DESTATIS
wissen. nutzen.

Spektrum Bundesstatistik

ROTIERENDE STICHPROBEN

Datenkumulation und Datenqualität

Walter Krug, Manfred Ehling und Nicole Ernst

Band 21

Statistisches Bundesamt

METZLER POESCHEL

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Herausgeber: Statistisches Bundesamt, Wiesbaden

Fachliche Informationen
zu dieser Veröffentlichung:

Gruppe I B,
Tel.: 06 11 / 75 29 03
Fax: 06 11 / 75 39 50
manfred.ehling@destatis.de

Allgemeine Informationen
zum Datenangebot:

Informationsservice,
Tel.: 06 11 / 75 24 05
Fax: 06 11 / 75 33 30
info@destatis.de
www.destatis.de

Veröffentlichungskalender
der Pressestelle:
www.destatis.de/presse/deutsch/cal.htm

Erschienen im August 2002

Preis: EUR 10,80 [D]

Bestellnummer: 1030521 - 02900

ISBN: 3-8246-0670-4

Recyclingpapier aus 100 % Altpapier.

Verlag: Metzler-Poeschel, Stuttgart

Verlagsauslieferung: SFG – Servicecenter Fachverlage GmbH
Postfach 43 43
72774 Reutlingen
Telefon: 0 70 71 / 93 53 50
Telefax: 0 70 71 / 93 53 35
www.s-f-g.com
destatis@s-f-g.com

Vorwort

Der vorliegende Band ist ein Bericht über den Workshop "Rolling Samples and Sampling in Time. Problems of Data Accumulation and Data Quality", der am 18. und 19. Juli 2001 an der Universität Trier stattfand und zu dem das Statistische Bundesamt eingeladen hatte.

Ausgehend von Überlegungen zur Neugestaltung der Haushaltsstichproben (N. Ernst), die neben dem Mikrozensus die jährlichen Haushaltsbudgeterhebungen und die bisher fünfjährlichen Einkommens- und Verbrauchsstichproben umfassen, wurden Stichprobendesigns diskutiert, die miteinander verbunden den bisherigen Gesamtaufwand und die Dauer der Erhebungen sowie die Belastung der Befragten reduzieren, ohne die Anforderungen an den Merkmalskatalog und die Genauigkeit der Daten zu beeinträchtigen.

Breiten Raum nahmen dazu die internationalen Erfahrungen in der amtlichen Statistik Polens (J. Kordos), Schwedens (J. Selén), Dänemarks (B. Møller) und Frankreichs (J. Dumais / Ph. Bertrand / B. Christian) ein, ergänzt durch die vom Statistischen Amt der Europäischen Gemeinschaften (Eurostat) beabsichtigte neue EU-SILC Erhebung (V. Verma). Der Schwerpunkt lag bei diesen Darstellungen auf den Stichprobenplänen, wechselnd zwischen rollierenden und sich überschneidenden Stichproben, um kontinuierliche Erhebungen zu erzeugen. In diesem Sinne wurde auch der deutsche Mikrozensus problematisiert (W. Bihler).

In dieser Veröffentlichung finden sich neben einer Zusammenfassung der Ergebnisdiskussion des Workshops auch die Folgerungen aus dieser Veranstaltung zur Weiterentwicklung der entsprechenden Haushaltsstichproben in der deutschen amtlichen Statistik (M. Ehling).

Wiesbaden, im Juli 2002

Prof. Dr. Walter Krug
Universität Trier

Dr. Manfred Ehling
Statistisches Bundesamt

Inhalt

Manfred Ehling*)

Considerations on Changing the System of Household Statistics in Germany

1 Introduction

The bodies of German official statistics are currently discussing different ways of **further developing their system of household statistics** to be able to satisfy future requirements.

The demands articulated by policy-makers, scientists and citizens for statistical information on households and individuals has undergone substantial changes in recent years. While in the past it was sufficient for official statistics to provide structural aggregate data in partly great subject-matter and regional detail on a variety of topics, nowadays there is a strong demand for **integrated information systems**. As in the past, those systems are expected to supply structural data. Additionally, however, they are to enable flexible evaluations of micro data on varying issues and a short-term collection of new variables that may be integrated into the information system and illustrate trends or changes, i.e. provide not only cross-sectional but also longitudinal information.

The existing system of household statistics was established in the 1950s and early 1960s. It mainly consists of the microcensus, including the EU Labour Force Survey, and the household budget surveys. Besides, surveys for special purposes are sporadically taken on the basis of Article 7 of the Federal Statistics Law (BStatG). Although the surveys have been adjusted over the years to changing demands, the **unconnected parallel existence** of microcensus, household budget surveys and ad hoc surveys has been left untouched. This structure is making it increasingly difficult for official statisticians to fulfil their tasks.

Official statistics must be able to answer **complex sets of questions.** They should, for instance, provide information on the transition from employment to retirement and its impact on the households' living conditions and income situation as well as on the pension system, supply data on the phenomena of poverty and social exclusion and the related dynamic processes, as well as produce results on the equalisation of burdens for families. On top of that, the bodies of official statistics must be able to carry out simulation studies to prepare future political actions. Surveys whose definitions and concepts have not been harmonised so that a joint presentation of their results is not or not fully possible, are badly suited to these requirements. Hence, it is necessary to set up an **overall system of household statistics** in order to obtain a consistent set of data in which results from different surveys complement each other. After achieving this goal, ad hoc surveys of specific topics may be based on a known data structure, which will significantly raise their informative value beyond that of unrelated surveys.

---

*) Dr. Manfred Ehling, Federal Statistical Office of Germany, Wiesbaden.

Apart from that, official statistics must be able to respond **flexibly and timely** to **new data requirements** of, for example, policy-makers or Eurostat. There are quite serious restrictions to that today. If certain preconditions are given, ad hoc surveys can be taken in accordance with and subject to the narrow limits defined by Article 7 BStatG without the need for a specific legal provision. But official statistics are lacking the infrastructure for a quick and cost-efficient execution of such surveys. At present, that infrastructure needs to be set up anew for each of those surveys. The microcensus is specified by law down to the smallest detail, which makes short-term changes of the survey programme impossible. Besides, there is absolutely no room to include new questions in the microcensus, which, on the contrary, is regarded as being overloaded. As for the household budget surveys, the rather general wording of the relevant legal act does offer considerable room for designing the survey programme, but this room can only be exploited to collect data in sufficient subject-matter and regional detail when the sample survey of income and expenditure is taken every five years. Hence, what we should aim at is to develop the system of household statistics in such a way that

- ad hoc surveys can be executed in a flexible, rapid and cost-efficient manner;
- the strain on the microcensus is relieved by creating alternative means of meeting data requirements;
- household budget information in sufficient subject-matter and regional detail is not only available every five years but annually if possible, and
- European statistical projects can be flexibly integrated into the national system.

The considerations on further development set forth in the following paragraph are not only aimed at the objectives described above, but are based on two other essential prerequisites:

- The overall system must be designed in such a way that the high methodological quality of the data is preserved.
- The currently available resources for household statistics must not be exceeded.

The primary goal of all considerations on further development is to ensure that the bodies of German official statistics will be able to perform their future tasks. In addition, official statistics are faced with the challenge of increasing the **proportion of external financing.** However, at present official statisticians are lacking the necessary tools to successfully obtain contracts for surveys commissioned, for example, by supreme federal authorities or Eurostat. A system of household statistics like the one described below, however, would enable the official statistical agencies to compete successfully on the growing market for externally financed projects where they have not been able to gain a foothold as yet. The statistical offices of the Federation and the Länder are pursuing two main lines of considerations:

- combining the sample survey of income and expenditure with the continuous household budget surveys to set up one annual household budget survey;
- setting up an access panel of official statistics.

2 Annual household budget survey

The sample survey of income and expenditure is taken every five years and was last held in 1998, covering about 70 000 households. The continuous household budget surveys collect information annually from some 6 000 households on the basis of a quarterly rotation schedule. The two surveys could be combined in an annual household budget survey covering the income and expenditures of 18 000 to 24 000 households. Such a survey would for the first time provide **yearly results** in sufficient subject-matter and regional detail (federal Länder), thereby considerably **improving the timeliness** of the data and substantially facilitating the **presentation of trends.** As far as the list of questions is concerned, no decisive changes are envisaged as compared to the existing household budget surveys.

The present considerations are based on applying a **rotation schedule** in the sample survey, that means, the selected households participate in the survey for one reference period and are replaced by other households after that period has ended. The use of new mathematical-statistical instruments to **cumulate results** from different reference years is to ensure that the **requirements of the users** of the former sample survey of income and expenditure can be **satisfied** even with the modified sample size. In the next presentation Nicole Ernst will discuss this proposal in more detail.

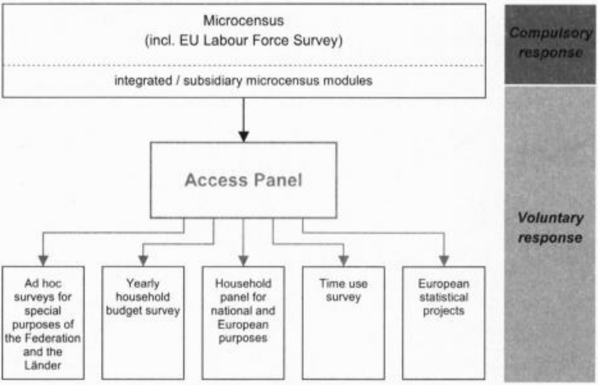
3 Access panel of official statistics

An **access panel of official statistics** could be set up for the purpose of efficiently carrying out household inquiries with voluntary response for which a substantially smaller sample size is required than for the microcensus. That access panel would serve as a **basis for drawing samples.** The panel would comprise those microcensus households that, after taking part in the usual four microcensus inquiries, are willing to participate in further surveys of official statistics and agree to have their addresses and the characteristics covered by the microcensus stored in the database of the access panel. These households would form the access panel. A household would remain in the panel until the household itself cancels its participation or is discharged by the statistical office. In both cases all the stored data would be deleted. If a household participated in the access panel for several years, it would repeatedly be requested to participate in voluntary surveys to be taken among access panel respondents. Through these continuous contacts the **basic household characteristics** stored in the access panel database would be **updated regularly**.

Figure 1 (see p. 8) illustrates the basic conception of an access panel of official statistics as a basis for household samples with voluntary response.

As part of an overall system, an access panel could feed all samples of households and individuals which do not require compulsory response and for which a markedly smaller sample size than that of the microcensus is considered to be sufficient. The spectrum could thus range from ad hoc surveys for special purposes of the Federation and the Länder via household budget and time use surveys to household panel surveys for national and European purposes and to other European statistical projects. Since all surveys would be based on the microcensus, a **harmonisation of the survey variables** could be achieved. Harmonised definitions, delimitations and breakdowns prevent that disparate data from different surveys are used to answer identical sets of questions – a circumstance that users of official statistics find very difficult to understand. Furthermore, that harmonisation facilitates the joint use of results from different surveys for settling complex problems. Ad hoc surveys could be reduced to the actually lacking variables because the results of the special surveys could be incorporated into the **consistent set of data** available from the other household statistics. By increasing the efficiency of official statistics in the execution of ad hoc surveys, it would be possible to relieve the load of **supplementary surveys** on the **microcensus and household budget surveys and** to counteract the risk of overloading them with new modules time and again.

Figure 1



An access panel could be expected to offer major advantages to official statistics. Since the access panel would consist of households willing to respond to statistical inquiries, on whose basic characteristics up-to-date information would regularly be available, it would be possible

- to provide survey results more flexibly and timely because the households for individual surveys could be recruited far more easily and rapidly. This would be true especially of samples that are designed to cover only a specific group of the population (e.g. low-income earners). Using the stored basic characteristics, households belonging to that group could be contacted directly without the current need to take a survey in advance simply to identify these households;
- to reduce the survey costs because higher degrees of coverage could be expected and, on the whole, fewer households would have to be contacted;

- to draw high-quality stratified samples by random selection, using the stored variables, which would improve the accuracy of the survey results correlating with the stratification variables;
- to minimise the burden on the respondents by utilising up-to-date information from the access panel in subsequent surveys or by not collecting the relevant data anew in those surveys but simply bringing them up-to-date. Besides, it can be assumed that households who are willing to respond would perceive the burden involved with the surveys to be lower than households that are not willing to participate. Consequently, a better acceptance of the surveys could be achieved.

On account of their advantages for timeliness, flexibility and cost efficiency, access panels have been used in commercial opinion research for a long time. An access panel of official statistics would, however, show two special features:

- the microcensus would provide the sampling frame for the access panel and
- microcensus information could be used to correct the volunteer bias.

The principle of random sampling that is applied in the selection of microcensus households would communicate itself to a microcensus-based access panel. However, it has to be expected that an access panel would mainly comprise households that, in general, show an increased readiness to participate in inquiries. As a result, non-response would not occur at random and the access panel itself could not be regarded as a genuine random sample. Consequently, all samples to be drawn from the access panel would also be affected by the volunteer bias. Drawing samples from a microcensus-based access panel would, however, provide the advantage that information would be available on households that participate neither in the access panel nor in the samples drawn from it. This information could be used to **correct the non-response bias**. Based on surveys with voluntary response, official statistics at present can just record the distribution of estimates of the total population willing to respond. By correcting the volunteer bias and taking into account the population that is not willing to respond, these estimates would come closer to the actual distribution and the **quality of the results in terms of their validity would increase.** The information available from the microcensus could be exploited far better for producing follow-up statistics on households and individuals, i.e. the **advantages** that are **inherent in the system** could be put to use.

4 Further action

The following legal provisions provide the current framework for household statistics:

- Law on the Execution of a Sample Survey of the Population and the Labour Market as well as of the Housing Situation of Households (Microcensus Law), responsible agency: Federal Ministry of the Interior;
- Law on Statistics on Household Budget Surveys, responsible agency: Federal Ministry for Family Affairs;

- Article 7 (Surveys for Special Purposes) of the Law on Statistics for Federal Purposes (Federal Statistics Law), in accordance with which, for example, the time use survey and the German sample for the European Household Panel could be taken, responsible agency: Federal Ministry of the Interior;
- Council Regulation (EEC) on the organization of an annual labour force sample survey in the Community, responsible agency: Federal Ministry of Labour.

First considerations on the future legal framework were aimed at laying down all household statistics in a single legal act. This was an obvious solution because the current microcensus law will expire in 2004. A revision of that law will presumably involve major changes because there are plans to change over to a continuing survey.

However, that idea has meanwhile been dropped in order to make a **clear distinction between the microcensus as a survey with compulsory response and the entirety of voluntary surveys.** The current considerations are based on regarding the latter surveys as an integrated whole and are aimed at drawing up a law on voluntary household surveys. Such a law could provide for

- an access panel of official statistics and
- a yearly household budget survey.

Furthermore, it should be envisaged to incorporate into that law a **flexible component** that is equivalent to the currently applicable Article 7 BStatG. What should be noted is that, from the point of view of official statistics, Article 7 still offers **too little flexibility**. The necessary changes are currently under discussion and should already be taken into account in drafting the law on voluntary household surveys. Accordingly, with a view to satisfying the data requirements of other prime users[1)], it should be abandoned that special-purpose household sample surveys can only be taken to meet information needs of supreme federal authorities. Besides, the limitation to a maximum number of respondents should be given up because reliable statistical findings on specific issues cannot be obtained on the basis of fixed arrangements. To warrant a timely provision of data for federal purposes, the law should lay down an option to dispense with Land results from the very outset and to assign the execution of corresponding surveys only to the Federal Statistical Office. Limits to the maximum permissible number of repetitive inquiries should be abandoned as well, because they would always be arbitrary.

There is general agreement that the revision of the Microcensus Law must be concluded by 2004 so that a microcensus can be taken in 2005. The fundamental revision of the system of household statistics with voluntary response is not subject to a specific deadline but it would be highly desirable for a new law to take effect by 2005.

---

1) Article 1 of the Federal Statistics Law lists the Federation, Länder incl. communities and local authorities, society, science and research as prime users of official statistics.

Nicole Ernst*)

Überlegungen zur Neugestaltung der Wirtschaftsrechnungen privater Haushalte

1 Anforderungen an jährliche Haushaltsbudgeterhebungen

Die Wirtschaftsrechnungen privater Haushalte fußten bis 1998 auf zwei unverbunden nebeneinander stehenden Erhebungen. Der Einkommens- und Verbrauchsstichprobe (EVS) und den Laufenden Wirtschaftsrechnungen (LWR). Erstere findet alle fünf Jahre statt und erhebt bei rund 70 000 Haushalten (1998) sowohl Einkommens- als auch Ausgabenvariablen in Verbindung mit soziökonomischen Strukturmerkmalen. Die zweite Erhebung findet jährlich statt, verfügt aber über einen geringeren Stichprobenumfang, der bis 1998 2 000 Haushalte mit drei bzw. fünf Haushaltstypen als Auswahlgrundlage umfasste. Neben einer Erhöhung des Stichprobenumfangs auf jährlich 6 000 Haushalte und einer Erweiterung der Auswahlgrundlage sind die Merkmalskataloge der beiden Erhebungen EVS und LWR seit 1998 aufeinander abgestimmt (vgl. Kaiser 2000, S. 773 ff.; Chlumsky, Ehling 1997, S. 460 f.).

Es ist geplant, diese beiden Erhebungen ab 2005 zu einer jährlichen Haushaltsbudgeterhebung mit 18 000 bis 24 000 Haushalten zusammenzuführen, was sowohl unter Berücksichtigung der bisherigen Anforderungen an die Erhebungen (z. B. regionale und sozioökonomische Tiefengliederung), als auch neuerer Erfordernisse (z. B. Aktualität, Kosten, Teilnahmebereitschaft) geschehen soll (vgl. Bechtold, Chlumsky, 2000, S. 85).

Bislang ist die Einkommens- und Verbrauchsstichprobe die einzige Erhebung der amtlichen Statistik, die tiefgegliederte Querschnittsdaten zu Einnahmen und Ausgaben privater Haushalte liefert. Darüber hinaus gibt sie als einzige Erhebung in diesem Kontext Aufschluss über sozioökonomische Strukturmerkmale. Auch die zukünftige Erhebung muss den genannten Anforderungen gerecht werden, da die Datennutzer aus Politik und Wissenschaft nicht allein ein repräsentatives Abbild typischer Haushalte fordern, sondern auch verlässliche Informationen über soziale Randgruppen wie beispielsweise Sozialhilfeempfänger oder Alleinerziehende benötigen. Ebenso muss auch weiterhin die Bereitstellung gesicherter regionaler Ergebnisse ermöglicht werden, was bisher ein ausreichend hoher Stichprobenumfang und die Quotierung der EVS u. a. nach Bundesländern zuließ. Es ist jedoch davon auszugehen, dass die zukünftige Erhebung aufgrund des geringeren jährlichen Stichprobenumfangs jene repräsentative Tiefengliederung zunächst nicht bieten kann.

Ferner muss der Datenbedarf der Volkswirtschaftlichen Gesamtrechnungen, der Preisstatistik und der Deutschen Bundesbank berücksichtigt werden. Dieser wurde bisher durch die jährlichen Laufenden Wirtschaftsrechnungen abgedeckt und wird voraussichtlich durch die Erhöhung des jährlichen Stichprobenumfangs auf einer verbesserten Datengrundlage basieren.

---

*) Nicole Ernst, Universität Trier.

Für die Nutzer der derzeitigen und zukünftigen Haushaltsbudgeterhebungen ist somit die Zuverlässigkeit der Ergebnisse mit einer der wichtigsten Aspekte. Einen entscheidenden Einfluss auf die Datenqualität haben neben dem Stichprobenumfang das Erhebungsdesign und die daraus entsprechend dem Datenbedarf abgeleiteten Schätzer. Beide Faktoren dürfen zukünftig keine Verschlechterung in der Qualität der Ergebnisse nach sich ziehen. Wäre dies der Fall, würden beispielsweise Entscheidungen wie die Festsetzung der Regelsätze für Sozialhilfeempfänger auf wesentlich ungenauerem Datenmaterial beruhen, was wiederum weitreichende Fehlentscheidungen (z. B. einer Über- oder Unterbewertung des Bedarfs zum laufenden Lebensunterhalt, der bisher mit Hilfe von EVS-Daten geschätzt wird) nach sich ziehen könnte. Mögliche Kosteneinsparungen beim Stichprobendesign würden dann an anderer Stelle ad absurdum geführt.

Neben den genannten Anforderungen gewinnt die Aktualität einkommensstatistischer Daten zunehmend an Bedeutung, was die neue Erhebung ebenfalls in verbesserter Form ermöglichen soll. Die Erhöhung des jährlichen Stichprobenumfangs von zur Zeit 6 000 Haushalten auf 18 000 bis 24 000 Haushalte ist dabei ein Schritt in die richtige Richtung.

Schließlich sollen Forderungen vom Statistischen Amt der Europäischen Gemeinschaften (Eurostat), dies würde insbesondere den anstehenden Datenbedarf für die Statistics on Income and Living conditions (EU-SILC) betreffen, wenn möglich bei der Umstrukturierung einbezogen werden.

Im Frühjahr 2000 hat das Statistische Bundesamt zwei mögliche Erhebungsdesigns für die zukünftigen Wirtschaftsrechnungen ausgearbeitet und der Arbeitsgruppe "Amtliche Statistik und sozioökonomische Fragestellungen" vorgestellt. Um den Tatsachen rückläufiger Teilnahmebereitschaft der Bevölkerung an Umfragen und steigenden monetären Restriktionen Rechnung zu tragen, machen sich die beiden Vorschläge die Idee der Stichprobenrotation zu Eigen. Aufgabe der Arbeitsgruppe ist es nun, zu entscheiden, welches der beiden Designs besser geeignet ist, den Anforderungskatalog zu erfüllen. Dabei sollen insbesondere die Fragen der Qualität der resultierenden Daten, gemessen anhand der Varianz der Schätzfunktion, die wiederum vom Erhebungsdesign abhängt, sowie die Abdeckung des Datenbedarfs als Entscheidungskriterium herangezogen werden.

2 Rotierende Stichproben

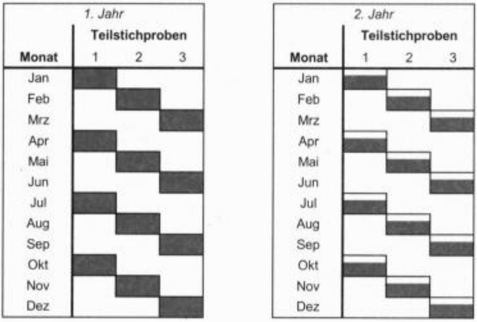
Die grundlegende Frage, die der Entwicklung eines Erhebungsdesigns voranzustellen ist, ist eigentlich die der Definition des Datenbedarfs. Es ist eingangs zu klären, ob Querschnitts- und/oder Längsschnittsdaten benötigt werden. Damit einhergehend ist festzulegen, was mit Hilfe der Stichprobenergebnisse geschätzt werden soll: Parameter oder Veränderungsraten. Da bereits zwei, insbesondere unter dem Gesichtspunkt des Kostendrucks, der schwindenden Teilnahmebereitschaft und der Aktualität, adäquate Designvorschläge existieren, muss der umgekehrte Weg beschritten werden: Es muss erörtert werden, ob die Stichprobendesigns "kompatibel" mit dem eingangs definierten Datenanforderungen sind.

Es können im weitesten Sinne drei Varianten unterschieden werden. In einer rollierenden Stichprobe werden überschneidungsfrei, d. h. ohne wiederholte Befragung zu einem weiteren Zeitpunkt, Teilstichproben nacheinander gezogen. Die zweite Möglichkeit ist die eigentliche Rotation, auch "sampling over time", "sampling on successive occasions with partial replacement of units" oder "sampling for a time series" genannt (vgl. Binder, Hidiroglou 1988, S. 187). Dabei wird ein Prozentsatz einer bereits befragten Teilstichprobe nochmals zum darauf folgenden oder späteren Zeitpunkt in die Erhebung einbezogen. Dieser Prozentsatz wird auch als "overlap" bezeichnet. Die dritte Variante ist das Panel, dessen overlap zwischen den Erhebungszeitpunkten 100% beträgt, alle Stichprobeneinheiten werden für die Befragung im folgenden Zeitpunkt übernommen. Die Unterscheidung dieser drei Varianten ist, wie sich bei der Beurteilung der Präzision später zeigen wird, von großer Bedeutung.

2.1 LWR-Modell

Der erste Vorschlag wird als LWR-Modell bezeichnet, da er auf dem Design der aktuellen Laufenden Wirtschaftsrechnungen aufbaut. Zwischen 4 500 und 6 000 Haushalte werden in drei Teilstichproben mit je 1 500 bis 2 000 Haushalten gegliedert. Diese werden nacheinander, entsprechend einem rollierenden Modell im ersten, zweiten bzw. dritten Monat des ersten Quartals eines Jahres befragt (vgl. Abbildung 1). Insgesamt nimmt jede Teilstichprobe jedoch vier mal in einem Jahr zum jeweils gleichen Zeitpunkt eines Quartals an der Erhebung teil. Auf Jahresebene handelt es sich aufgrund der wiederkehrenden Befragung der Teilnehmer um ein rotierendes Design bzw. Quartalspanel. Die Zahl der erhobenen jährlichen Datensätze beläuft sich auf 18 000 bis 24 000, wobei nur 4 500 bzw. 6 000 Haushalte angeworben werden mussten.

Abbildung 1
Verlauf des LWR-Modells im 1. und 2. Jahr

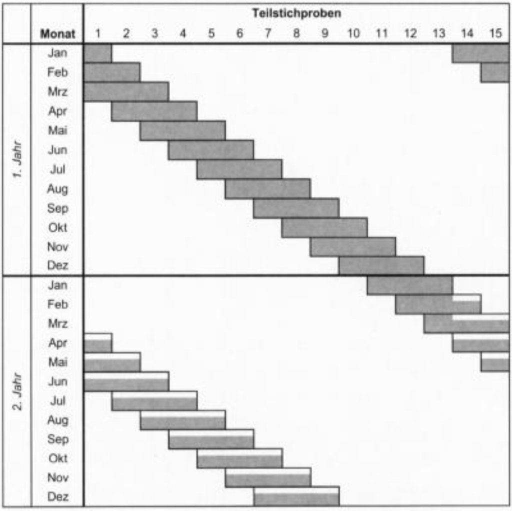


Nach einem Jahr wird mit dem sukzessiven jährlichen Austausch von Stichprobeneinheiten begonnen. Aus jeder Teilstichprobe wird jährlich ein Drittel (500 bis 667 Stichprobeneinheiten) durch neue Haushalte ersetzt, d. h. im zweiten Jahr nehmen nur noch Zweidrittel (66%) aus der Ursprungsstichprobe an der Erhebung teil. Im vierten Jahr basiert die Erhebung auf einer ganz neuen Stichprobe. Die Zahl der erhobenen Datensätze liegt jährlich stets zwischen 18 000 bis 24 000. Über fünf Jahre liegen schließlich zwischen 90 000 und 120 000 Datensätze, bei einer Zahl angeworbener Haushalte von 10 500 bis 14 000 vor.

2.2 Jahresrotation über 15 Monate

Als zweite Variante wird eine "Jahresrotation über 15 Monate" vorgeschlagen: Eine zu befragende Haushaltszahl von 22 500 wird in 15 Teilstichproben im Umfang von je 1 500 Haushalten gegliedert. Im ersten Monat werden drei Teilstichproben, d. h. insgesamt 4 500 Haushalte, befragt. In den folgenden 14 Monaten werden ebenfalls drei Teilstichproben befragt, wobei eine der bereits zuvor befragten durch eine neue, die vierte, ersetzt wird. Der monatliche overlap beträgt somit zwei Drittel. Nach genau einem Jahr, d. h. im Januar des zweiten Jahres, ist jede der 15 Teilstichproben mindestens einmal befragt worden.

Abbildung 2
Verlauf der Jahresrotation über 15 Monate im 1. und 2. Jahr



Im ersten Jahr liegen somit 54 000 Datensätze von 21 000 angeworbenen Haushalten vor. Ab Februar des zweiten Jahres wird die erste Teilstichprobe wieder einbezogen, nachdem ein Drittel (500 Haushalte) durch neue Haushalte ersetzt wurde. Zu dem monatlichen overlap von zwei Dritteln kommt nun noch ein jährlicher overlap von ebenfalls zwei Dritteln hinzu, so dass im zweiten Jahr nur noch 7 000 neue Haushalte gewonnen werden müssen. Schließlich werden in den darauf folgenden Jahren nur noch jeweils 6 000 neue Haushalte benötigt. Erst nach vier Jahren wird jede Teilstichprobe ausrotiert sein und die Gesamtstichprobe komplett durch eine neue ersetzt sein. Mit Hilfe dieses Stichprobendesigns können innerhalb von fünf Jahren insgesamt 270 000 Datensätze bei 46 000 Haushalten ermittelt werden.

3 Schätzer

Aus den eingangs aufgeführten Anforderungen an eine zukünftige Erhebung lassen sich erste benötigte Schätzer ableiten: Parameterschätzer in unterschiedlicher Tiefengliederung (beispielsweise nach bestimmten soziökonomischen Merkmalen oder Regionen) sowie Veränderungsschätzer (beispielsweise für EU-SILC).

Bei beiden vorgestellten Designs werden in jedem Monat bei einer oder mehreren Teilstichproben die relevanten Merkmale erhoben. Da weder eine noch zwei bis drei Teilstichproben einen ausreichend großen Stichprobenumfang für eine tiefgegliederte Schätzung zu einen Zeitpunkt bieten, ist von Seiten des Bundesamtes eine Datenkumulation vorgeschlagen worden. Unter Kumulation ist zu verstehen, dass alle Merkmalsausprägungen des interessierenden Merkmals unabhängig davon, zu welchen Zeitpunkten sie erhoben wurden, mit identischer Gewichtung von je 100% summiert werden. Dadurch können repräsentative Schätzer auch für Randgruppen geliefert werden, was zunächst die Annahme zulässt, dass sich auch die Präzision der Schätzung verbessert. Jedoch darf bei den vorliegenden Designs der overlap nicht unberücksichtigt bleiben, der eine Korrelation der Merkmalsausprägung identischer Stichprobeneinheiten über die Zeit impliziert.

