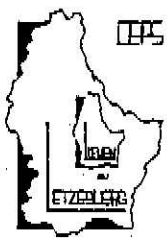
PANEL SOCIO-ECONOMIQUE

"LIEWEN ZU LETZEBUERG"

Document PSELL N° 17

ORGANISATION DER DATEN DES LUXEMBURGER HAUSHALTSPANELS

[Eingabe, Speicherung und Analyse von Paneldaten]



G. Schmaus

Document produit par le

CENTRE D'ETUDES DE POPULATIONS, DE PAUVRETE
ET DE POLITIQUES SOCIO-ECONOMIQUES

C.E.P.S./INSTEAD

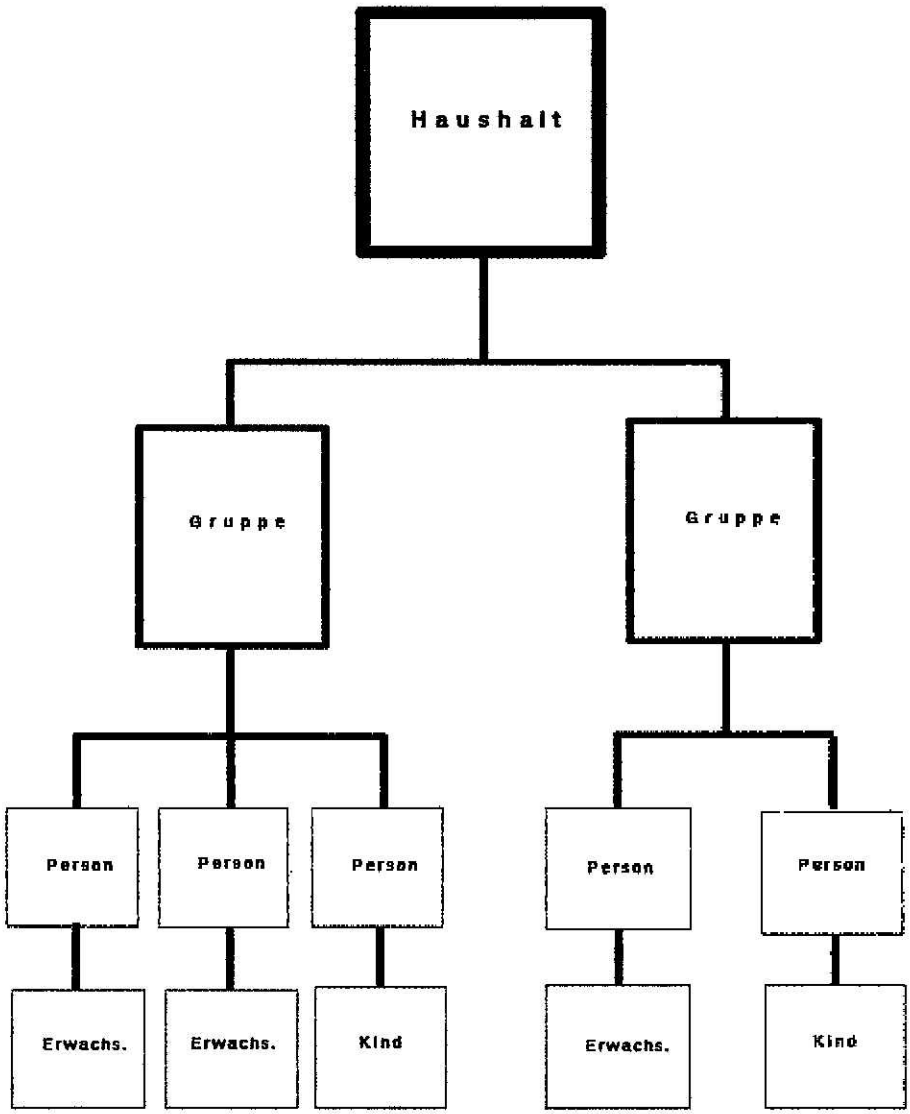
B.P. 65 L-7201 Walferdange
Tél. (352) 33 32 33 - 1

Président: Gaston Schaber

---

1 9 9 0

Uebersicht 1 : **Hierarchie der Paneldaten**



0. Problemstellung

Das sozio-ökonomische Panel "*Leben in Luxemburg*" ist eine Haushaltsstichprobe mit zahlreichen Variablen zu den Bereichen Demographie, Wohnungsausstattung, Gesundheit, Ausbildung und Einkommen. Der Sachverhalt, dass in einem Haushalt mehrere Personen miteinander leben und teilweise ihre Einkommen teilen, findet in der Fragebogengestaltung seinen Ausdruck. Es gibt Fragen auf Haushaltsebene, Fragen zu Einkommensgruppen und Fragen über einzelne Haushaltsmitglieder. Bei den Fragebogenteilen für einzelne Haushaltsmitglieder werden jeweils getrennte Fragen für Kinder und Erwachsene gestellt; ein Teil der Fragen wird an alle Personen im Haushalt gestellt.

In der Informatik spricht man bei einer solchen Datenstruktur von einem hierarchischen File. Das Panel besitzt die drei Hierarchiestufen Haushalte, Gruppen und Individuen, wobei die Individuen weiter in Kinder und Erwachsene unterteilt werden (Uebersicht 1). Aufgrund der mehrfachen Hierarchien ist das Panel ein komplexes File. Gleichzeitig handelt es sich auch um ein sehr grosses File in Bezug auf den Variablenumfang, weil auf jeder Hierarchieebene (Haushalte/Gruppen/Individuen) teilweise sehr detaillierte Fragen gestellt werden. Bei den Einkommensfragen werden Informationen für jeden Monat, differenziert nach bis zu 24 verschiedenen Einkommensarten erhoben.

Ein sozio-ökonomisches Panel zeichnet sich von anderen Querschnittsstichproben dadurch aus, dass der zugrunde liegende Fragebogenkatalog bei identischen Untersuchungseinheiten mehrmals hintereinander erhoben wird. Beim Luxemburger Panel wird jeweils jährlich eine Erhebung durchgeführt. Bis Ende 1988 liegen vier Wellen vor. Im Gegensatz zu einmaligen statistischen Erhebungen treten bei Panelstichproben Sonderfälle auf, die Berücksichtigung finden müssen. Dies sind zum einem die Fälle, in denen Personen sich weigern, an späteren Befragungswellen teilzunehmen oder aus Luxemburg fortziehen. Zum anderen treten Sonderfälle

auf, wenn Individuen sterben, neue Personen geboren werden. Eine zusätzliche Besonderheit ergibt sich aus dem Sachverhalt, dass aus einem Haushalt Personen ausscheiden können, die ihrerseits einen neuen Haushalt bilden können (sogenannter Splitoffs). Diese Sonderfälle müssen bei der Datenerhebung mitberücksichtigt werden und Vorkehrungen dafür getroffen werden, dass im Zeitvergleich auch eine Analyse bei denjenigen Personen, welche z.B. den Haushalt einmal gewechselt haben, möglich ist. Da im Panel gewisse Fragen nach Kindern und Erwachsenen unterschieden werden, ergeben sich weitere Schwierigkeiten, wenn die Lebensverläufe von Personen untersucht werden sollen, für die früher Kinder- und später altersgemäss Erwachsenenfragebögen vorliegen.

Jede einzelne Welle des Panel lässt sich wie eine Querschnittsstichprobe auswerten. Der eigentliche Wert einer Panelstichprobe ergibt sich aber erst dann, wenn die Wellen miteinander verbunden und ausgewertet werden. Um die Wellen in dieser Weise nutzen zu können, muss eine Zusammenführung (Merge/Match/Link) durchgeführt werden. Aufgrund der weiter oben erwähnten Sonderfälle (Splittoff, Tod und Geburt) ist dies ein komplizierter Prozess und erfordert weitere Arbeitsschritte als nur ein einfacher Merge von Computerfiles.

In diesem Arbeitspapier wird dargestellt, wie die Daten des Panel nach Durchführung des Interviews, nach der Phase der manuellen Ueberprüfung und der Vercodung auf der Grossrechenanlage des Centre Informatique de l'Etat (CIE) organisiert sind. Es wird erläutert, wie die Daten korrigiert und gemergt werden und wie sie für die statistische Auswertung bereitgestellt werden.

