurden dann nur jene Merkmalskombinationen herangezogen, für die sich deutliche Zusammenhänge ergaben. Die entsprechenden Auswertungen wurden in dem Statistikpaket R vorgenommen, weil dieses Programm gegenüber dem SPSS deutlich breitere Möglichkeiten zur Metaanalyse von Daten bietet.

Es wurde dabei folgende Vorgangsweise gewählt: (a) Vorab wurden Geschlecht, Alter (in verschiedenen Kategorisierungen), Haushaltsgröße, Stellung zur Referenzperson, Größe der Wohnung, höchste abgeschlossene Ausbildung, Wirtschaftszweig, Beruf und Bundesland als Stratumsvariablen festgelegt. (b) Die genannten Stratumsvariablen wurden daraufhin überprüft ob sie mit der Imputationsvariable korrelieren und/oder ob diese Variablen in einen Zusammenhang mit dem Response/Missing-Verhalten stehen.

Die Bedingungen für die Auswahl einer Stratumsvariable waren: (a) Zellbesetzungen größer fünf für mindestens 80% der Schichten der Stratumsvariablen und (b) ein Cramers-V größer 0.2 in mindestens einer der beiden überprüften Zusammenhänge. Die verbleibenden Stratumsvariablen wurden dann „händisch“ und nach Sinnhaftigkeit ausgewählt.

Abschließend ist zu diesem Punkt zu sagen, dass es hier noch zahlreiche Ausbaumöglichkeiten gibt. Unser Vorgehen mag einfach erscheinen, es sind jedoch zwei Dinge zu bedenken: Einerseits mussten wir in einem relativ kurzen Zeitraum zu passablen Ergebnissen für eine große Zahl von Variablen kommen. Das heißt, die Entwicklung komplexer Imputationsprozeduren für einzelne Variablen war schwer möglich und wurde nur dort durchgeführt, wo mit Hot-Deck keine befriedigenden Ergebnisse zu erzielen waren (Haushaltsrelationsvariablen). Andererseits gibt es nach unseren Recherchen auch in der statistischen Fachliteratur keine Vorschläge für ein standardisiertes Vorgehen bei der Auswahl von Stratumsvariablen für Hot-Deck-Prozeduren. Hier scheint es also auch an Grundlagenforschung zu mangeln.

Imputation von Haushaltszusammenhängen

Für die Imputation von Haushaltszusammenhängen eignet sich Hot-Deck nicht. Hier wird mittels eines **logischen Verfahrens** imputiert: Anhand der Grundparameter Alter, Geschlecht und Stellung zur Referenzperson jener Person, für die imputiert wird, und jener Personen, zu denen diese potenziell in Relation stehen kann, wird mittels unterschiedlicher Regeln versucht, bestimmte wichtige Relationen zuzuweisen. Als erstes wird geprüft, ob die betreffende Person einer anderen Person im Haushalt als Partner zuordenbar ist, danach, ob sie als Kind in Frage kommt, und als letztes, ob sie Elternteil von einer anderen Person sein könnte. Das Vorgehen ist ziemlich komplex und wird in einem eigenen Artikel zur Methodologie der Imputation und Ableitung von Haushalts- und

Familienrelationen gesondert beschrieben werden.

Distanzfunktion

Für einige wenige Fälle haben wir uns dafür entschieden, per **Distanzfunktion** vollständige Fälle zu ersetzen. In diesen Fällen, in denen einige wichtige Variablen nicht beantwortet sind bzw. die Beantwortungsrate deutlich unter dem Durchschnitt liegt, wurde so vorgegangen, weil hier eine sequenzielle Imputation über weite Strecken reine Artefakte produzieren würde und weil aufgrund der mangelnden Informationen die komplexe Struktur des Fragebogens durch die gewöhnliche Imputation wohl nur schlecht abgebildet werden könnte – es würde zu Inkonsistenzen kommen.

Die Distanzfunktion funktioniert folgendermaßen: Bestimmte Grundinformationen sind über jeden Befragten schon aus der Stichprobe vorhanden. Konkret verwenden wir Alter, Geschlecht, Wohnort und Staatsbürgerschaft. Aus dem Wohnort kann man zwei weitere wichtige Informationen generieren: (1) die räumliche Distanz, in der zwei Personen aus dem Datensatz zueinander leben und (2) eine aus verschiedenen Informationen über den Wohnort generierte ordinale Information über diesen Ort. Das ist eine Art spezieller „Verstädterungsgrad“, der aus den Variablen Einwohnerzahl, Erwerbstätigenzahl und Besiedlungsdichte errechnet wird. Die Distanzfunktion bestimmt für jede einzelne der betreffenden Personen jene Person im Datensatz, die ihr hinsichtlich der oben genannten Variablen am nächsten ist und übernimmt den kompletten Datensatz dieser Person. Im Idealfall werden hier – weil ja die Geo-Koordinaten der Haushalte verfügbar sind – die Daten des Nachbarn übernommen.