Vereinfacht sähe ein Mittelwertschätzer, der die Datensätze aus r monatlichen Teilstichproben vom Umfang n_r zu einem Schätzer zum Zeitpunkt r kumuliert und mittelt, wie folgt aus:

$$\hat{\mu}_r = \frac{\sum_{t=1}^{r} \bar{x}_t}{r} = \frac{\sum_{t=1}^{r}\sum_{i=1}^{n_r} x_{t,i}}{r \cdot n_r} \text{ mit} \qquad (1)$$

- $\bar{x}_t$ Mittelwert der Teilstichprobe zum Zeitpunkt t (vgl. Cassel, Granström, Lundquist, Selén 1997, S. 2 ff.)

Damit gehen in diesen Schätzer $r{\cdot}n_r$ Ausprägungen des interessierenden Merkmals X ein. Auf diese Art können beliebig viele Teilstichprobenergebnisse summiert werden und so auch unterjährige bzw. mehrjährige Schätzer in tiefer Gliederung gebildet werden.

4 Datenqualität

Da die Datenqualität u. a. als Entscheidungskriterium fungiert, müssen die Varianzen kumulierter Parameterschätzer ermittelt werden. Nach Kish (vgl. Kish, 1987, S. 191 f.). kann die Varianz für einen kumulierten Mittelwertschätzer über r Perioden approximativ wie folgt bestimmt werden:

$$Var(\hat{\mu}_r) = \frac{1}{r^2}\left(\frac{\sum_{t=1}^{r}\sigma_t^2}{n_r} + \frac{\sum_{t/s=1}^{r}\sigma_t\sigma_s}{n_r}P_{ts}\rho_{ts}\right) \quad t \neq s \qquad (2)$$

Unter der Annahme, dass die Varianzen der Teilstichproben σ_r^2 und die Teilstichprobenumfänge n_r identisch sind, ergibt sich der verkürzte Ausdruck:

$$Var(\hat{\mu}_r) = \frac{1}{r}\frac{\sigma^2}{n}\left(1 + \rho_{ts}\frac{P_{ts}}{r}\right) \quad t \neq s \qquad (3)$$

- P_{ts} bezeichnet die Höhe des Anteils der Überschneidung der Stichprobeneinheiten,
- ρ_{ts} ist der Korrelationskoeffizient für das erhobene Merkmal bei identischen Teilstichprobenelementen zwischen zwei Zeitpunkten.

Es zeigt sich, dass es wichtig ist, nach den eingangs definierten rollierenden bzw. rotierenden Varianten bei Verwendung kumulierter Schätzer zu unterscheiden. Bei fehlender Überschneidung der Teilstichprobeneinheiten ist die Korrelation ρ_{ts} annahmegemäß Null. Damit entspräche die Varianz und somit die Qualität des kumulierten Schätzers bei diesem Design dem eines Mittelwertschätzers bei einfacher Zufallsauswahl über $r \cdot n_r$ Elemente. Im konkreten Fall ergäbe sich dies nur bei der Anwendung innerhalb des LWR-Modells auf Quartalsebene. Sobald eine Überschneidung besteht, dies gilt sowohl beim LWR-Modell bei Kumulation über mehr als ein Quartal als auch bei der Jahresrotation über 15 Monate, ist davon auszugehen, dass $\rho \neq 0$ ist. Die Varianz wäre dann größer als bei einer einfachen Zufallsauswahl. Mit steigender Überlappung ist des Weiteren davon auszugehen, dass diese Korrelation steigt und damit die Schätzung für Querschnittsdaten unter Anwendung der Kumulation sich stets verschlechtert.

Bei einer Veränderungsmessung, wenn die Differenz eines Merkmals über zwei oder mehr Zeitpunkte geschätzt wird, entwickelt sich die Präzision gegenteilig. Die Varianz des Schätzers einer mittleren Veränderung über die Zeit berechnet sich approximativ entsprechend nach Kish wie folgt:

$$Var(\hat{\mu}_r) = r\frac{\sigma_r^2}{n}\left(1 - \rho_{ts}\frac{P_{ts}}{r}\right) \quad t \neq s \qquad (4)$$

Diese ersten Resultate werfen die Frage auf, ob die Designs mit der eingangs vorgestellten Zielsetzung in Einklang zu bringen sind. Die Idee der Kumulation ist offensichtlich nicht der richtige Weg, tiefgegliederte Daten basierenden auf rotierenden Designs zu erhalten. Und selbst wenn dies möglich wäre, bestünde noch immer die Frage, welche Aussagekraft eine Kumulation über fünf Jahre hätte.

Literaturhinweise

Bechtold, S.; Chlumsky, J. (2000): Amtliche Statistik und sozioökonomische Fragestellungen. Überlegungen zur Weiterentwicklung des bestehenden Systems der Haushaltsstichproben, in: Wirtschaft und Statistik, H. 2 , S. 81.

Binder, D. A.; Hidiroglou, M. A.: Sampling in time, in: *Krishnaiah, P. R.; Rao, C. R. (1988)*: Handbook of Statistics, Vol. 6, S. 187.

Cassel, C. M.; Granström, F.; Lundquist, P.; Selén, J. (1997): Cumulating Data from Household Budget Survey. Some Results for Model Based Calibration Techniques Applied to Swedish Data. Report financed by European Communities, LOT 23.

Chlumsky, J.; Ehling, M. (1997): Grundzüge des künftigen Konzepts der Wirtschaftsrechnungen der privaten Haushalte, in Wirtschaft und Statistik, H. 7, S. 455.

Kaiser, J. (2000): Die Statistik der laufenden Wirtschaftrechnungen in neu konzipierter Form, in: Wirtschaft und Statistik, H. 10, S. 773.

Kish, L. (1987): Sampling Design for Research.

Walter Krug / Manfred Ehling / Nicole Ernst*)

Ergebnis der Abschlussdiskussion

Ziel des Statistischen Bundesamtes ist es, repräsentative Ergebnisse auf Basis einer großen Stichprobe zu erhalten, wie es die Nutzer wünschen, wobei die Erhebung anhand vieler kleiner Stichproben durchgeführt wurde. Drei Fragen resultierten nach Ehling aus dieser Zielvorgabe als Grundlage für den Workshop: Erstens, welches ist das beste Stichprobendesign, das LWR-Modell, die Jahresrotation über 15 Monate oder ein völlig anderes Design? An zweiter Stelle ist zu erörtern, welches die beste Methode ist, um Ergebnisse aus mehreren Jahren zusammenzufassen, sofern ein rotierendes oder rollierendes Design gewählt wird? Insbesondere wäre zu klären, ob eine Kumulation in Betracht zu ziehen wäre. Schließlich ist zu überlegen, wie der Stichprobenfehler des adäquatesten Designs zu bestimmen wäre?

Nach Krug hat der Workshop gezeigt, dass die bisher vorgeschlagenen Designs, das LWR-Modell und die Jahresrotation über 15 Monate aus unterschiedlichen Gründen nicht verwendet werden sollten und ein Stichprobenplan aus der Kombination von rollierendem und rotierenden Design zu verwenden, wobei EU-SILC integriert werden sollte.

Entgegen der allgemeinen Tendenz in der amtlichen Statistik, die Reich erkennt, kontinuierliche Erhebungen zu entwickeln, die die Entwicklung bzw. Veränderung insbesondere individueller sozialer Parameter dokumentieren, und diese auf die Neugestaltung der Wirtschaftsrechnungen privater Haushalte zu übertragen, rät Verma, ein rollierendes Design einzuführen, d. h. ein Design ohne Überlappung der Stichprobeneinheiten. Dieses Design wäre gerade für Querschnittsanalysen zu bevorzugen. So könnte beispielsweise in Betracht gezogen werden, so Schadendorf, ob der bisherige Stichprobenumfang der ESV nicht über fünf Jahre rolliert werden könnte. Seiner Ansicht nach müssten bereits hier Kosten- und Aufwandsersparnisse realisiert werden können. Für dieses Design spricht, dass eine Erhebung über Einnahmen und Ausgaben im Vergleich zu einer Arbeitskräftestichprobe derart komplex und für die Teilnehmer belastungsintensiv sei, dass eine Rotation, die eine wiederholte Befragung der Erhebungseinheiten erfordert, als äußert gewagt und schwierig anzusehen wäre. Zu der außerordentlichen Belastung bei der Ermittlung des Konsumverhaltens kommt die Schwierigkeit der Interpretation von Konsumveränderungen hinzu, gibt Aasness zu bedenken. Es sei nicht eindeutig zu klären, ob eine mögliche individuell gemessene Konsumveränderung aus einer realen Änderung des Konsums oder aus einer Verhaltensänderung oder gar einem Wandel in der Qualität der Beantwortung resultiert. Eine Längsschnittanalyse sei somit für diesen Sachverhalt nicht in Erwägung zu ziehen.

Neben der geringeren Teilnehmerbelastung, ein sehr bedeutender Aspekt im Rahmen einer Erhebung, so Rendtel, spricht auch die Möglichkeit der effizienten, d. h. eine die Fehlervarianzen mindernde Anwendung der Kumulation, für ein rollierendes

*) Prof. Dr. Walter Krug, Universität Trier;
Dr. Manfred Ehling, Statistisches Bundesamt, Wiesbaden;
Nicole Ernst, Universität Trier.

Design. Der tiefgegliederte Datenbedarf der Nutzer kann dadurch adäquat abgedeckt werden. Zwei Dinge sind jedoch zu berücksichtigen: Die Erfahrungen in anderen Ländern haben gezeigt, dass Modelle durchaus akzeptable Ergebnisse liefern können. Es muss somit auch darüber nachgedacht werden, ob und auf welcher Ebene eventuell Modelle bei einer zukünftigen Erhebung Anwendung finden könnten. Des Weiteren haben die von Møller dargelegten dänischen Erfahrungen gezeigt, dass kumulierte Schätzer interpretatorische Probleme mit sich bringen können. In Abhängigkeit der Werte die kumuliert werden und der Wahl des Bezugszeitpunktes, ist ein Ergebnisvergleich zwischen aufeinander folgenden Jahresschätzern nicht ohne weiteres möglich.

Die endgültige Ausgestaltung des rollierenden Schemas sollte jedoch erst erfolgen, empfiehlt Verma, wenn die Prioritäten des Anforderungskataloges feststünden: Soll das Design mehr Gewicht auf die Abdeckung des aktuellen unterjährigen Datenbedarfs legen oder auf den repräsentativ tiefgegliederten jährlichen bzw. mehrjährigen Datenbedarf? Diesbezüglich merkte Mejers an, dass die europäischen Kommunen keine jährlichen Daten benötigten, sondern nur alle fünf Jahre Daten für die Preisindizes und deren Gewichte erfragt würden. Daher empfiehlt sie eine pragmatische Umsetzung des Anforderungskataloges.

Bezüglich des Datenbedarfs müsse auch darüber nachgedacht werden, die Belastung der Befragten bei Haushaltsbudgeterhebungen merklich zu senken. Dies könne über geänderte Tagebücher oder über eine Reduktion der Vielzahl von Einkommensvariablen laut Verma geschehen. Es habe sich in anderen Ländern gezeigt, dass nicht alle Variablen notwendig sind; sie könnten im Bedarfsfall durch zusätzliche kleinere Erhebungen ermittelt werden.

Die Entwicklung eines Access-Panels aus dem Mikrozensus wurde für den Fall als positiv bewertet, dass die Erhebungseinheiten nur aus ausrotierten Unterstichproben des Mikrozensus resultieren. So könne eine mögliche "Aushöhlung" des Mikrozensus durch die belastungsintensive Haushaltsbudgeterhebung vermieden werden argumentierte Verma. Schließlich sollten aus dem selben Grund EU-SILC und die Haushaltsbudgeterhebung getrennt, mit jeweils eigens auf den Datenbedarf abgestimmten Stichprobendesigns aus dem Access-Panel entwickelt werden.

Es sei jedoch zu beachten, so v. d. Heyden, dass die Teilnahmebereitschaft beim Access-Panel nicht zu hoch eingeschätzt werden sollte. Die Vermutung läge nahe, so gerade einmal 10% der ausrotierten Mikrozensus Haushalte bereit seine, auch zukünftig an Erhebungen teilzunehmen. Erste Test würden diese Vermutung durchaus zu lassen, daher sei das Access-Panel auch nur eine Lösung (Ehling). Bezüglich der Anwerbung der Access-Panel Teilnehmer wurde von Verma betont, dass eine persönliche Rekrutierung durch einen Interviewer die Bereitschaft zur Teilnahme merklich erhöhe.

Manfred Ehling*)

Weiterentwicklung der Wirtschaftsrechnungen und Integration neuer Datenanforderungen der Europäischen Union (EU)

Folgerungen aus dem Workshop[1)]

1 Einleitung

Im Rahmen des Workshops sind Modelle für rollierende bzw. rotierende Stichprobendesigns für eine jährliche Haushaltsbudgeterhebung vorgestellt worden. Ein zentrales Ergebnis der Veranstaltung stellt die Abkehr von den bisher diskutierten Stichprobenmodellen dar. Die Überlegungen gingen von großen Überlappungen der Stichproben eines Jahres und des darauf folgenden Jahres aus, die hinderlich für die Kumulierung der Ergebnisse mehrerer Jahre sind. Nach dem Workshop wurden Weiterentwicklungen vorgenommen, die nun Modelle mit jährlich komplett ausgetauschten Stichproben vorsehen, die eine Kumulation eines auf mehrere Jahre bezogenen Durchschnitts ermöglichen. Derartige Kumulierungsverfahren werden – wie im Rahmen des Workshops vorgestellt – erfolgreich in der amtlichen Statistik Dänemarks und Norwegens eingesetzt. Andererseits wurden Ideen aufgegriffen, die auf einer stärkeren Kontinuität der bisher laufenden Erhebungen in den Wirtschaftsrechnungen aufbauen.

Weiterhin plant die deutsche amtliche Statistik, die Daten für die Erhebung EU-Statistiken über Einkommen und Lebensbedingungen (EU-SILC) in das System der Wirtschaftsrechnungen der privaten Haushalte zu integrieren. Diskutiert werden derzeit verschiedene Modelle, mit denen eine Integration sowohl der Querschnitts- als auch der Längsschnittskomponente von EU-SILC möglich ist.

Mit EU-SILC sollen erstmals für das Jahr 2003 europaweit vergleichbare Querschnittsdaten zu den Themenbereichen "Demographie und Bildungsstand", "Soziale Ausgrenzung", "Einkommen", "Erwerbstätigkeit" und "Wohnen" bereitgestellt werden. Angestrebt wird ein Stichprobenumfang von 80 000 europäischen und 8 250 deutschen Haushalten. Grundsätzlich soll es möglich sein, bereits vorhandene nationale Datenquellen für die Bereitstellung der Informationen zu nutzen. Dies schließt auch ein, dass die Informationen aus verschiedenen Datenquellen stammen können, sofern sie auf Haushaltsebene verknüpft werden können.

Zusätzlich zu den Querschnittsdaten soll EU-SILC auch Längsschnittsdaten zu den Themen Einkommen und soziale Ausgrenzung bereitstellen, die in Form eines rotierenden Panels über einen Zeitraum von vier Jahren erhoben werden sollen. Für die Längsschnittsdaten müsste von Deutschland voraussichtlich ein Stichprobenumfang von 6 000 Haushalten bereitgestellt werden. Die Verknüpfung der Längsschnittsda-

---

*) Dr. Manfred Ehling, Statistisches Bundesamt, Wiesbaden.

1) Stand: Dezember 2001.

ten mit den Querschnittsdaten wird als wünschenswert, jedoch nicht als notwendig erachtet. Mit EU-SILC sollen folgende Erhebungsinhalte abgedeckt werden:

- Basisinformationen und Demographie
- Einkommen: Gesamteinkommen und ausführliche Gliederung nach Komponenten, jeweils Brutto- und Nettoangaben
- Zahlungsrückstände
- Nichtmonetäre Deprivationsindikatoren
- Physisches und soziales Umfeld
- Wohnungsmerkmale (z. B. Anzahl der Räume, Ausstattungsmerkmale)
- Bildungsstand
- Gesundheit
- Erwerbsbeteiligung
- Nur im Querschnittteil:
  - Mietausgaben
  - Wohnungskredite
  - Sonstige Kredite
- Nur im Längsschnittteil:
  - Erwerbsbiographie
  - Aktivitätenkalender

Die Planungen zu EU-SILC beziehen auch die Option ein, die neuen Erhebungsinhalte in die jeweiligen nationalen statistischen Systeme zu integrieren. Es muss also auf nationaler Ebene keine eigene, neue Erhebung eingeführt werden, sofern die Daten durch bestehende oder angepasste nationale Statistiken bereitgestellt werden können. Das Statistische Amt der Europäischen Gemeinschaften (Eurostat) plant das Instrument durch eine Verordnung des Europäischen Parlaments und des Rates zu regeln

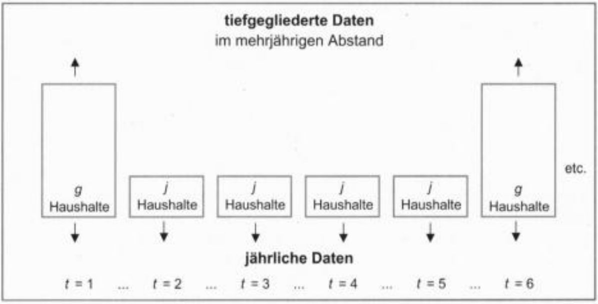
Im Rahmen dieses Beitrags soll der gegenwärtige Stand der Konzeption für eine Weiterentwicklung der Einkommens- und Verbrauchstichprobe (EVS) und der Laufenden Wirtschaftsrechnungen (LWR) sowie die Möglichkeiten der Integration von EU-SILC vorgestellt werden. Die Überlegungen stellen einen Werkstattbericht dar, in dem die drei Varianten skizziert werden, die aktuell in der amtlichen Statistik mit den Nutzern diskutiert werden. Es ist geplant, im zweiten Halbjahr 2002 die drei Varianten in Testerhebungen zu überprüfen. Die Testvarianten A und B zielen auf einen weitreichenden Umbau des derzeitigen Systems der Wirtschaftsrechnungen zu einer jährlichen Erhebung ab. Die Testvariante C baut auf einem Erhalt und Modifikationen dieses Systems auf. Zuerst sollen aber einige stichprobenmethodischen Grundsatzüberlegungen zur Kumulierung von Daten und zu alternierenden Stichproben diskutiert werden.

2 Kumulierung oder alternierende Stichproben

Beim Umbau der Haushaltsbudgeterhebungen sollen einerseits jährliche Ergebnisse produziert werden sowie andererseits, zumindest in mehrjährigem Abstand, ausreichend hohe Stichprobenumfänge zur Verfügung stehen, um tiefgegliederte Daten ausweisen zu können. Um beiden Anforderungen gerecht zu werden und den jährlichen Stichprobenumfang nicht zu hoch ansetzen zu müssen, können grundsätzlich zwei unterschiedliche Wege beschritten werden:

Bei einer alternierenden Stichprobe wird in regelmäßigen Abständen der jährliche Stichprobenumfang der Erhebung erhöht. Auf diese Weise könnte zum Beispiel alle fünf Jahre eine "große" Erhebung durchgeführt werden und daraus tiefgegliederte Ergebnisse bereitgestellt werden. In den übrigen Jahren werden kleinere Stichproben gezogen (vgl. Abbildung 1):

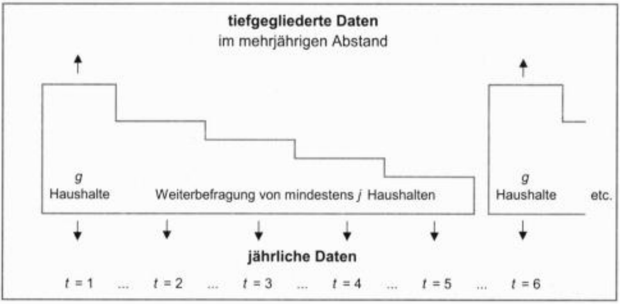
Abbildung 1
Alternierende Stichproben (a)



Eine Variante alternierender Stichproben besteht darin, einige der Haushalte aus der großen Stichprobe *g* weiterzubefragen und deren Befragungsergebnisse auch für die folgenden Stichproben *j* heranzuziehen (vgl. Abbildung 2, S. 23).

Um nach mehreren Jahren noch *j* Haushalte aus der großen Stichprobe weiterbefragen zu können, müssen zur Berücksichtigung der Panelmortalität anfänglich mehr als *j* Haushalte in das Panel übernommen werden. Alternativ wäre es möglich, nur *j* Haushalte aus der großen Stichprobe zu überhehmen und zum Ausgleich der Ausfälle in den folgenden Jahren neue Haushalte aufzunehmen. Im Vergleich zum hier dargestellten Design könnte dann auf einige Wiederholungsbefragungen verzichtet werden und es müssten dafür Neurekrutierungen vorgenommen werden. Im Hinblick auf die Kosten lohnt sich diese "Auffrischung" in einer Beispielrechnung nur dann, wenn eine Neurekrutierung einschließlich des Erstinterviews weniger als dreimal so teuer ist, wie eine Wiederholungsbefragung. Da diese Voraussetzung nicht unbedingt gegeben ist und das hier dargestellte Design einfach darzustellen ist, wird im folgenden bei der Variante (b) der alternierenden Stichprobe von diesem Schema ausgegangen. Ebenso wäre jedoch die Alternative – Übernahme von *j* Haushalten und Auffrischung der jährlichen Stichprobe – möglich.

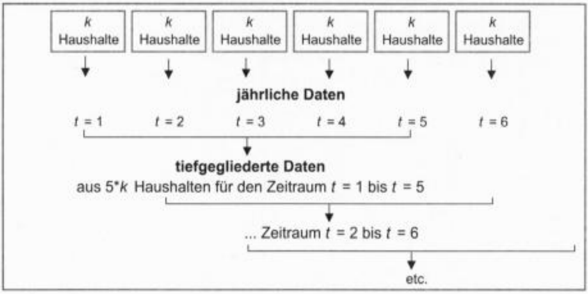
Abbildung 2
Alternierende Stichprobenumfänge (b)



Der Aufbau dieser Variante (b) ist der gegenwärtigen LWR und ihrer Verknüpfung mit der EVS sehr ähnlich. Die LWR-Teilnehmer werden gegenwärtig größtenteils aus der EVS übernommen und über längere Zeit hinweg weiterbefragt, um jährliche Ergebnisse zu erhalten. Ein Nachteil dieser Variante besteht darin, dass sich die Querschnittsrepräsentativität im Zeitablauf bis zur nächsten großen Erhebung zunehmend verschlechtert, da immer bestimmte Haushaltstypen ausscheiden werden (selektive Ausfälle). Dieser Nachteil lässt sich auch durch regelmäßige Auffrischungen nicht beseitigen, da der Umfang dieser Auffrischungen im Verhältnis zum jährlichen Stichprobenumfang relativ gering ist. Der Vorteil der Variante (b) gegenüber (a) besteht darin, dass insgesamt weniger Haushalte für die Befragung rekrutiert werden müssen.

Bei der zweiten Möglichkeit, der Kumulierung, werden zur Befriedigung des nationalen Bedarfs an tiefgegliederten Daten die Ergebnisse aus mehreren Jahrgängen der Haushaltsbudgeterhebung zusammengefasst, um auf diese Weise in jedem Jahrgang eine höhere Anzahl an Befragungsergebnissen unterschiedlicher Haushalte zu erhalten (vgl. Abbildung 3, S. 24). In methodischer Hinsicht können unterschiedliche Kumulierungsverfahren angewandt werden – auf Detailfragen zur Kumulierungsmethode soll hier jedoch nicht näher eingegangen werden (vgl. dazu die Beiträge in diesem Band von Ernst, S. 11 ff.; Verma, S. 35 ff.; Selén, S. 74 ff. und Møller, S. 88 ff.).

Abbildung 3
Kumulierung jährlicher Daten



Die Kumulierung weist gegenüber alternierenden Stichprobenumfängen folgende Vorteile auf:

- Die Erhebungsarbeiten fallen gleichmäßig über alle Jahre verteilt an. Hierdurch können Erhebungskosten eingespart werden, da personelle Ressourcen effizienter eingesetzt werden können.
- Im Vergleich zur Variante (a) der alternierenden Stichproben liegt die Zahl der jährlich erforderlichen Neurekrutierungen und auch die Anzahl durchzuführender Interviews bei der Kumulierung niedriger. Bei der Variante (b) ist die Zahl der Neurekrutierungen nicht höher als bei der Kumulierung, es müssen aber mehr Interviews durchgeführt werden.
- Die gleichmäßige Verteilung der Arbeit ermöglicht eine höhere Aktualität der tiefgegliederten Ergebnisse, da die Auswertungsarbeiten schneller durchgeführt werden können. Die jährlichen Stichproben können schon in den Jahren vor dem Kumulierungsjahr ausgewertet werden und die Zwischenergebnisse dann direkt in die Kumulierungsrechnung übernommen werden.
- Die Ergebnisse sind bei der Kumulierung im Durchschnitt aktueller, da sie sich immer auf den Durchschnitt zum Beispiel der letzten fünf Jahre beziehen. Bei alternierenden Stichproben können zwar anschließend an eine große Erhebung aktuelle Daten ausgewiesen werden, diese verlieren jedoch mit jedem Folgejahr bis zur nächsten großen Auswertung an Aktualität.

Gleichzeitig ist die Kumulierung mit folgenden Nachteilen verbunden:

- Die kumulierten Daten beziehen sich nie auf ein genau bestimmtes Jahr, sondern sind immer Durchschnittswerte für den Kumulierungszeitraum. Hieraus können sich Interpretationsprobleme ergeben.

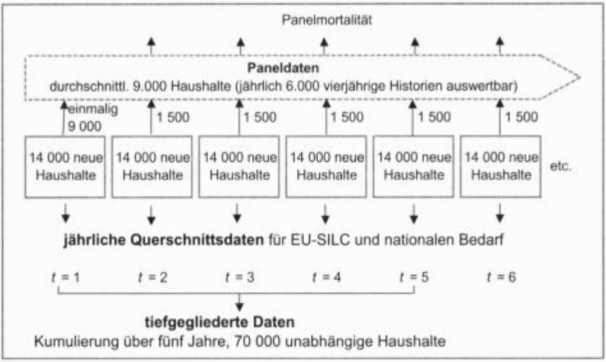
- Eine Kumulierung ist nur möglich, wenn Merkmalsabgrenzungen, Klassifikationen etc. über den Betrachtungszeitraum hinweg konstant gehalten werden.
- Die Stichprobe muss jährlich vollständig ausgetauscht werden, um eine überlappungsfreie Kumulierung zu ermöglichen. Überlappungsfrei bedeutet, dass bei einer Kumulierungsberechnung jede Stichprobeneinheit nur ein einziges Mal berücksichtigt wird. Eine überlappende Kumulierung, bei der die Interviews von mehreren Zeitpunkten derselben Einheit einbezogen werden, ist zwar aus methodischer Sicht möglich, sie führt jedoch zu weniger präzisen Ergebnissen und sollte daher vermieden werden.

Vorschläge für neue Erhebungsdesigns

3.1 Variante A: Grundstruktur einer integrierten Haushaltsbudgeterhebung

Ein integriertes Erhebungsdesign für die jährlichen Haushaltsbudgeterhebungen könnte folgende Grundstruktur aufweisen (vgl. Abbildung 4):

Abbildung 4
Vorschlag für die integrierten jährlichen Haushaltsbudgeterhebungen



1. Jährlich werden 14 000 Haushalte *neu* für die Haushaltsbudgeterhebungen angeworben, womit die Anforderungen aus dem Querschnittsteil von EU-SILC erfüllt wären.
2. Zur Befriedigung des nationalen Bedarfs an tiefgegliederten Daten werden die Befragungsergebnisse jeweils über *fünf* Jahre kumuliert, wobei jeder Haushalt nur ein einziges Mal in die Kumulierungsberechnung einbezogen werden sollte.

Um Längsschnittsdaten für EU-SILC zu erhalten, werden *einmalig* 9 000 derjenigen Haushalte, die für die Querschnittsdaten angeworben wurden, in ein Panel übernommen. In den Folgejahren wird zum Ausgleich der Panelmortalität das Panel *jährlich* aufgefrischt, indem jeweils 1 500 der für die Querschnittsdaten rekrutierten Haushalte in das Panel übernommen werden. Alle Haushalte des Panels müssen "weiterverfolgt", d.h. bei Umzug an ihrem neuen Wohnort weiterbefragt werden.

Mit diesem Erhebungsdesign wären sowohl die nationalen Datenbedürfnisse befriedigt als auch die Anforderungen aus EU-SILC erfüllt. Es müssten jährlich 14 000 Haushalte neu angeworben und 23 000 Befragungen durchgeführt werden.

Sowohl die Einkommen, als auch die Ausgaben sollten bei diesem Vorschlag per Anschreibung erhoben werden. Die übrigen Merkmale (soziale Ausgrenzung, Wohnbedingungen etc.) würden in einem Einführungsinterview erfragt werden. Eventuell können nach der Erstbefragung die Erhebungsmerkmale aus dem Einführungsinterview anhand von Selbstausfüllerbogen abgefragt werden.