Insbesondere zu dem Problembereich der Analyse von Paneldaten können hier nur Empfehlungen ausgesprochen werden. Die Praxis der Auswertungen wird zeigen, inwieweit Modifikationen und Verbesserungen in der Organisation und dem Retrieval der Daten in Zukunft notwendig werden.

1 Datenmanagement

Die Daten des Panels müssen so organisiert sein, dass die Dateneingabe, die Pflege der Daten und die Analyse der Daten erleichtert wird. Weitere Rahmenbedingungen müssen bei der Organisation erfüllt sein:

- Die Hierarchien der Paneldaten müssen abbildbar sein
- Die Wellen müssen sich mergen lassen
- Sonderfälle des Panels (Splitoff, Geburt, Tod) müssen leicht integrierbar sein.

Weiterhin sollen die Daten on-line zur Verfügung stehen. Um ad hoc Fehlertests und Korrekturen durchführen zu können, soll auf einzelne Records der Datei interaktiv zugegriffen werden können. Eine Analyse der Daten soll nicht an die Speicherart der Daten gebunden sein. Benutzerinterfaces zu Statistikpaketen wie SPSSX, LIMDEP etc. sollen möglich sein. Das Retrieval von Daten soll effizient sein. Wünschenswert ist auch ein System, welches benutzerfreundlich arbeitet und eine mächtige Abfragesprache beinhaltet.

In den Datenbanken des relationellen Typs können viele der aufgestellten Forderungen erfüllt werden. Das Panel wurde deshalb in der SQL-Datenbank von IBM (SQL/DS) gespeichert. Die Auswertung erfolgt derzeit vorwiegend durch das Statistikpaket SPSSX. Eine alleinige Speicherung der Daten als Systemfiles von Statistikpaketen ist nicht ausreichend, wenn mehrere Wellen des Panels gemergt und aggregiert werden müssen.

1.1 Organisation der Daten in der Datenbank

Die Datenbanken des relationellen Typs organisieren ihre Datenbestände intern in sogenannten Tabellen (Tables). Eine Tabelle besteht aus Zeilen und Spalten. Jede Spalte hat einen Namen und entspricht einer Variable in sequentiellen Files. Für jede Einheit gibt es in jeder Tabelle eine Zeile. Im Rahmen der Datenbankabfragen kann bei Bedarf auf jede Zeile und Spalte (gegebenenfalls auf mehrere Spalten gleichzeitig) gezielt zugegriffen werden. Um diesen Zugriff ausführen zu können, muss die Abfragesprache 'Structured Query Language' (SQL) verwendet werden.

Die Daten jeder Panelwelle sind in jeweils fünf unterschiedlichen Tabellen abgespeichert. Jede Hierarchie bzw. der entsprechende Fragebogenteil ist einer bestimmten Tabelle zugeordnet. Die Tabellen im SQL werden in der hier gewählten Organisationsform mit den Anfangsbuchstaben M,G,I,C,D gekennzeichnet. Das Jahr der Welle wird zusätzlich zu den Anfangsbuchstaben zur Unterscheidung der einzelnen Wellen angehängt.

Zuordnung der Fragebogenteile zu den SQL-Tabellen:

Haushaltsfragen	====>	Tabelle M
Familien Tableau	====>	Tabelle I
Zugänge		
Abgänge		
Gruppenfragen	====>	Tabelle G
Kinderfragen	====>	Tabelle C
Erwachsenenfragen	====>	Tabelle D

Beispiel:

Haushaltsfragen für das Jahr 1985 ===> Tabelle M85
Haushaltsfragen für das Jahr 1986 ===> Tabelle M86

Die SQL-Tabelle 'I' (Familien Tableau) hat zentrale Bedeutung für jede Welle. Variablen aus dieser Tabelle erlauben die Zuordnung von Personen zu Gruppen. Dieser Tabelle ist auch zu entnehmen, ob für ein Individuum ein Kinder- oder Erwachsenenfragebogen vorliegt. Somit ermöglicht das Familien Tableau die Tabellen für die Gruppen, die Kinder und die Erwachsenen miteinander zu verbinden. In der Tabelle für das Familien Tableau befinden sich auch die Diagnostikvariablen. Die Diagnostikvariablen geben Auskunft über die Historie von Personen (Zugänge, Abgänge), die zwischen zwei Wellen ausgeschieden oder neu hinzugekommen sind. Für alle Personen dieses Typs ist festgehalten, weshalb sie ausgeschieden sind (z.B. Tod, Verlassen des Haushalts (Splitoff), für alle neuen Personen, weshalb sie in den Haushalt eingetreten sind (z.B. Zuzug, Geburt).

Alle fünf Tabellen für eine Welle enthalten mindestens eine Identifikationsvariable: Dies ist die Haushaltsnummer. Zusätzlich dazu (Ausnahme Tabelle M) ist mindestens eine weitere Identifikationsvariable abgespeichert. Für die Tabelle 'I', 'C' und 'D' (Gruppen/Familien Tableau/Kinder/Erwachsene) sind dies die Matrikelnummern der Individuen, bei der Tabelle 'G' ist dies die Gruppennummer. Diese Identifikatoren sind notwendig, wenn auf Informationen aus mehreren Tabellen in einer Welle zugegriffen werden soll. In der Sprachregelung der Datenbank SQL spricht man hier vom 'Joinen' von Tabellen. Die Identifikatoren der Tabellen sind die sogenannten 'Join-Bedingungen'.

Die Haushaltsnummern und die Matrikelnummern der Individuen gewährleisten, dass identische Individuen und Personen zusammengeführt werden. Diverse Prüfroutinen müssen in der Phase der Dateneingabe und der Fehlerbereinigung gestartet werden, um die inhaltliche Identität von Haushalten und Personen entsprechend ihrer Nummer zu überprüfen.

Um die Performance der Abfragen aus dem SQL zu steigern, sind für jede Tabelle Indexe vereinbart worden. Dies sind in der Regel die Identifikatoren, die in den sogenannten 'Join-Bedingungen' eingehen.

1.2 Nomenklatur der Variablen

Alle Namen für die Variablen des Panels und damit auch alle Variablen einer entsprechenden SQL Tabelle werden entsprechend einer Namenssystematik vergeben. Im Variablennamen wird festgehalten, auf welche Hierarchie und auf welche Welle sich die Variable bezieht. Der Variablenname gibt auch Information darüber, ob es sich um eine Link- oder um eine normale Variable handelt und ob es sich um eine sogenannte Vektorvariable handelt.

Die jeweils erste Stelle des Variablennamens gibt das Niveau der Variablen in der Hierarchie wieder.

M ===> Haushalt
G ===> Gruppe
I ===> Individuum

Die nächsten zwei Stellen bezeichnen das Jahr der Welle.

85 ===> 1985
86 ===> 1986
87 ===> 1987

Die weiteren Stellen des Variablennamens werden verwendet, um die Variablen innerhalb ihrer Hierarchie unterscheiden zu können. In der Regel werden dazu Nummern verwendet. Ein kleiner Teil der Variablen, welche zum Linken und Identifizieren Verwendung finden, besitzt vor der Nummer noch ein 'L' (L=Link).

Beispiele für Variablennamen:

M87057	===>	Anzahl der Autos im Haushalt 1987
M86L01	===>	Haushaltsnummer des Haushalts 1986 (Link-variable)
I85L09	===>	Geschlecht des Individuums 1985 (Link-variable)
G87227	===>	Kreditbetrag, den die Gruppe zahlen muss 1987

In Sonderfällen wird an das Ende des Variablennamens ein Buchstabe angehängt. Hiermit wird gekennzeichnet, wenn die Definition der Variablen geändert wurde.

Bestimmte Informationen, die für jeden Monat eines Jahres erfragt werden, werden im Fragebogen des Panels in Listenform erhoben. So wird z.B. in einer Tabelle festgehalten, welches Einkommen aus welcher Einkommensart - differenziert nach 24 Einkommensarten - erzielt wurde. Informationen diesen Typs werden transformiert in Vektorform abgespeichert. In der Nomenklatur der Variablen werden diese Vektoren durch zwei Buchstaben gekennzeichnet. Ein dritter Buchstabe wird vor den Vektornamen gesetzt und identifiziert das Referenzjahr (X = Vorjahr, Y = laufendes Jahr). An den Vektornamen wird eine Zahl angehängt, welche den Referenzmonat kennzeichnet (1 = Januar, 2 = Februar, 12 = Dezember).