Dass durch die Koordinaten räumliche Distanzen berechnet werden können, ist für eine Distanzfunktions-Imputation sehr gut brauchbar. Auf diese Weise können regionale Besonderheiten in die Imputation eingehen, auf die sonst kaum eingegangen werden kann, weil die regionale Gliederung auf Gemeindeebene wegen des hohen Diffenenzierungsgrads kaum für die Hot-Deck-Imputation taugt. Gerade für Phänomene des Arbeitsmarkts ist aber die Wichtigkeit der regionalen Komponente nicht zu unterschätzen.

Das beschriebene Vorgehen war vor allem im ersten Quartal notwendig, weil der Mikrozensus zu diesem Zeitpunkt noch mit Umstellungsschwierigkeiten und demgemäß auch mit der Datenqualität zu kämpfen hatte. Im ersten Quartal gab es noch etwa 2300 Distanzfunktionsfälle. Wie man in Abbildung 3 sieht, ist diese Zahl jedoch im Laufe des Jahres stark zurückgegangen: im dritten Quartal waren es nur noch 700 und ab dem vierten Quartal nur mehr etwa 350. An Höhe und Entwicklung dieser Zahlen sind sowohl Umstellungsschwierigkeiten als auch positive Entwicklung der Erhebungs- und Datenqualität in der ersten Zeit nach der Umstellung ablesbar.

(d) Im Moment wird gerade an der Entwicklung und Implementierung von **Cross-Wave-Methoden** für die Imputation gearbeitet[8]: Da diese Methoden prozesstechnisch schwierig umzusetzen sind, werden sie nur bei wichtigen Variablen (Arbeitssuche, Erwerbstätigkeit, Bildung) oder bei Variablen mit hohen Non-Response-Raten (Wohnungskosten) zum Einsatz kommen. Grundsätzlich geht es dabei um die Nutzung von Werten aus anderen Erhebungswellen für die Imputation.

Im einfachsten Fall werden die entsprechenden Werte des Vorquartals direkt übernommen. In einer komplexeren Spielart, die gerade entwickelt wird und auch zum Einsatz

[8]Eine Einführung liefert Kalton and Lepkowsky (1983, S. 171 ff.)

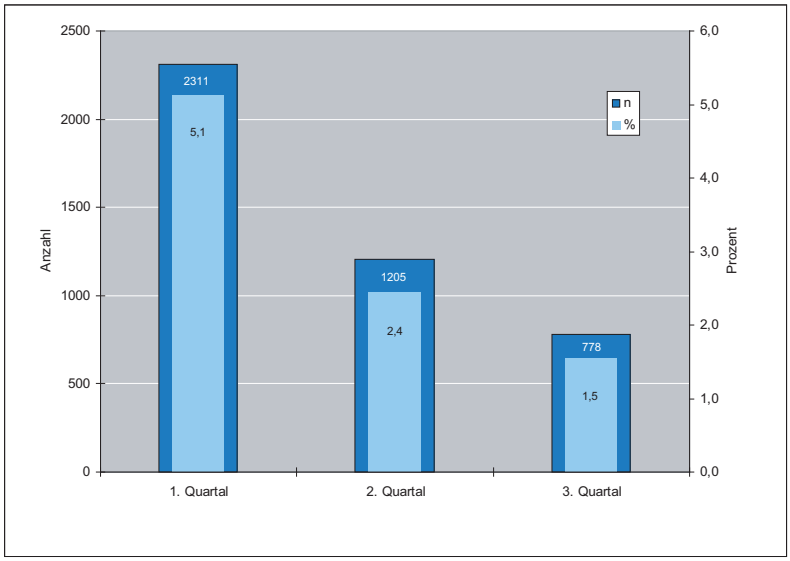

Abbildung 3: Distanzfunktionsfälle in den ersten drei Quartalen des Jahres 2004

kommen soll, werden die Vorquartalswerte als Stratum für eine Hot-Deck-Imputation benutzt. Es handelt sich damit um ein *stochastisches Cross-Wave-Verfahren*; die Wahrscheinlichkeit, mit der ein Wert sich über ein Quartal hinweg verändert, wird in der Imputation berücksichtigt. Die Implementierung dieser Imputationsmethoden in einen automatischen Datenaufarbeitungsprozess erfordert eine strikt organisierte und gut funktionierende Ablaufstruktur des gesamten Datenproduktionsprozesses. Aus diesem Grund muss mit der Einbindung dieser Methoden auch gewartet werden, bis die Datensätze einiger Quartale vorliegen.