Bei der Anschreibung werden die, während des Berichtsmonats anfallenden Einkommen direkt in das Haushaltsbuch eingetragen. Im Vergleich zur Erhebung von Einkommensangaben per Interview, in dem retrospektive Fragen zum Einkommen des Vorjahres oder der letzten 12 Monate gestellt werden müssten, dürften insbesondere einmalige Einkünfte, die im Berichtszeitraum anfallen, bei der Anschreibung besser erfasst werden. Die Anschreibungsmethode lässt den Befragten außerdem Zeit, eigene Unterlagen durchzusehen. Falls für EU-SILC die Einkommen der Vorperiode zusammen mit den übrigen Variablen der *aktuellen* Periode ausgewertet werden müssen (so die Vorgehensweise beim ECHP), so müssten die übrigen Merkmale für EU-SILC in einem *Schlussinterview* am Ende des Jahres oder am Anfang des Folgejahres erfragt werden.

Das bei den gegenwärtigen Haushaltsbudgeterhebungen angewandte Prinzip der *rotierenden* Anschreibung kann für die Befriedigung des Datenbedarfs aus EU-SILC nicht übernommen werden: Für EU-SILC werden die Einkommensangaben der Haushalte auf der Mikroebene benötigt, d.h. für jeden einzelnen Haushalt müssen Angaben für ein vollständiges Jahr vorliegen. Die gegenwärtige Praxis, fehlende Monate der Haushalte anhand von Angaben anderer Haushalte zu ergänzen, ist daher nicht möglich. Die Rotationsmodelle der gegenwärtigen EVS und LWR könnten demnach nicht direkt übernommen werden. Das gilt insbesondere für das Rotationsschema der EVS, bei der ein Haushalt nur für ein einziges Quartal des Jahres anschreibt. Günstiger ist in dieser Hinsicht die gegenwärtige LWR, bei der sich die Haushalte im Jahresverlauf mit der Anschreibung abwechseln und ein Haushalt in allen vier Quartalen jeweils für einen Monat anschreibt – aber auch hier würden die restlichen beiden Monate des Quartals fehlen.

Eine vollständige Anschreibung des Einkommens über alle 12 Monate eines Jahres dürfte jedoch nicht in Frage kommen, weil die Befragtenbelastung, die Auswertungskosten und die Auswertungsdauer sehr hoch wären. Die Anschreibung der Einkommen über 12 Monate wurde ursprünglich in den Einkommens- und Verbrauchsstichproben eingesetzt und unter anderem um eine höhere Aktualität der

Ergebnisse zu erreichen, aber auch wegen ansteigenden Antwortausfällen nach der EVS 1993 aufgegeben.

Ein indirekter Weg der Datenbereitstellung kann darin bestehen, fehlende Einkommensmonate unter Zuhilfenahme der Einkommensangaben von Haushalten derselben Schicht zu imputieren, die an anderen Monaten anschreiben. Damit wären zwar formal die Anforderungen aus EU-SILC erfüllt, für jeden Haushalt vollständige Jahresangaben bereitzustellen – inhaltlich wären jedoch keine zusätzlichen Informationen gegenüber der gegenwärtigen Praxis der Durchschnittsbildung über die Erhebungsmonate gewonnen. Bei diesem Ausweg wäre daher die Kritik der Datennutzer an den amtlichen Zahlen bereits jetzt absehbar.

Um dennoch die Vorteile der Anschreibung zu nutzen, könnte diese mit retrospektiven Abfragen zurückliegender Daten kombiniert werden. Zum Beispiel könnten alle Haushalte jeweils im letzten Monat eines Quartals (März, Juni, September und Dezember) ihr aktuelles Einkommen anschreiben und gleichzeitig jeweils die beiden fehlenden Vormonate retrospektiv ergänzen. Im Prinzip würden die Haushalte dann viermal im Jahr ihr Quartalseinkommen angeben, womit folgende Vorteile für die Datenqualität verbunden wären:

- Die Haushalte bräuchten sich über Informationen über weit zurückliegende Zeiträume nicht zu erinnern und einmalige Einkommen oder auch Einkommensausfälle der Haushalte würden trotzdem relativ gut erfasst. Die Reichweite der retrospektiven Fragen beträgt nur 2 bis 3 Monate und ist somit deutlich kürzer als bei der reinen Interviewbefragung.
- Das Prinzip der Anschreibung würde bei diesem Vorschlag (dann allerdings ohne Rotation) beibehalten. Das bedeutet, dass die Befragten mehr Zeit für die Informationsbeschaffung haben, um zum Beispiel Steuerunterlagen oder Gehaltszettel heranzuziehen, und es können daher genauere Ergebnisse erwartet werden.
- Zusammen mit den Quartalseinkommen könnten auch gleichzeitig grundlegende Quartalsangaben zur Erwerbstätigkeit erhoben werden. Dadurch würden die Einkommensangaben besser mit den Angaben zur Erwerbstätigkeit übereinstimmen. Weiterhin könnten die so erzeugten Informationen unmittelbar für einen weiteren Datenbedarf aus EU-SILC, nämlich die Angaben zu beruflichen Übergängen, verwendet werden.

Es würde sich auch die Möglichkeit bieten, analog zur Verteilung der Einkommens- und Erwerbstätigkeitsbefragung über das Jahr, auch die restlichen Fragen des Einkommensmoduls (soziale Ausgrenzung und Wohnbedingungen) auf vier Interviewzeitpunkte eines Jahres zu verteilen. Das hätte den Vorteil, dass diese Themengebiete nach und nach erfragt würden und somit die Länge des einzelnen Quartalsfragebogens reduziert werden könnte. Diese Option kommt jedoch eher nur für Selbstausfüllerfragebögen in Betracht, da bei der Interviewbefragung sich die Wegekosten vervierfachen würden.

Um die Befragtenbelastung in Grenzen zu halten, sollten die Ausgaben nur in der Gliederungstiefe und der Form angeschrieben werden, die gegenwärtig bei den EVS angewandt werden. Das bedeutet insbesondere, dass auf die offene Anschreibung und die Feinanschreibung soweit möglich verzichtet werden sollte. Die Frage, ob die von der amtlichen Statistik benötigten Informationen (Preisstatistik, VGR) nur durch die Beibehaltung der gegenwärtigen Feinanschreibung bereitgestellt werden können oder ob ein Fremdbezug der benötigten Daten möglich wäre, wird gegenwärtig geklärt. Eventuell kann auch durch die Ergänzung um Informationen zu den Nahrungs- und Genussmitteln durch Daten eines privaten Marktforschungsinstituts (GfK oder A.C. Nielsen) der Aufwand für die Feinaufzeichnung soweit reduziert werden, dass sie beibehalten und von allen Haushalten gleichermaßen durchgeführt werden könnte. Dies hätte zum einen den Vorteil, dass sich die Haushalte nicht auf verschiedene Anschreibungstiefen einstellen müssten, zum anderen hätte die Preisstatistik deutlich validere Ergebnisse zu erwarten als mit den derzeitigen LWR. Sowohl die veränderte Datenerhebung als auch der Zukauf von Daten müssten vor einer Entscheidung für dieses Vorgehen sorgfältig getestet werden.

Da bei den Ausgaben keine vollständigen Jahresangaben auf der Mikroebene benötigt werden, kann die unterjährige Rotation bei der Anschreibung beibehalten werden. Es stünden dann – wie bisher auch – Durchschnittsangaben für die Ausgabendaten einzelner Schichten zur Verfügung, die Mittelwerte sowohl über die Anschreibungsmonate als auch über die Stichprobeneinheiten einer Schicht darstellen. Im gegenwärtigen LWR-Modell schreiben drei verschiedene Unterstichproben jeweils für einen Monat im Quartal an, während im EVS-Modell in vier Unterstichproben die Ausgaben jeweils für ein ganzes Quartal aufgezeichnet werden. Das LWR-Modell hat damit den Vorteil, dass Verzerrungen aufgrund von Schwankungen des Verbrauchsverhaltens besser ausgeglichen werden. Beim EVS-Modell müssen die Haushalte nur für insgesamt drei Monate anschreiben, was mit einer geringeren Belastung verbunden ist. Die Entscheidung über das anzuwendende Rotationsmodell ist noch offen.

Insgesamt sollten folgende Befragungsmethoden zum Einsatz kommen:

1. Einkommen:
   - Anschreibung an vier Monaten des Jahres: März, Juni, September, Dezember.
   - Retrospektive Ergänzung der übrigen Monate.
2. Ausgaben:
   - Anschreibung über drei Monate (EVS-Modell) oder an vier Monaten des Jahres (LWR-Modell).
   - Bildung von Durchschnitten über die Monate und die Stichprobeneinheiten.
   - Detaillierungsgrad der Ausgaben-Anschreibung so wie in den gegenwärtigen EVS.
   - Prüfung des Zukaufs feingegliederter Ausgabendaten.
3. Fragen zur Sozialen Ausgrenzung und weitere Merkmale:
   - Beantwortung von Fragen in einem Einführungsinterview, bei Wiederholungsbefragungen evtl. per Selbstausfüllerbogen.
4. Daten zum Längsschnittteil von EU-SILC:
   - Praktisch keine zusätzlichen Fragen notwendig.

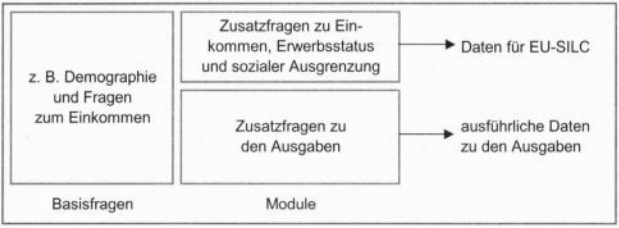
3.2 Variante B: Modularer Aufbau der integrierten Haushaltsbudgeterhebung

Die in den vorigen Abschnitten dargestellte Grundstruktur ist für sich genommen ein tragfähiges Erhebungsdesign für die zukünftigen integrierten Haushaltsbudgeterhebungen. Als mögliche Variante könnte dieses Erhebungsdesign um einen modularen Aufbau ergänzt werden, bei dem den Befragten entweder Einkommens- oder Ausgabenfragen vorgelegt werden. Damit könnten folgende Ziele verfolgt werden:

- Entlastung der Befragten.
- Höhere Responserate bei der Erhebung der Einkommen und damit Verbesserung der Repräsentativität der Einkommensdaten.
- Mehr Spielraum bei der Ausgabenerhebung (insbesondere für die Feinanschreibung) und damit genauere Ausgabendaten.

Beim modularen Aufbau würde der Datenbedarf für EU-SILC aus dem Einkommensmodul gespeist, das dann auch Fragen zum Erwerbsstatus und zur sozialen Ausgrenzung enthalten müsste (vgl. Abbildung 5). Die auf nationaler Ebene benötigten Daten zu den Ausgaben würden durch das Ausgabenmodul bereitgestellt. Tiefgegliederte Angaben zu den Einkommen für den nationalen Bedarf können aus beiden Modulen zusammen bezogen werden. Mit der modularen Struktur könnte gewährleistet werden, dass die Befragtenbelastung nicht über das bereits jetzt sehr hohe Belastungsniveau bei den bisherigen Haushaltsbudgeterhebungen hinausgeht.

Abbildung 5
Modularer Aufbau der Haushaltsbudgeterhebung



Der modulare Aufbau hat nicht nur zur Folge, dass die Befragtenbelastung auf zwei unterschiedliche Module aufgeteilt wird. Zusätzlich besteht die Möglichkeit, dass die Ergebnisse aus einzelnen Modulen anhand der Ergebnisse aus den Basisfragen justiert werden könnten, da diese mit höheren Stichprobenumfängen erhoben werden. Falls die Basisfragen mit den Zusatzfragen eines Moduls stark korreliert sind, kann anhand der Justierungsmöglichkeit gegebenenfalls die Varianz der Ergebnisse in diesem Modul verringert werden, beziehungsweise könnte bei gleichbleibender

Varianz der Stichprobenumfang innerhalb des Moduls etwas gesenkt werden. Darüber hinaus können die Informationen aus den Basisfragen möglicherweise dazu verwendet werden, den Nonresponse-Bias des belastungsintensiven Ausgabenmoduls zu korrigieren.[2)]

Da bei der modularen Erhebungsstruktur eine möglichst geringe Befragtenbelastung im Einkommensteil von hoher Bedeutung ist, sollten die Basisfragen jedoch möglichst nicht über das Erhebungsprogramm des Einkommensmoduls hinausgehen. Daher wäre es denkbar, dass die Basisfragen nur aus Einkommensfragen und demographischen Merkmalen bestehen, die sowohl im Einkommens- als auch im Ausgabenmodul auf die gleiche Weise erfragt werden. Eine spezielle Justierung und Korrektur des Nonresponse würde dann nur beim belastungsintensiven Ausgabenmodul anhand der mit höheren Stichprobenumfängen vorliegenden Einkommensergebnisse vorgenommen. Die Ausgabendaten könnten dabei auch anhand der Einkommensangaben aus den Basisfragen plausibilisiert werden. Die Ausgabendaten könnten nach wie vor mit den Einkommensdaten aus den Basisfragen verknüpft werden – umgekehrt wäre im Einkommensmodul eine solche Verknüpfung allerdings nicht mehr möglich.

Schließlich besteht die Möglichkeit, bei der Kumulierung von Einkommensangaben nicht nur auf die Daten des Einkommensmoduls, sondern auch auf die Einkommensangaben des Ausgabenmoduls zurückzugreifen. Dieser Weg könnte bei der Erstellung von tiefgegliederten Einkommensergebnissen genutzt werden und würde Einsparungen beim Stichprobenumfang des Einkommensmoduls ermöglichen.

Bei der modularen Erhebungsstruktur sollten die Einkommen per Interview erfragt werden, da das Interview erwartungsgemäß mit einer geringeren Befragtenbelastung verbunden ist. Dabei können retrospektive Fragen entweder zum vergangenen Kalenderjahr oder den vorausgehenden 12 Monaten gestellt werden.

Die Erhebung der Einkommen per Anschreibung in Kombination mit retrospektiven Fragen zu den beiden vorausgehenden Monaten hat zwar den Vorteil, dass im Allgemeinen genauere Einzelergebnisse erwartet werden können, als bei der Interviewerhebung (vgl. Abbildung 6, S. 31). Zur Beurteilung der Qualitätseffekte der Befragungsmethoden müssen jedoch auch solche Einflüsse Berücksichtigung finden, die sich indirekt auf die Qualität der Daten auswirken. Der Vorteil der genaueren Einzelergebnisse bei der Anschreibung ist letztlich dem Qualitätsverlust gegenüber zu stellen, der aufgrund einer hohen Befragtenbelastung und dem damit verbundenen Nonresponsebias eintritt. Bei niedrigen Beteiligungsquoten können auch sehr zuverlässige Ergebnisse der einzelnen Interviews nicht mehr als repräsentativ für die Grundgesamtheit bezeichnet werden, denn erfahrungsgemäß verweigern gerade solche Stichprobeneinheiten die Teilnahme, die über besonders niedrige oder besonders hohe Einkommen verfügen. Dieses Problem kann auch durch die Anwendung einer Quotenstichprobe nicht umgangen werden, da hier bei hoher Teilnahme-

---

2) Vgl. zur Verwendung von Zusatzinformationen aus gemeinsamen Basisfragen für die Hochrechnung: *Hofmans, M. G.*: Innovative weighting in POLS. Making use of core questions, Hrsg. Netherlands Official Statistics.

verweigerung prinzipiell dieselben Selektionsmechanismen wirken, wie bei der Zufallsauswahl.

Abbildung 6
Befragungsmethoden für das Einkommen

	Befragtenbelastung	**Qualität der Einzelergebnisse**
Interview mit retrospektiven Fragen	Relativ geringe Belastung, weil: • Motivation und Hilfe durch Interviewer • Befragtenführung und Filterfragen • begrenzte Zeitdauer der Belastung weil nur ein Interview je Jahr mit max. einer Stunde	Tendenziell ungenauere Einzelergebnisse, weil: • Sichtung von Unterlagen nicht möglich • Sondereinkünfte oder Einkommensänderungen können wegen langen Retrospektiv-Zeiträumen leicht vergessen werden
Kombination aus Anschreibung und retrospektiven Fragen	Höhere Belastung, weil: • keine Unterstützung durch Interviewer • Angaben viermal je Jahr erfragt werden	Relativ genaue Einzelergebnisse weil: • Hinzuziehen von Unterlagen möglich • Sondereinkünfte werden retrospektiv für die letzten beiden Monate erfasst

Vom Prinzip her muss daher entschieden werden, ob sehr genaue Informationen über einen Teil der Bevölkerung gesammelt werden sollen, oder ob weniger genaue Daten erhoben werden sollen, die für möglichst viele Haushalte repräsentativ sind. Werden Teile der Bevölkerung nicht mit einbezogen, so entstehen Erfassungsfehler, während ungenaue Einzelangaben zu inhaltlichen Fehlern der Einkommensmessung führen.

Wird im Einkommensmodul die Interviewerhebung angewandt, so dürfte die Befragtenbelastung in diesem Modul mit der Belastung im ECHP vergleichbar sein. In der deutschen Stichprobe zum ECHP wurden Teilnahmequoten von 49% erreicht[3)]. Diese Responserate könnte daher in einem Einkommensmodul, bei dem die Einkünfte per Interview erfragt werden, ebenfalls erwartet werden. Wird hingegen *keine* modulare Struktur gewählt und werden die Einkommen zusammen mit den Ausgaben angeschrieben, so wäre die Befragtenbelastung mit der in den Laufenden Wirtschaftsrechnungen vergleichbar. In der Testerhebung zur Neukonzeption der Laufenden Wirtschaftsrechnungen wurden im zufälligen Auswahlverfahren Responseraten von 9 bis 15% festgestellt[4)]. Die unterschiedlichen Zahlen zeigen, dass beim modularen Aufbau mit Interviewerhebung der Einkommen im Einkommensmodul ei-

---

3) *Bechtold, S.; Meyer, K. (1996)*: Das europäische Haushaltspanel, in: Wirtschaft und Statistik, H. 5, S. 299.

4) Vgl. *Kühnen, C. (2001)*: Das Stichprobenverfahren der Einkommens- und Verbrauchsstichprobe 1998, in: Statistisches Bundesamt (Hrsg.): Methodenberichte, H. 1, S. 8 f.

ne *deutlich* höhere Beteiligungsbereitschaft zu erwarten wäre und somit der Erfassungsfehler hier erheblich niedriger liegen dürfte.

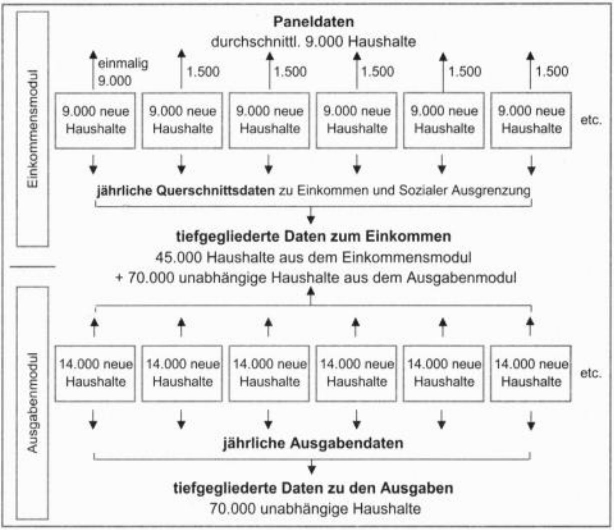
Die Entscheidung, ob bei der Messung des Einkommens die Erfassungsfehler oder die inhaltlichen Fehler stärker zu gewichten sind, müsste vor dem Hintergrund des Verwendungszwecks der Daten gefällt werden. Dabei kann davon ausgegangen werden, dass die Angaben zum Einkommen sowohl auf nationaler, als auch auf internationaler Ebene hauptsächlich für die Messung der Einkommensverteilung benötigt werden. Für den europäischen Datenbedarf ist zusätzlich zu berücksichtigen, dass die Verteilung nicht nur möglichst genau gemessen wird, sondern auch eine möglichst gute internationale Vergleichbarkeit der Ergebnisse gegeben sein muss. Für Verteilungsmessungen müssen möglichst die Einkommen *aller* Haushalte erfasst werden, die Einkommen müssen möglichst *genau* aufgezeichnet werden und es müssen möglichst alle *Einkommensarten* berücksichtigt werden. Besonders wichtig sind dabei Fehler, die zu systematischen Verzerrungen am oberen oder unteren Ende der Verteilungshierarchie führen: Werden gerade die besonders reichen und besonders armen Haushalte nicht mit einbezogen, so wird die Ungleichheit der Einkommensverteilung unterschätzt. Werden die Einkommen bestimmter Haushaltsgruppen ungenau angegeben, zum Beispiel die Einkommen der Selbständigenhaushalte systematisch zu niedrig ausgewiesen, so sind ebenfalls Verteilungsverzerrungen zu erwarten. Schließlich führt das Auslassen bestimmter Einkommensarten dann zu falschen Ergebnissen bei der Einkommensverteilung, wenn diese Einkommensart überwiegend bei besonders reichen oder armen Haushalten vorkommt. Ungenauigkeiten, die bei allen Haushalten in prozentual etwa gleicher Höhe und gleicher Richtung auftreten, oder die zufällig über alle Einkommensklassen verteilt sind, haben hingegen keinen großen Einfluss auf die Verteilungsergebnisse.

Ein empirischer Vergleich jeweils der Erfassungs- und Inhaltsfehler zwischen Interview- und Anschreibungserhebungen steht noch aus. Es ist jedoch sehr wahrscheinlich, dass sich bei Fragen der Einkommensverteilung Erfassungsfehler stärker auswirken, als Fehler bei der Genauigkeit der Einzelergebnisse.

Die Ausgaben können praktisch nur auf dem Wege der Anschreibung erhoben werden. Eine Entlastung der Befragten könnte hier durch die Vorgabe von Ausgabenkategorien und den Verzicht auf die Feinanschreibung erreicht werden. Allerdings scheint für die Auswertung der Ausgabendaten die Erfassung eine weniger wichtige Rolle zu spielen, während die Genauigkeit der Einzelergebnisse in den Vordergrund tritt. Die modulare Struktur erlaubt es, in den Modulen jeweils unterschiedliche Strategien zu verfolgen, d.h. bei den Ausgaben eine hohe Ergebnisgenauigkeit in den Vordergrund zu stellen, während bei den Einkommen die Befragungsbeteiligung vorrangig bleiben kann. Daher wäre es möglich, im Ausgabenmodul die Feinanschreibung der Ausgaben beizubehalten und eine entsprechend hohe Belastung in Kauf zu nehmen.

Das Erhebungsdesign beim modularen Aufbau ist in Abbildung 7 (vgl. S. 33) skizziert. Bei dieser Variante müssen jährlich 23 000 neue Haushalte rekrutiert und 32 000 Befragungen durchgeführt werden.

Abbildung 7
Vorschlag für die modulare Erhebungsstruktur



Zusammenfassend können für die modulare Struktur folgende Befragungsmethoden genannt werden:

1. Einkommen:
   - Erhebung per Interview.
   - Retrospektiv zum letzen Kalenderjahr oder den vorausgehenden 12 Monaten.
2. Ausgaben:
   - Anschreibung über drei Monate (EVS-Modell) oder an vier Monaten des Jahres (LWR-Modell).
   - Bildung von Durchschnitten über die Monate und die Stichprobeneinheiten.
   - Detaillierungsgrad etwa so wie in den gegenwärtigen LWR.
3. Fragen zur Sozialen Ausgrenzung und weitere Merkmale:
   - Beantwortung von Fragen im Interview des Einkommensmoduls.

4. Daten zum Längsschnittteil von EU-SILC:
    - Praktisch keine zusätzlichen Fragen benötigt.

Bei der modularen Erhebungsstruktur bräuchten für den Ausgabenteil keine Panelinformationen erhoben werden. Die benötigte Interviewzahl wird im *Ausgabenmodul* praktisch durch den nationalen Bedarf an tiefgegliederten Ausgabendaten vorgegeben. Geht man hier von 70 000 unabhängigen Interviews alle fünf Jahre aus, so müssten jährlich 14 000 Haushalte neu rekrutiert werden.

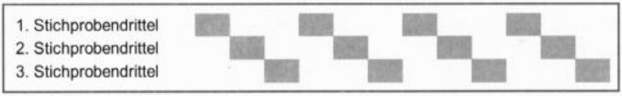
Für den nationalen Bedarf an tiefgegliederten Ergebnissen zum *Einkommen* stehen sowohl die Ergebnisse des Einkommens – als auch die Ergebnisse des Ausgabenmoduls zur Verfügung, in dem die Einkommen auch erfragt werden. Die Vorgabe für den benötigten minimalen Stichprobenumfang resultiert in diesem Modul daher aus dem Querschnittsteil für EU-SILC. Unter Berücksichtigung von Designeffekten werden hierfür jährlich etwa 9 000 Haushalte benötigt.

3.3 Variante C: Integration von EU-SILC in die bestehenden Laufenden Wirtschaftsrechnungen

Im Rahmen dieser Testvariante werden die Datenanforderungen aus EU-SILC in das bestehende System der Wirtschaftsrechnungen aus LWR und EVS integriert. Hauptdatenbasis bilden die LWR, deren Stichprobenumfang erhöht werden müsste.

Im Rahmen der LWR werden jährlich 6 000 private Haushalte nach Einnahmen und Ausgaben befragt. Die LWR sind als Quartalspanel strukturiert, bei dem ein Drittel der Haushalte jeweils im ersten Monat eines Quartals, ein weiteres Drittel im zweiten Monat eines Quartals und das letzte Drittel im dritten Monat eines Quartals befragt wird. In den Befragungsmonaten schreiben die Haushalte ihre Ausgaben als Klartextangaben an und machen darüber hinaus u. a. Angaben zu Einnahmen und zur Wohnsituation. EU-SILC soll über ergänzende Fragebogen in die LWR integriert werden.

Abbildung 8
Erhebungsablauf: Integration EU-SILC in die LWR



Vijay Verma*)

Proposals for a survey structure for those countries beginning a new EU-SILC survey

1 Introduction

As explained in the in the Eurostat document *Draft Regulation on the collection of Statistics on Income and Living Conditions in the Community (EU-SILC) – general structure and first thoughts*, EU-SILC envisages the creation of one or more micro-data base(s) in each country to be used for the follow-up and monitoring of income and social exclusion at the EU and national levels. This requires the data to be comparable, accurate and timely.

1.1 EU-SILC Data sources

At the same time, EU-SILC has to be *flexible* in terms of both data sources and survey design. The use of existing data sources, whether surveys or registers, is strongly encourage, so long as a high degree of comparability can be guaranteed. Depending on the country, micro-data could come from:

1. one existing national source (survey or register);
2. two or more existing national sources (surveys and/or registers) directly linkable at micro-level;
3. one or more existing national sources combined with a new survey – all of them directly linkable at micro-level;
4. a new harmonised survey (or survey system) to meet all EU-SILC requirements.

This document describes some *basic aspects of the structure* of the EU-SILC survey, specifically in situations (4) where a *new harmonised survey (or survey system)* to meet all EU-SILC requirements is to be developed. To a considerable extent, the discussion is relevant to other situations as well. For instance, given that EU-SILC will also have to collect some qualitative/subjective data (on perception, self-assessment ...), even those member states that plan to use administrative registers will have to complete this information by survey-based data. This may involve the development of a new survey instrument (situation 3). Or a new survey may be developed for one of the two (cross-sectional or longitudinal) components, while an existing national survey is used for the other components. Similar considerations may also arise in the more general situation involving the adaptation of existing national surveys for providing EU-SILC data (situations 1 or 2).

To discuss the survey and sampling structure, it is necessary to be clear about the substantive objectives and scope of EU-SILC. As formulated by Eurostat, the main consideration is that EU-SILC may involve different types of data: cross-sectional versus longitudinal; and national versus regional.

*) Prof. Vijay Verma, ORC Macro International, Social Research, London.

For EU-SILC, the sample selection and implementation procedures must follow certain *common standards* so as to ensure that a representative sample of households *and* persons is obtained in each country. This document is not concerned with general sampling standards or sample design issues involved in EU-SILC. The objective is to discuss aspects of the *structure of the sample over time* in view of the dual, cross-sectional and longitudinal, data requirements of EU-SILC. Regional (subnational) requirements are also discussed briefly.

1.2 Cross-sectional versus longitudinal data

Data are required in both cross-sectional (pertaining to a given time or time period) and longitudinal (pertaining to individual level changes over time) dimensions. This will involve the provision of cross-sectional levels as well as estimates of net changes or trends.

However, the first and clear priority is to be given to the delivery of comparable, timely and high quality *cross-sectional* data. Longitudinal data will be limited to income information and a limited set of critical qualitative, non-monetary variables of deprivation, aimed at identifying the incidence and dynamic processes of persistence of poverty and social exclusion among subgroups in the population. The longitudinal component will also be more limited in sample size compared to the primary, cross-sectional component. Furthermore, for any given set of individuals, micro-level changes will be followed up only for a limited duration, such as a period of three to four years.

Both cross-sectional and longitudinal micro-data sets need to be updated on an *annual* basis. The two types of data can come from separate sources, i.e., the longitudinal dataset does not need to be "linkable" with the cross-sectional dataset at the micro-level. Of course, such linkage is not precluded, and would normally be possible when the two types of data come from the same source.

Targeted panels (e.g. panel of young people to study transitions from education to work) may be considered, but as special operations *not interfering with the basic structure*.

1.3 National versus regional estimates

The main concern will be the production of income and social exclusion indicators (whether cross-sectional or longitudinal) at the *national level*, of course with the needed breakdown by demographic, socio-economic and other (non-geographical) classes within each country. Regional (subnational) breakdown will be given a clearly secondary priority, and may or may not be included in the main EU-SILC regulation. If included, regional estimates will be limited to main cross-sectional indicators of the income distribution (e.g. average/median income, poverty rates...). Longitudinal indicators at the regional level are clearly out side the scope of EU-SILC.