Beispiele für Variablennamen von Vektoren:

I87XAB5 ===> a) Vektornamen: AB
b) Information für das Vorjahr = 1986: X
c) Information für Monat Mai: 5

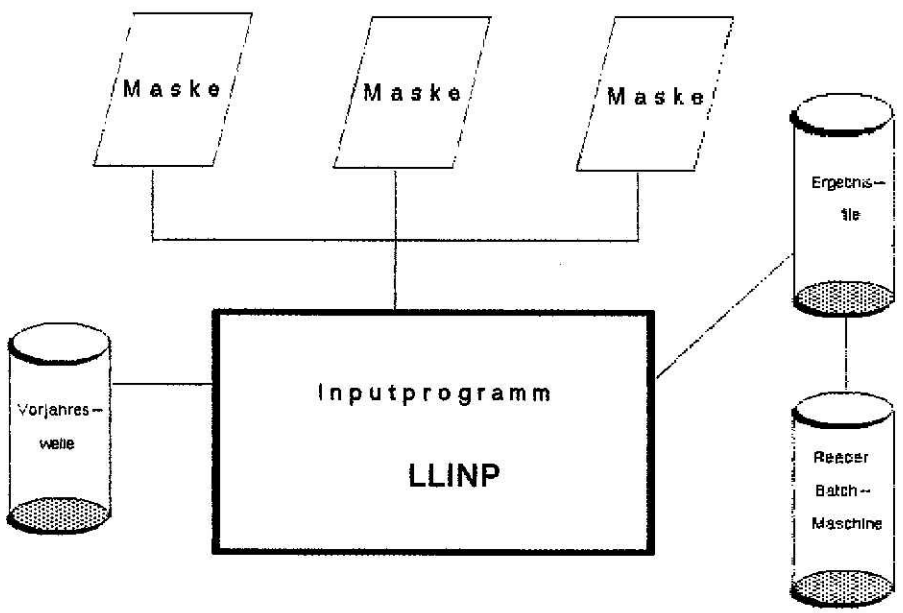
I87YAB1 ===> a) Vektornamen: AB
b) Information für laufendes Jahr = 1987: Y
c) Information für Monat Januar: 1

Aus den Variablen- und Vektornamen für Haushalte und Gruppen kann sofort erkannt werden, in welchen SQL-Tabelle die einzelnen Variablen abgespeichert sind. Bei den Variablen für die einzelnen Personen ist dies nicht immer direkt ersichtlich. Eine Zuordnung zu den Tabellen des SQL ist aber möglich, da die einzelnen Fragebögen des Panels mit nur wenigen Ausnahmen streng bestimmten Tabellen zugeordnet sind.

2 Inputprogramm für Paneldaten

Die Fragebögen für jede Welle des Panels werden mit Hilfe eines interaktiven Programms eingegeben (Uebersicht 2). Auf dem Bildschirm des Terminals erscheinen in sukzessiv aufeinander folgenden Bildschirmmasken die einzelnen Fragebogenteile. Das Outlay orientiert sich an den Fragebögen. Die Bildschirmmasken für die Eingaben werden dynamisch aufgerufen. Der Datentypist kann so nicht vergessen, Teile des Fragebogens einzugeben oder er erhält Hinweise, wenn Teile des Fragebogens fehlen oder mit Fehlern behaftet sind.

Uebersicht 2 **Inputprogramm**



Das Programm erleichert die Dateneingabe in den einzelnen Bildschirmmasken erheblich. Teilstücke des Fragebogens werden graphisch auf dem Bildschirm in schematischer Form dargestellt. Der Cursor des Bildschirms bewegt sich automatisch an die Stelle, an der als nächstes eine Eingabe erwartet wird. Innerhalb jedes Bildschirms werden umfangreiche Fehlertests durchgeführt. So wird überprüft, ob einzelne Felder leer bleiben dürfen, ob sie numerische Felder enthalten müssen und ob der numerische Wert in einem bestimmten Bereich liegt. Weiterhin werden die Beziehungen zwischen Variablen in einer Bildschirmmaske getestet. Dies sei an einem Beispiel gezeigt. Liegt z.B. ein Eigentümerhaushalt vor, so muss das Feld für die Miete frei bleiben und das Feld mit der Hypothekentilgung gefüllt sein.

Durch die Fehlertests können nur formale Fehler gefunden werden. Der Ursprung des Fehlers kann schon im Fragebogen vorhanden sein oder kann aus Tippfehlern bestehen. Beide Arten von Fehlern können durch das Programm erkannt werden.

Eingabefehler werden mit einer Fehlermeldung sofort angezeigt. Der Cursor zeigt auf die fehlerhafte Eingabe. Das Programm erfordert eine korrekte Eingabe, anderenfalls kann der Fragebogen nicht weiter eingegeben werden.

Die Reihenfolge der einzelnen Fragebogenteile ist die gleiche, wie sie der Interviewer beim eigentlichen Interview im Haushalt angewendet hat. Zusätzlich zu den erhobenen Variablen werden zuerst die vercodeten Variablen eingegeben. Die Reihenfolge der Bildschirmmasken ist demnach die folgende:

(1) Vercodete Variablen ('Fiche-Précodage')
    a) Diagnostikvariablen
    b) weitere Variablen
       z.B. Haushaltstypisierungen

(2) Allgemeine Fragen zum Interview
    ('Fiche-Enquêteur')

(3) Familien Tableau (Tableau familial)

(4) Zu- und Abgänge
    a) Zugänge
    b) Abgänge

(5) Haushaltsfragen

(6) Gruppenfragen

(7) Personenfragen
    a) Kinder
    b) Erwachsene

Als erste Dateneingabe erfolgt die Matrix mit den Diagnostikvariablen. In dieser Matrix sind alle Personen der Vorjahreswelle mit ihren Matrikelnummern gelistet mit dem Vermerk, ob und welche Personen hinzugekommen sind und welche Personen ausgeschieden sind. Das Programm überprüft diese Angaben mit den schon vorhandenen Angaben aus der Vorjahreswelle. Wird eine Identität festgestellt, ist damit gewährleistet, dass der Link in Bezug auf Haushalte und Personen mit der Vorjahreswelle fehlerfrei möglich ist. Wird keine Identität festgestellt, verweigert das Programm weitere Eingaben, bis eine solche hergestellt wird.

Die Diagnostikvariablen entscheiden darüber, welche Personen im Familien Tableau aufgeführt sein müssen. Alle Personen, welche in der vorigen und der aktuellen Eingabewelle vorhanden sind und alle Personen, welche neu in den Haushalt eingetreten sind, müssen einen Eintrag im Familien Tableau enthalten. Die Diagnostikvariablen determinieren auch, für welche Personen die Fragebögen für die Zugänge und für die Abgänge eingegeben werden müssen.

Nach Eingabe der weiteren vercodeten Variablen werden für jedes Haushaltsmitglied die Variablen aus dem Familien Tableau eingespeichert. Unter anderem wird eingegeben, zu welcher Einkommensgruppe die Individuen gehören und ob ein Kinder- oder Erwachsenenfragebogen ausgefüllt worden ist. Diese Variablen entscheiden darüber, wie oft nachfolgend Bildschirmmasken für Gruppen, Kinder und Erwachsene angezeigt und ausgefüllt werden müssen.

Nachdem eventuell die Fragebögen für Zugänge und Abgänge eingeben worden sind, erscheinen die Fragen für die Gruppen auf dem Bildschirm. Wenn ein Haushalt mehrere Gruppen umfasst, werden, sooft wie Gruppen vorhanden sind, weitere Gruppenfragebögen eingegeben. Für jede Gruppe wird erfragt, welche Mitglieder sie umfasst. Diese Eingaben werden mit den Angaben im Familien Tableau überprüft und eventuelle Fehler angezeigt.

