Das Datenmanagement in EU-SILC - von der Befragung zu Sozialindikatoren

RICHARD HEUBERGER
NADJA LAMEI

Der Artikel gibt einen Einblick in das Datenmanagement der Erhebung EU-SILC, einer jährlichen Haushaltsbefragung zu Einkommen und Lebensbedingungen. Nicht nur der Fragebogen, in dem alle Einkommensbestandteile erhoben werden, auch die Aufarbeitung dieser Daten ist komplex. Deshalb wurde ein automatisiertes System entwickelt, das sämtliche Schritte der Bearbeitung der Rohdaten von der Plausibilisierung über Imputationen und Tabellenerstellung bis zu den fertigen Indikatoren umfasst. Dabei wird im Speziellen auf Effizienz, Anpassungsfähigkeit und Transparenz Wert gelegt. Schließlich sind die einzelnen Einkommensbestandteile der Personen zu einem Jahreseinkommen auf Haushaltsebene zu aggregieren. Das Haushaltseinkommen wiederum ist die wichtigste Grundlage zur Berechnung von Indikatoren zur sozialen Lage der Haushalte, wie zum Beispiel der Armutsgefährdung, und muss daher einer ständigen Qualitätskontrolle unterliegen.

Die Erhebung EU-SILC

EU-SILC (Statistics on Income and Living Conditions) ist eine Statistik über Einkommen und Lebensbedingungen von Privathaushalten in Europa und bildet eine wichtige Grundlage für die Europäische Sozialstatistik.[1]) Zentrale Themen sind Einkommen, Beschäftigung, Wohnen und viele andere Bereiche einschließlich subjektiver Fragen zu Gesundheit und finanzieller Lage, die es erlauben, die Lebenssituation von Menschen in Privathaushalten abzubilden. EU-SILC ist auch Quelle zur Erhebung der vom Europäischen Rat verabschiedeten Laeken-Indikatoren zur Messung von Armut und sozialer Ausgrenzung.

In Österreich wurde EU-SILC erstmals 2003 als einmalige Querschnitterhebung von der STATISTIK AUSTRIA durchgeführt. Mit 2004 begann eine integrierte Längs- und Querschnitterhebung, das heißt, jeweils rund drei Viertel der über 4.500 Haushalte werden auch im Folgejahr wieder befragt, ein Viertel der Stichprobe kommt jährlich neu hinzu.

Den teilnehmenden Ländern steht bei EU-SILC sowohl die Datenquelle als auch die Art der Datenerhebung frei. In EU-SILC gibt es daher keinen gleichlautenden Fragebogen in allen EU-Mitgliedstaaten, sondern per Verordnung verbindlich definierte Zielvariablen (*primary target variables*). Die grundsätzliche Struktur der Befragung besteht aus einem Haushalts- und einem Personenregister, einem Haushaltsfragebogen, einem Fragenblock zur Kinderbetreuung (ab 2004) und Personenfragebögen für alle im Haushalt lebenden Personen ab 16 Jahren. In Österreich wird die Erhebung mittels CAPI-Fragebogen (*computer assisted personal interviewing*) durchgeführt. Diese Befragungstechnik weist gegenüber PAPI (*paper and pencil interviewing*) gewichtige Vorteile auf, wie die automatische Filterführung, die elektronische Datenerfassung (keine zusätzlichen Fehler durch Dateneingabe) und die Möglichkeit der integrierten Durchführung von Checks. Gerade in einem derart umfangreichen und gleichzeitig stark untergliederten Fragebogen hat sich die Plausibilisierung und Konsistenzüberprüfung von Register-, Haushalts- und Personenfragebögen als sehr effizient erwiesen. Kommen die ersten Befragungsdaten ins Haus - die Erhebung wird derzeit ausgelagert durchgeführt - werden weitere Kontrollmechanismen aktiviert. Falls nötig werden Datenmängel und Inkonsistenzen an das Erhebungsinstitut zurückgemeldet und Nachrecherchen gestartet. Es soll so möglichst viel Information aus dem Feld geholt und möglichst wenig durch Nachbearbeitungsprozesse zustandekommen.

Das Datenmanagementsystem - Anforderungen und Aufbau

Eine komplexe Erhebung braucht ein „einfaches" Datenmanagement

Die Aufgabe, jährlich und in möglichst kurzem Abstand zur Erhebung vollständig geplauste, imputierte und gewichtete Datensätze an EUROSTAT zu senden sowie die Indikatoren

[1]) Allgemeine Informationen zu EU-SILC bietet der in Heft 3/2005, S. 224 ff., erschienene Artikel bzw. die Website der STATISTIK AUSTRIA (http://www.statistik.at/fachbereich_03/eusilc_txt.shtml). Aktuelle Ergebnisse sind unter http://www.statistik.at/fachbereich_03/eusilc_ergebnis.shtml zu finden.

zum sozialen Zusammenhalt zu berechnen, macht es notwendig, diese Arbeitsschritte so weit wie möglich zu standardisieren und systematisieren. Die Entwicklung inhaltlicher Aufarbeitungsschritte soll die Mikro- und Makro-Plausibilität der Datensätze und die Qualität garantieren. Im Fragebogen und damit auch in der Aufarbeitung ist zudem ständig auf Änderungen bei Sozialleistungen zu reagieren. Außerdem werden durch die *Längsschnittkomponente* nach und nach zusätzliche Informationen verfügbar, die eine Plausibilisierung erleichtern.