If included, regional indicators are likely to be updated less frequently (say, every 3-4 years) than those at the national level. The regional statistics are not required to be based on same EU-SILC data as the main national statistics, or to be linkable with the latter at the micro-level. It fact, it would be sufficient for EU-SILC for regional statistics to be provided by Member States in the aggregated form. In any case, major departures in the size, distribution or structure of the national sample for the purpose of producing regional estimates are to be avoided.

1.4 Main focus: national cross-sectional micro-data on income and social exclusion, updated annually

Hence the key priority of EU-SILC is to deliver comparable, timely and high quality cross-sectional micro-data annually at the national level within each Member State, covering household and personal income and main indicators of social exclusion. In particular, EU-SILC is to be the reference source of comparative income distribution statistics at EU level. This, of course, does not preclude the possibility of adding ad-hoc modules to EU-SILC to investigate particular areas of policy interest in more detail if and when required. The quality of the main, national cross-sectional, data should not be compromised by the effects of the longitudinal component (such as attrition), nor of course by the need to produce regional (subnational) figures.

1.5 Sampling standards

The EU-SILC data will be based on *nationally representative probability sample* to permit detailed analysis by population subgroups. As to the survey units, it is envisaged that EU-SILC micro-dataset will contain statistical information on *all members of a given household*, with detailed information on members aged 15+. Precision (sampling error) requirements will be mainly in the form of the *minimum effective sample size* required per country. For sample survey-based systems, these may be in the range 4 000 (for the smallest countries) to 10 000 households for the largest (or twice those numbers in terms of the number of adults covered). Substantially larger sample sizes would generally be possible where register data are used. In this latter case, some non-income variables will still come from surveys, and for those variables, the sample sizes may be smaller (such as 2 500 – 6 000 households per country). Samples for longitudinal data may also be similarly small.

It is *neither possible nor necessary to standardise the sample design and procedures*, whether the data come from registers, surveys, or a combination of the two. For instance, depending on the sampling frame available in each country, the sampling unit may be the person, the household, or an address. However, the sample selection and implementation procedures must follow certain *common standards* so as to ensure that a representative sample of households *and* persons is obtained in each country.

2 The cross-sectional sample

2.1 Analysis units: households and persons

In terms of the units involved, three types of data are involved EU-SILC, as in ECHP:

(1) variables measured at the household level;

(2) variables measured at the personal level, but aggregated to construct household-level variables (such as income); and

(3) person-level variables.

Given the primary importance of cross-sectional national-level estimates in EU-SILC, objective (1) is served best by having an equal probability ("self-weighting") sample of *households*. Variations in selection probabilities – by region, household size, or whatever – mostly results in reduced sampling efficiency. Similarly, objective (3) is served best by having a equal probability ("self-weighting") sample of *persons*. Objective (2) links the household and person levels. It requires that all persons in any sample household be included in the sample, which imparts the same probability of selection to a household as to any of its members.

Consequently, EU-SILC primary objective is served best by having an equal probability ("self-weighting") sample of households, taking all persons in a selected household into the sample for the personal interview, and hence obtaining an equal probability ("self-weighting") sample of persons as well.

2.2 Sampling over time

Annual cross-sectional estimates can be produced from

(1) independent samples from year to year; or

(2) retaining the same sample from one year to the next; or

(3) a rotational design – i.e. a combination of the above two – rotating a part of the sample from one year to the next and retaining the other part unchanged.

Cross-sectional estimates for a single year are essentially unaffected by the pattern of rotation (theoretically, modest improvements may be achieved with partial overlaps using special estimation procedures). The appropriate pattern of rotation is determined by the compromise between two objectives:

(a) cumulation of data over time, so as to achieve increased sample size, which favours maximum rotation i.e. independent samples (1); and

(b) the measurement of change over time, which favours maximum overlap (2).

The effect of departures from these patterns depends on the correlation over time. The relationships may be expressed in the following simple forms.

2.2.1 Cumulation

If Var_0 is the variance which would be achieved with the aggregation of two independent samples (of the same design and size), then with overlapping samples the increased variance is approximately:

$$Var = Var_0 \cdot (1 + P \cdot R),$$

where P is the overlap (0-1), and R is the correlation coefficient for a particular statistic from the two samples.

For instance with $P = 0{,}75$ (3/4 overlap from one year to the next) and taking R (the correlation coefficient between successive years for some main variables of interest in the overlapping part of the sample) as 0.7, variance in estimating the two-year aggregate will be increased roughly by 50% $(1 + P \cdot R = 1{,}5)$ due to the sample overlap.

2.2.2 Net change

If Var_0 is the variance which would be achieved for the net difference between estimates from the two independent samples (of the same design and size), then with overlapping samples the reduced variance is

$$Var = Var_0 \cdot (1 + P \cdot R).$$

For instance with $P = 0{,}75$ R (the correlation coefficient between successive years for some main variables of interest in the overlapping part of the sample) as 0.7, as in the above example, variance in estimating change will be roughly halved $(1 - P \cdot R = 0{,}5)$ due to the sample overlap.

In the case of EU-SILC, the measurement of trends (changes over time) is likely to be clearly more important than cumulating data over years, favouring large overlaps P from one year to the next. There are practical limitation in continuing with the same sample, i.e. having large overlaps (P close to 1,0) from year to year.

2.2.3 Illustration

Consider two successive years with partially overlapping samples. For the cross-sectional sample for each year to be separately representative requires each of the following three parts to be a representative sample: (i) the dropped part to be representative of the population at year 1; (ii) the added part to be representative of the population at year 2; and (iii) the overlapping part to be representative of the population at both times.

Normally, the above is achieved in practice by selecting the total sample in the form of a number of *replications*. The scheme is illustrated in Figure 1 (see p. 40). Each replication is in itself a representative sample, typically with the same design (structure, stratification, allocation etc) as the full sample, differing from the latter only in sample size. From one year to the next, some of the replications are retained, while others are dropped and replaced by new replications depending on the extent of overlap desired.

Figure 1
Replications

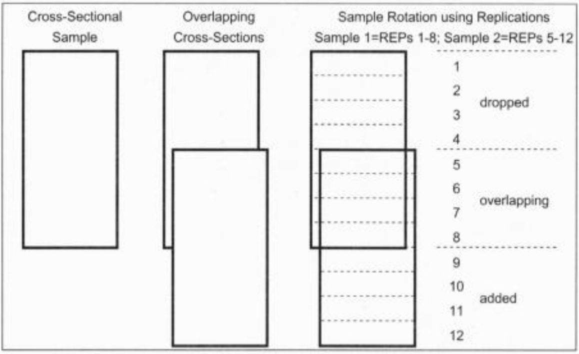
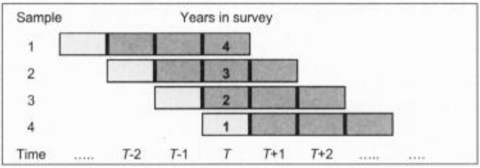


Figure 2 illustrates a simple rotational design (once the system is fully established). The sample for any one year consists of 4 replications, which have been in the survey for 1 – 4 years (as shown for "Time=*T* " in the figure). Any particular replication remains in the survey for 4 years; each year one of the 4 replications from the previous year is dropped and a new one added, giving a 75% overlap from one year to the next. For surveys two years part, the overlap is 50%; it is reduced to 25% for surveys three years apart, and to zero for longer intervals. Generally with *n* replications, each kept in the survey for *n* rounds, the overlap between rounds declines linearly as the interval separating them increases. For two surveys *i* intervals apart the overlap is $(n-i)/n$, up to $i=(n-1)$, after which $(i \geq n$ the overlap becomes zero.

Figure 2
Successive Panels of Limited Duration

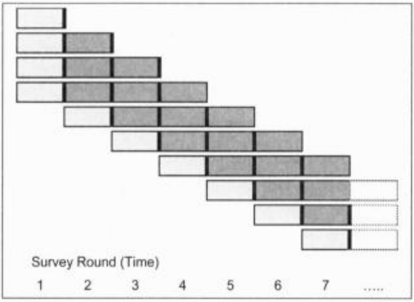


For EU-SILC, such a "linear" rotation pattern is the simplest and most appropriate in so far as the main interest is in monitoring year-to-year changes. More complex patterns can be introduced to vary the degree of sample overlaps and how that changes over time, as for instance is done in some labour force surveys (such as in Finland, USA etc), but they are unlikely to be of interest for EU-SILC.

2.3 Starting the rotation pattern

Figure 3 illustrates how a rotation pattern may be started from year 1. To obtain the full sample with 4 replications for the first year, it is necessary to begin with all the 4 replications. These replications are treated differently over time. One of these is dropped immediately after the first year, the second is retained for only 2 years, the third for 3 years, and only the fourth is retained for the full 4 years. The pattern becomes "normal" from year 2 onwards: each year one new replication is introduced and retained for 4 years. Alternative starting schemes are possible, but this illustration is perhaps the simplest one.

Figure 3
Rotation Pattern from Year 1



2.4 Special re-weighting to measure change more precisely

In principle, the gain in precision in estimating change can be enhanced by giving more weight to the overlapping part and correspondingly less weight to the rotating part in the estimation procedure. (This can be done because each of the two parts is a representative sample in itself.) With optimally determined weights, the theoretical gain can be shown to be

$$Var = Var_0 \cdot [(1-R)/(1-Q \cdot R)]$$

where $Q = 1 - P$. With $P = 0{,}75$ and $R = 0{,}7$ for instance, the variance is reduced to just over a third due to the sample overlap, compared with only halving of the variance in the previous illustration without reweighting. Hence the advantage of such special reweighting in estimating changes can be very substantial.

Similarly for cumulation over years, the loss in precision in estimating aggregates can be reduced by giving more weight to the rotating part and correspondingly less weight to the overlapping part in the estimation procedure. With optimally determined weights, the theoretical loss can be shown to be

$$Var = Var_0 \cdot [(1+R)/(1+Q \cdot R)].$$

The reweighting is much less effective in the case of cumulation: with $P = 0{,}75$ and $R = 0{,}7$ for instance, the inflation in variance due to overlapping is reduced only to 1,45, from 1,5 in the previous illustration without this special reweighting. In any case, this is not a major concern in EU-SILC, where the main interest is in estimating annual changes rather than aggregating the data over years.

2.5 Sample overlap in the context of a cross-sectional survey

Finally, it is important to be clear about what "overlapping" means in the context of a purely cross-sectional survey. The basic requirements are:

(1) that the data for the two years from the overlapping part are correlated; and

(2) to the extent possible, the overlapping part is representative of population at both times.

Correlation does not necessarily require the samples to be identical in terms of the ultimate units (persons, or even households). A common procedure (as for instance used in most labour force surveys with rotational designs) is to have a common sample of addresses or dwelling units. In each survey, all households and persons found at those addresses are taken into the sample. Households and persons which move are *not* followed up to their new location, as the sample is defined by the *locations* (addresses) originally selected. For the measurement of net changes, it is *not* necessary to link the information for individual units over time: such linking is required only if the objective is to analyse gross or longitudinal changes at the micro level (as in panel surveys). With multi-stage sampling designs (e.g. the selection of areas followed by the selection of addresses) it is in fact not essential that the overlap be in terms of the ultimate sampling units (addresses); correlation, albeit reduced, can also be obtained by having common higher stage units (same areas but independent samples of addresses within each).

3 The longitudinal sample

3.1 Basic structure of the combined cross-sectional and longitudinal design

As noted above, EU-SILC requires a limited longitudinal component. The required longitudinal data are limited to income and selected social exclusion indicators. The sample size will be most likely substantially smaller than the cross-sectional component. Furthermore, for any given set of individuals, micro-level changes will be followed up only for a limited duration, such as over a period of three to four years.

While in principle, the cross-sectional and longitudinal data can come from separate sources (as they need not to be "linkable" at the micro-level), it will clearly be a great advantage to incorporate both types of data requirements in a single, integrated system. This should certainly be the case if a new EU-SILC survey is being developed.

Figures 2 and 3 above also illustrate the type of structure which may be suitable for meeting the combined cross-sectional and longitudinal requirements. Figure 2 illustrates the system once established, supplemented by Figure 3 displaying how the system may be started from EU-SILC year 1.

At the beginning, a cross-sectionally representative sample of households is selected. It is divided into say 4 subsamples, each by itself representative of the whole population and similar in structure to the whole sample (except for sample size). One subsample is purely cross-sectional and is not followed up after the first round. Respondents in the second subsample are requested to participate in the panel for 2 years, in the third subsample for 3 years, and in the fourth for 4 years. From year 2 onwards, one new panel is introduced each year, with request for participation for 4 years.

In any one year, the sample consists of 4 subsamples, which together constitute the cross-sectional sample. In year 1 they are all new samples; in all subsequent years, only one is new sample. In year 2, three are panels in the second year; in year 3, one is a panel in the second year and two in the third year; in subsequent years, one is a panel for the second year, one for the third year, and one for the fourth (final) year.

3.2 Tracing rules and related aspects

While the above structure looks similar to that for a purely cross-sectional survey described in the previous section, it is important to appreciate differences in the follow-up ("tracing") rules between the two. These differences have important practical consequences.

In the purely cross-sectional survey, the follow-up normally consists of revisiting the same sample addresses, without following-up moving households and persons to their new location. However, in the combined cross-sectional and longitudinal structure, it is necessary to *follow-up individuals once selected*. The tracing rules and procedures can be the same as those used for the ECHP, for instance. Each panel begins with a probability sample of households (irrespective of the actual selection mechanism – the sampling units may be address, households or persons, so long as they yield a valid sample of households). A household interview and a roster (register) listing all members and their basic characteristics is completed. All current residents of sample households are included in the sample as "sample persons", and those aged 15+ are interviewed in detail using a personal questionnaire. Every sample person is then followed-up for the duration of the panel, to a new location if necessary. (Additional rules may be adopted to drop some persons, such as those moving to an institution or outside the EU, those giving a final refusal to continue in the survey, those not traced for two or more successive years, etc, as done in the case of the ECHP.) The sample is augmented over time to include as "sample persons" children born to women in the original sample. Original sample children as they achieve the age of 15 become eligible for the detailed personal interview. Furthermore, all residents aged 15+ in the current household of any original sample person are interviewed, including any non-sample persons not in the original sample. The "weight-sharing" approach can be used to include these non-sample persons into cross-sectional analyses. In other words, the structure, tracing rules, weighting procedures etc can be the same as those for the first (four) ECHP waves.

3.3 Consequences

What are the consequences of the more complex follow-up procedures in the combined design, compared to the simpler ones for a purely cross-sectional survey?

1. The main administrative complication is the need to *identify and trace individuals over* time. As in the ECHP, the success of the panel critically depends on uniquely and correctly identifying every person ever coming into the survey and linking his/her information over time. The only difference is the limited duration of EU-SILC panels compared to the ECHP.

2. The major technical complications arise from the more complex weighting, editing and imputation procedures to ensure longitudinal completeness, consistency and representatiness of the data, and also to produce valid cross-sectional estimates making full use of all the available data. The procedures developed for ECHP can be directly imported for this purpose, with some further elaboration to deal with the fact that EU-SILC at any one time will involve several panels each with a different starting point.

3. A true panel of individuals is likely to suffer from somewhat higher attrition rates, compared to repeated enumeration of simply the same addresses in a rotational cross-sectional survey.

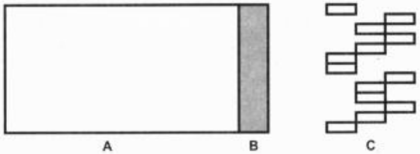
4. The main additional fieldwork cost arises from the need to trace movers. Assuming that on the average 2 – 3% of the population move address during any one year, the EU-SILC survey at any one time should not have more than 5% or so movers. (The four subsamples in it are subject to 0, 1, 2 and 3 years of moving, i.e. an average of 1,5 years.) Even though the proportion of movers is small, it can be taxing in terms of complexity, cost and duration of the fieldwork required.

5. The important point to note is that a vast majority (perhaps 95% or more) of households and persons will be re-enumerated at their original address, i.e. just as in a purely cross-sectional survey with sample overlaps. It is this fact which makes the incorporation of a panel component into a cross-sectional survey cost-effective: much of the longitudinal component is in fact enumerated as a by-product of the cross-sectional enumeration.

3.4 A useful option: doubling the enumeration of movers

The samples resulting from the different tracing rules in a purely cross-sectional sample (section 2.6) and a combined cross-sectional longitudinal design (section 3.2) are illustrated in Figure 4.

Figure 4
Cross-Sectional (A+B) and Longitudinal (A+C) Samples at Year 2



Stationary	A	Set of persons residents at the same address at both years
Movers-in	B	New set of person, found at year 2 at the original (year 1) sample addresses
Movers-out	C	Set of persons at the original year 1 sample addresses, who have moved to some new location at year 2

Set (A+C) provides a valid longitudinal sample for years 1 – 2 (with C enumerated at their original addresses at year 1 and at their new location at year 2)

At year 2, there are in fact three valid cross-sectional samples: (A+B); (A+C); and hence also their combination (A+B)+(A+C), which is the same thing as [A+(B+C)/2], meaning a sample consisting of set A, plus B and C each with half their normal weight.

Each of the three alternatives have their advantages.

- Set (A+B) does not involve follow-up of movers to new addresses, and hence is likely to be completed earlier. The cross-sectional results can be produced in a more timely fashion, not affected by the normal delay in completing the longitudinal component (C).
- Set (A+C) has the convenience of consistency in that in this case the longitudinal and the year 2 cross-sectional samples have the same base (A+C).

The recommended set [A+(B+C)/2] uses all the available information, which increases precision. The advantage is more than simply increasing the sample size. By bringing in both out-movers and in-movers, it doubles the available sample size for groups likely to be of special interest.

3.5 Adjusting the relative sample sizes of the two components

The model described by Figure 2 is too rigid in one respect: it assumes a fixed relationship between the cross-sectional and longitudinal sample sizes. The relative size of the panel component can be increased (reduced) only by increasing (reducing) its duration. However, that duration (such as 3 – 4 year in EU-SILC) is given by the survey's substantive objectives, and is not a parameter which can be chosen on the basis of sampling considerations.

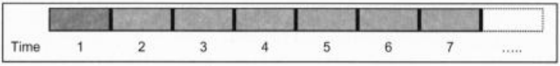
Greater flexibility can be achieved by supplementing the above structure by one or the other of the following.

3.5.1 Split panel

This refers to addition to the basic structure (Figure 2) of a panel component of unlimited duration, of the type illustrated in Figure 5. This increases the available sample size of the panel part. (The term "split panel" was introduced by Leslie Kish to describe such an arrangement.) The size of the split panel can be chosen flexibly to obtain a panel component of the required size.

Of course, the addition of a split panel of unlimited duration brings in new considerations and possibilities, beyond the stated basic (minimum) requirements of EU-SILC.

Figure 5
"Split Panel"

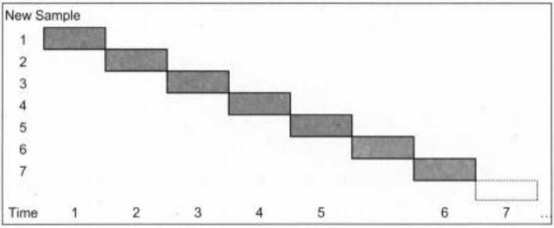


3.6 Cross-sectional booster

By contrast, the size of the cross-sectional component can be increased by adding to the basic structure (Figure 2) a fully rotational cross-sectional booster of the type illustrated in Figure 6. Again, the size of the cross-sectional booster can be chosen flexibly to obtain a cross-sectional component of the required size.

The Canadian SLID provides an example of a rotational design with a large cross-sectional booster added annually.

Figure 6
"Cross-Sectional Booster"



3.6.1 Illustration

Consider a rotational design with *r* replications or subsamples, each of size *s*. In the basic model, each subsample is retained in the survey for *r* years.

In any round,

- the cross-sectional sample is $n_1 = r \cdot s$;
- the longitudinal sample linked over two years is of size $n_2 = (r-1) \cdot s$ (since all but the newly introduced panel provide such linkage with the previous year);
- the longitudinal sample linked over three years is of size $n_3 = (r-2) \cdot s$ (since all but the two most recently introduced panels provide such linkage with year $y-2$);
- that linked over four years is of size $n_4 = (r-3) \cdot s$; and so on.

With the addition of a split panel of size *p*, each of the above is essentially increased by *p*, so that the longitudinal to cross-sectional sample size ratio, such as n_{i+1} / n_1 is increased from

$$\frac{n_{i+1}}{n_1} = \frac{r-i}{r}$$

to

$$\frac{n_{i+1}}{n_1} = \frac{r-i+(p/s)}{r+(p/s)}.$$

With the addition of a cross-sectional booster of size *x*, the available cross-sectional sample is increased by *x* without affecting the longitudinal components. The longitudinal to cross-sectional sample size ratio is therefore reduced from:

$$\frac{n_{i+1}}{n_1} = \frac{r-i}{r}$$

to

$$\frac{n_{i+1}}{n_1} = \frac{r-i}{r+(x/s)}.$$

3.7 Targeted panels

According to the draft Regulation, EU-SILC envisages the possibility that *targeted panels* (e.g. panel of young people to study transitions from education to work) may be considered, but as special operations not interfering with the basic structure.

The main limitation of the basic structure discussed above for this purpose is likely to be the smallness of the sample size available for studying special subgroups in the population. Pending a more detailed consideration of the issues, it can be noted that *cumulation of data over time* may be one simple method of increasing the available sample sizes. This is illustrated in the following subsection.

3.8 Cumulation of longitudinal data: an illustration

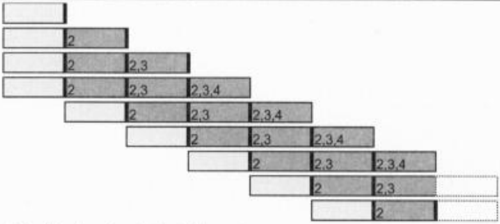
Consider first year to year transitions, with the design of Figures 2 – 3, with *r* subsamples for instance.

Each year starting with year 2, (*r*-1) subsamples provide observations of *year-to-year transitions*. These can be cumulated over time to obtain

$$(r-1)\cdot(y-1)$$

subsamples proving observations of year-to-year transitions over the years 1 to *y*. The resulting analysis provides an average picture of such transitions over the *y* years. Figure 7 (based on Figure 3) provides an illustration with $r = 4$.

Figure 7
Cumulation of Longitudinal Observations



Cumulative Number of Longitudinal Observations

1. Subsamples Providing Cross-Sectional Data								
4	8	12	16	20	24	28	.....	4·Y
2. Subsamples Providing Year-to-Year Transitions								
0	3	6	9	12	15	18	.....	3·(Y-1)
3. Subsamples Providing 3 Years Longitudinal Observations								
0	0	2	4	6	8	10	.....	2·(Y-2)
4. Subsamples Providing 4 Years Longitudinal Observations								
0	0	0	1	2	3	4	.....	(Y-3)
Survey Year								
1	2	3	4	5	6	7	.....	Y

Note: The numbers in the cells of the diagram indicate the type(s) of observations provided by the subsample. The above numbers may be multiplied by the subsample size to obtain the cumulated number of observations.

Similarly, each year starting with year 3, $(r-2)$ subsamples provide a set of longitudinal observations each covering a *three year period*. These can be cumulated over time up to survey year *y* to obtain

$$(r-2)\cdot(y-2)$$

subsamples proving observations each covering 3 years. The resulting analysis provides an average picture of such observations over the *y* years.

As for sets of longitudinal observations *each covering a 4-year period*, each year starting with year 4 provides $(r-3)$ subsamples for the purpose. These can be cumulated over time up to survey year y to obtain

$$(r-3)\cdot(y-4)$$

subsamples proving observations each covering 4 years.

4 Regional data

4.1 Cumulation over time

In so far as the main survey-based system is to be used for the generation of regional (subnational) statistics, the main constraint is likely to be the limited sample size. More reliable (but less frequent) regional statistics may be produced by aggregating the data over time (e.g. 3 – 4 years).

It can be seen from Figure 7 that the total number of observations (sample size) cumulated over y survey years is

$$t = y\cdot[r\cdot s + p + x]$$

where

r is the number of subsamples in the basic design (e.g. $r = 4$ in Fig. 2);

s is the sample size per subsample (replication);

p is the sample size of the split panel, if any;

x is the size of the annual cross-sectional booster, if any; and

y is the number of years for which the data are cumulated.

The *effective increase in sample size is smaller than the above* because of the high positive correlation between repeated observations on the same units due to overlapping samples. The degree of overlaps can be easily estimated from the rotation patterns described above. The magnitude of their effect on the efficiency of the cumulation depends on the size of the correlation coefficients involved, which can be estimated only from empirical data. The ECHP is a potential source of such data.

4.2 Differential sampling rates by region

When the production of regional estimates is important, it is a common practice to vary the sampling rates across regions to obtain appropriate regional sample sizes. Typically, small regions are over-sampled (i.e. sampled at above-average rates) to obtain adequate sample sizes; and large regions are under-sampled (i.e. sampled at below average rates) to avoid unnecessarily large sample sizes. The sample sizes are larger for larger regions, but less than proportionate to their population sizes. For example, allocation in proportion to square-root of the population size has sometimes been used.

Variation in sampling rates across regions requires weighting the data appropriately for national level estimates. The effect of such sample weights is generally to inflate variances and reduce the overall efficiency of the design as concerns the national level estimates. Furthermore, since such weights are usually uncorrelated with population variances, *this effect tends to be persistent uniformly across estimates for diverse variables and population subclasses, including estimates of differentials and trends.*

The increase in variance depends on the variability in the selection probabilities or the resulting design weights, and is approximated as:

$$D_w^2 = (1 + cv^2(w_i)) \text{ , or}$$

$$D_w = \sqrt{1 + cv^2(w_i)} \text{ ,}$$

where $cv(w_i)$ is the coefficient of variation of the weights. The first expression approximates the factor by which sampling variances are inflated (i.e. *the effective sample size is reduced)*. The second expression gives the corresponding factor by which standard errors and confidence intervals are inflated.

In so far as the primary objective in EU-SILC is to produce estimates at the national, rather than at the regional, level, it is desirable to avoid the introduction of different sampling rates by region – unless the overall sample size can be increased to compensate for the associated loss in precision.

Vijay Verma*)

The case for a Continuous Household Budget Survey

Mr Bo Møller in a memorandum to Eurostat has described the new structure of the Danish HBS. Essentially, the idea consists of a survey of modest size conducted on a continuous basis, data from which can be *cumulated over years* to achieve more adequate sample sizes. This is an interesting and exciting development, and an important question is the extent to which this model can be adopted by other countries who at present do not conduct annual surveys. Furthermore, there are also countries who already conduct annual (or even more frequent) surveys, but with rather small sample sizes. To what extent can the data from individual rounds be cumulated in such cases?

In this section some technical observations are made on operational, design and analysis aspects of the Continuous Household Budget Survey model. Essential elements of the model may be described in terms of the following illustration.

Suppose that in place of conducting one survey of, say, 5.000 households every five years, the survey is conducted on a continuous basis with a representative sample of 1.000 households every year. During the year the workload is also distributed more or less uniformly, e.g. enumerating around 80 – 100 households per month. The work can be conducted by a small team of interviewers (e.g. 8 or so) deployed permanently for the task. With a continuous flow of data from the field, data preparation and processing also becomes an ongoing operation. The sample is designed such that the information can be efficiently cumulated over time to achieve sufficient sample sizes, and the results are reported on a regular (annual) basis in the form of "moving averages" over a number of most recent years.

1 Advantages of the model

Mr Møller has surmised the potential advantages very well. The main operational advantage spring from the relatively moderate but regular workload. The number of survey staff required is reduced, who can then be better (and more cheaply) trained and supervised. Engagement on a permanent basis also tends to generate a sense of loyalty, responsibility and satisfaction, and is conducive to building more enduring survey capability.

There can be major substantive advantages as well. The results can be regularly updated and reported in a more timely manner. From the point of view of the EU, the reporting from different Member States can be better synchronised.

2 Possible disadvantages or problems

Lack of flexibility can be a possible disadvantage of a continuous survey. Major redesigns are more easily accommodated in ad hoc surveys separated by long intervals. By contrast, in a continuous survey it is necessary to carefully regulate and

---

*) Prof. Vijay Verma, ORC Macro International, Social Research, London.

control changes in content, design and procedures. This is for operational as well as substantive reasons.

In any case, sudden or major changes have to be avoided so as not to disrupt the system. This requires that the survey is started only after very careful preparation and testing of all procedures. It is very difficult to rectify the situation once the survey is "locked into" a particular design. In this sense, the situation is similar to that in a panel survey.

3 Some basic sampling considerations

To cumulate the survey results over time, it is necessary that the sample be representative simultaneously over space and time. This means that for annual surveys for instance, the sample for each year separately should be representative of the whole country. Actually, it is desirable to divide the year into shorter (such as half-monthly, monthly, or at least quarterly) periods, each with a separately representative sample of the country.

The second basic requirement is that the annual samples should be independently selected, so as to avoid positive covariance and permit efficient cumulation over years. If a multi-stage sampling design is used, the samples for different periods should ideally use different, independently selected primary sampling units.

Generally, it would be most efficient to use equal, or at least fairly similar sample sizes from one period (year) to the next. In practice, some variation in the achieved sample sizes from year to year cannot be avoided. Proper procedures will be required to weight the annual samples for the production of cumulated estimates covering a number of years.