Entsprechend der Reihenfolge aus dem Familien Tableau werden im nächsten Schritt die Fragebögen für die Kinder und die Erwachsenen auf dem Bildschirm angezeigt. Bei den Erwachsenen wird ein Teil der Fragen noch weiter danach unterteilt, ob es sich um einen Erwerbstätigen, Arbeitssuchenden, Teilzeitbeschäftigten oder um einen Nichterwerbstätigen handelt.

Eine wichtige Aufgabe des Programms besteht auch darin die Konsistenz der Eingaben für die Hierarchie des Haushaltes und seine Unterteilung in Gruppen, Personen, Kinder und Erwachsene sicherzustellen. Dies geschieht durch die dynamischen Panels. Die Organisationsstruktur der dynamischen Panels gewährleistet ein Maximum an fehlerfreien Eingaben. Dieses soll an zwei Beispielen gezeigt werden.

a) Sind die Diagnostikvariablen falsch eingetragen worden, wurde z.B. eine Person irrtümlich als alte Person (Personen bestehend in neuer und voriger Welle) und nicht als Abgang gekennzeichnet, fordert das Programm

Eingabe für eine Person im Familien Tableau. Angaben für diese Person fehlen im Fragebogen und die Person, welche die Dateneingabe vornimmt, kann diesen Fehler erkennen und korrigieren.

b) Liegt eine fehlerhafte Eintragung im Familien Tableau vor, wurde z.B. eine Person irrtümlich als Kind gekennzeichnet, so wird ein Kinderfragebogen für diese Person zur Eingabe angezeigt, ein solcher Fragebogen für diese Person ist nicht vorhanden, und die Person für die Dateneingabe erkennt den Fehler.

Ein Teil der Variablen für die Gruppen und die Individuen enthält Angaben für zwölf Monate in einem Jahreszeitraum. Diese zwölf Monatsangaben werden durch einen Algorithmus in eine Variable komprimiert. Ein solches Vorgehen bei diesen Monatsvariablen ist aus folgenden Gründen zweckmässig:

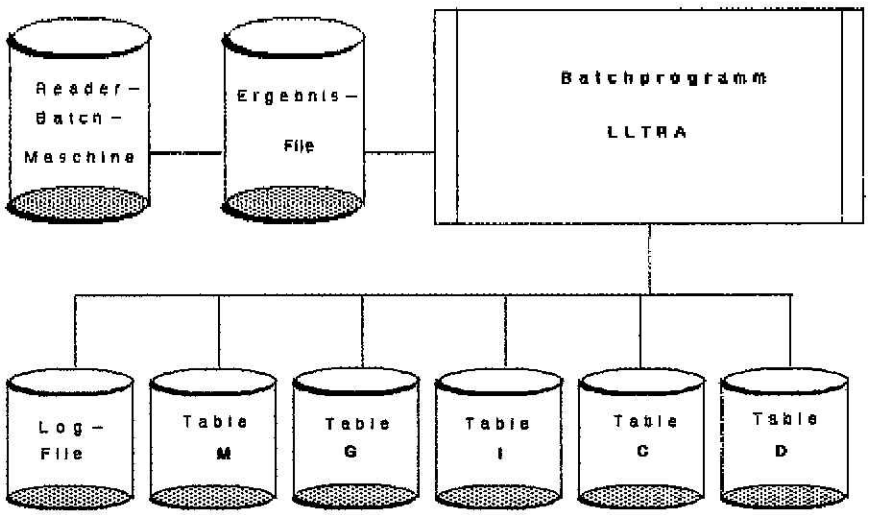
a) Kompression spart Speicherplatz

b) Anzahl der Variablen, die im SQL gespeichert werden, reduziert sich merklich (Ersparnis pro Welle circa 500 Variablen).

Das Paneleingabeprogramm erzeugt für jeden Haushalt jeweils ein separates Ergebnisfile. Das File besteht aus mehreren Teilen, die durch Identifikationsmarken abgetrennt sind. Es enthält einen Haushaltssatz, einen oder mehrere Sätze für die Gruppen und einen oder mehrere Sätze für das Familien Tableau, die Kinder und die Erwachsenen.

Am Ende des Eingabeprozesses wird dieses File von jeder Eingabemaschine an die Batchmaschine geschickt. In der Batchmaschine läuft ein Programm, das die einzelnen eingebenen Fragebögen sammelt (Uebersicht 3). Ueber ihren virtuellen Reader liest die Batchmaschine einen Fragebogen ein und splittet diesen entsprechend den fünf SQL-Tabellen in fünf Teilstücke auf. Diese Teilstücke werden an die fünf Zwischenfiles angehängt, welche zum Laden der Tabellen in SQL Verwendung finden. Zu Kontrollzwecken erzeugt das Programm in der Batchmaschine ein Logfile. Das Logfile zeichnet auf, welche Fragebögen eingegeben worden sind und hält fest, ob darunter fehlerhafte Fragebögen gefunden worden sind.

Uebersicht 3: **Batchprogramm LLTRA**



3 Zwischenprüfung der Daten

Während der Phase der Dateneingabe und bevor die Daten endgültig in die SQL-Datenbank abgespeichert werden, gibt es weitere Prüfschritte:

a) Die fünf Zwischenfiles werden zu Testzwecken in das SPSSX eingelesen. Dort können die Daten fallweise gelistet werden und manuell mit den Originaldaten in den Fragebögen verglichen werden. Zur weiteren Kontrolle werden im SPSSX einfache statistische Auswertungen durchgeführt. Für jede Variable werden das Minimum/Maximum und die Durchschnitte bestimmt und die minimalen und die maximalen Werte mit den zulässigen Wertebereichen verglichen. Systematische Fehler im Eingabeprozess können so leicht gefunden und korrigiert werden.

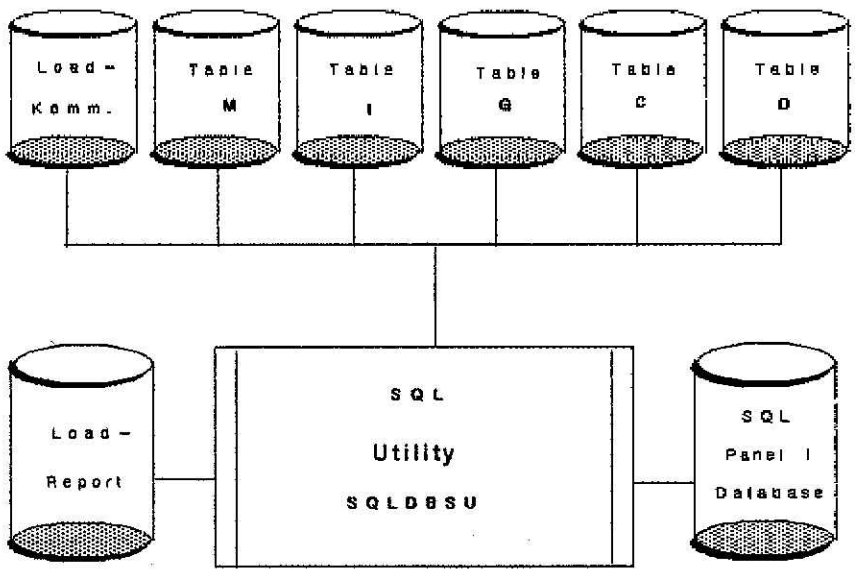
b) Es wird auch überprüft, ob in den Zwischenfiles doppelte Sätze (Haushalte,-Gruppen,-Personen) auftreten. Werden solche Fälle gefunden, werden die entsprechenden Sätze gelöscht, bevor der Ladevorgang in das SQL vorgenommen wird.

c) Die Konsistenz der Hierarchie wird in einem eigenständigen Fortranprogramm überprüft. Unter Konsistenz wird hier verstanden, dass keine Widersprüche zwischen den einzelnen Sätzen in der Hierarchie auftreten. So müssen die Angaben im Familien Tableau über Zugehörigkeit zu Gruppen und die Angaben, ob ein Kinder- oder Erwachsenenfragebogen vorliegen sollte, mit den Angaben bei den Gruppen, den Kinder- und Erwachsenenfragebögen übereinstimmen. Weiterhin dürfen keine Personen mehrfach auftreten. Jede Person darf nur einer Gruppe zugeordnet sein und für keine Person darf sowohl ein Kinder- als auch ein Erwachsenenfragebogen ausgefüllt sein.