Die Besonderheit von EU-SILC ist die Betonung der *Haushaltszusammenhänge*, die auch in den darauf basierenden Indikatoren, wie der Armutsgefährdung, zum Ausdruck kommt. Aus diesem Grund reicht es nicht aus, in sich konsistente Personenfragebögen zu haben. Alle Kontroll- und Plausibilisierungsschritte auf Personenebene müssen in Relation zu den anderen Personen des Haushalts erfolgen und aufeinander abgestimmt sein. So ist - um ein einfaches Beispiel zu geben - eine allein lebende Pensionistin, die angibt über keinerlei Einkünfte zu verfügen, unplausibel; lebt sie aber in einem Haushalt mit einem Partner oder mit der Familie der Kinder, ist die Situation eine andere, was bei der Feldkontrolle berücksichtigt werden muss.

Diese einzelnen Anforderungen gilt es „unter einen Hut" zu bringen: Die Entwicklung des Datenmanagementsystems hat also ein weitestmöglich standardisierter, aber gleichzeitig offener und erweiterbarer Prozess zu sein. Gerade die inhaltliche Komplexität in der Aufarbeitung setzt Übersichtlichkeit und Nachvollziehbarkeit bei der Durchführung der notwendigen Prozessschritte voraus. Die Vorteile des hier vorgestellten Datenmanagementsystems, mit Hilfe dessen die Rohdaten zu „fertigen" Datenfiles verarbeitet werden, liegen in folgenden Punkten:

- eine hohe Standardisierung der Datenchecks, Editierungsarbeiten, Imputationen und der anschließenden Auswertungsschritte,
- Anpassungsfähigkeit und Möglichkeit, neue „Module" zu integrieren,
- alle Veränderungen an den Daten sind in Syntax dokumentiert und der Prozess daher transparent.

Es versteht sich, dass das gesamte Datenmanagementsystem seit seinem ersten Aufsetzen so etwas wie eine „ewige Baustelle" ist - neue Teile kommen hinzu, unbrauchbar gewordene Bausteine werden entfernt oder „renoviert". Jedoch stellt gerade das eine permanente Qualitätskontrolle des Aufarbeitungsprozesses sicher.

Die gesamte Programmierung erfolgt im Fachbereich, was deshalb von Vorteil ist, weil alle Überprüfungen und Entscheidungen im Hinblick auf einen inhaltlich konsistenten Mikrodatenbestand erfolgen müssen. Die Festlegung und Formalisierung dieser inhaltlichen Checks geht daher über reine Programmierarbeit hinaus.

Die einzelnen Module

Aus der Logik der Befragung und der externen und internen Erfordernisse ergeben sich folgende Bearbeitungsschritte (eine nähere inhaltliche Beschreibung vor allem im Bezug auf die Einkommen folgt im nächsten Kapitel):

1. Check der Rohdaten und Rückmeldungen an das Erhebungsinstitut,
2. Editieren von Einkommensvariablen,
3. Editieren von Nicht-Einkommensvariablen,[2])
4. Imputationen,
5. Gewichtung inklusive Non-Response Analyse,
6. Berechnung der EUROSTAT-Zielvariablen,
7. Berechnung von zentralen Auswertungsindikatoren und der Laeken-Indikatoren,
8. Erstellung von Publikationstabellen und
9. Erstellung von Datenfiles, die an externe Nutzer weitergegeben werden können.

Die genannten Programmblöcke sind inhaltlich verbunden und in dieser Reihenfolge auszuführen, stehen aber programmtechnisch relativ unabhängig nebeneinander. Änderungen in einem Modul haben inhaltliche Auswirkungen auf die folgenden Module und die Endresultate, sollen aber keine programmtechnischen Änderungen in den folgenden Modulen zur Folge haben. Auch die Abfolge der Module ist nicht so starr, wie sie auf den ersten Blick scheint. Es kann zum Beispiel notwendig erscheinen, die EUROSTAT-Zielvariablen zu Prüfzwecken früher als hier dargestellt zu berechnen, Haushaltseinkommen zu berechnen, wenn noch gar nicht alle Einkommensvariablen imputiert sind usw. Das ist möglich und wird auch vielfach genutzt.

Die Herstellung eines vollständigen und plausiblen Datensatzes

Plausibilisierung

Um zu einem auswertbaren Datensatz zu gelangen, müssen die Daten plausibilisiert, vervollständigt und gewichtet werden. Die Plausibilisierung bezeichnet hier all jene Arbeitsschritte der Prüfung und Kontrolle, die notwendig sind, um die Daten stimmig und widerspruchsfrei zu machen. Versucht wird also, Fehler und Irrtümer, die bei der Erhebung, Dateneingabe oder Verarbeitung entstanden sein können, zu entdecken und durch plausible Werte zu korrigieren.

Die ersten Kontrollen der Daten finden bereits während der Feldphase von EU-SILC statt. Das Feldinstitut liefert - neben den zweiwöchigen Feldberichten über den Fortgang der Feld-

[2]) Die getrennte Behandlung von Einkommensvariablen und Variablen anderen Inhalts ergibt sich aus dem zentralen Stellenwert der Einkommen in SILC, der Komplexität der Einkommensaufbereitung (siehe die inhaltlichen Beschreibungen im Folgenden) und aus der Anforderung, dass die Einkommensdaten vollständig geplaust und imputiert an EUROSTAT geliefert werden müssen.

arbeit - in regelmäßigen Abständen den vorläufigen Datenbestand. Die Kontrolle dieser vorläufigen Datenbestände ermöglicht es, auf Fehler (Probleme des Fragebogens, Probleme mit bestimmten Interviewern u.ä.) frühzeitig zu reagieren und, falls nötig, bei den Haushalten nachzurecherchieren. Dies gewährleistet gemeinsam mit den integrierten Checks des CAPI-Fragebogens, dass die vom Erhebungsinstitut gelieferten Enddaten von hoher Qualität sind.