2.6 Die abschließenden Prozeduren

Nachdem die beschriebenen Prozeduren abgeschlossen sind, werden in einem letzten Arbeitsschritt noch die Gewichte eingebunden, die verschiedenen abgeleiteten Variablen erstellt und diverse User-Files erzeugt. Bei diesem Prozessschritt muss im Moment noch manchmal händisch eingegriffen werden. Seine vollständige Automatisierung erfordert die erst teilweise vorliegende Spezifikation der internen und externen Bedarfsstrukturen.

3 Resumee

Der Mikrozensus ist mit seinen 30 Jahren Laufzeit ein Survey mit großer Tradition. Er ist aber wohl auch ein wenig in die Jahre gekommen. Aller Veränderungsresistenz von Tradition zum Trotz wurde Anfang vorigen Jahres die Chance einer grundlegenden Neuorientierung ergriffen – mit allen Vorteilen und Problemen, die das mit sich bringt.

Zu den großen strukturellen Neuerungen auf Ebene des Datenmanagements zählt, dass die Programmierung der Datenaufarbeitung, wie auch die laufende Wartung und Verbesserung, nicht mehr am Großrechner durchgeführt werden, sondern in dem weit verbreiteten Statistikpaket SPSS am PC. Auf diese Weise kommt es zu einer direkteren

Zusammenarbeit zwischen Datenmanagement und Datenanalyse; das beeinflusst die Datenqualität und die Produktivität langfristig positiv. Syntax für Plausanweisungen oder Variablenableitungen kann nun zum Beispiel in vielen Fällen von den jeweiligen Fachstatistikern selbst geschrieben werden. Das ist Angesichts der deutlichen Erweiterung des Grundprogramms auch notwendig.

Die auffälligste – und anfangs sicher ein wenig ungewöhnliche – Neuerung für den Nutzer ist die neue Struktur der Variablen, die eine einfachere Handhabung des Datensatzes ermöglicht. Im internen Standardrecodefile – das mit bestimmten Einschränkungen auch der Öffentlichkeit zugänglich gemacht wird – ist für eine größtmögliche Kontinuität und Klarheit der Datenstruktur gesorgt. Es ist damit auch eine wichtige Grundlage für Längsschnittanalysen vorhanden.

Mit mittlerweile sechs abgeschlossenen Quartalen ist der neue Mikrozensus relativ jung. Erhebung und Aufarbeitung weitestgehend sehr modern organisiert und auf dem letzten Stand der Technik. Die Vorgaben Automatisierbarkeit, Transparenz und Ausbaufähigkeit konnten umgesetzt werden. Man kann aber noch nicht davon ausgehen, dass schon alle erdenklichen Probleme beseitigt sind. Durch die transparente Struktur des Datenmanagements und seine enge Einbindung in den Bereich der fachstatistischen Datenanalyse ist aber eine Grundlage für eine kontinuierliche Weiterentwicklung und Verbesserung der Datenaufarbeitung und damit der Daten selbst geschaffen.

Literatur

Barnard, J., Rubin, D. B., and Schenker, N. (1998). Multiple Imputation Methods. In P. Armitage and T. Colton (Eds.), *Encyclopedia of Biostatistics.* New York: Wiley.

Bartunek, E. (1994). Die Arbeitskräfteerhebung von EUROSTAT. *Statistische Nachrichten*, *11*.

Kalton, G., and Lepkowsky, J. M. (1983). Cross-Wave Item Imputation. In M. David (Ed.), *Lessons of the Income Survey Development Program (ISDP).* New York: Social Science Research Council.

Kytir, J., and Stadler, B. (2004). Die kontinuierliche Arbeitskräfteerhebung im Rahmen des neuen Mikrozensus. vom „alten“ zum „neuen“ Mikrozensus. *Statistische Nachrichten*, *6*.

Lessler, J. T., and Kalsbeck, W. D. (1992). *Nonsampling Errors in Surveys*. New York: Wiley.

Author's address:

Mag. Winfried Moser
Statistik Austria
Direktion Bevölkerung - Analyse und Prognose
Guglgasse 13
A-1110 Wien

Tel. +43 (01)71128 7039
Fax +43(01)71128 7455
E-mail: winfried.moser@statistik.gv.at