4 Speicherung der Daten im SQL

Nach der Zwischenprüfung der Daten wird die neue Panelwelle in die entsprechenden Tabellen des SQL gespeichert (Uebersicht 3a). Für jede Tabelle steht ein separates Ladefile zur Verfügung. Nach dem Laden der Daten müssen aus Effienzgründen für das SQL sogenannte Indexe definiert werden. Je nach Tabelle werden unterschiedliche Indexe vereinbart:

Uebersicht 3a : **Laden der Paneldaten**



Indexe für SQL-Tabellen:

Tabelle	Index
M	Haushaltsnummer
I	Matrikelnummer
	Haushalts- und Matrikelnummer
G	Haushalts- und Gruppennummer
C	Matrikelnummer
	Haushalts- und Matrikelnummer
D	Matrikelnummer
	Haushalts- und Matrikelnummer

Das Instrumentarium der Indexe wird im SQL primär dazu verwendet, um die Abfragen effizienter zu gestalten. Sekundär können sie dazu benutzt werden, um zu überprüfen, ob durch bisher unentdeckt gebliebene Fehler doppelte Sätze (Sätze mit gleicher Identifikationsnummer) in die Tabellen geladen wurden.

Nach Bilden der Indexe werden mit den Tabellen des SQL weitere Tests durchgeführt. Der eine Teil der Tests ist querschnittsbezogen, der andere Teil ist längsschnittsbezogen.

4.1 Querschnittstests

Mit den Querschnittstests wird die eindeutige Hierarchie im Haushalt überprüft. Sie ist gewährleistet, wenn sich alle Einheiten einer Tabelle mit denjenigen der entsprechenden Tabelle linken lassen. Für diesen Zweck werden folgende Tests durchgeführt.

Liste der Querschnittstests:

a) Alle Gruppen in Tabelle G.. müssen sich einem Haushalt in Tabelle M.. zuordnen lassen.

b) Alle Personen in Tabelle I.. müssen sich einer Gruppe in Tabelle G.. zuordnen lassen.

c) Alle Personen in Tabelle I.. müssen sich einem Haushalt in Tabelle M.. zuordnen lassen.

d) Alle Kinder aus Tabelle C.. müssen in Tabelle I.. gelistet sein.

e) Alle Erwachsenen aus Tabelle D.. müssen in Tabelle I.. gelistet sein.

f) Kein Kind aus C.. darf zusätzlich noch einmal in Tabelle D.. auftreten.

g) Kein Erwachsender aus D.. darf zusätzlich noch einmal in Tabelle C.. auftreten.

4.2 Längsschnittstests

Ziel des Längsschnittsvergleichs ist es, festzustellen, ob jeder Haushalt, jede Person und jede Gruppe einer neuen Welle sich mit den entsprechenden Einheiten der Vorjahreswelle verbinden lassen. Die Methode, die hier angewendet wird, geht vom Gedanken aus, dass, wenn sich jede neue Welle eindeutig mit der Vorjahreswelle linken lässt, das Linken mit allen anderen Vorjahreswellen damit auch gewährleistet ist.

Liste der Längsschnittstest:

a) Alle Haushalte der neuen Welle in der Tabelle M.. müssen einen Eintrag in der entsprechenden Tabelle M.. der Vorjahreswelle haben.

b) Alle 'alten' Personen (Personen bestehend in neuer und vorheriger Welle), identifiziert durch die Diagnostikvariablen in Tabelle I.., müssen einen Eintrag in der Tabelle I.. der Vorjahreswelle haben.

c) Alle 'alten' Kinder aus Tabelle C.., identifiziert durch die Diagnostikvariablen in Tabelle I.., müssen einen Eintrag in der Tabelle I.. der Vorjahreswelle haben.

d) Alle 'alten' Erwachsenen aus Tabelle D..,identifiziert durch die Diagnostikvariablen in Tabelle I.., müssen einen Eintrag in der Tabelle I.. der Vorjahreswelle haben.

e) Personen mit Kinderfragebögen in der Vorjahreswelle können einen Erwachsenenfragebogen in der aktuellen Welle besitzen, jedoch ist ein Wechsel von Erwachsenenfragebögen in der Vorjahreswelle zu einem Kinderfragebogen der neuen Welle nicht zulässig.

Bei dem zeitlichen Längsschnittstest werden auch komplexere Fälle mitberücksichtigt. Die Komplexität ergibt sich daraus, dass in jeder Welle neue Personen hinzukommen oder Personen ausscheiden. Das gleiche gilt auch für Haushalte, von denen sich Splitoffs gebildet haben. Die Abfragen und Tests werden dadurch erheblich umfangreicher als für die Normalfälle.

Besonderes Augenmerk muss auch auf die inhaltliche Qualität des Merge gelegt werden. Tests, die weiter oben beschrieben worden sind, konnten nur die formale Qualität des Merge gewährleisten. Zu dem Problembereich der inhaltlichen Mergequalität werden zwei Programme ausserhalb des SQL eingesetzt.

a) In einem Programm wird überprüft, ob jeder Haushalt, der formal aufgrund der Haushaltsnummer gemergt wurde, auch inhaltlich der gleiche Haushalt ist. Hierfür werden nicht die Haushaltsvariablen herangezogen. Diese Variablen können sich von Welle zu Welle ändern. Als Indiz für einen inhaltlich identischen Haushalt kann vielmehr der Sachverhalt gelten, dass es mindestens eine Person im Haushalt geben muss, die in beiden Wellen dem gemergten Haushalt angehört hat. Diese Person wird für den Test nicht weiter spezifiziert. Ein anderer möglicher Test, ob der Haushaltsvorstand oder die Ehefrau in beiden Wellen vorhanden ist, ist nicht bei allen Haushalten sinnvoll. Von Welle zu Welle können Personen aus unterschiedlichen Gründen (auch der Haushaltsvorstand und die Ehefrau) ausscheiden, der Haushalt als solches bleibt aber bestehen.

   Werden Diskrepanzen bei gemergten Haushalten festgestellt, so können ergänzend die Originalfragebögen aus dem Haushalt zur Klärung der Frage herangezogen werden, ob es sich wirklich um identische Haushalte handelt oder nicht.

b) In dem weiteren Programm wird nachgeprüft, ob jede Person mit gleicher Matrikelnummer auch inhaltlich die gleiche Person ist. Für diesen Vergleich werden primär das Geschlecht und die Geburtsdaten herangezogen. Stimmen diese beiden Variablen überein, so ist ein fehlerhafter Merge nahezu ausgeschlossen. Bei fehlender Uebereinstimmung werden die Daten dieser Person ausgedruckt und unter Benutzung der Originalfragebögen die Unstimmigkeiten aufgeklärt und abgeändert.

5 Datenaufbereitungen und Analysen

Nach Speicherung und Ueberprüfung der Daten in der Datenbank erfolgt die Phase der Datenaufbereitung, bevor anschliessend mit den Daten die Analysen durchgeführt werden können. Unter Datenaufbereitung wird hier verstanden, dass die sogenannten 'Missing Values' durch Schätzwerte (zu mindestens teilweise) ersetzt werden, Fehler korrigiert und, um die Analyse zu erleichtern, dass aus den bestehenden Variablen zusätzlich 'neue' Variablen generiert werden.

5.1 Probleme bei Missing Values

Missing Values treten auf, wenn Befragte Fragen nicht beantworten können oder die Antwort verweigern. Dabei ist zwischen qualitativen und quantitativen Variablen zu unterscheiden. Insbesondere 'Missing Values' bei quantitativen Variablen behindern die Analyse dieser Variablen stark. Im Panel wird das Einkommen der Individuen in 24 unterschiedlichen Einkommensarten erhoben. Das Gesamteinkommen einer Person und das Gesamteinkommen des Haushalts, aggregiert über alle Personen, kann im SPSSX nicht ermittelt werden, wenn nur eine Teilkomponente, (d.h. nur eine Variable unter vielen anderen) einen Missing Value enthält. Die Ersetzung dieser Missing Values durch eine Null ist nicht adäquat, wenn man aus anderen Variablen weiss, dass ein Wert grösser als Null eingesetzt werden muss.