Fehler und Inkonsistenzen in den Daten lassen sich aber dennoch nicht vollständig vermeiden. Diese sind etwa auf Fehlinterpretationen der Befragten oder Interviewer und Interviewerinnen zurückzuführen (z.B. bei der korrekten Bezeichnung öffentlicher Leistungen). Probleme treten dabei vorrangig bei der Erfassung der Einkommen der Personen und Haushalte auf. Diese Probleme müssen behoben werden, um zu einem plausiblen und konsistenten Datensatz zu gelangen; dies deshalb, weil EU-SILC zur Berechnung verschiedenster Indikatoren zu Armut und Lebensbedingungen verwendet wird und Inkonsistenzen etwa bei der Erfassung der Einkommenssituation des Haushalts leicht zu Verzerrungen bei der Erfassung der sozialen Lage des Haushalts führen können. Die bereits in der Feldphase stattfindenden Kontrollen und Korrekturen sorgen dafür, dass diese nachträgliche Bearbeitung der Datensätze nur noch einen Bruchteil der befragten Haushalte und Personen betrifft.[3])

Die Plausibilisierung der Daten lässt sich in mehrfacher Hinsicht unterteilen:
(1) Hinsichtlich des zeitlichen Bezugs: *Längsschnittplausibilisierungen* verwenden die Daten *vorangegangener Befragungen derselben Erhebungseinheiten* (Haushalte und Personen), um die Angaben des jeweils aktuellen Jahres zu überprüfen, *Querschnittplausibilisierungen* verwenden für diese Prüfung die Angaben der *jeweils aktuellen Erhebungswelle*. Die Möglichkeit der Längsschnittplausibilisierung ist bei EU-SILC für jene Erhebungseinheiten der Stichprobe gegeben, die bereits in vorangegangenen Wellen an der Befragung teilgenommen haben; für die Erhebungseinheiten der jeweiligen Erstbefragung sind nur Querschnittplausibilisierungen möglich.

(2) Plausibilisierungen können auch danach unterschieden werden, was bzw. welche Variable plausibel gemacht werden soll. Bei EU-SILC kann dabei zwischen *Einkommensvariablen* und *Nichteinkommensvariablen* unterschieden werden. Bei der Plausibilisierung der Nichteinkommensvariablen werden etwa Angaben zu Kindern (Anzahl der Pflichtschul- und Vorschulstunden) oder die Angaben zur Wohnsituation (Angaben zu Wohnkosten und Wohnnutzfläche) geprüft und nötigenfalls korrigiert.

[3]) Nachträgliche Korrekturen der Einkommensbeträge sind beispielsweise nur bei maximal 3% der jeweiligen Einkommensbezüge notwendig, und bei den meisten Einkommenskomponenten muss gar keine nachträgliche Korrektur vorgenommen werden.

Bei der Plausibilisierung von Einkommensvariablen sind wiederum mehrere Arbeitsschritte notwendig. Einkommen werden in EU-SILC sehr detailliert erfasst, d.h. zu einer Einkommenskomponente müssen die Fragen bzw. Variablen zum Bezug, zur Anzahl der Monate und zum Betrag der Einkommenskomponente berücksichtigt werden.

Die Plausibilisierungen werden für jede Gruppe der Nichteinkommensvariablen bzw. jeder Einkommenskomponente (hier wiederum getrennt nach Bezugsfragen, Fragen nach der Anzahl der Monate und nach der Höhe des Betrags) einzeln in die Steuerungssyntax integriert. Dies gewährleistet die Überschaubarkeit des Prozesses und dessen Anpassungsfähigkeit. Einzelne Plausibilisierungsschritte können damit leicht an geänderte Rahmenbedingungen und Erfordernisse angepasst werden.

Vervollständigung

Sind die Daten hinreichend plausibilisiert, kann daran gegangen werden, den Datensatz zu vervollständigen, d.h. fehlende Variablenwerte zu ergänzen. Grund für fehlende Variablenwerte sind vorrangig fehlende Angaben durch die Befragten selbst (non-response), also weil die Befragten nicht auf die jeweilige Frage antworten können oder möchten. Aber auch im Zuge der Plausibilisierungen können diese Lücken im Datensatz erzeugt werden: Werden einzelne Angaben als unplausibel klassifiziert (etwa sehr hohe Notstandshilfe-Einkommen), so werden diese Variablenwerte zunächst gelöscht - und müssen wiederum geschätzt werden. Da die Erhebung der Einkommenssituation das wesentlichste Ziel von EU-SILC ist, sollen lt. Verordnung alle fehlenden Angaben zu Einkommen ersetzt bzw. ergänzt werden.

Die Vervollständigung des Datensatzes - oder auch Imputation der Werte im Sinne der Schätzung nicht vorhandener Angaben - geschieht in EU-SILC auf drei verschiedene Arten:
1. deduktive Verfahren,
2. deterministische Verfahren,
3. stochastische Verfahren.

Deduktive Verfahren sind solche, bei denen der fehlende Wert einer Variable logisch abgeleitet werden kann. Das heißt, dass der Wert entweder aus anderen Variablenwerten des Datensatzes abgeleitet oder aus anderen, externen Quellen (etwa gesetzlichen Regelungen) erschlossen werden kann. Beispiel für die erste Art der Deduktion wäre etwa die Berechnung der Familienbeihilfe. Dabei wird aus den Angaben des Haushalts- bzw. Personenfragebogens die Anzahl der Kinder, für die ein Bezug der Familienbeihilfe besteht, und die Höhe der Familienbeihilfe berechnet. Beispiel für die zweite Art der Deduktion wäre etwa der Fall, wenn die Angabe zur Höhe des Kinderbetreuungsgeldes fehlt.[4]) Hier kann der Wert aus den gesetzlichen Regelungen entnommen bzw. berechnet werden.