A good procedure appears to be as follows. Weight each annual sample to be representative of the mid-year population of the year concerned (taking into account selection probabilities, response rates, external control totals etc.), and then put together the annual sample estimates with weights *in proportion to their corresponding mid-year populations* to produce cumulative results. In so far as the population does not change much over a few years, the above implies giving equal weights to the annual estimates in putting together the results.

4 Reference period

Cumulating data over several years raises no special problems. In a "normal" survey conducted over a one-year period as well, data with different recording and reference periods cover different times over and preceding the survey year.

At the aggregate (total sample) level the time periods covered may be expressed as follows. This is illustrated in the diagram appended at the end of this document.[1)]

1) *Verma, V.; Gabilondo, L. G. (1993)*: Family Budget Surveys in the EC: Methodology and Recommendations, Statistical Office of the European Community, Series 3E, Luxembourg.

Assume that data are collected with the sample uniformly distributed over Y years, and for a particular set of items, with a *moving reference period* of X years preceding the survey interview. For instance, for major expenditures (such as purchase of motor vehicles) the reference period may be $X = 1$ year preceding the survey; for items such as clothing, we may have a reference period of six months ($X = 0.5$ years); while for items recorded on a continuous basis in a diary, we have $X = 0$; and so on depending on the country questionnaire. For a single-year survey, $Y = 1$, while with data cumulated over three years in a continuous survey, we have $Y = 3$.

For the sample as a whole, data collected with a moving reference period of X years then pertain to the time period $P = X + Y$ years preceding the interview. The quantum (volume) of the information collected is distributed symmetrically, centred at the point $P_0 = (X + Y)/2 - 1$ years before the *beginning* of the most recent survey year, or $P_M = (X + Y)/2 - 0{,}5$ before the *mid-point* of the most recent survey year.

Figure 1
Time period to which the data pertain with cumulation over Y >1 years

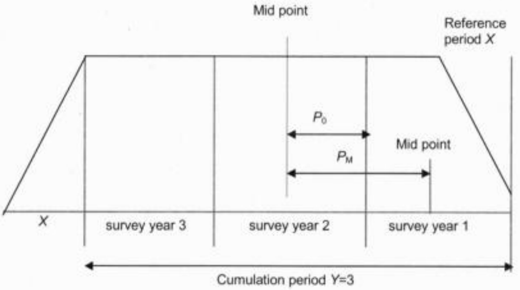
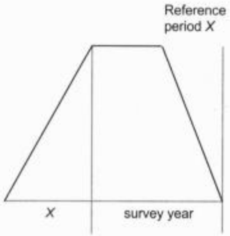


Figure 2
Time period to which the data pertain in "normal" survey (Y =1 year)



It can be seen that with different values of the reference period for different types of items, the situation with cumulation over a number of years ($Y > 1$) is similar in form to that with a conventional survey over a single year ($Y = 1$). Actually, with increasing Y, the period covered becomes less sensitive to differences in reference periods for different types of items in the same survey. This greater uniformity of the time-periods covered for different types of items is in fact an *advantage* of increasing Y (the period of cumulation).

For a *fixed reference period* such as the preceding calendar year, the situation is similar but simpler. For a conventional one-year survey, the period covered is centred at the middle of the calendar year. We have: for $Y = 1$: $P_M = 1$; $P_0 = 1/2$.

More generally, with cumulation over $Y>1$ years we have, $P_M = (P+1)/2$; $P_0 = P/2$, giving, for instance, P_M = mid-point of the second year when cumulated over $Y = 3$, and P_M = beginning of the second year when cumulated over $Y = 4$ years.

5 Price adjustment

In any HBS, prices (whether actual or imputed) for all items of consumption or expenditure need to be adjusted in accordance with the periods they refer to. At the individual level, the relevant price is the one prevailing at the mid-point of the reference period for the item concerned.

Exactly the same procedure as that for a single-year survey applies to any continuous survey involving cumulation over a number of years.

In adjusting prices, it is important to note that *a single, common adjustment factor – reflecting the overall consumer price index for private households – applies to all types of items and all categories of households.* (Using different adjustment factors for different categories will fail to reflect changes in the structure of consumption in terms of values.) This fact considerably simplifies the adjustment process.

On the other hand, if certain items of consumption such as imputed rent are obtained from an external source and refer to a different period than the reference period of the survey, price adjustments to bring them in line with the survey period have to be made using appropriate quantity, price and quality indicators *specific to the item concerned.*

6 Application of the Continuous Survey Model

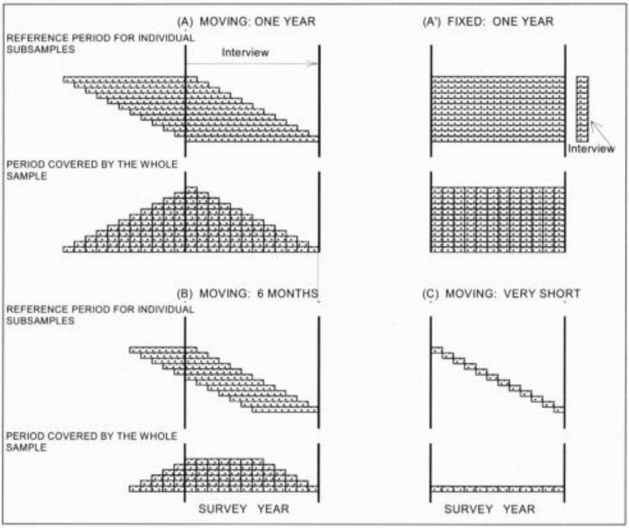
It is obvious that this model has a number of potential advantages, and presents no special technical problems. It is an attractive model of survey structure for countries which at present are able (or willing) to conduct HBS's of sufficient size only periodically, once every several years.

7 Cumulation in existing annual surveys

A number of Member States already conduct annual household budget surveys. In some of these (e.g. Italy, U.K.), the sample size is large enough to permit annual reporting with sufficient precision. However, in some others (e.g. Spain, Netherlands), the annual samples are rather small, and cumulation over several consecutive years may be desirable in the reporting of certain results.

From the point of view of sample design, the most important requirement is having non-overlapping or independent annual samples which can be cumulated efficiently to build-up a large effective sample size. Interviewing within the same sampling areas from one round to the next, and even more so repeated interviews with the same households, as done in the Spanish continuous survey for instance, is not efficient for this purpose.

Table 1
Effect of type and Length of Reference Period



The table shows the effect of the type and length of the reference period on the period covered in the whole survey. The type refers to whether the reference period is "moving" or "fixed". A fixed reference period means a period of the same duration and timing for all units in the sample, irrespective of the timing of their interviewing.

With a moving reference period the duration is the same, but the actual time covered differs from one unit or subsample to another depending on the timing of its interview. With interviewing distributed over the whole survey year, the period covered for the sample as a whole includes the survey year, and extends beyond it into the preceding year depending on the length of the moving reference period. With the reference period also of one year, the period covered in the survey extends over two years with a triangular density distribution (diagram A). As the length of the moving reference period is reduced, the coverage becomes more concentrated in the actual survey year, and more rectangular in shape (diagrams B and C). With a fixed reference period, the coverage is always confined to the fixed survey year, and is rectangular in shape (diagram A') irrespective of the length of the fixed reference period.

Another effect of the length is on the "quantity" of the time covered, as indicated by the area of the diagram representing the period covered by the whole sample. In a sense, this is the effective sample size. This quantity does not depend on the type of the reference period but only on its length: for example, it is the same for diagrams (A) and (A'). It is reduced in proportion to the length of the reference period (as in diagrams B and C).

Jan Kordos*)

Estimation Problems and Data Quality in Rotation Samples

1 Abstract

The author presents his experiences with rotation sampling while designing sample surveys at the Polish Central Statistical Office, and working abroad as an expert at the Food and Agriculture Organization of the United Nations (FAO) and a World Bank consultant. The first half of the paper reviews designs of the surveys across time, depending on different objectives, focussing on partial rotation of sub-samples, and next considers estimation problems and data quality issues generally. His attention is concentrated on Polish household surveys, and mainly on Household Budget Survey (HBS) and Labour Force survey (LFS). He outlines designing the Polish HBS and LFS with focus on rotation sampling and their redesigning over time. Some results of nonresponse rates for the Polish HBS and LFS in last years are presented. An example of estimation issue of the Current Population Survey in USA (i.e. construction of composite estimator, using rotation groups and results from the previous year, are presented). Concluding results are given at the end.

2 Introduction

There have been many articles published, as well as books and chapters, about surveys over time, survey designs, estimation problems and data quality (e.g.: Bailar, 1975, 1979; Brackstone, 1999; Duncan, Kalton, 1987; Eckler, 1955; Kalton, Citro, 1993; Kish, 1965; Kordos, 1967; Lyberg, 1997; Patterson, 1950; Särndal et al., 1993; Wilks, 1940; Woodruff, 1963). It is impossible to present in this short paper even the most important problems connected with sample surveys across time. I confine myself to some aspects of rotation sampling while designing household sample surveys in Poland and abroad. First, I introduce some issues connected with designing of sample surveys over time depending on different objectives. Some issues connected to rotation sampling used as a compromise in solving different problems in estimation of parameters for various objectives are outlined. Data quality issues in surveys over time are very complicated, and I consider only some aspects connected with nonresponse rates illustrating them with data from the Polish household surveys. Let us start with the objectives of the surveys across time.

3 Objectives for surveys over time

Changes in population characteristics and composition over time lead to a variety of objectives for surveys across time. These objectives include the following (Kalton, Citro, 1993):

(a) Estimating population parameters at distinct time points.

*) Prof. Dr. Jan Kordos, Central Statistical Office of Poland, Warsaw.

(b) Estimating average values of population parameters.

(c) Estimating net change.

(d) Estimating gross and other components of individual change.

(e) Aggregating data for individuals over time.

(f) Collecting data on events occurring in specified time period.

(g) Cumulating samples over time.

(h) Keeping contact with members of a rare population identified at a point of time.

To estimate these parameters for different objectives from sample surveys there are various problems connected with sample design, size of sample, allocation of sampling units in space and time, collection of required data from these units, quality of obtained information, etc. Some of these problems may be partly solved by applying rotation sampling (i.e. changing subsamples of units at different periods of time). Let us consider some aspects of designing surveys over time.

4 Survey designs over time

A number of survey designs have been developed to provide the data needed to address the above mentioned objectives. These designs are (Duncan, Kalton, 1987; Kalton, Citro, 1993):

1. *Repeated survey*: a series of separate cross-sectional surveys conducted at different time points.
2. *Panel survey*: collecting of the survey data for the same sample elements at different points of time.
3. *Repeated panel survey*: is made up of a series panel surveys each of fixed duration:
   - *with no overlap* (one panel may starts only as the previous one ends).
   - *with overlap* (with two or more panels covering of the same period of time).
4. *Rotating panel survey*: is equivalent to a repeated panel survey with overlap. Both limit the length of a panel, and have two or more panels in the field at the same time. Rotating panel surveys are widely used to provide a series of cross-sectional estimates and estimates of net change, whereas repeated panel surveys with overlaps also have a major focus on longitudinal measures. In consequence, repeated panel surveys tend to have longer duration and have fewer panels in operation at any given time than rotating panel surveys.
5. *Split panel survey*: is a combination of a panel survey and a repeated survey or rotating panel survey (it was used in Polish HBS in 1982 – 1998).
6. *Repeated survey with overlap*: Like a repeated survey, and overlapping survey is a series of cross-sectional surveys conducted at different time points, and is designed to provide overlaps.

The choice of design in a particular case depends on the desired objectives to be satisfied. Some designs are better than others for some objectives but less suitable for other objectives. Some designs cannot fulfil certain objectives at all (for a detailed discussion see Duncan, Kalton, 1987 and Kalton, Citro, 1993).

As it has been mentioned above, there are problems with the choice of sample design to satisfy several objectives in one survey. For example, to estimate totals or averages at each time point, it is better to select a repeated sample survey (i.e. a new sample at each time point). However, to estimate changes between time points, a panel survey is better. The key advantages of the panel design are its abilities to measure gross change, and also aggregate data for individuals over time. A great analytic potential provided by the measurement of individual changes is the major reason for using a panel design. Very often in one sample survey several objectives should be reached. In such a case, a compromise may be achieved by applying rotation sampling.

A benefit of rotating and overlapping panel surveys is that they enable comparison of estimates for the same time period obtained from different panels. Such comparisons clearly exemplify the presence of what is termed "rotation bias group" in the U.S. and Canadian Labour Force Surveys (Bailar, 1975,1979; Kalton, Citro, 1993). Rotation group bias may reflect nonresponse bias and conditioning effects. In analyses comparing the overlapping 1985, 1986 and 1987 effects in the Survey of Income and Program Participation panel, Pennel and Lepkowski (1992) found few differences in the results from the different panels.

Rotating panel surveys are primarily concerned with estimating current levels and net change (objectives (a) and (c)). As such, sample elements are usually retained in the panel for only short periods. A special feature of rotating panel surveys is the potential to use composite estimation to improve precision of both cross-sectional estimates and estimates of net change for an alternative method of using past information in forming estimates from a rotating panel design. Like rotating panel surveys, overlapping surveys are primarily concerned with estimating current levels and net change. They can also provide some limited information on gross change and aggregations over time.

In comparing alternative designs for surveys across time, the costs of the designs need to be considered. For instance, panel surveys avoid the costs repeated sample selections incur with repeated surveys, but they involve costs of tracking and tracing mobile sample members to continue co-operating in the panel. If two designs can each satisfy the survey objectives, the relative costs for given levels of precision for the survey estimates need to be examined.

5 Estimation problems

An objective of a survey over time is usually to estimate more than one population parameter. Very often several objectives given in section 3 are to be achieved. Two of many important choices that must be made in a survey are:

1. The choice of sampling design and a sample selection scheme to implement the design.
2. The choice of a formula (an estimator) by which to calculate an estimate of a given parameter of interest.

The two choices are not independent of each other. For example, the choice of estimator will usually depend on the choice of sampling design. It is necessary to use an appropriate strategy.

A strategy is the combination of a sampling design and an estimator. For a given parameter, the general aim is to find the best possible strategy, that is, one that estimates the parameter as accurately as possible. As we will see below, different sample designs are needed to find the best possible strategy.

In the above sections we note five objectives and six designs. Most periodic studies have several purposes and thus we should face – not necessarily solve – the difficult problems of multipurpose designs. Actually, objectives from (a) to (c) can be met with any of the six listed designs, but with some increase in the variances or in costs. But estimates of gross and individual change need panels, and objective (e) need some changes. Often reasonable compromises become possible – to the degree that purposes can be defined. The chief variation in these designs concerns the amount (and kind) of overlaps between periods (Kish, 1987, pp. 159 – 169).

Table 1
Purposes and Designs for Periodic Samples

Purposes	Designs	Rotation Scheme
A. Current levels	A. Partial overlaps $0< P < 1$	abc – cde – efg
B. Cumulating	B. Nonoverlaps $P = 0$	aaa – bbb – ccc
C. Net changes (means)	C. Complete overlaps $P = 1$	aaa – aaa – aaa
D. Gross changes (individual)	D. Panels	Same elements
E. Multipurpose, time series	E. Combination	
	F. Master frames	

Source: Kish, 1987, p. 160

The rotation scheme of complete overlaps shows, with *aaa – aaa,* that the periods have all common parts; the non-overlap with *aaa – bbb* shows none, and the partial overlap *abc – cde – efg* shows *c* and *e* as one-third overlaps between succeeding periods only.

Here we concentrate on the effects of varying proportions P in diverse designs on different purposes; in complete overlaps $P = 1$, in nonoverlaps $P = 0$, and in partial overlaps $0< P < 1$. The purposes are discussed in terms of variances for estimated means, because means (and percentages, rates, proportions) are both the most used and the simplest estimates to be treated. Effects on other estimates will not be entirely different but they are too many, diverse, and difficult to be explored here.

Effects on the variances of means from different proportions P can be treated clearly in this brief section. Other questions of biases, of feasibility, of costs are often even more important, but also more difficult. They are treated in all sampling books. We assume here for simplicity that the period samples are of the same size or of the same sampling fraction; but changes in sizes, fractions, and designs are possible. We start below with current levels and partial overlaps.

5.1 Current levels and partial overlaps

Variances of current estimates are the same for complete overlaps $P = 1$ and nonoverlaps $P = 0$; they can be expressed briefly for means as $Deft^2\, S^2/n$, where $Deft^2$ is the effect of the sample design on either the element variance S^2 or the sample size n (Kish, 1987, p. 162 – 163).

That simple formula also holds for *simple* means from partial overlaps ($0 < P < 1$). But statistics based on them can untilize the overlap P for a reduction of the variance with a complex mean: With help of the correlation R^2 between surveys within the sample overlap P, the portion $(1 - P)$ of the *preceding* sample is combined with the current mean to improve it. The variances are reduced by the factor

$$\frac{1-(1-P)R^2}{1-(1-P)^2R^2}$$ (Cochran, 1977, secs. 12.11 – 12.12).

The actual gains unfortunately tend to be modest in most practical situations; the maximal reduction in variance, utilizing optimal proportions P and optimal weights, is in the ratio $[1+((1- R^2)^{0.5})]/2$. The reductions increase to about 33 percent only for very high R^2 values, seldom seen in practice; for $R = 0.9$, for example, $[1+((1- R^2)^{0.5})]/2 = 0.72$. This ratio is obtained either with the optimal $P = 0.30$ or with $P = 1/3$. For $R = 0.6$ that ratio becomes 0.9, only 10 percent reduction of the variance. It is worth stressing that in a long series the complex mean from the preceding sample can already benefit from reductions from its predecessors, and that using a longer series provides further slight reductions. Fortunately, for other purposes of repeated surveys statistical theory is more productive as well as simpler.

5.2 Net changes of Means and Overlaps

Net change refers to the difference $d = \bar{x}_1 - \bar{x}_2$ of means between two periods; whereas *gross change* deals with the total changes of individuals, some of which remain hidden because they cancel in the nest change of means. Measuring net changes are common and important aims of surveys, and they are also related to other uses of the data. Perhaps the most common forms are differences in dichotomies, denoted by proportions $d = p_1 - p_2$, and in similar rates and ratios.

The variance of $(\bar{x} - \bar{y})$ can be greatly reduced when the pairs of variables have high positive correlation R in overlapping samples. We will discuss here several cases shortly.

1. The variance of mean differences are reduced by factors (1 – R) in complete overlaps; this is the extreme (with P = 1) of the factors (1 – PR) that may be obtained from the partial overlaps. Hence for minimizing variance $Var(\bar{x}-\bar{y})$ complete overlaps would be best. But partial overlaps are used in practice: (a) for reasons of feasibility, to reduce burdens, fatigue, and biases of respondents; and to reduce variances of other statistics in multipurpose designs. It is simple to think of the variances as $(2S^2/n)$ for differences between pairs of samples of size n without overlaps; as $(2S^2/n)(1-R)$ with complete overlaps, and $(2S^2/n)(1-PR)$ with partial overlaps P. The S^2/n assumes simple random sampling and for complex samples the design effect $Deft^2$ should be included. But for differences and changes the factors $Deft^2$ tend to be smaller (closer to 1) than for the means. Therefore reductions obtained from overlaps in complex samples, where $Deft^2$ are large for single means, may even be considerably greater than indicated by the factors (1 – PR).
2. We may obtain almost the full reductions of complete overlaps even from partial overlaps by using improved estimators of the differences. These estimators are useful when circumstances may prevent complete overlaps but still permit partial overlaps. In those estimators the overlap portion P gets larger weights than the nonoverlaps portion 1- P = Q, by the factor 1/(1 – R), because elements in the overlap contribute much less to the variance. This improved estimator of the difference is (Kish, 1965, table 12.4.III):

$$\hat{D}(\bar{y}-\bar{x}) = [P(\bar{y}-\bar{x})_p + Q(1-R)(\bar{y}-\bar{x})]/(1-QR)$$

Its variance may be expressed, for two srs samples of size n, as:

$$Var([\hat{D}(\bar{y}-\bar{x})] = \frac{(1-R)S^2}{(1-QR)n}$$

The factor (1 – R)/(1 – QR) approaches (1 – R) for high values of R and for higher values of P say P = 2/3. High values of R are common for stable characteristics that can be well measured.

5.3 Composite estimation

A number of parameters of interest are of the form (Särndal et al., 1993, p. 371):

$$t = \varphi t_z + \psi t_y \qquad (1)$$

where ϕ and ψ are constants,

$t_z = \sum_U z_k$ - an estimator of the total at the first wave,

$t_y = \sum_U y_k$ - an estimator of the total at the second wave.

For example

a) $\phi = 0$, $\psi = 1$ leads to $t = t_y$, the current total, which is a parameter of level;

b) $\phi = -1$, $\psi = 1$ leads to $t = t_y - t_z$, the absolute change;

c) $\phi = 1$, $\psi = 1$ leads to $t = t_y + t_z$, which is the sum of the totals over two occasions.

Choosing a design for the second wave

We have more information than at the first wave: for every surveyed unit at the first wave we know the value Z_k. For the new design, we can consider:

- no overlapping,
- complete overlap, or
- partial overlap with the first wave.

We can show that different parameters have optimal sampling designs at the second wave.

It is intuitively clear that there are cases in which the information from the first wave may be used to improve the estimation. Hence, we opt for partial overlap. At the second wave, two independent samples are drawn, one *matched sample* and one *unmatched sample.*

Estimating the Current Total

Using the values from two sub-samples (the matched sample and unmatched), we can obtain two estimators of current total:

$$\hat{t}_1 \text{ and } \hat{t}_2$$

Both estimators $\hat{t}_1$ and $\hat{t}_2$ are unbiased for the current total. By linear combination we obtain the new unbiased estimator

$$\hat{t}_y = w_1\hat{t}_1 + w_2\hat{t}_2 \qquad (2)$$

where w_1 and w_2 are nonnegative constant weights to be determined and such that

$$w_1 + w_2 = 1.$$

We call $\hat{t}_y$ a composite estimator; it combines the matched sample estimator with the unmatched sample estimator. Below a composite estimator for the Current Population Surveys is given to estimate a number of unemployed persons in a given month using a composite estimation from the previous months, and adjustment for rotation group bias.

5.4 An Example of CPS Composite Estimation

A CPS composite estimation example (Bailar, 1975, pp. 23 – 30):

$$x_{t'} = (1-K)x_t + K(x_{t'_{-1}} + d_{t,t-1}) + Ad_t \qquad (3)$$

Where

x_t: simple weighted estimate for current month,

$x_{t'-1}$: composite estimate for previous month,

$d_{t,t-1}$: estimate of change between months from sample present in both months

d_t: adjustment for rotation group bias.

$K = 0.4$ and $A = 0.2$ are chosen as approximately optimal for reducing variances across a variety of estimates. Prior to 1985, $K = 0.5$ and $A = 0$. The d_t term was added to bring the composite and simple weighted estimates closer together.

For example, to estimate the number of unemployed persons in April 1996, we need a composite unemployment estimate for March 1996 ($x_{t'-1}$ = 7 700), unemployment estimate for April 1996 ($x_{t'}$ = 7 197), an estimate of change between months from sample present in both months ($d_{t,t'-1}$ = -717) and ($d_{t'}$ = 70).

Finally, we get composite estimate of unemployed persons in surveyed month (April):

$$x_t = (0.6 \cdot 7\,197) + 0.4(7\,700 - 717) + (0.2 \cdot 70) = 7\,125 \text{ (thousand)}$$

It is a good example how to use available data from the sample and previous estimates to improve precision of the current estimate.

6 Application of rotation sampling in surveys of the Polish CSO

We started applying of rotation sampling in statistical surveys in 1960s. At the beginning, we studied such articles as Wilks (1940), Jessen (1942), Patterson (1950), Eckler (1953) and Woodruff (1963). In 1967 I published a review paper on rotation method in sampling surveys (Kordos, 1967).

The following surveys were carried out with application of rotation sampling:

1) Rotation method in morbidity survey (1967 – 1968).
2) Rotation method in post letters and parcels survey (1961, repeated).
3) Time use survey (1967, 1976, 1984, 1996).
4) Experiments of HBS by rotation method (1968 – 1969, 1981).
5) Survey of workers starting first job (1971).
6) Epilepsy survey in Warsaw (1971).
7) Household Budget Survey (since 1982+).
8) Labour Force Survey (since 1992 +).

Some results from the Polish HBS and LFS are presented below.

6.1 Some design features of the Polish Household Budget Surveys

Methodology of HBS in Poland was described in detail in my paper (Kordos, 1996). The design features with rotation sampling of the HBS in Poland in 1982 – 2000 is given in table 2. Below a sample design in years 1996-2000 is described shortly.

Table 2
Some Design Features of Household Budget Surveys in Poland in 1982-2000

Period of the survey	*Method*	*Size of sample, coverage*	*Sampling frame*	*Sample design*	*Nonresponse rates*
The pilot survey in 1981; in 1982 two subsamples: first with the old method (territorial), second with the new method (quarterly rotation)					
1982 – 1992	Rotation method Quarterly rotation, 4-year split panel (66.6%)	About 10 000 hhs in 1982 21 600 hhs in 1982 – 1985 32 400 hhs in 1986 – 1991 (9 652 hhs in 1992)	First stage: census enumeration areas for population census Second stage: dwellings	First stage: PPS (number of dwellings) Second stage: the same number of dwellings	29.0 - 39.5 % in 1982 – 1987 34.9 - 41.6 % in 1988 – 1992
In 1992 one subsample of hhs took part in the survey according to old method in the same year one subsample of hhs participated - according to the new method					
1992 – 1999	Sequential sampling Monthly rotation, 4-year split panel (50.0%)	About 16 000 hhs in 1992 All types of private hhs (32 400 hhs) in 1993 – 1999	First stage: census enumeration areas for population census Second stage: dwellings	First stage: PPS: (to number of dwellings in PSJ) Second stage: the same number of dwellings	23.2 - 49.4 % in 1992 – 1999
2000 – 2001	Monthly rotation, 2-year panel: one new subsample each year	All types of private hhs 32 400 hhs)	As above	As above	49.2 % in 2000

hhs – households Source: Kordos, 1997, table 1, p. 1124 and other CSO publications

The sample was selected at a two-stage stratified sampling. The primary sampling units were statistical districts or clusters of districts covering at least 250 dwellings. The file was created for the needs of the Population and Housing Census and which in addition constituted a sampling frame for sample surveys. It was updated on an annual basis and the updating covers: an increase of the dwelling stock due to the completion of new buildings, a decrease of the dwelling stock due to the demolition and changes in the boundaries of districts due to changes in the administrative division of the country. For each district the sampling frame contains information on the addresses and estimated data on the number of population and number of dwellings.

In order to use this file as a sampling frame for a selection of a sample for the HBS, statistical districts covering less than 250 dwellings are combined with neighbouring districts (i.e. those districts which have neighbouring numbers). In this way the sampling frame was created for the first-stage units each of which covered at least 250 dwellings. The PSUs created in this way were stratified by voivodship (regions) and within each voivodship two strata: urban and rural were created. In total 98 strata were created.

Sampling of the first stage units

When selecting sample design for the selection of a sample for the HBS an assumption was made that the selected sample should be more less self-weighting. The PSUs were selected with the probability proportional to the estimated number of dwellings in a PSU. 1 350 PSUs were selected in the sample (i.e. two subsamples of 675 PSUs each).

The primary sampling units were selected separately in each stratum. The procedure of systematic sampling was applied after systematisation of the units (PSUs selected with PPS using the Hartley-Rao method). The sample selected in a given stratum was divided at random into two subsamples, the first of which was used in the survey since 1 January 1996 and the second since 1 January 1997. The first subsample was used in surveys in the years 1996 – 1999 and the second in the years 1997 – 2000.

Sampling at the second stage – sampling of dwellings

When sampling units at the second stage the following organisational and methodological approach was applied:

a) the sample consists of two parts: permanent and replaceable, replaced each year,

b) in the survey the model of monthly rotation is applied (i.e. in each month a new subsample is used),

c) in a given PSU each month 2 dwellings are selected in which all households are surveyed,

d) sampling of dwellings for the permanent sample is done at the beginning (i.e. for 4 years),

e) sampling of dwellings for the replaceable sample is done twice;

   - for the first subsample:
     for the years 1996 – 1997 in 1995,
     for the years 1998 – 1999 in 1997;
   - for the second subsample:
     for the years 1997 – 1998 in 1996,
     for the years 1999 – 2000 in 1998;

f) due to the application of the sequential sampling in case of non-response, a reserve sample of dwellings is selected which are ordered at random.

Due to the application of the above mentioned approach, the sampling of dwellings was done in the following way:

- dwellings in a given PSU were ordered at random,
- the permanent part of the sample covered the first 12 dwellings,
- additional dwellings were selected to the replaceable part of the sample, i.e. two times 12 dwellings and after 2 years again two times 12 dwellings,
- the consecutive *n* dwellings ordered at random constituted a subsample of dwellings for sequential sampling, where n is less or equal 150.

6.2 Some design features of LFS

Rotation chart for the LFS sample selection is given in table 3.

Table 3
Rotation chart for the LFS in 1999 – 2004

Sub-sample No.	Years and quarters																							
	1999				2000				2001				2002				2003				2004			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1				x																				
2				x	x																			
3				-	x	x																		
4				-	-	x	x																	
5				x	-	-	x	x																
6				x	x	-	-	x	x															
7					x	x	-	-	x	x														
8						x	x	-	-	x	x													
9							x	x	-	-	x	x												
10								x	x	-	-	x	x											
11									x	x	-	-	x	x										
12										x	x	-	-	x	x									
13											x	x	-	-	x	x								
14												x	x	-	-	x	x							
15													x	x	-	-	x	x						
16														x	x	-	-	x	x					
17															x	x	-	-	x	x				
18																x	x	-	-	x	x			
19																	x	x	-	-	x	x		
20																		x	x	-	-	x	x	
21																			x	x	-	-	x	x
22																				x	x	-	-	x
23																					x	x	-	-
24																						x	x	-
25																							x	x
26																								x

Sub-sample 1, 2, 3, 4, 5 – sub-sample with shorten duration.
Sub-sample 6 and further – a new sub-samples with normal duration.