In diesem Fall hilft nur eine plausible Schätzung für den Missing Value weiter. Zur Schätzung können verschiedene Verfahren verwendet werden. Bei dem Regressionsansatz werden aus den Daten von denjenigen

Einheiten, die keine Missing Values aufweisen, Regressionskoeffizienten geschätzt. Diese Regressionsergebnisse werden dazu benutzt, die Missing Values zu ersetzen. Je nach Differenziertheit des Ansatzes sind sehr grobe oder sehr feine Schätzungen möglich. Alternativ dazu ist der Tabellenansatz zu sehen. Zum einen können bei diesem Ansatz aus den Daten von denjenigen Einheiten, die keine Missing Values aufwiesen, mehr oder weniger differenzierte Durchschnitte austabelliert werden und diese als Schätzwerte eingesetzt werden. Zum anderen können externe Informationen in Tabellenform als Schätzwerte eingehen.

Unabhängig von der eingesetzten Methode treten Probleme auf, wenn in einer Welle ein Wert für eine Variable geschätzt werden musste, in einer anderen Welle Angaben des Haushalts vorliegen. Im Falle von Einkommensvariablen können sich so unplausible Einkommensprünge in der Längsschnittsanalyse ergeben. Um solche Fälle leichter identifizieren zu können, werden Variablen, die geschätzt worden sind, mit einem Flag versehen werden. Im Prinzip müsste jede Schätzung einer Welle korrigiert werden, wenn Variablen in späteren Wellen auf eine frühere fehlerhafte Schätzung hindeuten.

Missing Values bei qualitativen Merkmalen behindern die Analyse nicht so stark wie bei den quantitativen Merkmalen. Wo immer möglich, sollten diese ersetzt werden. Eine Panelstichprobe besitzt im Vergleich zu Querschnittsstichproben zum Teil hierfür günstige Korrekturmöglichkeiten. Fehlt in einer Welle eine Angabe, so kann diese aus einer anderen Welle entnommen werden.

5.2 Probleme bei fehlerbehafteten Variablen

Ein Teil der Variablen wird in jeder Welle erhoben. Für diese Variablen ergibt sich über die Redundanz die Möglichkeit eventuelle Fehler in Vorjahreswellen identifizieren zu können. Die Schwierigkeit besteht hier darin, herauszufinden, ob eine Veränderung von Welle zu Welle wirklich eingetreten ist oder ob sie nur einen Fehler anzeigt, der korrigiert werden muss. Für einen kleinen Teil der Variablen (so z.B. Geschlecht, Familienstand und Geburtsdatum von Individuen, Haushaltszusammensetzung) besteht von Welle zu Welle jedoch die Möglichkeit, Fehler in vorhergehenden Wellen zu korrigieren und damit die Daten kohärenter zu gestalten.

5.3 Probleme bei neuen Variablen

Aus den bestehenden Variablen werden aus vielfältigen Gründen für
die Analyse 'neue' (zusätzliche) Variablen gebildet:

a) Recodierung von Variablen (z.B. Altersklassen)

b) Bildung von Typisierungsvariablen mit der Hilfe von anderen Variablen

c) Addition und Subtraktion von Variablen zur Bildung von Einkommensbegriffen

d) Aggregation von Variablen auf eine höhere Hierarchiestufe

Bei umfangreichen Analysen kann die Anzahl der neu zu bildenden Variablen sehr gross werden. Es stellt sich hier die Frage, ob alle, oder wenn nicht, welche der neu gebildeten Variablen in der Datenbank

abgespeichert werden sollen. Als Alternative zum Abspeichern kann die jeweilige Neuberechnung von Variablen angesehen werden. Das Speichern von neuen Variablen benötigt zusätzlichen Speicherplatz, das Neuberechnen kostet zusätzliche Computerzeit. Allgemeingültig kann die Frage (Speicherung oder Neuberechnung) nicht beantwortet werden. Deshalb wird im Einzelfall entschieden, welche Variablen gespeichert und welche Variablen jeweils neu berechnet werden.

Mit dem neuen Release (3.0) von SPSSX stehen zwei Befehle zur Verfügung, die es ermöglichen, das Neuberechnen von Variablen organisatorisch zu bewältigen. Es sind dies die Möglichkeit, von einem externen File SPSSX-Kommandos in ein SPSSX-Programm dynamisch einzufügen (INCLUDES) und die Möglichkeit, mit MAKROS zu arbeiten. Mit diesem Instrumentarium können für Gruppen von Variablen jeweils externe Includes/Makros bei der Analyse zur Verfügung gestellt werden. Diese Includes/Makros werden auf einer Minidisk allen Benutzern zur Verfügung gestellt. Da diese nur einmal für alle Benutzer existieren, können sie zentral gepflegt, und wenn notwendig, korrigiert werden.

5.4 Update Möglichkeiten

Der Datenbestand des Panel wird aus verschiedenen Gründen laufend geändert und erweitert. Hierfür sind hauptsächlich Einzelkorrekturen an bestehenden Variablen und Generierung von 'neuen' Variablen verantwortlich. Die Einzelkorrekturen könnten ad hoc interaktiv am Terminalbildschirm vorgenommen werden (ISQL). Diese Vorgehensweise empfiehlt sich nicht, wenn die durchgeführten Aenderungen dokumentiert werden sollen. Die Einzelkorrekturen sollten über einen Editor in ein Kommandofile für die SQLDBSU Utility eingegeben und ausgeführt werden. Dieses Kommandofile sollte permanent erhalten bleiben; es dokumentiert die durchgeführten Aenderungen und erlaubt beim Neuladen einer Tabelle die Aenderungen zu replizieren. Neue Variablen können in die Datenbank eingebracht werden, indem bestehende Tabellen

erweitert oder dafür eigene Tabellen eingerichtet werden. Welche der beiden Methoden besser oder schlechter ist, kann nur pragmatisch entschieden werden. Ist die Anzahl der neuen Variablen relativ klein, so könnte man die bestehenden Tabellen erweitern. Ist die Anzahl der neuen Variablen relativ gross, sollte für diese Variablen eine neue Tabelle vorgesehen werden. Für diese Vorgehensweise spricht, dass die 'neuen' von den Variablen aus den Fragebögen eindeutig getrennt sind. Müssen Teile der neuen Variablen korrigiert werden, so kann vorher die ganze Tabelle entfernt werden und alle neuen Variablen erneut geladen werden. Das alternative Updaten der neuen Variablen wäre, wenn sie zusammen bei der Tabelle mit der alten Variable gespeichert wären, erheblich aufwendiger.

6 Analysen

Nach Dateneingabe, Datenladen und Datenkorrektur erfolgt die Phase der Auswertungen. Bei hierarchischen Datensätzen und bei Paneldaten existieren eine Vielzahl von unterschiedlichen Auswertungsmöglichkeiten:

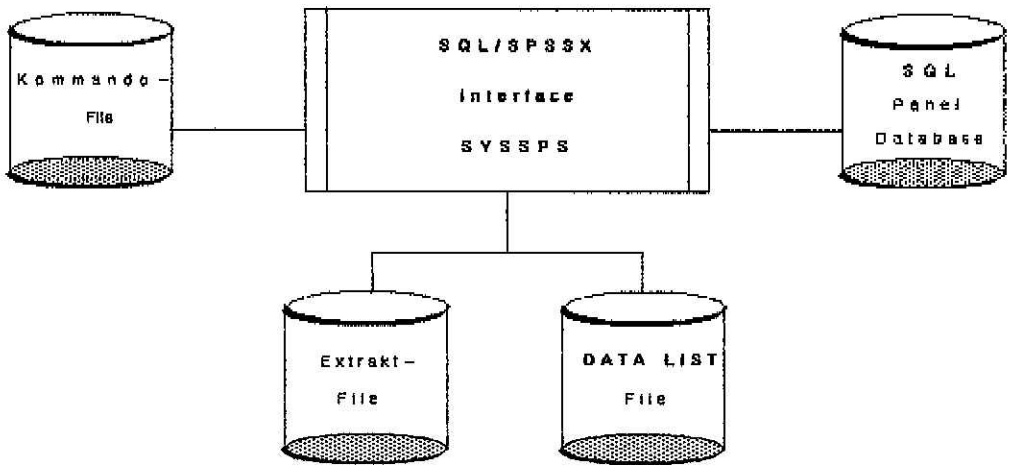
Die Auswertungen können separat für Haushalte, Gruppen und für Individuen durchgeführt werden können. Werden Informationen von mehreren Hierarchieebenen benötigt, so müssen die Daten aus niedrigeren Hierarchieebenen auf ein höheres Niveau aggregiert oder die Daten aus einer höheren Hierarchieebene auf eine niedrigere Hierarchieebene transformiert werden. Weiterhin können die Auswertungen sowohl querschnitts- als auch längsschnittsorientiert erfolgen.