[4]) Die Höhe des bezogenen Kinderbetreuungsgeldes wurde nur 2004 erfragt; ab 2005 wird nur noch erfragt, ob ein Bezug besteht oder nicht.

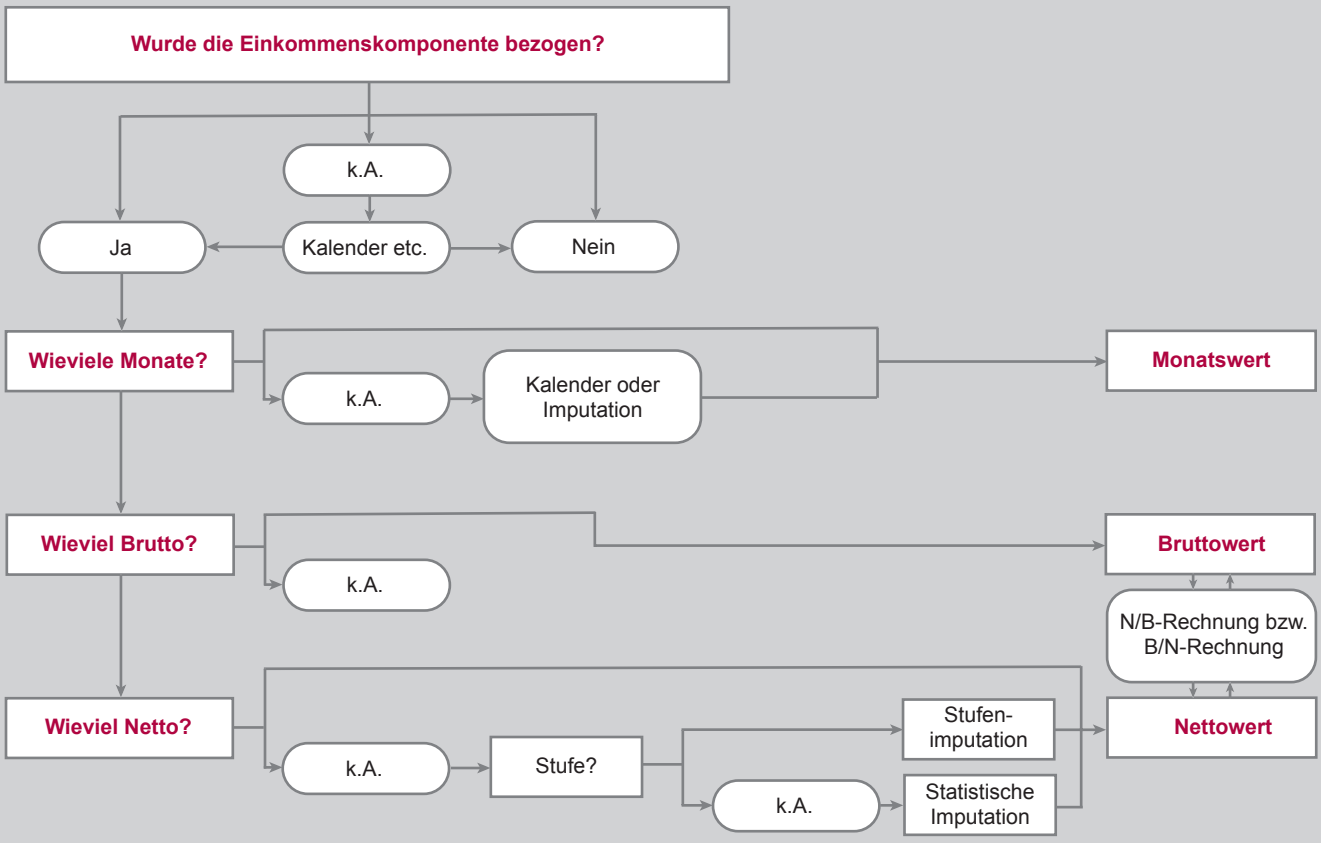
Deterministische und stochastische Verfahren unterscheiden sich nur hinsichtlich der Frage, ob das jeweilige Verfahren der Ersetzung des fehlenden Wertes einen Zufallsterm inkludiert (stochastisch) oder nicht (deterministisch). Zur Anwendung kommen in EU-SILC verschiedene Verfahren: Längsschnittimputationen, auf Regressionsmodellen basierende Verfahren, „Stufenimputationen" und Medianimputationen.

Fehlende Angaben zum Einkommen können in EU-SILC aus drei Gründen auftreten: (1) Die Angabe darüber, ob eine Einkommenskomponente bezogen wurde, fehlt; (2) die Angabe über die Dauer des Bezugs fehlt; (3) die Angabe über die Höhe des bezogenen Betrags fehlt. Eine Übersicht über die Vorgehensweise gibt das *Ablaufdiagramm*. Fehlt die Angabe darüber, ob eine Einkommenskomponente bezogen worden ist, so wird zunächst versucht, die Angabe aus anderen Daten des Datensatzes abzuleiten. Herangezogen werden kann beispielsweise der Beschäftigungskalender. Sind Angaben aus dem Vorjahr vorhanden, so können diese als Proxy für die fehlende Angabe des aktuellen Jahres genommen werden. Ist es nicht möglich, aus anderen Variablen des Datensatzes (bzw. der vorjährigen Datensätze) abzuleiten, ob die Einkommenskomponente bezogen wurde oder nicht, so wird davon ausgegangen, dass sie nicht bezogen wurde. Bei fehlenden Angaben über die Anzahl der Monate, in denen eine Einkommenskomponente bezogen wurde, wird zunächst wiederum versucht, diese Information aus anderen Variablen des Datensatzes zu erschließen. Hierbei wird insbesondere der Beschäftigungskalender verwendet bzw. mit der Bezugsdauer anderer Komponenten verglichen. Kann auch hierdurch nicht die Bezugsdauer ermittelt werden, so wird ein Zufallswert imputiert.[5])