The Labour Force Survey on the economic activity of the population was implemented in Poland for the first time in May 1992 and was repeated on the quarterly basis. It was prepared according to the ILO recommendations and could be treated as a modern statistical survey (Szarkowski, Witkowski, 1994). In each quarter usually about 20 000 households and members of those households aged 15 and above were surveyed. The results of the survey were used in the broad scope for the evaluation of the situation on the labour market and the size of unemployment. Occasionally modules on selected social topics were included into the survey which considerably extends the use of the results in social and economic analyses.

It should be stressed that since the fourth quarter of 1999 the LFS has been carried out as a continuous survey. The quarterly sample currently amounts to 24.440 dwellings. It was constructed in such a way that every one of 13 weekly samples is not only the same size but also has the same structure. The selection of quarterly samples is performed according to the rotation system: 2 (2) 2 (i.e. two quarters in the sample, two quarters out of the sample and two quarters again in the sample). The results are processed and published quarterly. Results of the survey conducted through this method enable presentation of situation on the labour market during the whole quarter (Central Statistical Office, 2000).

7 Data quality issues

Survey data quality is a concept with many dimensions linked with each others. In theory, all dimensions of data quality are very important, but in practice, it is usually not possible to place high importance on all dimensions. Thus, with fixed financial resources, an emphasis on one dimension will result in a decrease in emphasis on another. More emphasis on accuracy can lead to less emphasis on timeliness and accessibility; or an emphasis on timeliness may result in early/preliminary release data of significantly lower accuracy. Each dimension is important to an end user, but each user may differ in identifying the most important priorities for a data collection program.

An extensive literature exists and continues to grow on the topic of survey data quality (e.g. Bailar, 1975, 1979; Brackstone, 1999; Kordos, 1988; Lyberg et al. 1997) and its management in national statistical agencies (Brackstone 1999). Definitions of the concept proliferate, but cluster around the idea that the characteristics of the product under development meet or exceed the stated or implied needs of the user.

For a sample survey over time, very important are the following sources of errors:

(1) nonresponse losses,

(2) time-in-sample, or conditioning, effects,

(3) recall errors, including the seam effect, and other nonsampling errors.

We confine here to nonresponse rates in HBS and LFS.

7.1 Nonresponse rates for household surveys

A great deal of research and attention has been devoted to nonresponse errors (for reviews see Groves 1989; Groves, Couper 1998).

There are two main types of nonresponse error: unit nonresponse and item nonresponse. Unit nonresponse is a complete failure to obtain data from a sample unit, whether the sample unit is a household or person within a household. Despite best efforts during data collection, some level of unit nonresponse is likely to occur. Weighting techniques are often used to minimize the effect of unit nonresponse error. However, to the extent that the underlying assumptions are not fully met, such as, for example, the sample units are not missing at random, nonresponse error may still affect estimates derived from the data.

A variety of reasons exist for unit nonresponse, and they may vary depending on the mode of the survey and the survey design. Unit nonresponse in an interviewer-administered personal visit household survey can occur because no one is home (noncontact), refusal to participate, or inability to participate due to language barriers or cognitive or physical incapacity to respond.

It is important to identify and measure the different reasons and components of nonresponse because different levels of nonresponse bias may be associated with different reasons for nonresponse. For example, noncontacts, those who were never given the opportunity to choose whether or not to participate in the survey, may have very different characteristics than refusals, who were contacted but chose not to participate, and both may differ from survey respondents on some survey variables. Very different trends over time may exist for some of these components, and these trends should be monitored.

Longitudinal survey designs commonly have panel nonresponse, which is unit nonresponse to one or more waves of the survey. In long-term longitudinal surveys, panel nonresponse is often permanent and results in attrition of cases from the sample (i.e., very few cases drop out one time and return the next). In this case, the effective level of nonresponse increases over time.

Nonresponse errors are likely to be present to some degree in any survey, and they may affect the quality of the data and estimates produced from the survey. The exact amount of nonresponse error on a specific estimate is almost never known. However, there are a number of indicators of the quality of those data that can be extremely useful to analysts, customers, and consumers in evaluating the quality of the data and their usefulness.

Frequently the response rate (or its complement, the unit nonresponse rate) is used as an overall indicator of the quality of the data in the survey. There is also a variety of other more specific indicators that provide insights into the quality of the data and the data collection process. Specifically, rates of refusals and noncontacts can be useful when computed for the entire data collection as well as at the level of field administrative units or interviewers. Also, address not locatable, postmaster returns, and undetermined eligibility (e.g., answering machine-eligibility unknown) can be important in monitoring and evaluating mail and telephone surveys.

Some similar quality indicators can be calculated for item nonresponse rates. These provide a more micro-level data quality indicator for a specific estimate or set of estimates, but also may be informative about the data collection process when computed at different levels.

We would like to demonstrate that nonresponse rates depend on method of data collection and length of reporting period. In 1992 in the Polish HBS two methods of data collection were used: one sub-sample with the old method (three months keeping records) and one sub-sample with the new method (one month). It is possible to compare nonresponse rates for the two methods. One can observe an impact on the nonresponse rates caused by two methods of data collection. The results are given in Table 4.

Table 4
Nonresponse rates for the HBS using old and new methods by quarters in 1992

Methods of data collection	Nonresponse rates in quarters				
	Total	I	II	III	IV
	%				
Three months	38.4	37.1	37.7	39.8	38.8
One month	22.6	17.2	22.3	26.1	24.1

Source: Kordos, 1995, table 3, p. 801

As it can be seen from the above table, nonresponse rates are smaller for about 15 percentage points for a new method when data on incomes and expenditures are recorded for one month only.

Table 5
Nonresponse Rates in the Polish HBS and LFS in 1992 – 2000

Years	Type of the survey	
	HBS	LFS
	%	
1992	23.2	5.0
1993	27.6	5.6
1994	25.3	9.1
1995	29.1	10.2
1996	31.4	10.5
1997	34.3	11.5
1998	40.7	13.8
1999	49.4	19.7
2000	49.2	23.5

Sources: Kordos, 1995 and CSO publications

The nonresponse issues in the Polish HBS and LFS were widely discussed in the journal *Statistics in Transition* (Kordos, 1995). Here we would like to draw the attention to the current issues related to the most recent surveys and compare nonresponse rates for HBS and LFS in 1992 – 2000.

8 Concluding remarks

I presented here only general ideas connected with estimation problems and data quality for sample surveys over time, focussing on rotation sampling. We used rotation sampling in Poland in different surveys, and mainly in HBS and LFS. We consider that our sample surveys should be further improved and adjusted to the new needs of the transformation of economy and the requirement of Eurostat. There is also an issue of the integration of household surveys in Poland. A lot of methodological experience in conducting such surveys was gained at that time but the change of the economic system requires a different approach. It seems that there is a possibility to integrate household surveys taking into account the needs of the users and of international organisations and to minimise the costs of surveys at the same time.

References

Bailar, B. A. (1975): The Effect of Rotation Group Bias on Estimates from Panel Surveys, in: Journal of the American Statistical Association, Vol. 70, pp. 23 – 30.

Bailar, B. A (1979): Rotation Sample Biases and Their Effects on Estimates of Change, in: Bulletin of the International Statistical Institute, Vol. 48, No. 2, pp. 385 – 407.

Brackstone, G. (1999): Managing Data Quality in a Statistical Agency in: Survey Methodology. 25(2), pp. 139–149.

Central Statistical Office (1999): Metodyka badania budzetow gospodarstw domowych (Methodology of Household Budget Survey), Zeszyty Metodologiczne, Głowny Urząd Statystyczny, Warszawa, p.146.

Central Statistical Office (2000): Labour Force Survey in Poland – I Quarter 2000, Informacje i Opracowania Statystyczne, Głowny Urząd Statystyczny, Warszawa.

Cochran, W. G. (1977): Sampling Techniques, New York.

Duncan, G. J.; Kalton, G. (1987): Issues of Design and Analysis of Surveys across Time, in: International Statistical Review, Vol. 55. No. 1, pp. 97 – 117.

Eckler, A.R. (1955): Rotation Sampling, in: Annals of Mathematical Statistics, Vol. 26, pp. 664 – 685.

Ghangurde, P. D. (1982): Rotation Group bias in the LFS estimates, in: Survey Methodology, Vol. 8, pp. 86 – 101.

Groves, R. M. (1989): Survey Errors and Survey Costs. New York.

Groves, R. M.; Couper, M. P. (1998): Nonresponse in Household Interview Surveys. New York.

Jessen, R. J. (1942): Statistical Investigation of a Sample Survey for Obtaining Farm Facts, in: Iowa Agricultural Station Research Bulletin, Vol. 304, pp. 54 – 59.

Kalton, G.; Citro, C. F.(1993): Panel Surveys: Adding the Fourth Dimension, in: Survey Methodology, Vol. 19, No. 2, pp. 205 – 215.

Kish, L. (1965): Survey Sampling, New York.

Kish, L. (1987): Statistical Design for Research, New York.

Kordos, J. (1967): Metoda rotacyjna w badaniach reprezentacyjnych (Rotation Method in Sampling Surveys), in: Przegląd Statystyczny, Vol. XIV, z. 4, pp. 373 – 394.

Kordos, J. (1982): Metoda rotacyjna w badaniach budzetow rodzinnych w Polsce (Rotation Method in Household Budget Surveys in Poland), in: Wiadomosci Statystyczne, No. 9.

Kordos, J. (1985): Towards an Integrated System of Household Surveys in Poland, in: Bulletin of the International Statistical Institute, (invited paper), Vol. 51, Amsterdam, Book 2, pp. 13 -18.

Kordos, J. (1988): Jakosc danych statystycznych (Quality of Statistical Data), PWE, Warszawa.

Kordos, J. (1995): Nonresponse Problems in Polish Household Surveys, in: Statistics in Transition, Vol. 2, No. 5, pp. 789 – 812.

Kordos, J. (1996): Forty Years of the Household Budget Surveys in Poland, in: Statistics in Transition, Vol. 2, Nr 7, pp. 1119 – 1138.

Kordos, J. (1997): Metody monitorowania i analizy procesów społeczno-ekonomicznych i demograficznych na podstawie badań panelowych (Methods of monitoring and analysis of socio-economic and demographic processes on the basis of panel surveys), in: Roczniki Kolegium Analiz Ekonomicznych, No. 5/1997, pp. 9 – 29.

Lednicki, B. (1982): Schemat losowania i metoda estymacji w rotacyjnym badaniu budżetów gospodarstw domowych (Sampling Design and Estimation Method in Household Budget Rotation Survey), in: Wiadomosci Statystyczne, No. 9.

Lyberg, L.; Biemer, P.; Collins, M.; Deleeuw, E.; Dippo, C.; Schwarz, N.; Trewin, D. (eds., 1997): Survey Measurement and Process Quality. New York.

Patterson, H. D. (1950): Sampling on Successive Occasions with Partial Replacement of Units, in: Journal of the Royal Statistical Society, Ser. B, Vol. 12, pp. 241 – 255.

Pennel, D.; Lepowski, J. M. (1992): Panel conditioning effects in the Survey of Income and Program Participation. Proceedings of the Section on Survey Research Methods, in: American Statistical Association, pp. 432 – 456.

Särndal, C.- E., Swensson, B.,Wretman, J. (1993): Model Assisted Survey Sampling, New York, Berlin, London,.

Szarkowski, A.; Witkowski, J. (1994): The Polish Labour Force Survey, in: Statistics in Transition, Vol. 1, No. 4, pp.467 – 483.

Wilks, S. S. (1940): Reprezentative Sampling and Poll Reliability, in: Public Opinion Quarterly, 4.

Woodruff, R. S. (1963): The Use of Rotating Samples in the Census Bureau's Monthly Surveys, in: Journal of the American Statistical Association, Vol. 58.

Jan Selén*)

Some comments on the use of cumulated surveys and a presentation of a model-based calibration

1 Introduction

My presentation is based on the results of a development project at Statistics Sweden. This project was partly sponsored by Eurostat[1)] with main results summarised in Cassel, Granström, Lundquist and Selén (1997). The presentation draws upon this paper including two applications[2)]. Some comments on cumulation are added.

The project workgroup focussed on some constant problems of household budget surveys; these costly and respondent-demanding surveys with a high non-response. The impetus was that a yearly design for the Swedish HBS was desirable; earlier the survey was launched intermittently with a fairly large sample, now the scenario was a smaller yearly survey.

The workgroup considered possible estimation improvements in this setting, improvements by cumulation, calibration or consumer models, the precise meaning of which I will shortly explain. By improvements was meant increased precision in the estimation of average expenditures for groups of households on main categories of goods and services.

In the following I will first discuss cumulation. A short presentation of the regression estimator and calibration follows before the model-based calibration is introduced. Estimator properties are examined and results for the Swedish household budget survey (HBS) are presented.

2 Cumulation

The simplest example of cumulation of data is the basic cumulation method (BCM) where different survey samples just are lumped together and regarded as one large sample, for the HBS such cumulation of data from adjacent years has actually been suggested.

If we compare this to stratified sampling and estimation where the stratified design is ignored, our intuition is that BCM is appropriate only under restrictive assumptions. In Cassel at al (1997) restrictions are derived; thus BCM can provide a good approximation to an estimator accounting for the sample structure, if there is no change in levels and no change in population nor in sample size over time. Consequently BCM with its simple variance estimation cannot be recommended.

A rotating panel design is a complex approach. Then some elements are retained from one year to the next and re-observed while other elements are new. There is

---

*) Dr. Jan Selén, Statistics Sweden, Stockholm.

1) Lot 23, SUP.COM 96.

2) The details are given in two unpublished mimeos of Statistics Sweden: "The Swedish HBS" and "Experiments from an experiment of applying model based calibration to the Finnish HBS".

overlapping between the samples for different years. From the point of view of cumulation overlapping is not favourable while on the other hand variances for estimates of change and for current values are reduced. Results are summarised in Kish (1987, 1999), who suggests that an overlapping of about 1/3 can be a reasonable compromise if these different estimation needs are considered.

To illustrate the effects of overlapping and cumulation a simple data generator may be helpful. Let us form observations X according to

$$X_{ijk} = t_i + c \cdot j + e_{ijk},\ (j = 1,\dots,n;\ k = 1,\dots,m)$$

Here subscript i may be taken to indicate different surveys or years; t_i then is a year effect. The constant c is a scale factor for the variance of X and the e:s are randomly generated standard normal deviates. Data are generated in m groups of n elements each. By different specifications of which groups to include for a specific year, different levels of overlapping are obtained. Suppose for example that there are $m = 5$ groups and three years, $I = 0, 1, 2$, and that the groups $k = 1, 2, 3$ are included for year 0, $k = 3, 4, 5$ for year 1 and $k = 3, 6, 7$ for year 2; then the elements in $k = 3$ are observed every year while the other groups are included for a specific year only.

In table 1 results are shown for three years and 3 to 9 groups with 100 observations each. Groups are included differently for the different columns according to the specifications in the first row. Standard errors for ratio-estimated means are accurately computed by CLAN97 (Andersson; Nordberg 1997) and given under the assumption that the population size is 10 millions each year.

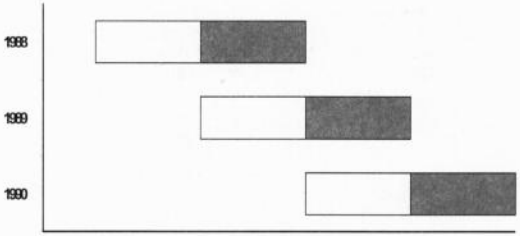
Table 1
Cumulation and overlapping.
Standard errors for averages by ratio estimators, simulations

		i, (year)	*k (groups included year i)*			
		0	**123**	**123**	**123**	**123**
		1	**456**	**345**	**234**	**123**
		2	**789**	**3 67**	**23 5**	**123**
			Standard error, average			
c						
1			.960	1.250	1.480	1.670
.1			.100	.130	0150	0170
.01			.034	.034	.036	.038
			Standard error, average, BCM			
c	**Δ*t***					
1	**0**		.96	.96	.96	.96
1	**10**		1.00	1.00	1.00	1.00
1	**100**		2.89	2.89	2.89	2.89
			Standard error, yearly change			
			2.4	1.9	1.4	0.1

In the table we see a gradual increase in standard errors for cumulated data, accompanying an increased overlapping of elements. Results for BCM are unaffected by overlapping as the complication is ignored, but sensible for differences between years ($\Delta t = t_i - t_j$), see the middle rows. The positive effect of overlapping for the measurement of yearly change is documented in the last row.

Figure 1 shows the rotating panel design for the Swedish household income survey 1988 to 1990. About half of the 10 000 households each year are retained from the previous year while the rest are newcomers.

Figure 1
The rotating design of the Swedish household income survey 1988 – 1990 (white area for new sample, shadowed area for retained sample)



In table 2 we see a yearly increase in the average disposable income, amounting to about 5 percent in constant prices and using ratio estimators. A cumulation of data, accounting for overlapping, results in a standard error for the total average about 33 percent lower than the average for each year (from .80 to .55). The BCM alternative, somewhat surprisingly, understates this standard error by over 20 percent (.44 instead of .55).

Table 2
Disposable income, 1989 prices, thousands, all households (ca 10 000 per year), the Swedish household income survey

	Mean	Se
1988	110	.80
1989	115	.80
1990	124	.80
Cumul. rotating panel	116	.55
Basic cumulation	116	.44

3 Generalised regression estimation

The main emphasis in Statistics Sweden's development project was the introduction of model-based calibration. Before the approach is presented I will make a recapitulation of the regression estimator and of calibration in survey sampling.

Let the N elements in the population U be identified by subscript k;

$$U = \{1,...,k...,N\}$$

Denote the variable under study, say consumption, by z and suppose that we wish to estimate the total

$$T_z = \sum_1^N z_k$$

from a sample s selected from U with probability $p(s)$ and whose elements have the inclusion probabilities π_k. The standard Horvitz-Thompson estimator t_z is defined as

$$t_z = \sum_{k \in s} \left(\frac{1}{\pi_k} \right) z_k = \sum_{k \in s} d_k z_k$$

where the inverse probability d_k is the (expansion) weight for element k. Suppose now that there exists some auxiliary information x_k with a known total

$$T_x = \sum_1^N x_k$$

This is utilised in the generalised regression estimator (GREG)

$$t_{GR}(z) = t_z + (T_x - t_x)' \hat{B}_{zx}$$

where t_X is the Horvitz-Thompson estimator of T_x from the sample, and $\hat{B}_{zx}$ is the regression coefficient

$$\hat{B}_{zx} = \left(\frac{\sum_{k \in s} d_k x_k z_k}{\sum_{k \in s} d_k x_k^2} \right)$$

If x and z are correlated then this is exploited in GREG such that the precision of $t_{GR}(z)$ is better than that for the simpler t_z.

4 Calibration

The original idea in calibration is to modify the weights so that known totals are reproduced for the sample. More precisely for one known total T_x we calibrate by constructing weights w_k such that

$$t_x^{'} = \sum_{k \in s} w_k x_k = T_x ,$$

where the new weights w_k are as close as possible to the old weights d_k.

Minimising a quadratic distance measure

$$\sum_{k \in s} (w_k - d_k)^2 / d_k ,$$

the solution giving the new weights is

$$w_k = d_k \left(1 + \frac{(T_x - t_x)}{t_s} x_k \right),$$

where

$$t_s = \sum_{k \in s} d_k x_k^2$$

Deville and Särndal (1992) showed the asymptotic equivalence of calibration to the generalised regression estimator, Cassel, Särndal and Wretman (1976), thereby providing a way to establish the properties of calibration estimators.

There are some distinct advantages with calibration. Auxiliary information is incorporated in the weights. The obtained estimates are "consistent" with known information; usually this applies to totals. There is a correction for non-response, if non-response is related to the auxiliary information, and there is a reduction of variances for totals if variables are correlated.

Among the problems of the approach is that negative weights may appear. This is against our intuition. Modifications are possible at the expense of a more complicated procedure. Variances in ratios for the estimation of e.g. averages may increase for a calibrated estimator.

5 Model based calibration

The modification of the GREG estimator suggested in Cassel et al. (1997) is obtained if an estimator or predictor t_x^* of T_x is used in the GREG estimator, thus

$$t_c(z) = t_z + (t_x^* - t_x)' \hat{B}_{zx}$$

There are many different alternatives for this extension, depending on the definition t_x^* but here I will recapitulate the main alternatives in Cassel et al. (1997):

- $t_x^* = R_m t_y$, where a ratio R_m is given by a supporting model, for example an autoregressive model, or in our application a consumer model; the AIDS model below
- $t_x^* = \hat{R} t_y$ an estimate of the average ratio obtained from the current sample and constructed using a so called separate ratio estimator.

A basic idea behind these choices is that we try to borrow support from the estimate of a major variable t_y in the sample, a variable hopefully measured with better precision, such as total expenditure compared to expenditure in a special category. The specifications also rest on the assumption that the proportion of expenditure for households of a certain type is on the average approximately the same; in the models we also assume that the differences over time can be modelled.

For more detail let x_k denote expenditure for a certain category, for example for clothing, and let y_k denote total expenditure, all for household k. Form the ratio (x_k / y_k), which is the proportion of total expenditure spent on clothing by household k. If we consider households of the same type, for example cohabitants with 2 children, we may expect the proportion of expenditure spent on clothing to be roughly equal within this group, at least over a longer period. Then we may also expect the that the average $\hat{R}$, where

$$\hat{R} = \left(\sum_s \left(\frac{x_k}{y_k} \right) / \pi_k \right) \Big/ \left(\sum_s 1/\pi_k \right)$$

is of the separate ratio type, carries auxiliary information. Thus we form t_x^* as

$$t_x^* = \sum_s \left(\frac{y_k \hat{R}}{\pi_k} \right) = \hat{R} \sum_s \left(\frac{y_k}{\pi_k} \right) = \hat{R} t_y .$$

Let us continue with the properties of the estimator. The expected value of $t_c(z)$ generally is

$$E(t_c(z)) = E\left(t_z + (t_x^* - t_x)' \hat{B}_{zx}\right) \approx T_z + \left(E(t_x^*) - T_x\right) B_{zx}$$

where the latter product indicate a possible bias. If the model is true such that the expression within the parentheses is zero then there is no bias; therefore the estimator is model dependent. B_{zx} = 0 also results in a zero bias.

Specifically for $t_x^* = \hat{R} t_y$ we have

$$E(t_x^*) = T_y \left(\frac{1}{N} \sum_1^N \left(\frac{x_k}{y_k} \right) \right) = T_y \overline{R}$$

where $\overline{R}$ is the average proportion. The bias then becomes

$$\left(T_y \overline{R} - T_x\right)' B_{zx}$$

which is small if $x_k \approx \overline{R} y_k$. The variance of $t_c(z)$, $V(t_c(z))$, is written

$$V\left(t_z - (t_x - t_x^*)' \hat{B}_{zx}\right) = V\left(\sum_s \left\{ \frac{z_k - (x_k - x_k^*)' \hat{B}_{zx}}{\pi_k} \right\} \right) = V\left(\sum_s \left(\frac{e^*}{\pi_k} \right) \right)$$

where

$$e_k^* = (z_k - (x_k - x_k^*)\hat{B}_{zx})$$

and

$$x_k^* = \begin{cases} \hat{R}y_k \\ R_m y_k \end{cases}$$ for the two main alternatives above respectively.

Under the approximations $\hat{B}_{zx} \approx B_{zx}$ and $\hat{R} \approx \overline{R}$ we get

$$e_k^* = (z_k - (x_k - \overline{R}y_k)B_{zx})$$

An alternative way of writing the estimator is

$$t_c(z) = \sum_s \left(\frac{e_k^*}{\pi_k}\right) = \sum_U I_k \left(\frac{e_k^*}{\pi_k}\right)$$

where $\sum_U$ denotes summation over the entire population and I_k is an indicator which indicates whether k is included in the sample or not.

To estimate the variance it should be possible to use the expression

$$\hat{V}(t_c(z)) = \sum\sum \frac{1}{\pi_{kl}}\left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1\right)e_k^* e_l^*$$

Note finally that when the variable under study z is identical to the auxiliary variable x we get

$$t_c(z) = t_x + (t_x^* - t_x)1 = t_x^*$$

and under the model $t_x^* = R_m t_y$ we find that

$$E(t_c(z)) = R_m T_y$$
$$V(t_c(z)) = R_m^2 V(t_y)$$

thus the variance decreases considerably if the ratio is small.

The information from the model assimilated in the weights. In an extended calibration we may include both register based information and model based information at the same time.

6 A model of consumer demand

There are different alternatives for a model-based calibration regarding consumption as regards the model choice. In Cassel et. al. (1997) the AIDS (Almost Ideal Demand System) model was tried. This is an elaborate model based on utility maximisation which initially was specified for individual households observed over time, Deaton and Muellbauer (1980). In Deaton (1985) the model is specified for cross sections of domain means and thus interpretable as a modelling of the behaviour of an average household in a domain. The model can be used to predict the proportion of expenditure spent on certain goods. The model is specified as a system of equations with typical member

$$P_{gdt} = \alpha_g + \partial_d + \sum_{g_0} \gamma_{g_0} I_{g_0 t} + \beta_g \bar{x}^{*}_{dt} + \varepsilon^{*}_{gdt} \qquad (1)$$

where P_{gdt} is the proportion of expenditure spent on category g for domain or type of household d at time t. Subscripts g and g_0 denote categories of expenditure and

$$\bar{x}^{*}_{dt} = \frac{1}{n_{dt}} \sum_{h \in d} \log\left(x_{ht} / y_{ht} I^{*}_{t}\right)$$

where $x_{h,t}$ is the total amount of expenditure and $y_{h,t}$ is the number of consumer units for household h at time t. The overall price index I^{*}_{t} is in our application approximated with Stones index at time t; $\log I^{*}_{t} = \sum_g \bar{P}_{gt} \log I_{gt}$. Further notations I_{gt} or $I_{g_0 t}$ is the value of the price index for category g (or g_0) at time t, α_g, β_g, and γ_{g_0} are parameters, ∂_d is a domain effect and $\varepsilon^{*}_{gdt} = 1/n_{dt} \sum_{h \in d} \varepsilon_h$ is the residual.

Even though the model is defined on a domain level it can be deduced from the micro level assuming that the parameters α, β and γ are the same for households within the domain d. Using dummy variables for the domains it can be estimated on the entire sample of households. For the estimation restrictions on α, β and γ are required, see Deaton and Muellbauer (1980).

Table 3
Household types in the analyses

SFNOC64	Single woman without children -64 years
SMNOC64	Single men without children -64 years
SNOC65_	Single persons without children 65- years
CONOC64	Cohabitant households without children -64 years
CONOC65_	Cohabitant households without children 65- years
CO1CH	Cohabitant households with 1 child
CO2CH	Cohabitant households with 2 children
CO3_CH	Cohabitant households with more than 2 children
ONOCH	Other households without children
OCH	Other households with children

The AIDS-model (1) is estimated for Swedish HBS data using the three surveys[3)] 1985, 1988 and 1992. Seven categories of expenditure were identified; food, clothing and footwear, furniture and household articles, transportation, recreation and cultural services, spirits and tobacco and other expenditures. In the study we used

3) Samples drawn from the register of the total population. The initial sample sizes were about 6 000 households. The response rates were about 63 percent in all surveys.

month as time-unit, hence T=36. The domains ∂_d were type of household, specified in 10 categories according to table 3.

Thus 359 aggregated observations were available for the analysis. Some of these consisted of rather few sampled households; at one time-point an empty type of household occurred and some observations were based on less than ten initial households.

The residuals in (1) are aggregated means for the domains at time t. We assume that

$$Var\left(\varepsilon^{\cdot}_{gdt}\right) = \frac{\sigma^2}{n_{dt}} \quad d = 1,2,\ldots,10 \text{ and } t = 1,2,\ldots,36$$

$$Cov\left(\varepsilon^{\cdot}_{1d_1t}, \varepsilon^{\cdot}_{g_2d_2t}\right) = 0 \text{ when } g_1 \neq g_2 \text{ or } d_1 \neq d_2$$

where n_{dt} is the number of households in domain d at time t. Heteroscedasticity caused by the aggregation was taken into account.

For the estimation we also need the total expenditure for the households and price indexes for the expenditure categories as well as for the total expenditure.