Die Analysen können mit unterschiedlichen Statistikpaketen durchgeführt werden. Als Software stehen derzeit SPSSX, LIMDEP, MDSX und SPADNA zur Verfügung. Die Statistikpakete unterscheiden sich u.a. im Datenhandling voneinander. Drei dieser Pakete (LIMDEP, MDSX, SPADNA) besitzen keine eigentlichen Datenbankfunktionen wie z.B. Mergen und Aggregieren von Datensätzen. Das SPSSX jedoch beinhaltet zum Teil sehr differenzierte Möglichkeiten, um komplexe Strukturen zu verarbeiten. Insbesondere sind in diesem Zusammenhang hier die Funktionen 'Match Files' und 'Aggregate Files' wichtig. Als Systeminput erwarten alle diese Pakete in der Regel formatierte rektanguläre Files ('Flatform files').

Diese unterschiedlichen Auswertungsmethoden in Bezug auf die Datenstruktur und auf die Statistikpakete erfordern eine flexible Verbindung zwischen SQL und Anwendungspaket. Hierfür steht ein Interface zur Verfügung (Uebersicht 4). Das Interface stellt die Daten aus der SQL-Datenbank in einem formatierten rektangulären Extraktfile ('Flatform File') zur

Verfügung. Das Extraktfile lässt sich in das SPSSX einlesen. Die notwendigen SPSSX-Kommandos zum Einlesen dieser Daten werden vom Interface erzeugt und stehen in einem separaten File zur Verfügung. Im SPSSX können mit diesem Extraktfile sofort statistische Auswertungen durchgeführt werden, oder aber ein SPSSX Systemfile erzeugt werden. Ebenso kann dieses Extraktfile als Dateninput für die anderen Auswertungspakete dienen.

Uebersicht 4: **SQL/SPSSX Interface**



Dem Interface muss mitgeteilt werden, welche Art von Extraktfile gewünscht wird und welche Variablen ausgewählt werden sollen. Dies erfolgt durch ein Kommandofile mit notwendigen SQL-Kommandos. Das Kommandofile muss folgende Informationen enthalten:

a) SELECT-KLAUSEL: welche Variablen (Kolumnen) extrahiert werden sollen

b) FROM-KLAUSEL: welche Tabellen benutzt werden sollen

c) WHERE-KLAUSEL: wie diese Tabellen miteinander verbunden werden sollen

d) welche Sätze (Zeilen) entnommen werden sollen (optional)

e) ORDER BY-KLAUSEL: ob das File sortiert werden soll (optional)

Mit Hilfe dieser Kommandos können Extraktfiles für Analysezwecke erstellt werden. In dem einfachsten Fall werden Teile oder die ganze Tabelle aus der Paneldatenbank entladen. Folgender Befehl kopiert z.B. alle Variablen aus dem Familien Tableau 1986 in das Extraktfile:

```
SELECT * FROM I86
```

Bei umfangreichen Analysen und insbesondere für Längsschnittsanalysen werden komplexere Extraktfiles benötigt, welche Informationen aus zwei und mehr SQL-Tabellen enthalten.

Das Extraktfile des Interface kann ganz unterschiedlichen Inhalt
besitzen:

1) Daten aus einer Tabelle (einfacher Querschnitt)

2) Daten aus mehreren Tabellen der gleiche Welle (hierarchischer Querschnitt)

3) Daten aus mehreren Tabellen unterschiedlicher Wellen (einfacher Längsschnitt)

4) Daten aus mehreren Tabellen unterschiedlicher Wellen (hierarchischer Längsschnitt)

Die Satzstrukturen der einzelnen Extraktfiles unterscheiden sich nur durch die Variablen voneinander.

Beispiele für Satzstrukturen von Extraktfiles:

1) File mit Variablen für Individuen aus der Welle 85:

   I85001, I85002

2) File mit Variablen für Individuen aus der Welle 85, zu denen Haushaltsvariablen expandiert worden sind:

   I85001, I85002, M85010, M85011

3) File mit Variablen für Individuen aus der Welle 85 und 86:

   I85001, I85002, I86001, I86002

4) File mit Variablen für Individuen aus der Welle 85 und 86, zu denen Haushaltsvariablen expandiert worden sind:

   I85001, I85002, I86001, I86002, M85010, M85011, M86010, M86011

Um solche komplexen Extraktfiles zu erhalten, müssen die Tabellen des SQL 'gejoint'(zusammengeführt) oder die Files im SPSSX gematcht werden. Das 'Joinen' von Tabellen oder Files lässt mehrere Möglichkeiten zu:

a) Paralleler Match (z.B. Verbinden von neuen und alten Variablen einer Welle)

b) Nicht paralleler Match (z.B. Personeninformationen aus zwei und mehr Wellen)

c) Variablen expandieren (z.B. Haushaltsvariablen den Personen zuweisen)

d) Sätze aggregieren (z.B. Einkommen von Gruppenebene auf Haushaltsebene addieren)

Es stellt sich hier die Frage, welche Fileoperationen im SQL und welche im SPSSX vorgenommen werden sollten. Für das 'Verbinden' von Tabellen gibt es zwei Realisierungsmöglichkeiten: Bei der ersten Methode wird aus jeder SQL-Tabelle ein separates Extraktfile erzeugt. Wenn Informationen aus zwei und mehr Tabellen benutzt werden sollen, müssen die entsprechenden Files im SPSSX transformiert werden. Bei der zweiten Methode wird nur jeweils ein Zwischenfile erzeugt, die entsprechenden Tabellen werden im SQL gejoint (verbunden). Empirische Tests haben ergeben, daß beide Methoden Vor- und Nachteile besitzen. Im SPSSX stehen eine Vielzahl von unterschiedlichen Transformationsmöglichkeiten zur Verfügung. Dies ermöglicht komplexe Files ohne jede praktische Begrenzung der Variablenanzahl zu erstellen. Als Nachteil beim SPSSX für Panelauswertungen ist auszusehen, daß je nach Komplexität mehrstufige Verfahren mit einer größeren Anzahl von Hilfsfiles notwendig werden.

Bei der zweiten Methode(SQL) lässt sich diese Art von Transformationen einfacher realisieren. Mit relativ wenigen, wenn auch komplexen FROM und WHERE Bedingungen, lässt sich im SQL ein Extraktfile mit Informationen aus mehreren Tabellen erzeugen. Entsprechende SQL-Kommandos transformieren die Tabellen so, dass keine zusätzlichen Hilfsfiles notwendig werden. Das Extraktfile wird im SQL in einem einstufigen Verfahren erstellt. Jedoch ist bei dieser Methode als Nachteil zu beachten, dass die Anzahl der selektierten Variablen nicht zu gross wird. Je nach Komplexität lassen sich nur circa 300 Variablen in das Extraktfile übernehmen. Weiterhin ist es nur schwer möglich, aus dem SQL Daten aus einer sehr großen Auswahl von Tabellen gleichwertig zu entladen. Es kann deshalb die Notwendigkeit bestehen, bei Extraktfiles mit sehr vielen Variablen nach der ersten Methode (Transformierung im SPSSX) zu verfahren.