Fehlt die Angabe über den Betrag einer Einkommenskomponente, ist die Vorgehensweise komplexer. Prinzipiell haben die Befragten mehrere Möglichkeiten, Angaben zum Betrag einer Einkommenskomponente zu machen: Entweder wird der (exakte) Brutto- und/oder Nettobetrag angegeben, oder es wird eine Einkommensstufe angegeben.[6])

[5]) Vgl. zu dieser Vorgehensweise auch die Dokumentation zum ECHP Doc.Pan 164/00 „Imputation of income in the ECHP".
[6]) Hierzu werden den Befragten in der Befragungssituation Karten mit den Einkommensstufen vorgelegt, um sie zu unterstützen, sich an die Einkommenshöhe zu erinnern. Außerdem werden hierdurch Befragte, die vor der Angabe eines exakten Betrages zurückscheuen, dazu animiert, dennoch eine Angabe zumindest zur ungefähren Höhe des Einkommens zu geben.

Der Umgang mit fehlender Einkommensinformation



k.A.= keine Angabe. - B/N= Brutto/Netto. - N/B= Netto/Brutto.

Grundsätzlich werden in der Befragung zunächst Brutto- und dann Nettowerte erfragt. In Einzelfällen werden nur Nettobeträge erfragt. Nur Nettobeträge werden imputiert, fehlende Bruttowerte werden durch die Netto-Brutto-Rechnung ermittelt. Wird nur der Bruttobetrag, aber nicht der Nettobetrag angegeben, so wird der Nettobetrag mittels Brutto-Nettorechnung ermittelt. Wird eine Einkommensstufe angegeben, so wird ein Zufallswert aus der Verteilung der empirischen Werte der jeweiligen Einkommensstufe als Schätzwert für die Betragshöhe eingesetzt („Stufenimputation"). Die Ermittlung der Verteilung innerhalb der Stufe erfolgt automatisch mittels eines Makros, sodass auch Veränderungen der Einteilung der Stufen zwischen den Erhebungsjahren einfach integriert werden können.

Wird hingegen überhaupt keine Angabe zur Einkommenshöhe gegeben, so wird der fehlende Wert imputiert. Dabei wird zunächst - wenn Brutto- und Nettobeträge erfragt werden - der jeweilige Nettowert imputiert und der korrespondierende Bruttowert mittels Netto-Brutto-Rechnung berechnet. Die Imputation von Beträgen erfolgt auf zwei grundsätzlich verschiedene Arten: durch Längsschnitt- und durch Querschnittimputationen. *Längsschnittimputationen* verwenden zur Schätzung des fehlenden Werts in der aktuellen Welle Informationen aus den vorangegangenen Wellen, *Querschnittimputationen* nutzen die Informationen der jeweils aktuellen Welle. Insgesamt müssen nur verhältnismäßig wenige Werte solcherart imputiert werden: Beispielsweise wurden 2005 bei Unselbständigen-Einkommen etwa 3% imputiert, wobei ein Großteil durch Längsschnittimputationen erfolgte (ca. 3/4 aller zu imputierenden Unselbständigen-Einkommen).

Die Längsschnittimputation in EU-SILC basiert auf der von Roderick Little und Hong-Lin Su vorgeschlagenen Methode der „row-and-column"-Imputation.[7]) Diese Methode hat den Vorteil, dass sie relativ einfach zu implementieren ist und dass für die einzelnen Variablen keine gesonderten Annahmen getroffen werden müssen. Die Programmierung erfolgt wiederum mittels eines SPSS-Programm-Makros. Diese Methode ermittelt einerseits die Relation der Vorjahreswerte der Fälle mit fehlenden Werten zu den Vorjahreswerten der Fälle mit vollständiger Information (Spalteneffekt) und andererseits die Relation der jeweiligen Welle zu allen anderen Wellen. Erstere Relation wird als Spalteneffekt und letztere als Reiheneffekt bezeichnet. Diese Effekte werden zur Sortierung des Datensatzes verwendet, und der dem Fall mit dem fehlenden Wert nächstliegende Fall wird als Spender herangezogen. Der Spenderwert wird mit einem Residuum versehen, um die Reduktion der Varianz, die durch die Imputation ohne Varianz entstehen würde, zu vermeiden.

Liefert die Längsschnittimputation keinen Schätzwert für fehlende Werte (weil es keine Vorjahreswerte gibt), so werden Querschnittimputationen verwendet. Hier wird versucht, für jede Einkommenskomponente auf der Basis linearer Regressionsmodelle Schätzwerte zu berechnen, die mit einem stochastischen Residuum versehen werden. Diese Modelle werden vorab entwickelt. Um sicherzustellen, dass auch bei fehlenden Werten einer Prädiktorvariablen Schätzwerte berechnet werden können, werden für jede Einkommenskomponente mehrere Modelle spezifiziert. Die Prädiktorvariablen werden dabei aufgrund ihrer inhaltlichen Relevanz und ihrer Vorhersagekraft ausgewählt. Kann für eine Variable kein angemessenes Regressionsmodell spezifiziert werden (etwa weil zu wenige empirische Werte zur Verfügung stehen), so werden die Schätzwerte auf der Basis des Medians berechnet. Diese werden wiederum mit einem stochastischen Störterm versehen.