Table 4
Parameter estimates for the AIDS-model estimated on the Swedish HBS

	Food		Clothing/ footwear		Furniture/ Household articles		Transport		Recreation/ cultural services		Spirits/ tobacco		Other expenditures	
R^2	0.75		0.54		0.42		0.54		0.51		0.64		0.75	
Root MSE	0.081		0.074		0.056		0.111		0.100		0.039		0.124	
Dep mean	0.239		0.066		0.043		0.122		0.142		0.029		0.358	
Variables	Param. est.	Std. error	Param. est.	Std. error	Param. est.	Std. error	Param. est.	Std. error	Param. est.	Std. error	Param. est.	Std. error	Param. est.	Std. error
α_g	1.042	0.055	-0.197	0.047	-0.248	0.043	-0.232	0.072	-0.183	0.065	0.079	0.028	0.738	0.073
γ_1	0.014	0.071	-0.019	0.016	0.208	0.085	0.043	0.063	-0.060	0.046	-0.055	0.024	-0.130	0.045
γ_2	-0.055	0.024	-0.012	0.009	0.0004	0.027	-0.012	0.038	0.064	0.026	0.042	0.019	-0.027	0.025
γ_3	-0.019	0.016	0.089	0.012	0.061	0.015	-0.006	0.024	-0.093	0.023	-0.012	0.009	-0.020	0.017
γ_4	0.208	0.085	0.061	0.015	-0.415	0.121	0.300	0.074	-0.014	0.045	0.0004	0.027	-0.139	0.047
γ_5	0.043	0.063	-0.007	0.024	0.299	0.074	-0.411	0.129	-0.114	0.066	-0.012	0.038	0.202	0.082
γ_6	-0.060	0.046	-0.093	0.023	-0.014	0.045	-0.114	0.066	0.224	0.073	0.064	0.026	-0.006	0.046
γ_7	-0.130	0.045	-0.020	0.017	-0.139	0.047	0.202	0.082	-0.006	0.046	-0.027	0.025	0.121	0.059
β_g	-0.132	0.009	0.045	0.008	0.042	0.007	0.063	0.012	0.057	0.011	-0.006	0.005	-0.069	0.012
SFNOC64	0.030	0.004	0.028	0.003	-0.0001	0.003	-0.052	0.005	0.020	0.005	-0.001	0.002	0.035	0.005
SMNOC64	-0.016	0.003	-0.018	0.003	-0.019	0.002	0.008	0.005	0.026	0.004	0.021	0.002	-0.002	0.005
SNOC65_	-0.026	0.006	0.004	0.005	-0.001	0.004	-0.052	0.008	0.009	0.007	-0.014	0.003	0.081	0.008
CONOC65_	-0.001	0.004	0.005	0.004	0.016	0.003	-0.008	0.006	0.009	0.005	-0.012	0.002	-0.009	0.006
CO1CH	-0.014	0.003	0.008	0.003	0.003	0.002	0.001	0.004	-0.018	0.004	-0.008	0.002	0.028	0.004
CO2CH	-0.029	0.004	0.019	0.003	0.003	0.003	0.001	0.005	-0.011	0.004	-0.014	0.002	0.031	0.005
CO3_CH	-0.033	0.005	0.032	0.004	0.008	0.004	0.014	0.007	-0.011	0.006	-0.016	0.003	0.006	0.007
ONOCH	-0.018	0.004	0.017	0.003	-0.001	0.003	0.022	0.005	0.011	0.005	-0.006	0.002	-0.026	0.005
OCH	-0.024	0.004	0.032	0.004	0.003	0.003	0.003	0.006	0.003	0.005	-0.011	0.002	-0.005	0.006

In Table 4 below estimation results are shown. For some categories of expenditures such as for food, spirits and tobacco and other expenditures the coefficient of determination is high. As a reference group we selected the group consisting of cohabitant households without children -64 years. The other household types are compared to the reference group. A negative value of a parameter indicates a lower proportion of the specific group of expenditure and a positive value indicates a higher proportion. The standard errors can be used for testing the hypothesis that the parameter estimates are zero.

The system was estimated with the SAS procedure SYSLIN (ver 6.10). To take the restrictions on the parameters within and between the equations in consideration we used the SRESTRICT function. We also used the WEIGHT command to avoid heteroscedasticity due to the aggregation.

This specific model may not the best choice when the purpose is to analyse the household consumption behaviour. Our more modest purpose with the estimation is to obtain estimated proportions

$$\hat{P}_{gdt} = \hat{\alpha}_g + \hat{\partial}_d + \sum_{g_0} \hat{\gamma}_{g_0} I_{g_0 t} + \hat{\beta}_g \overline{x}^{*}_{dt}$$

which we can use as auxiliary information in the calibration. In section 5 the auxiliary information is specified as totals, T_x. We then have to transform the estimated proportions. One possibility is

$$t_x^{*} = R_m t_y ,$$

where the ratio $R_m = \hat{P}_{gdt}$ is given by the supporting model and t_y is the estimated total expenditure. For the 1992 HBS this would mean 12 months × 7 groups of expenditures × 10 types of households = 840 auxiliary variables. To reduce this set we decided to aggregate over the year, thus $R_m = \hat{\overline{P}}_{gd}$. Consequently 7 groups of expenditures × 10 types of households = 70 auxiliary variables were used in the calibration. Compared to an ordinary least squares regression, where a smaller number of explanatory variables often is desirable, this could be still regarded as too many variables. But the purpose is to examine survey estimators when an econometric model provides auxiliary information. For a consumption model that satisfies econometric requirements, we are able to produce consistent consumption proportions for different kinds of households and goods.

7 Calibration results

Finally we are ready to evaluate the different calibration alternatives. Results for the Swedish 1992 HBS are compared using four methods for the estimation of average expenditures, namely

- unadjusted ratio estimates;
- CRD, traditional calibration using register data (population size by age and sex, number of households in four geographical regions, number of owner-occupied dwellings and other dwellings);

- CMD, calibration using model data or model-based calibration (given the results in section 6);
- CMRD, calibration using both model and register data.

For all details see Cassel at al. (1997). The unadjusted ratio estimates are based on simple random sampling assumptions; these have been found to well approximate the actual HBS design, the survey consists on random samples of individuals. No account for the about 37 percent non-response is taken in the unadjusted ratio estimates which are a base for a comparison of the three calibration estimators

Results are presented as ratios in figures 2, 3 and 4. The ratios are the results of either calibration method divided by the corresponding result for the unadjusted ratio estimates. Thus the methods are only compared to each other, and not to any "true values". In figure 2 the ratios between the averages for all households for different expenditure categories are compared, as well as coefficient of variance ratios. The ratios of averages show very small differences for the estimation alternatives; the impression is that there is no non-response bias since the ratios are close to one, or rather that the auxiliary information is unable to correct for non-response effects. There is no sign of any additional bias following from the model-based calibration.

The CV-ratios are, however, different. While the differences for total expenditure and some categories such as food, household services and non-durables are small, we seem to obtain gains for other expenditures. Interestingly, only one of those expenditure categories with small differences was used explicitly in the AIDS-model, that is, the auxiliary model-based information does usually not concern these categories specifically. For five out of the ten consumption categories in figure 2, the CV-ratios are at least halved for the model-based calibrations, CMD or CMRD. The differences between these two model-based alternatives are small. There seems to be little or no gain of including register information in the calibration both as regards the unadjusted estimator, the CRD-ratios are close to one, and the CMRD-ratios are close to the CMD-ratios, usually.

In figure 3 and 4 CV-ratios are given for some different household types. The pattern is very similar to what we saw for all households, gains for the same expenditure categories of about the same size for the model-based calibrations. There are two notable exceptions in figure 4. For non-durable goods the model-based estimates show a CV-gain for single persons age 65 and above, while for household services the model-based estimates are inferior with CV-ratios larger than one for cohabitant households age 65 and above. This is different from the results for all and for other household types, also those not shown. There thus is no guarantee of an improvement even if this was observed for a higher level of aggregation.

8 Discussion

In the paper by Cassel et al. (1997) model-based calibration is suggested to improve precision. I have presented the approach and an accompanying application to the Swedish HBS 1992. The auxiliary information obtained from a rather elaborate model of consumer demand seems to have the potential of improving precision in estimates of average expenditure. The same model-based calibration procedure has also been applied on the Finnish HBS. Lacking some information on the survey design and on price indices, the explanatory power for the equations in the AIDS-model was lower and the gains from the calibration smaller than for the Swedish case. Still a gain was observed for most expenditure categories. A drawback of a model-based calibration is the reliance on the model, which is assumed to be "true". More research is needed on the effects of models of different quality before more definite recommendations on the use of this kind of calibration are possible. The combination of rotating or rolling samples and calibration remains to be examined.

References

Andersson, C.; Nordberg, L. (1998): A User's Guide to CLAN 97 – a SAS-program for computation of point and standard error estimates in sample surveys, Statistics Sweden.

Cassel, C. M.; Granström, F.; Lundquist, P; Selén, S. (1997): Cumulating Data from Household Budget Survey. Some Results for Model Based Calibration Techniques Applied to Swedish Data, Report financed by European Communities, LOT 23.

Cassel, C. M.; Särndal, C. E.; Wretman, J.H. (1976): Some results on generalized difference estimation and generalized regression estimation for finite populations, in: Biometrika, Vol. 63, pp. 615 – 620.

Deaton, A. (1985): Panel data from time series of cross-sections, in: Journal of Econometrics, Vol. 30, pp. 109 – 126.

Deaton, A.; Muellbauer, J. (1980): An almost ideal demand system, in: American Economic Review, Vol. 70, pp. 312 – 126.

Deville, J.-C.; Särndal, C.-E. (1992): Calibration estimators in survey sampling, in: Journal of the American Statistical Association, 87, pp. 376 – 382.

Kish, L (1987): Statistical Research Design. Wiley.

Kish, L (1999): Cumulating/combining population surveys, in: Survey Methodology Vol. 25, 2, pp. 125 – 138.

Appendix

Figure 2
Ratios of averages and CV's, different calibrations and expenditures, all

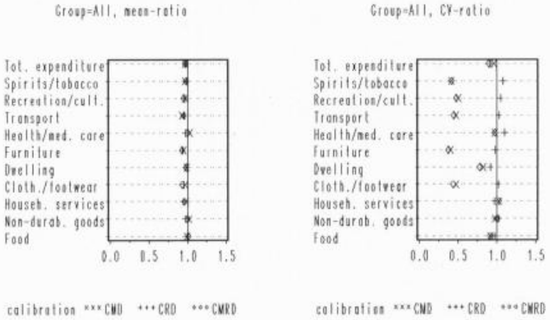


Figure 3
Ratios of CV's, different calibrations and expenditure categories, cohabitant households, one child and single women age 64 and below

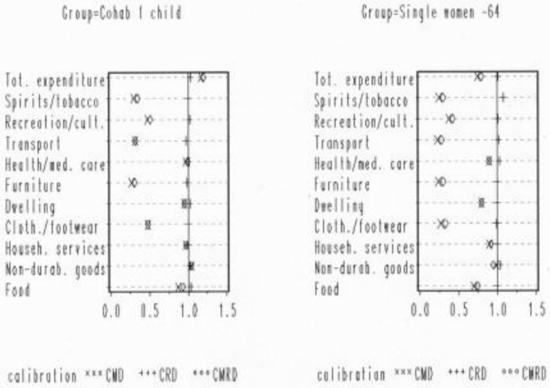
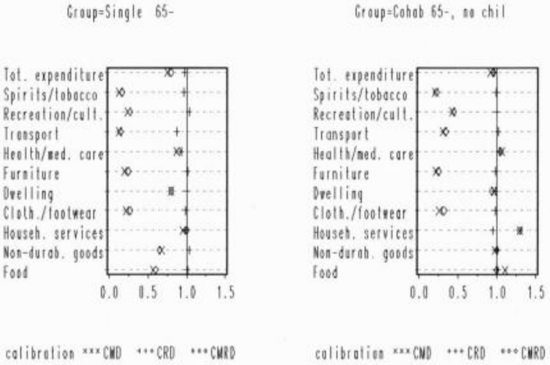


Figure 4
Ratios of CV's, different calibrations and expenditure categories, single households and cohabitant households 65 and above, no children



Bo Møller*)

Accumulating data from different years

The methods used in the Danish Household Budget Survey

This paper will give a short description on why and how data from different years are accumulated in the Danish Household Budget Survey (HBS). The general idea of using data from several years to form the basis of the HBS is not a Danish one – the method was originally invented in Statistics Norway. The practical solutions, etc., however, are not the same in the 2 countries.

Up till 1987 the HBS in Denmark was conducted as in many other countries, that is a so-called big survey was conducted every 5th or 6th year. When using the word "big" the size of Denmark should be taken into account. On the one hand it is not so important to have a big sample to enable e.g. regional analysis, as there is not so big differences of regional character in this small country. On the other hand being a small country also yields small budgets for statistics – which prevent the use of very big samples. Actually the sample size net has been around 3 000 households in the old as in the new survey as well.

The results of the old form of survey were not felt satisfactory for several reasons:

- The users wanted more updated figures.
- The production time of the survey were to long.
- The quality was felt not to be good enough due to high and growing non-response and to other non-sampling errors.
- It was difficult to compare figures from subsequent surveys as changes in methods etc., were always implemented.

On this background the survey was methodological changed a lot from 1994. One of the changes was the transition to having a permanent or continuous survey, which is the topic of this paper.

The HBS is continuous in 2 ways:

- The households are visited evenly spread round the hole year (to cover seasonal changes).
- Households are visited each year.

A continuous survey has in the opinion of Statistics Denmark a lot of advantages where the most important are:

- The survey results are updated yearly, which is very important for the users – that are internal or external users.

*) Bo Møller, Statistics Denmark, Copenhagen.

- It is much easier to implement and evaluate changes in methodology, etc., if you have a continuous survey. In the old system, where the time span between 2 surveys was about 5 years, it was impossible to separate the impact of changes in methods, etc., from "true" changes in consumption, etc.
- When you have a continuous survey it is easier to reuse EDP-programmes and to make necessary changes compared to the old system, where EDP-programmes often had to be totally rewritten from one survey to the next. This shortens the production time.
- The staff in the central bureau is working with the survey on a permanent basis – this implies that the staff is always well trained which can give a better quality.
- The continuous survey gives a better distribution of the workload round the year, which in the long run can save some resources and shorten the production time of the final statistics.
- As for the bureau staff the continuous survey also ensures that the interviewers are always skilled and trained, which is crucial for the quality of the data collected in the field.

Altogether these advantages will lead to higher quality and more usable results produced quicker and by relatively less use of resources.

The HBS is an expensive survey. Thus the number of households in the survey is decided not from what is the statistical optimum but from the funds available. The result of the budgetary considerations is that each year the sample will be of about 1.000 households net – that is after non-response. The gross sample is each year 1.500 households.

This rather small sample is felt to be too small to produce reliable figures. The solution on these problems has been the accumulation of data from several years. In the Danish HBS data collected in 3 subsequent years are accumulated. A kind of moving average is constructed – that is that in each year's update the oldest year is removed and a new year is included in the survey. The most recent survey was the 1997 – 1999 surveys, and the next will therefore be the 1998 – 2000 survey. It should be stressed that this method implies that data from a specific household is used 3 times for 3 different updates – but the household is not revisited. There is no panel element in the Danish HBS.

A few words on the data sources and collection will be necessary. The Danish HBS has 3 main data sources:

- *An interview with the households.*

  This interview is conducted as CAPI. The interview is divided into 2 parts. The interview gives information on fixed payments, on expenditure on durables, etc.

- *The diaries kept by the households for 2 weeks.*

The diaries are open – that is the households have to write down with their own words all their expenses. Afterwards the diaries are coded and checked in the bureau.

- *The administrative registers.*

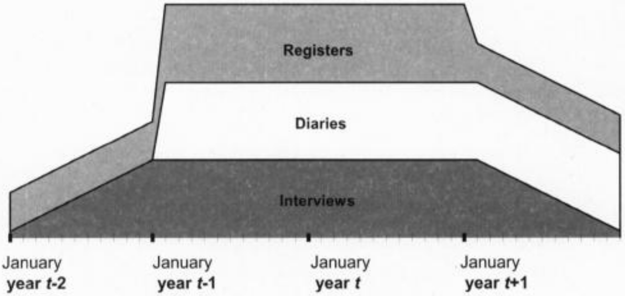
  Especially the tax-based register on incomes and taxes is used intensively. The data in the register is matched direct with the persons in the survey, as Statistics Denmark has access to the personal identification number of the persons in the survey.

The different data sources have different reference periods. For the interview the reference period is the 12 month previous to the date of the interview. For the diaries the reference period is the 14 days starting the day after the first visit at the households. For the register based information the reference period is the calendar year. For the households who are visited in the first half of a year the register version from the year before is normally used, and for households from the second half of the year the register version of that year is normally used.

Even though the data in the survey is collected during 3 years the data in the HBS in fact refers to four different years, as some of the data collected in the first year refers to the year before. One of these years should be chosen as the reference year for the published results, as it is important for the users to be able to compare the HBS-results with other kind of statistics.

The data coverage for the 4 different years is schematically shown in this graph.

Figure 1
Data Coverage



It can be seen that for the **years *t*-1** and ***t*** then amount of data or information is bigger than in the 2 other involved years. Therefore Statistics Denmark think one of these 2 years should be chosen as reference year – and we have chosen the **year *t*.**

It can be mentioned that in the Norwegian HBS the last year – that is the **year t+1** – is the reference year. By doing this the Norwegian figures seems to be more updated – but Statistics Denmark regards this as having a more "cosmetic" character.

Depending on what the accumulated figures are intended to show, different ways of accumulation could be used. The 3 main methods having each its own "philosophy" are:

1. *Accumulating without any adjustment for changes in prices or quantities.*

   This method will give a true picture of the average consumption in the 3-years period. If the actual sample size structure is however different the 3 years of course a kind of reweighting should be used before the accumulation – but comments on reweighting schemes are outside the scope of this paper.

2. *Accumulating after having inflated/deflated with some price indices.*

   The idea here is still to show the 3 years average, but taking account of the price changes. That is the result will be a true 3-years average in constant prices. If such a method is used it should further be considerate which price indices should be used: It could be the most detailed series or it could be just the average price index for total consumption. The latter implies that the structure of consumption (in amounts) is constant, while the former method will change the structure.

3. *Accumulating after having adjusted for changes in prises and in quantities as well.*

   The main idea here is to show estimates not on the 3-years average, but to give estimates of how the households would have spend, etc., if they had all been in the survey in the reference year.

Statistics Denmark has the feeling that the users of the statistics normally are not interested in the 3-years averages. They want estimates for a single year that can make it possible to compare to other statistics, etc. Therefore method 3 is used in Denmark (while e.g. Norway use the method 2).

No experiments on real data have been made on the impact of using the one or the other of the 3 main methods – but it is believed that in "normal" years the differences in the estimated total consumption figures will be rather marginal.

In more "abnormal" years – that is years where things changes up- and downward – the differences will be bigger.

To be able to simultaneously to take account of changes in prices and in quantities, the basis of the adjustments in Statistics Denmark is *the most detailed quarterly national account figures on private consumption in current prises.* Changes in these figures reflect on the same time the effect of the real changes in prises and in consumed quantities.

The indices created on this basis are used not just for consumption but also for incomes, taxes, etc. For these non-consumption parts of the survey the index reflecting the total private consumption is used.

A necessary precondition for the method used in Denmark is, of course, that national account figures are available on a quarterly basis, and that the figures are assessed to have the sufficient quality.

If national account figures do not exist on a quarterly basis, it should be considered to use the yearly figures in stead by in some way interpolate the "missing" values in the quarters (or months or what so ever).

A main problem when using accumulated data (independent of the method of accumulation) is that it is not possible to give valid information on changes in the short term – that is from one year to the next. A little example will demonstrate the problem:

Table 1

1996	1997	1998	1999
100	110	105	110

HBS 1996 – 1998 = (100 + 110 + 105)/ 3 = 105
HBS 1997 – 1999 = (110 + 105 + 110)/ 3 = 108

It can be seen that the average value (it could be total consumption or a value for a single consumption item) is growing in the new HBS update of the estimates compared to the version from the year before. But from the underlying figures it can be seen that the actual change was a fall from 1997 to 1998.

The shape of the changes causes this result from one year to the next. If everything is stable or growing with a constant rate no problems will arise.

To this comes the problem that just 1/3 of the households are replaced at each update. This implies that if some changes happens in consumption, etc., these will be captured "slowly".

As a consequence Statistics Denmark in its publications, etc., never compare one year's updated figures with the version from the year before. We always go at least 3 years back. And further we always stresses that the HBS is a very good survey when you want information on structures, on consumption in different household types, etc., but if the need is for statistics on short-term changes in economy other statistics should be preferred – it could be national accounts, the VAT-based statistics on retail sale or something else.

The general method used is evaluated to be able to give good estimates for total consumption and for aggregated parts of consumption as the adjustment by the national account figures "ensures" the right levels, etc.

But if you go below the levels where consumption figures can be found in the national account, severe problems can arise. As an example can be mentioned situations where new commodities comes into the market – that could e. g. be mobile phones. The adjustment with the national account figures on "Communications", which is a lot broader than just mobile phones, will not "capture" the where steep rise of expenditures on mobile phones the recent years.

A solution on this kind of problems could in theory be to make some adjustment "by hand". But a main problem would be that if you "by hand" raise the consumption on one item you would on the same time have to lower some other parts of consumption – but which? So in the Danish HBS no such additional adjustments are maid.

Another true example from Denmark is the following: From 1998 a new "tax-like" compulsory pension scheme was introduced. For incomes from labour normally 1 pct. is collected by the tax authorities and put into a pension fund.

The problem was now how and when to incorporate this in the survey. If you just incorporated it as it really came to power, it would in the 1996 – 1998 updated version of the survey just affect 1/3 of the households and then give wrong impressions on the size and impact of the pension payments.

The practical solution was the following: In the 1996 – 1998 – version this new "tax" was deliberately "hidden" among other kind of pension contributions. As the amount was rather small as just 1/3 of the households were affected, the problem was not regarded to be very big.

In the 1997 – 1999 – version now 2/3 of the households were affected by the payment, and the reference year was now the year, where the system was actual introduced. Therefore it was felt necessary to show this new pension scheme separately. But to avoid showing misleading low figures (just 2/3 of the actual level in 1998) a pension contribution in 1997 was imputed – that is, Statistics Denmark calculated what the contribution in 1997 would have been if the system had been effective this year.

Also this kind of estimations is not without problems. If adjustments are made one place in the HBS system, you will have to directly or implicit to make another adjustment elsewhere. In this case it was implicitly assumed, that the imputed pension contribution in 1997 was totally financed by a smaller unknown net saving in the households from 1997, but of course the true case could also be that households, that from 1998 actually pay this new pension contribution choose to lower their consumption instead.

A general problem arises on micro level in all cases where a specific household has a consumption pattern different from the national average – and they always have!

Table 2

	Year *t*-1	Year *t* Average consuming household	Year *t* Household with non-average consumption
Income	100	110	110
Consumption	90	99	105
Saving	10	11	5
Saving in percent of income	10	10	4.5

By the adjustment with the average national account figures for total consumption from **year *t*-1** to **year *t*** the structure is the same for households with average consumption, that is consumption and saving has an unchanged proportion of income.

But for the household that consumes more than average on goods, for which the national consumption is growing faster than average, the structure will be changed. The adjusted consumption will be higher, and hence the net saving will be to low, as net savings is calculated as the residual between incomes and uses of income.

No solution on this problem on micro level has been found – but the hope is that the problem for groups of households will be averaged out.

In Statistics Denmark the incomes and taxes, etc., in the survey are adjusted according to the change in average consumption in national accounts. This can be regarded as a little odd – a more "logical" way of adjusting is to use an index that reflects changes in incomes as such. But if that were done a new problem would be the result:

Table 3

	Year *t*-1	**Year *t***
Income	100	110
Consumption	90	95
Saving	10	5

In this example income according to other statistics has a higher rate of growth than average consumption. If we in this case adjusted according to the growth of incomes in society, a too low saving would be imputed – so to keep the structure unchanged one have to adjust also incomes, taxes, etc., by the consumption figures.

The first problem to solve is to decide which quarter should be used to adjust according to the quarterly national account figures. For the diaries there is no problem as we exactly know the dates where the diaries where kept. For the register-based information it is also straightforward, as wee know which calendar year the register refers to.

For the information based on the interview we just know, that the expense have been made during the last 12 months counted back from the day of the interview. The only solution is to assume that the expense was maid exactly half a year before the day of the interview. For the parts of the interview covering regular payments this will not give big problems, but for the other parts – e.g. the question on buying of cars – problems can arise, especially if the consumption on the specific item is fluctuating, as it is for cars.

The household data used for each updated version of the survey consists in 2 parts: The one part is the "old" households, that where also used for last year's update. For these households the information already exits on a yearly basis and adjusted to a common reference year – the year before the new reference year. For these households all information is now just readjusted by exactly 1 year.

For the other part of the households, the "new" households, the adjustment has to be done separately for each household depending on the date of the interview and the dates of the diary period.

The classification of consumption is the same (apart from very minor points) in the HBS and in the national accounts – the COICOP. Therefore the adjustment is rather simple:

- From the national accounts a file with detailed figures on consumption per quarter in current prices is received.
- The file is reformatted to show the deflation factor from each quarter to the average level in the reference year.
- The quarter to be basis for the deflation is constructed for each record in the HBS-file as described above.
- In a file of all codes used in the HBS it is marked whether this type of record should in fact be adjusted. This is necessary in the Danish survey where also all other data is stored in the same file – that is including data on household size, number of cars, etc. Of course these records should not be adjusted.
- The 3 files are matched by code and the deflation factor is multiplied with each HBS-record in all the cases, where an adjustment should be done.

Imputed rent of owner occupied dwellings, including secondary dwellings, is recalculated separately – and the values are used as consumption and as income as well.

After consumption and income have been adjusted the net saving is recalculated.

The amount of indirect taxes in consumption, etc., is recalculated on the basis of the tax rules, etc., of the reference year. Thereby it is implicit assumed that possible changes in tax rates, etc., do not influence the consumption pattern – an assumption that is obviously wrong, but no solutions have been found.

As mentioned the publication of the results never include comparisons with figures from the version from the year before. Always comparisons are made to version at least 3 years old.

In the publication it is strongly stressed that the HBS is a very fine data source for analysis of structures – that could be structures in consumption patter, etc., or it could be structural differences for different household types. But if statistics on short time changes is needed, other statistical sources should be preferred.

Michael Wiedenbeck*)

Never jointly observed variables: can they be analysed by means of fused data

1 Abstract

The problem of the use of consecutive cross-sections for longitudinal analyses is treated as a specific case of multivariate analyses, where the data are several samples of lower-dimensional marginal distributions. If for variables X, Y and Z samples of the joint distribution of (X,Z) and (Y,Z) are given, in applied research often data are linked in accordance with their Z-values. This technique is called data fusion. Its application implements conditional independence, given Z, in the sample of fused data. Estimation of the joint distribution of X, Y and Z by the criterium of maximising the Shannon entropy leads to solutions which can be interpreted as generating distributions of samples of fused data.

2 Introduction

Longitudinal information as it is contained in panel data can help to improve the estimation of cross-sectional distributions. But can we draw inferences on longitudinal distributions from consecutive cross-sectional data? In general, this has to be denied. Time-dependent data are nothing but a peculiar type of multivariate observations, and it is obvious that in principle for each observation all components have to be collected from one single unit. Samples separately collected at different points in time therefore contain in general no information about developments in time on the microlevel, but they may contain longitudinal information on a more aggregate level or about the population as a whole.

Nevertheless, there are situations where multivariate analyses of variables which are not jointly observed are performed. Well-known examples are found in media and marketing research. Here a typical question for planning the timing of commercial breaks is: which combinations of certain comsumption behavior occur frequently with which habits of watching TV? Because rather detailed knowledge is needed, the collection of all the relevant information by a single source design is often impossible for practical reasons. Instead a design with two sources can be applied where one sample contains data of consumption behavior and the other sample the data of TV-habits. Additionally both samples include common variables. After the collection of the data, units from different samples can be joined in accordance with those variables. It is then assumed that the virtual data generated by this fusion are realistic, and that samples of them can be used at least approximately as substitutes of samples of jointly observed variables. Comparisons of results based on single source data with results based on fused data are mainly available within the framework of methodological investigations. They are usually limited to small sets of variables.

*) Michael Wiedenbeck, Center for Survey Research and Methodology (ZUMA), Mannheim.

3 Conditional independence and interpretation of fused data

A justification for data fusion is that identity of units of single source data can be substituted by equality of units in fused data, where equality is achieved only within certain limits of accuracy. But the appropriateness of fused data depends on the degree by which the true common distribution of consumption behavior and TV-habits can be approximated by a distribution satisfying conditional independence given the fusion criterium. Fusion yields data which are close to a sample where all variables are jointly observed, only if the true common distribution itself satisfies the condition of conditional independence. A judgement of this is usually based only on plausibility considerations. When the individuals are conceived as classified according to the fusion criterium, the conditional independence has a special meaning: within those classes the observations of one of the never jointly observed variables can be exchanged arbitrarily between individuals. In the consumption/TV-habits example, for instance, some socio-demographic and life-style variables are used as fusion variables. If then for empirical or theoretical reasons, no association between the consumption behavior and the habits of watching TV is known or thought to be plausible within classes, then an arbitrary or random combination of consumption and TV habits would deliver descriptions appropriate for real individuals. We then could feel justified to adopt conditional independence as a working hypothesis, and it would make sense to collect data from different sources to perform data fusion.

4 Maximum entropy estimation

As mentioned before, fusion of data is motivated by the idea that "identity" can be substituted by "equality" according to fusion variables. The objective of data fusion can be viewed as the analysis of an unknown joint distribution $[X,Y]$ of two multivariate variables X and Y. The data are samples from distributions $[X,Z]$ and $[Y,Z]$, where Z is a multivariate covariate. Usual procedures of estimation of $[X,Y]$ cannot be applied without making strong assumptions on $[X,Y]$ as interactions between X and Y are not directly represented by the data and cannot be identified.

This point can be illustrated by an artificial example, where it is assumed that $[X,Y,Z]$ is normal with mean μ and variance Σ. Furthermore, we assume that two samples are given, one consisting of independent observations from the marginal distribution $[X,Z]$ and the other one of independent observations from the marginal distribution $[Y,Z]$. Then the likelihood of the joint samples depends on means and variances of X, Y and Z and covariances of X and Z and of Y and Z. But it does not depend on the covariance of X and Y. Therefore it cannot be estimated by maximum likelihood estimation. By data fusion one avoids that kind of estimation problems at the price of implementing conditional independence, given Z, which in data fusion is used as a criterium for linking units. Another example is the common distribution of a set of categorical variables: in general, marginal distributions do not determine the full distribution, and hence samples from marginal distributions are not sufficient for a unique estimation.