Probleme können sich ergaben, wenn mit der zweiten Methode Datensätze aggregiert werden sollen. Die SQL-Abfragesprache enthält nur vier Aggregations-funktionen, im SPSSX sind circa 19 unterschiedliche Funktionen vorhanden. Es ist deshalb bei Aggregations-vorgängen günstiger, ein gemischtes Verfahren anzuwenden:

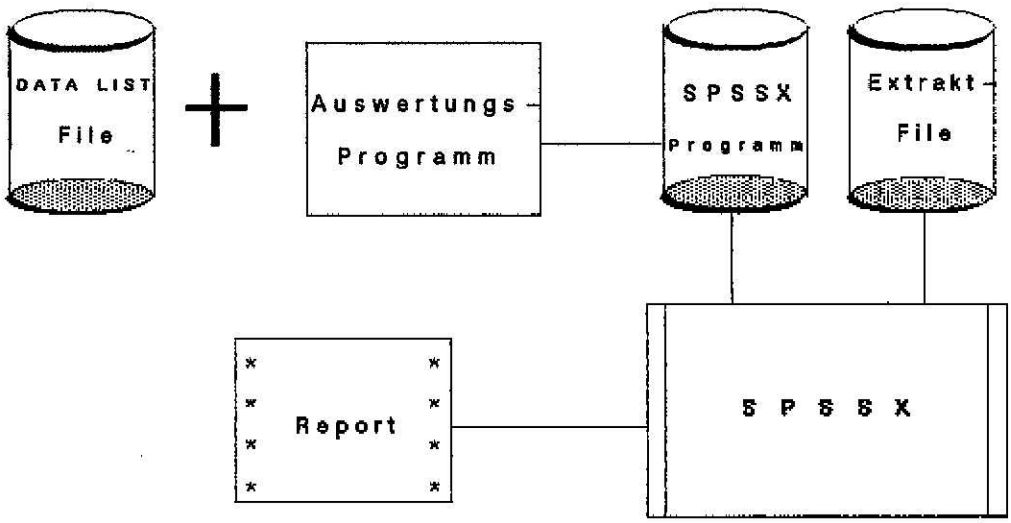
Werden z.B. Informationen aus der Haushalts- und der Personenebene auf der Gruppenebene benötigt, so sollte im SQL ein Personenfile erzeugt werden, in welches die Haushalts- und die Gruppenvariablen expandiert werden. Im SPSSX wird dieses Personenfile auf die Gruppenebene aggregiert.

In dem Kommandofile für das Interface müssen diejenigen Variablen gelistet werden, welche extrahiert werden sollen. Wird mehr als eine Tabelle angesprochen, so muss dem SQL mitgeteilt werden, wie diese Tabellen miteinander verbunden werden sollen. Um das Verbinden der Tabellen im SQL durchführen zu können, sind genaue Kenntnisse notwendig, wie die Tabellen organisiert sind und welche Identifikatoren benutzt werden müssen. Unerfahrene oder neue Benutzer besitzen dieses Wissen nicht. Für diese Benutzer steht eine kleine Programmbibliothek mit entsprechenden SQL-Kommandos zur Verfügung. Die einzelnen Teilmodule enthalten die FROM-KLAUSELN und die WHERE-KLAUSELN für Standardauswertungen bei Querschnitts- und Längsschnittsauswertungen. Der Benutzer, der mit den SQL-Paneldaten arbeiten will, kann diese Befehle in das Kommandofile für das Interface übernehmen. Er braucht in der Regel nur noch die SELECT-VARIABLEN spezifizieren.

Das SQL enthält die Möglichkeit, sogenannte Views zu vereinbaren. Views sind imaginäre Tabellen, die sich durch Kombination existierender Tabellen ergeben. Views wären damit eine Alternative zu der Methode mit der Programmbibliothek. Die maximale Anzahl der Variablen in den VIEWS ist durch das SQL auf 140 Variable beschränkt. Da in den meisten Analysen jedoch mehr Variablen gleichzeitig benötigt werden, kann das Instrumentarium der Views bei den spezifischen Panelauswertungen nicht verwendet werden.

Das Interface erzeugt Extraktfiles, welche direkt in das SPSSX eingelesen werden. Dort können die statistischen Auswertungen durchgeführt werden (Uebersicht 5). Eventuell sind Aggregationsvorgänge vorzuschalten.

Uebersicht 5: **SPSSX Auswertungsprogramm**



Strategische überlegungen sprechen dafür, jeweils nur kleine Zwischenfiles zu erzeugen, die in der Regel auch nur temporär gehalten werden. Aus Effizienzgründen (Rechenzeit und Speicherplatz) erscheint es nicht immer sinnvoll, jeweils alle Variablen einer oder mehrerer Tabellen über das Interface zu entladen. Zweckmässig ist es, nur diejenigen Variablen zu extrahieren, welche später in der Analyse benötigt werden. Je nach Typ der Analyse können die Interface-Files temporär oder permanent gehalten werden. Für ad hoc Tabellierungen sind temporäre Files adäquat. Für die Dauer einer grösseren Studie kann es sinnvoll sein, Auswahlfiles zu erzeugen, die in Form von SPSSX-Systemfiles

gespeichert werden. Dies ist insbesondere dann vorteilhaft, wenn eventuelle Match- und Aggregationsvorgänge zur Erzeugung notwendig waren. Nach Abschluss der Studie können diese Files wieder gelöscht werden. Im Gegensatz zu normalen Querschnitsstichproben hält die Phase der Fehlerbereinigungen bei Paneldaten kontinuierlich an. Deshalb werden die Daten im SQL ständig upgedated. Die Analysen sollen jeweils mit dem upgedateten Datenbestand durchgeführt werden. Dies ist nur dann möglich, wenn die Daten oder Teile davon nicht parallel im SQL und als SPSSX-Files gehalten werden.

Mit jeder weiteren Welle kann die Möglichkeit bestehen, die Analysen, welche für die Vorjahreswellen durchgeführt worden sind, zu wiederholen. Für diesen Zweck muss das Auswertungsprogramm allgemein gehalten werden. Allgemein gehalten heisst hier, dass das Jahr der Welle nicht explizit im Variablennamen enthalten sein darf. Jeder Variablenname sollte das Jahr der Welle als einen Parameter enthalten. Je nach Welle (Querschnitt) oder Untersuchungszeitraum (Längsschnitt) kann diesem Parameter neue Werte zugewiesen werden. Eine solche Möglichkeit besteht im SPSSX durch das Anwenden von Makros. Für jede Analyse kann ein allgemeines Makro geschrieben werden. Als Inputparameter dient nur das Jahr der Welle und/oder der Untersuchungszeitraum.

ANHANG :

Verzeichnis der Computer-Programme:

1) Inputprogramm:

   a) Erstellung des Kontrollfiles aus der Vorjahreswelle: LLPREP (Fortran)
   b) Dateneingabeprogramm: LLINP (REXX,ISPF)
   c) Batchprogramm: LLTRA (REXX,Fortran)

2) Programme zur Zwischenprüfung der Daten:

   a) Ueberprüfung des Haushaltsfiles (M): LLPRA (Fortran)
   b) Ueberprüfung des Personenfiles (I): LLPRB (Fortran)
   c) Ueberprüfung des Kinder- und des Erwachsenenfiles (C,D): LLPRC
   d) Ueberprüfung des Gruppenfiles (G): LLPRD (Fortran)
   e) Ueberprüfung der Hierarchie: LLPRE (Fortran)

3) Laden der Daten in das SQL (SQLDBSU)

Tabellen	kreieren:	laden:
Haushalte:	CREATM	LOADM
Personen:	CREATI	LOADI
Gruppen:	CREATG	LOADG
Kinder:	CREATC	LOADC
Erwachsene:	CREATD	LOADD

4) Ueberprüfung der Daten im SQL:

   a) Querschnittstests: ISQL-Kommandos
   b) Längsschnittstests: ISQL-Kommandos
   c) Test auf identische Haushalte: CONS2 (Fortran)
   d) Test auf identische Personen: CONTEST1 (SPSSX)

5) SQL Interface:

   SYSSPS (SQLDBSU/REXX)

Erläuterungen:

REXX: 'Restructured Executer Language' ist eine interpretative Sprache von IBM auf der Kommandoebene

ISPF: 'Interactive System Productivity Facility' ist eine IBM-Software, die das flexible Programmieren von Bildschirmmasken ermöglicht

ISQL: 'Interactive Structered Query Language' ist die dialogorientierte Abfragesprache der SQL-Datenbank

SQLDBSU: 'SQL Database Service Utility' ist die batchorientierte Abfrage- und Manipulationsmöglichkeit der SQL-Datenbank