Vom Datensatz zur Auswertung

Die Gewichtung ist einstweilen nicht in de SPSS-Steuerungssyntax integriert und wird unabhängig davon durchgeführt. Die Berechnung der Gewichtungsvariablen folgt den Vorgaben von EUROSTAT und besteht aus drei Schritten: Die Berechnung der Designgewichte dient dazu, den Effekt des Designs der Stichprobe auf die Auswahlwahrscheinlichkeit der Elemente der Grundgesamtheit auszugleichen. Das Ziel der Non-Response-Gewichtung ist es, den Bias, der durch den Ausfall von Haushalten aus der Stichprobe entsteht, zu korrigieren. Die Anpassungsgewichtung, die auf der Basis des Produkts des Designgewichts und der Non-Response-Gewichte durchgeführt wird, dient der Erhöhung der Genauigkeit der Daten und der Hochrechnung.

Aus den nunmehr plausibilisierten, vervollständigten und gewichteten Datensätzen werden Zielvariablen sowie Auswertungsvariablen und Indikatoren erstellt. Ein umfassender Kontroll- und Auswertungsprozess soll die Qualität der Daten sicherstellen. Die beiden Arbeitsschritte Kontrolle und Auswertung lassen sich dabei nicht voneinander trennen. Die Ergebnisse werden mit den vorjährigen Ergebnissen aus EU-SILC verglichen, vor allem im Hinblick auf Laeken-Indikatoren und Tabellen über Einkommen, Armut und soziale Ausgrenzung. Ebenso werden aber auch umfangreiche Vergleiche mit externen Datensätzen durchgeführt (mit Daten der Lohnsteuerstatistik, der VGR, des Mikrozensus, etc.). Auswertungen und Indikatoren sind in EU-SILC somit nicht allein das Ziel, sondern werden auch als unerlässliches Mittel der Qualitätssicherung eingesetzt. Bevor eine Tabelle oder

[7]) Little, Roderick J.A./Su, Hong-Lin (1989); „Item Non-response in Panel Surveys". In: Kasprzyk/Duncan/Singh; Panel Surveys. New York; John Wiley, 1989. Vgl. auch Butrica, Barbara (1996); „Imputation Methods for Filling in Missing Values in the PSID-GSOEP Equivalent File 1980-1994", Working Paper, DIW, und die Dokumentation des ECHP in Österreich „ECHP Imputationen Schlüsselvariablen 2001", ICCR International.

eine Zahl veröffentlicht werden kann, muss sie von allen Seiten durchleuchtet werden. Auch EUROSTAT leistet hierbei wertvolle Arbeit, indem sowohl die nationalen Mikrodaten eingehend geprüft, als auch Indikatoren nachgerechnet und zwischen den Ländern verglichen und alle Unklarheiten an die nationalen Statistikämter rückgemeldet werden.

Als weitere Maßnahme zur laufenden Verbesserung von EU-SILC werden die Mikrodaten interessierten Nutzern in einer vollständig anonymisierten und teilweise aggregierten Form - auch die Aufbereitung dieser sog. Nutzerdaten ist in die Struktur integriert - gegen ein geringes Entgelt für eigene Auswertungen zur Verfügung gestellt. Feedback zu den Metainformationen der Erhebung (Qualitätsberichterstattung, Fragebögen etc.) sowie zu den Daten selbst kann den Prozess der Erhebung und Datenaufbereitung weiter verbessern.

Exkurs: Die technische Umsetzung

Programmierung

Sämtliche Schritte werden in SPSS-Syntax-Programmen am PC umgesetzt und gleichzeitig so dokumentiert. Die einzelnen oben genannten Module werden durch eine Haupt-Steuerungssyntax aufgerufen; sie selbst sind wiederum in einzelne Arbeitsschritte untergliedert, für die es jeweils eine eigene Syntax gibt. Oft erfolgt die Untergliederung nach den einzelnen Einkommenskomponenten, also beispielsweise gibt es eine eigene Syntax für Querschnittimputationen bei Unselbständigen-Einkommen, Pensionen, Arbeitslosenleistungen, Krankenleistungen usw. So bleibt das System trotz seiner Größe und Komplexität überschaubar.

Besonders nützliche Features in SPSS sind Makros, die eine noch weitergehende Automatisierung und Vereinheitlichung ermöglichen. *Makros* sind Befehle, die ihrerseits Befehlssyntax erstellen und anwenden. So werden beispielsweise längere Pfadnamen durch einen einfachen Platzhalter ersetzt, in unserem Fall etwa *!data05* für einen mehrgliedrigen Pfad zum Datendateienordner P:\SILC\EU-SILC-2005\Daten. Oder: Es werden in einem Check die Relationen der Bruttoeinkommen zu den Nettoeinkommen überprüft (siehe unten).

Um die Qualität und Systematik der Programme bei der Bearbeitung durch mehrere Personen zu garantieren, müssen gewisse Standards bei der Programmierung beachtet werden:
- Werden die gemeinsamen Dateien und Pfade verwendet? Hilfsfiles sind nur temporär auf der eigenen Festplatte abzuspeichern, hingegen müssen alle Files, die der Dokumentation dienen oder weiterverarbeitet werden müssen, auf dem Netzwerklaufwerk zugänglich sein.
- Ist die Syntax ausreichend beschriftet (Titel, Funktion, Ersteller, Datum) und kommentiert, d.h. ist es für andere möglich nachzuvollziehen, was getan wird und warum? (Erwartete) Änderungen in Folgejahren bzw. vorgenommene Veränderungen gegenüber Vorjahren sind bestmöglich zu dokumentieren.
- Sind alle Variablen nötig und sinnvoll? Es sollen keine Variablen doppelt gebildet werden, reine Hilfsvariablen müssen wieder gelöscht werden.

Konvention der Variablennamen

Damit ein einmal erstelltes Programm auch mehrere Jahre bzw. Erhebungen lang angewendet werden kann, ist es notwendig, dass Variablennamen und -inhalte nahezu unveränderlich sind. Verbesserungen, Ergänzungen und Modifikationen sind aber in jedem Fragebogen, selbst in Längsschnitterhebungen, unabdingbar.[8])

Unterschieden werden Erhebungs- und abgeleitete Variablen. *Erhebungsvariablen* sind die tatsächlich befragten bzw. anders in der Erhebung gesammelte Informationen, *abgeleitete Variablen* sind im Nachhinein berechnete Variablen für Plaus-, Auswertungs- oder Kennzeichnungszwecke sowie die EUROSTAT-Zielvariablen.[9]) Die Erhebungsvariablen sind im Gegensatz zu den Auswertungsvariablen nicht mit sprechenden Namen bezeichnet, sondern mit Buchstaben und Ziffern, die aber ihrerseits - bei Kenntnis der Namenskonventionen, die im Folgenden erläutert werden soll - einen Sinn erhalten.

Programmtechnische Voraussetzungen, wie alle Variablennamen mit einem alphanumerischen Feld zu beginnen und maximal acht Zeichen zu vergeben, bilden die Basis.[10]) Die letztendliche Entscheidung, wie die Variablennamen gebildet werden, wurde aber anhand der Variableninhalte getroffen. Den zentralen Fragebogenteil stellt die Erhebung der Einkommen dar, wobei sich mit geringen Unterschieden je nach Einkommenskomponente die Struktur der Fragen wiederholt:
1. Wird eine Einkommenskomponente bezogen, ja oder nein?
2. Wenn ja: Wie oft, wie viele Monate lang,…?
3. Wie hoch ist der Bruttobetrag?
4. Wie hoch ist der Nettobetrag?
5. Wenn weder Brutto- noch Nettobetrag genannt wurden: Bitte um Einschätzung des Einkommens entlang einer Skala mit vorgegebenen Kategorien (Interviewer zeigt Karte).

[8]) Veränderungen von Codierungen oder neue Variablen werden in einer Liste festgehalten, die jeweils den Status quo mit dem vorangegangenen Erhebungsjahr vergleicht. So kann im Zuge der jährlichen Modifikationen der Programme darauf Rücksicht genommen werden.
[9]) EUROSTAT sieht Zielvariablen vor, die nur teilweise genau den österreichischen Erhebungsvariablen entsprechen; meist ist die Abfrage viel detaillierter als gefordert, und die Zielvariablen werden erst nach Plaus, Imputation und Aggregierung dieser Detailvariablen berechnet.
[10]) Obwohl aktuelle SPSS-Versionen schon längere Variablennamen zulassen, wird aus Kompatibilitäts-Gründen diese Regel bei den Standard-Variablen eingehalten.

Diese typische Abfolge und die Notwendigkeit auf den ersten Blick (also ohne im Codebook nachblättern zu müssen) zu erkennen, um welche Variable entsprechend den oben angeführten Fragen es sich handelt, wird in folgender Namenskonvention sichtbar *(Übersicht 1)*:

Bildung der Variablennamen — Übersicht 1

Besetzung und Bedeutung der einzelnen Stellen								
Stelle 1:	Bezeichnung des Datenfiles (D, R, H, K, oder P)							
Stellen 2-4:	Nummerierung der Frage entsprechend dem Fragebogen							
Stellen 5-6:	Nummerierung der einzelnen Items einer Frage							
Stelle 7:	Information über Variablentyp							
	0= keine Einkommensvariable							
	1-5= Einkommensfrage:							
	1= Bezug 2= Dauer des Bezugs 3= Brutto 4= Netto 5= Einkommensstufe							
Stelle 8:	kann frei belegt werden: z.B: F für Flagvariablen							
Beispiel P057024:								
Stelle	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Variable (Beispiel)	P	0	5	7	0	2	4	F
P.....	Die Variable stammt aus dem Personenfragebogen.							
P**057**...	Es ist die 57. Frage im Personfragebogen.							
P057**02**...	Es ist das zweite Item der Frage 57 .							
P057024	Es handelt sich um die Frage nach dem Nettoeinkommen.							
P057024**F**	Es handelt sich um die Flagvariable der Variablen.							

Zunächst ist für jede Variable erkennbar, aus welchem Fragebogenteil bzw. Datenfile sie stammt: *D* steht hierbei für das Haushaltsregister aller Haushalte, auch der nicht gültig befragten; *R* kennzeichnet das Personenregister mit Basisinformationen über alle Personen in befragten Haushalten; *H* entspricht dem Haushaltsfragebogen; *K* dem Kinderfragebogen für alle unter 16-Jährigen; und *P* ist der Personenfragebogen, der allen Personen ab 16 Jahren vorgelegt wird.

Beispiel für Anwendung der Variablennamen in einem Makro — Übersicht 2

define brunet (!pos !cmd).	←	Makrodefinition
!do **!a** !in (!1).	←	!a ist Platzhalter für die Bruttovariablen, diese werden eingegeben.
!let **!b** = !concat(!substr(!a,1,6),**4**).	←	!b ist Platzhalter für die Nettovariablen, die in den Stellen 1-6 den Bruttovariablen entsprechen und dann die Ziffer 4 haben. Diese werden nicht eingegeben sondern mit der Formel aufgerufen.
compute nebrpro = $sysmis.	←	Bildung einer Prüfvariablen
do if !a gt 0 and !b gt 0.	←	Prüfung wird durchgeführt, wenn Brutto- und Nettovariable einen Betrag größer 0 enthalten.
if !a lt !b nebrpro = 1.	←	Wenn Brutto- kleiner Nettobetrag, dann wird die Prüfvariable auf 1 gesetzt.
if !b lt 0.4 * !a nebrpro = 2.	←	Wenn Netto kleiner als 40% von Brutto ist, dann wird die Prüfvariable auf 2 gesetzt.
if !a eq !b nebrpro = 3.	←	Wenn Brutto gleich Netto ist, wird die Prüfvariable auf 3 gesetzt.
end if.	←	Ende der Prüfung
temp. sel if nebrpro >= 1. list var pid nebrpro !a !b int.	←	Alle Fälle, in denen die Prüfvariable einen Wert größer gleich 1 aufweist werden in einem Listing ausgegeben.
!doend. exe.		
!enddefine.	←	Ende der Makrodefinition

Die Stellen 2 bis 4 (Fragennummer) bzw. 5 bis 6 (Nummerierung, wenn eine Frage in mehrere Unterfragen gegliedert ist) erleichtern die Verknüpfung von Fragebogen und Codebook, indem sie die gleiche Nummerierung für die Fragen wie die zugehörigen Variablen vorsehen. Wesentlich für die effiziente Programmierung sind die Stellen 7 und 8. Das oben bereits genannte Beispiel der Überprüfung von Brutto-Netto-Relationen macht dies deutlich *(Übersicht 2)*. Der Konvention der Variablennamen folgend, haben alle Bruttoeinkommensbeträge die Endziffer 3 und die Nettoeinkommensbeträge die Endziffer 4. Zur Durchführung der Prüfung ist nur mehr der Aufruf des Makros „brunet“ unter Angabe der Bruttovariablen nötig.

Flags als Wegweiser durchs System

Neben der Kennzeichnung der Variableninhalte durch standardisierte Benennungen hat es sich auch als notwendig erwiesen, zu kennzeichnen, was mit den Einkommensvariablen im Zuge ihrer Bearbeitung passiert. Denn sowohl im Einzelfall als auch auf Ebene der Variablen ist es wichtig zu wissen, ob ein Wert direkt aus der Befragung kommt, ob er verändert oder imputiert wurde. So kann beispielsweise für jede Einkommensvariable der Anteil der imputierten Werte ausgewertet werden. Zwar werden nach Durchlauf jedes der oben genannten Module Zwischendatenfiles abgespeichert, mit Hilfe derer sich diese Information ebenfalls rekonstruieren ließe; dies wäre aber während des Datendurchlaufs vielfach zu aufwendig, sodass eine Kennzeichnung im gleichen File sinnvoller erscheint.

Praktisch umgesetzt wird diese Kennzeichnung mit Flag-Variablen. Jeder Einkommensvariable wird ein „Fähnchen“ in Form einer zweiten Variable des gleichen Namens, jedoch mit dem Suffix „F“ zur Seite gestellt, das Metainformationen zur eigentlich interessierenden Variablen enthält. Folgende Codes werden zur Zeit vergeben:

-2 nicht zutreffend
-1 keine Angabe und nicht imputiert
1 Betrag lt. Befragung
2 Betrag aus Stufe
3 Betrag durch Brutto-Netto- oder Netto-Brutto-Rechnung
4 Betrag durch deduktives Verfahren eingesetzt
5 Betrag imputiert (Längsschnitt)
6 Betrag imputiert (Querschnitt)
7 Betrag nachträglich korrigiert
8 Betrag aus Angabe eines Monatseinkommens (bei Jahreseinkommensvariablen anzuwenden).

Eine fehlende Angabe ist zu Beginn des Prozesses mit -1 gekennzeichnet, sollte sich aber im Zuge des Datendurchlaufs in einen Wert größer 1 verwandeln, je nachdem welche Möglichkeiten zur Ersetzung des fehlenden Wertes gegeben sind. Freilich sind auch hier die Flags jeweils nur ein

Abbild des momentanen Standes. Diese Information ist aber für die Standardprozeduren und zur Dokumentation ausreichend.

Resümee

Das Datenmanagement von EU-SILC, wie eben vorgestellt, wurde anhand der Erhebung EU-SILC 2004 implementiert und fand bzw. findet auch bei EU-SILC 2005 und 2006 in bereits mehrfach adaptierter und erweiterter Form Anwendung. Neben der programmtechnischen Verbesserung steht in den nächsten Jahren die inhaltliche Weiterentwicklung vor allem in Bezug auf den Längsschnitt im Vordergrund (im Frühjahr 2007 sind erstmals Längsschnittfiles an EUROSTAT zu schicken). Auch die Dokumentation der Datenaufbereitungsarbeiten in textlicher Form zusätzlich zur Syntax sowie die Beschreibung der Auswertungsvariablen stellen noch zusätzliche wichtige Aufgaben dar.

Summary

This paper gives an insight into the development of survey data in EU-SILC from the time a question is asked in a household and keyed into the CAPI-laptop to the final statistical analysis. Four years of EU-SILC in Austria so far and the requirement to produce consistent micro-datasets, fully checked, imputed and weighted have brought up numerous challenges. We try to convey how some of them were met in a data management perspective. Streamlining the data editing process whilst at the same time improving data quality is the foremost goal of this system. It is implemented in SPSS command syntax and set up in a modular way. Variable names have to follow a certain logic, flag variables act as guides through the system and certain standards of programming have to be met. Recoding of illogical values, imputation - cross-sectional as well as longitudinal methods are used - and weighting are major tasks to arrive at a data-set ready for analysis and calculating indicators. Quality management foresees extensive comparisons to other EU-SILC years and external data in order to gain reliable results. Although much work has to be done to further improve the data management system it is an already functioning system and has stood the test for EU-SILC 2004 and 2005.