On the other hand, there is some information on the covariance of X and Y contained in the samples of $[X,Z]$ and $[Y,Z]$. If we want to use this information we have to consider the common distribution of X,Y and Z jointly, (i.e. $[X,Y,Z]$). In general, this information, however, will not fill out the gap completely. For, if we take any estimation $[X,Y,Z]^\wedge$ of $[X,Y,Z]$, there will in general be many alternative estimates which fit the data, i.e. the samples of $[X,Z]$ and $[Y,Z]$, with the same accuracy as $[X,Y,Z]^\wedge$.

For the uniqueness of a solution we need restrictions. In case we don't have any prior knowledge about $[X,Y,Z]$ itself, restrictions of a general form are appropriate which do not implement any information not contained in the data. Therefore it is plausible to choose among all estimates which fit the data the one containing the least amount of information. In statistical theory the classical entropy measure of Shannon, also known as Kullback-Leibler distance, is often applied as a measure of information. As an example (Golan et al., 1996, p. 1) can be mentioned, who treat estimation based on the maximum entropy criterium in situations "when the underlying sampling model is incompletely or incorrectly known and the data are limited, partial or incomplete".

When we apply Kullback-Leibler distance to the described situation, we can show the following (see the Proposition in the Appendix): given any estimates $[X,Z]^\wedge$ and $[Y,Z]^\wedge$ of $[X,Z]$ and $[Y,Z]$, where the derived estimates of the marginal distribution $[Z]$ are equal, there exists always a unique estimate $[X,Y,Z]^*$, for which the marginal distributions of (X,Z) and (Y,Z) coincide with $[X,Z]^\wedge$ and $[Y,Z]^\wedge$. Furthermore, for X, Y, and Z being jointly distributed as $[X,Y,Z]^*$, X and Y are conditionally independent given Z.

This opens a new view at samples created by fusion of data. Because fusion implements conditional independence given Z, fused data can be interpreted as sample of $[X,Y,Z]^*$ whose bivariate marginals of (X,Z) and (Y,Z) coincide with $[X,Z]^\wedge$ and $[Y,Z]^\wedge$, respectively. The maximum entropy solution $[X,Y,Z]^*$ is of minimum information, and it can be shown that for its marginal bivariate distribution $[X,Y]^*$ information is minimised, too: $[X,Y]^*$ maximises the entropy for all distributions $[X,Y]$, which are derived as marginal distributions of a distribution $[X,Y,Z]$ consistent with $[X,Z]^\wedge$ and $[Y,Z]^\wedge$. Therefore, the results of an analysis of $[X,Y]^*$ which is essentially equivalent to an analysis based on the fused data will use all information contained in the given data, in the sense of the entropy information measure.

5 Concluding remarks

In general, an empirical validation of data fusion for large scale data is nearly impossible. It has been performed only for very small numbers of components of vector-valued variables X, Y and Z, which can be jointly observed. The full set of parameters of the original distributions $[X,Y,Z]$ or $[X,Y]$ can be compared with the corresponding parameters of – estimated – distributions of the fused data only if they are known in advance (i.e. by simulation studies) or if they can be estimated from samples of $[X,Y,Z]$ or $[X,Y]$, which are not available for those types of variables where data fusion is typically applied. Raessler and Fleischer (1998) have performed simulation studies and discuss the effects of data fusion on estimates of covariances and correlations between X and Y when $[X,Y,Z]$ is trivariate normal or lognormal. One of their main findings is that in the case of a normal distribution of (X,Y,Z) the covariance of X and Y can be estimated without bias if and only if the expected value of the conditional covariance of X and Y given Z equals 0. This is a special case of the result presented here.

If one thinks of data fusion for consecutive cross-sectional data as a means to the generation of quasi-panel data, one faces an additional problem. It may be very difficult to find appropriate fusion variables because those variables must not depend on time. Otherwise it will be not clear what type of units are grouped together and in which sense we call them "equal".

References

Golan, A.; Judge, G.; Miller, D. (1996): Maximum Entropy Econometrics. Robust Estimation with Limited Data, New York.

Luenberger, D. G. (1969): Optimization by Vector Space Methods, New York.

Raessler, S.; Fleischer, K. (1998): Aspects Concerning Data Fusion Techniques, in: Nonresponse in Survey Research, ZUMA-Nachrichten Spezial Nr. 4, eds. Koch, A. and Porst, R., Mannheim: ZUMA, pp. 317 – 333.

Rogers, W. L. (1984): An Evaluation of Statistical Matching, in: Journal of Business and Economic Statistics 2, pp. 91 – 102.

Appendix

Proposition (Existence and uniqueness of probability densities of multivariate random variables, which maximize the entropy measure under the restriction of fixed marginal distributions for two sets of components):

Let X, Y and Z be (multivariate) random variables with bivariate marginal densities $f_1 = f_1$ (y,z) for (X,Z) and $f_2 = f_2$ (y,z), where the marginal densities f_1(z) and f_2(z) coincide almost everywhere (a.e.). Then (1), there exists a unique distribution $[X,Y,Z]^*$ of (X,Y,Z) with density $g^* = g^*(x,y,z)$ among all distributions of (X,Y,Z), whose marginal densities of (X,Z) and of (Y,Z) coincide a.e. with f_1 resp. f_2, that has maximal entropy $\phi(g) = -\int \ln(g^*)dg^* = \max_g - \int \ln(g)dg$, where dg is equal to the probability measure $g(x,y,z)dxdydz$.

Further, (2), for g^* the following conditional independence of X and Y given Z holds $g^*(x,y,z) = g^*(x/z)g^*(y/z)g^*(z)$ (a.e.).

Wolf Bihler*)

Problems of Converting the Microcensus into a Continuous Survey

1 Introduction

The Microcensus (MZ) is the largest and most important population sample survey in Germany. The MZ provides data in great detail on the population structure, the economic and social situation of the population as a whole and families and households, employment, job search and education. The EU Labour Force Survey (LFS) is integrated into the MZ as a subsample.

At present, the whole sample focuses on a single week (usually the last week of April). This reference week is seen as a typical week, but the estimated totals are not annual means, and quarterly estimates are not available. In order to obtain estimates of quarterly and annual means, we intend to change the MZ into a continuous survey throughout the year, beginning in 2005. In converting the MZ into a continuous survey, we will have to consider two important requirements:

- The estimates of annual means must not be less precise than the estimates for the survey week at present.
- The number of interviews per year should not increase.

Converting the MZ into a continuous survey will have an impact on the following three items: the rotation scheme, the construction of interviewer areas and the estimation method. Before dealing with these items, we outline shortly the present MZ sample design.

2 The MZ Sample Design

The MZ is a one-stage stratified random sample. The sampling units are areas with about 9 flats on average. Depending on the size of the building, a sampling unit can be a single building, a small number of – usually neighbouring – buildings or a part of a building. The sample is stratified by regions in a detailed breakdown and by size classes of the building. The sampling fraction is 1%, and the inclusion probabilities are the same for all units. All households and persons living in a given area are included in the sampling unit.

Every year a quarter of the sample rotates. This means that a household remains in the sample for four years. Please note that the MZ is not a panel in the strict sense: households or persons moving away from the sampling area are no longer surveyed in the following year unless they move into another selected unit.

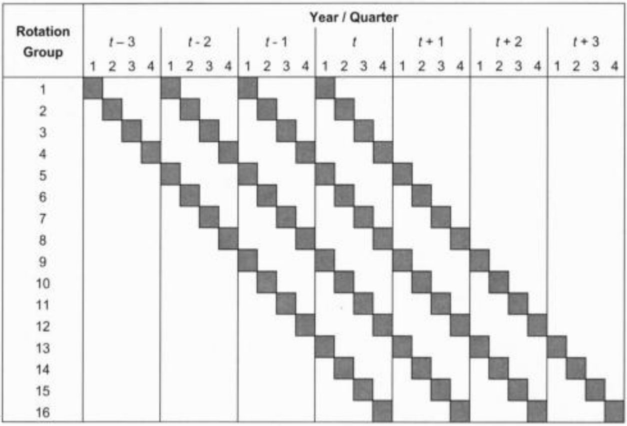
---

*) Wolf Bihler, Federal Statistical Office of Germany, Wiesbaden.

3 Rotation Scheme

In a continuous survey, the rotation scheme is very important: Should successive years and successive quarters overlap? The present 75% overlap between two successive years has proved to be good, and we will keep this overlap. The crucial question is whether successive quarters should overlap. We examined two potential rotation schemes: *Rotation scheme A*: 75% overlap between two successive years; no overlap between two successive quarters; *Rotation Scheme B*: 75% overlap between two successive years; 50% overlap between two successive quarters.

Figure 1
Rotation scheme A



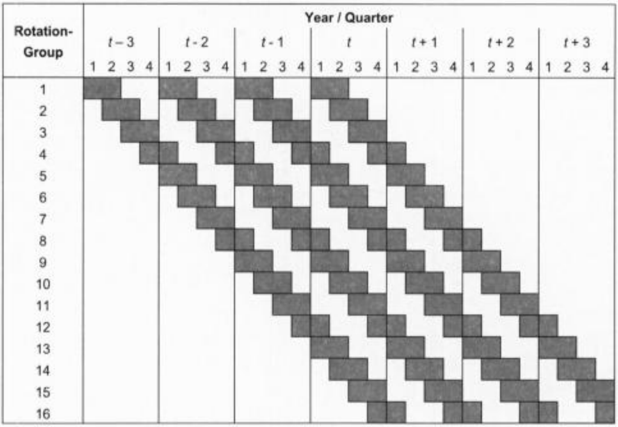
If the number of interviews per year is the same in both schemes, the number of different households per year under scheme B is about 0.53*n*, where n is the number of different households per year under scheme A. This means that the coefficient of variation for annual estimates under scheme B can increase to roughly

$$\sqrt{1/0.53} \triangleq 37\ \%$$

The increase varies for the different variables: for variables where the variance within the households is very small compared to the variance between the households, the increase would be near this upper limit. For variables varying to a greater extent over time within the households the increase would be smaller. We expect most of the MZ variables to be of the first type. As we cannot increase the number of interviews per year and the annual estimates must not become less precise, we have decided to choose rotation scheme A. According to Grohmann[1], the precision of the annual estimates remains approximately the same, when the yearly sample is divided into four quarters.

Under scheme A, the estimates of changes between quarters will be sufficiently precise, however, we have to raise the LFS sampling fraction to 1% in order to fulfill the EU regulation requirements concerning the precision of quarterly changes. Therefore the LFS will no longer be a subsample.

Figure 2
Rotation scheme B



4 Construction of Interviewer Packages

An interviewer has the capacity to work in more than one sampling area: on average he works in about 5.5 sampling areas. The sampling areas are allocated to the interviewers in order to minimise the interviewer journeys. The set of sampling areas which an interviewer works in is called an "interviewer package". At present, all the sampling areas of a year are allocated for the reference period to the interviewers together. Because of the large number of sampling areas (sampling fraction 1%), the distances to be covered by the interviewers are relatively small. In a continuous survey, we will have a smaller number of sampling areas to allocate to the interviewers for a specified time period. Therefore the distances between the sampling areas to

1) *Grohmann, H. (1991)*: Zum Übergang vom "Berichtswochenkonzept" zum "Konzept unterjähriger Erhebungen" beim Mikrozensus, Statistisches Bundesamt.

be covered by an interviewer will become larger (especially in rural areas) and the travelling costs will increase. Drawing random samples for each week of the year (sampling fraction 1/52%) and allocating this volume to the interviewers would not be practicable and produce too high costs. On the other hand, constructing interviewer packages based on the whole sample for a year and then dividing them into time periods (weeks, months or quarters) would cause a clustered sample with less precision of the estimates. Therefore we did not consider these options in detail. However, there are two more promising options between these extremes:

Interviewer packages based on monthly volumes: A random sample of the sampling areas is drawn for each month (sampling fraction 1/12%). The interviewer packages are constructed based on the selected sampling areas. Below the level of months, the sample is not randomly split into weeks.

Interviewer packages based on quarterly volumes: A random sample of the sampling areas is drawn for each quarter (sampling fraction 0.25%). The interviewer packages are constructed based on the selected sampling areas. Then the interviewer packages of a quarter are randomly split into three monthly parts.

As for interviewer packages based on quarterly volumes, the ways to be covered by the interviewers would be shorter than with packages based on monthly volumes. However, they would be longer than in the present situation (yearly basis).

We expect the coefficients of variation to be larger when the interviewer packages are based on quarterly volumes, because the sample is two-staged and more clustered. We analysed the coefficients of variation of the two methods by a simulation: We split the 1999 MZ sample data randomly into quarters and months. For the second-quarter subsample in some "Bundesländer" (NUTS 1), we built interviewer packages for the method based on quarterly volumes as if we already had a continuous survey in 1999. (For the method based on monthly volumes, the construction of interviewer packages was not necessary for variance estimation).

For simplification, we didn't consider here stratification and poststratification; their effects might be nearly the same in both methods. For the method based on monthly volumes, we used the classical variance formula to obtain the Horvitz-Thompson estimator for a certain month (index for the month is omitted):

$$\hat{Y}_{MB} = \frac{N}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} y_{ij} \tag{1}$$

$$\hat{V}\left(\hat{Y}_{MB}\right) = N^2 \frac{1-f}{n} S_1^2 \tag{2}$$

The method based on quarterly volumes is a two-stage sampling technique with sampling areas as primary sampling units and interviewer packages as secondary sampling units. We take a ratio estimator which uses the number of monthly sampling areas. As described above, the estimator refers to a certain month:

$$\hat{Y}_{QB} = \frac{N}{n'} \sum_{j=1}^{m} \sum_{i=1}^{n_j} y_{ij} \tag{3}$$

$$\hat{V}\left(\hat{Y}_{QB}\right) = N^2 \frac{1-f_1}{n} S_1^2 + \left(\frac{N}{n}\right)^2 M^2 \frac{1-f_2}{m} S_2^2 \tag{4}$$

where

N :	number of sampling areas per year in the population;
n :	sample size (sampling areas);
$f = n/N = 1/12\,\%$:	sampling fraction when the interviewer packages are based on monthly volumes;
$f_1 = n/N = 0.25\%$:	1st stage sampling fraction when the interviewer packages are based on quarterly volumes;
M :	number of interviewer packages per quarter (when based on quarterly volumes);
m :	number of interviewer packages per month;
$f_2 = m/M = 1/3$:	2nd stage sampling fraction when the interviewer packages are based on quarterly volumes;
y_{ij} :	value of the variable of interest in sampling area i of the interviewer package j;
n_j :	number of sampling areas in interviewer package j;
$n' = \sum_{j=1}^{M} n_j$:	sample size per month;
S_1^2 :	estimated variance between the sampling areas $S_1^2 = \frac{1}{n-1}\sum_{j=1}^{M}\sum_{i=1}^{n_j}\left(y_{ij} - \bar{y}\right)^2, \bar{y} = \frac{1}{n}\sum_{j=1}^{M}\sum_{i=1}^{n_j} y_{ij}$;
S_2^2 :	conditional variance between the interviewer packages $S_2^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(u_j - \bar{u}\right)^2, \; u_j = \sum_{i=1}^{n_j} y_{ij} - \bar{y} \cdot n_j$.

For variance calculation, the data of all months are used together. The months are fictitious, since we have no real monthly data. The coefficients of variation refer to a typical average month. For a certain month the coefficients of variation may differ only slightly, because the monthly samples have the same design and about the same size. In order to get variances for estimated quarterly means, we simply multiplied the monthly variances by three.

Some results are shown in the following tables 1 and 2. As a rule, the coefficients of variation are clearly larger in the method based on quarterly volumes than in the method based on monthly volumes, but the differences vary between the variables and the Bundesländer. Because of these results we have decided to use interviewer packages based on monthly volumes.

Table 1
Microcensus 1999, One Quarter
Coefficient of Variation
with Interviewer Packages Based on Monthly and Quarterly Volumes

No.	Variable of Interest	Estimated Quarter Mean	Coefficient of Variation		
			Based on Monthly Volumes	Based on Quarterly Volumes	Change[1)]
		Number in 1 000	%		
		1	2	3	4
		Hesse			
1	Population, total	5 551.2	6.4	7.9	22.2
2	Foreigners	613.1	20.1	21.6	7.5
3	Persons in employment	2 516.8	7.2	8.6	19.2
4	Unemployed persons	227.1	16.4	16.5	0.9
5	Male persons in employment	1 410.3	7.7	9.3	20.8
6	Female persons in employment	129.9	21.5	21.2	-1.6
7	Male unemployed persons	1 106.5	7.8	8.8	11.7
8	Female unemployed persons	97.2	20.9	21.0	0.4
9	Main source of livelihood: support by relation	1 712.2	9.4	11.4	21.5
10	Main source of livelihood: pension	1 146.8	9.4	10.3	10.3
11	Income under 600 DM	438.5	15.2	16.6	8.9
12	Income from 1000 to under 1800	754.2	9.3	10.6	13.1
13	Income from 1800 to under 3000	1 272.5	8.0	8.7	8.6

1) (Col. 3 – Col. 2) / Col. 2 * 100

Below the level of months, the weeks are not randomly selected. The interviewers determine the week to do their work efficiently. A pretest showed that it is appropriate to create two random groups of interviewer packages: one group starts with the field work at the beginning of the month, the other at the middle of the month. The interviewers of both groups ask for the variables of the previous week.

Table 2
Microcensus 1999, One Quarter
Coefficient of Variation
with Interviewer Packages Based on Monthly and Quarterly Volumes

No.	Variable of Interest	Estimated Quarter Mean	Coefficient of Variation		
			Based on Monthly Volumes	Based on Quarterly Volumes	Change[1)]
		Number in 1 000	%		
		1	2	3	4
		Saxony			
1	Population, total	3 970.4	7.1	8.5	20.4
2	Foreigners	32.5	50.8	54.7	7.7
3	Persons in employment	1 748.2	8.6	9.8	14.5
4	Unemployed persons	360.6	12.2	14.5	18.9
5	Male persons in employment	940.2	9.0	10.5	16.3
6	Female persons in employment	162.6	15.9	17.0	6.8
7	Male unemployed persons	808.0	9.3	10.1	9.6
8	Female unemployed persons	198.0	14.6	17.7	20.9
9	Main source of livelihood: support by relation	777.8	11.7	13.1	11.7
10	Main source of livelihood: pension	1 084.2	9.9	11.5	16.3
11	Income under 600 DM	285.1	14.0	14.3	1.9
12	Income from 1000 to under 1800 DM	1 197.5	8.5	10.4	21.3
13	Income from 1800 to under 3000 DM	1 131.6	8.2	9.2	12.7

1) (Col. 3 – Col. 2) / Col. 2 * 100

5 Estimation

The change into a continuous survey has a big effect on the estimation technique. We have not tackled the estimation topic yet. Here we only outline some problems we will have to deal with. There are some requirements we will have to pay attention to:

- Reliable annual estimates for regions are very important.
- Quarterly estimates should be consistent with annual estimates.
- The estimation program will not be run in the national office, but in every Bundesland office.

Our current estimation technique is characterised by poststratification to margins of the Current Population Statistics: 133 so-called regional adjustment cells in combination with sex and nationality (German/non German). In a continuous survey, we will have the month (possibly week or quarter) as an additional auxiliary variable and consequently many more cells in poststratification. To avoid the problem of small or empty cells, we intend to use a calibration technique, for instance a generalised regression estimator. Then we can adjust, for instance, to margins of the regional adjustment cells and separately to the common distribution of sex, nationality and month.

At present, first results are issued nine months after the reference week. There is enough time for reminders, and the nonresponse rate is only about 3% (the MZ is a mandatory survey). In a continuous survey, we will require the results twelve weeks after the end of the respective quarter. The reference week of a household which does not answer in time will be postponed. If the week is postponed to the next quarter, the household will probably be treated as not responding during the quarter it was selected for. A nonresponse rate in this sense will be clearly higher than 3%. On the other hand, the postponements from the previous quarter could be used to enlarge the number of eligible households of a certain quarter.

Jean Dumais / Philippe Bertrand / Barbara Christian*)

Sampling and Estimation Issues in the Re-Engineered Census of Population

Redesigning the programme of the population census, INSEE

1 Introduction

Over the last decades, a number of Occidental countries have reduced their public expenditures. In some of those countries, the national statistical agencies have growing difficulties to justify and raise the budget necessary to the complete enumeration of the population. Moreover, with the prevalence of administrative files, citizens fail to understand the continuing need for a decennial census. In France, the last census operations were conducted in 1968, 1975, 1982, 1990 and 1999.

In many countries, especially in northern Europe, population registers and interconnectivity of administrative files appear as a viable alternative to traditional census taking (Laihonen, 2000; Borchsenius, 2000]. Given these environmental changes, France has chosen to rethink the way census operations should be carried out. As population registers are legally impossible in France, and dwelling registers are hardly possible, census field work remains necessary.

The plan detailed here follows up on ideas developed by Kish (1981, 1990) and Horvitz (1986); a first plan was laid out by Deville and Jacod (1996). The basic ideas are the observation, each year, of a fraction of the population, using sampling methods and modelling of administrative information. As a consequence, costs and workload are spread over time, response burden is lightened, publication of data is annualized.

This redesign must be accomplished within a set of constraints:

- Data quality, as it pertains to reliability and precision, must be maintained; relevance should be improved with the annual statistical outputs; the proposed plan must be robust, in the statistical sense and in a more environmental sense also (resilient to weather hazards, labour disputes, etc.);
- Budget cannot be augmented; decentralisation of field, perhaps data grooming, operations, increased usage of administrative data, specialization of field staff should improve performance and help keep the operation within the budget that used to be allocated to a census;
- Confidentiality of individual information must be guaranteed, whatever the level of decentralization, and whatever transmission circuits are chosen;
- Acceptance of the census redesign by the local authorities is key to its success as the communes become one of the key players in the realisation of the census.

---

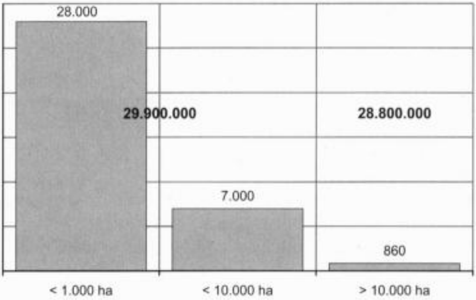
*) Jean Dumais, Philippe Bertrand, Barbara Christian, National Institute of Statistics and Economic Studies (INSEE), Lyon.

2 Sampling Strategy

The budget allocated to conducting a traditional census was enough to collect, process, analyse and disseminate the information gathered from the whole population (some 57 million people in 1999), and such budget was voted every 7 or 8 years. Hence, the budget for the annual campaign is set at 1/7 of a traditional census operation budget.

The next consideration is the time period over which the collection can be spread. Obviously a 10-year cycle (as in Switzerland, the UK or the USA) is not an option as it is longer than the previous 7- or 8-year cycle. A 5-year cycle, as recommended by the UN and used, for example, in Australia and Canada, appears adequate; a shorter cycle would prove inefficient or too costly.

Figure 1



Thus, the annual budget is 1/7 of 57 million individual's forms, that is 8,1 million individual's forms per year. Now, France comprises 860 large communes accounting for nearly half of the population, and some 35.000 mid-size and small communes accounting for the rest. In small and mid-size communes, organising a sample survey of dwellings and persons would be statistically and economically inefficient: a traditional census is the reasonable option. In larger communes, where the population is at least 10 000, a sample survey of dwellings and persons could be organised with some assurance of statistical quality. Still, given the importance of these large communes, a sample of the communes is not an option; they must all be included. With the 5-year cycle in mind, one comes to the following sampling and rotation scheme: every year, 1/5 of the small communes would be enumerated, and, working out the arithmetic, 40% of 1/5 of the population of the large communes. After the five years, every small commune would have been enumerated, and 40% of the population of the larger communes would have been surveyed.

The various fifths, called rotation groups, are built using equal probability "balanced sampling" (Deville, Tillé, 1999). These groups are built to maintain stability of age-sex groups estimates over time; correlated variables are also quite stable. The rotation groups comprise communes in the domain of "mid-size and small communes", and buildings in the domain of "large communes".

Figure 2
Total number of men; Rhône-Alpes, 1990

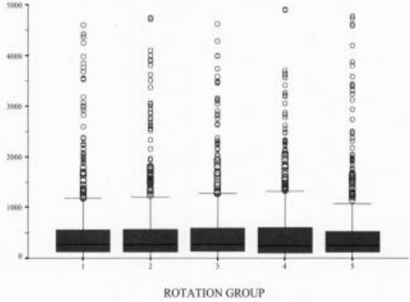
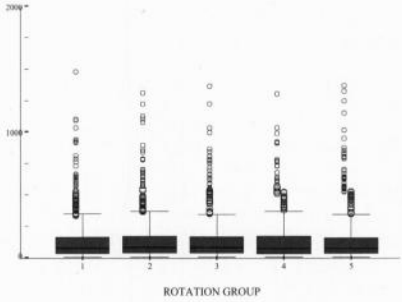


Figure 3
Households with 2 cars; Rhône-Alpes, 1990



The preceding box – and – whiskers plots show how interchangeable the rotation groups are. The variable "men" is not a balancing variable per se, it is the sum of 5 balancing variables, and thus should be quite well balanced over the rotation groups. On the other hands the variable "Households with 2 or more cars" is neither a balancing variable nor the combination of some of them. Still, it is quite remarkable how the groups are interchangeable.

Moreover, tests of stability over time were conducted with Rhône-Alpes. First, rotation groups were created for the small communes of Rhône-Alpes using census 1982 data; as seen in the table below, with perfect balance, the expected relative sampling error (CV) is zero. For estimation domains, like a "département" or an urban area, the relative error is that of a simple random sample of the same size.

Table 1

CV (%)	Rhône-Alpes, N = 2 815, E(n) = 563		Département Rhône, N = 274, E(n) = 55		Urban Area Grenoble, N = 25, E(n) = 5	
Number of persons...	census 82	census 90	census 82	census 90	census 82	census 90
0 – 19	0	1.3	16.0	15.9	45.9	45.1
20 – 39	0	1.1	16.4	16.0	46.3	45.6
40 – 59	0	0.9	15.7	15.7	43.8	44.4
60 – 74	0	0.5	15.3	15.2	45.2	44.0
75 – 95	0	0.8	16.1	16.2	50.9	47.3
Male	0	0.9	15.6	15.4	44.8	44.5
Female	0	0.8	15.9	15.7	45.1	44.5
Total	0	0.9	15.7	15.6	45.0	44.5
Foreigners	6.2	6.0	26.2	24.8	55.0	51.2
Migrants	1.5	2.3	17.2	16.7	45.9	45.0
Unemployed	2.2	2.2	18.7	18.4	47.5	47.9
Employed	0.5	1.1	15.7	15.5	45.1	44.9

Then, the expected precision was calculated for those rotation groups using 1990 census data; as shown above, the loss of precision is minimal after almost ten years.

Sampling in larger communes is slightly different. First, all larger communes are in sample every year. Then, two main options are currently under study. The first option supposes the sampling of individual dwellings from the list of dwellings compiled from a listing of the buildings selected in the rotation group. This allows for lesser clustering effect and fits within the allotted budget. The second option is to survey all the dwellings found in a sample of the buildings of a rotation group. Alternative ways of implementing either option are under study.

All these options make extensive use of an address register (AR), updated annually for construction and demolition.

Other two-stage strategies have been examined, but none found to be as precise as balanced sampling.

3 Data Collection

Let's consider the calendar of events for the year of reference *Y*. First, rotation groups of addresses have been formed and maintained to ensure maximum relevance and coverage. Address register maintenance is planned for the early Fall.

With a two-stage sample design in larger communes, listing would happen in November-December *Y*-1; the samples of swellings would be drawn in mid- or late December, in time to prepare interviewers allotments. Collection would happen in January and February *Y*, lasting 4 weeks in small communes and up to 6 weeks in larger communes. dwelling sample. The collection mode is drop-off – pick-up, as was done in previous census operations.

The census questionnaires are (at the time of printing) almost definitive and are quite similar to those used for Census 99. Of course, some modifications had to be made to the question wording so that the rotating scheme could be accounted for (especially for secondary residences and migration).

Plans to use the census collection as a vehicle for piggy-back surveys are being studied. Also, plans to use the census as a first stage screening process for supplementary surveys are considered. Thirdly, plans to use census annual samples as a master sample for the social and household survey programmes at INSEE will be drawn during the coming years.

4 Estimation

Two main series of estimates will derived form the census data collection. Firstly, an advanced series, called "global estimates", will be issued at the end of the reference year *Y* using solely data from the collection campaign *Y*. Secondly, the series of "detailed estimates" will be issued at the end of year *Y* for reference year *Y*-2.

The global estimates should be published at rather broad geographical (national and 22 regions) levels. AS well, census estimates for the smaller communes visited during *Y* should be released. The global estimates are typical Horvitz-Thompson expansion estimates. A measure of precision for these estimates will be made available, either through the tabulation system itself, or, more likely, as a set of look-up tables.

There are no plans to issue global estimates at the sub-stratum level as such estimates are deemed to have moderate or low statistical precision. For example, the Département Drôme (one of the 7 comprising the region Rhône-Alpes) is made up of 366 small communes and 4 large; thus the annual sample, drawn at the region level, will contain, by design, the 4 large communes and an average 72 small ones. The table below shows the relative error of some census estimates with one year of data; by contrast, if the 5-year cycle is completed, sampling error among the small communes is virtually removed, and the variation left is that of the second phase sample of dwellings in the large communes. Dramatic gains in precision can be made.

Table 2
Drôme, 1990 annual sample = 4 large, 72/366 Small

Variable	1990 Census Value	Relative sampling error	
		1Yr (CV%)	5 Yr (CV%)
Number of male	191 100	14.4	0.14
Number of persons 0 – 19	105 700	14.3	0.28
Number of persons 40 – 59	93 400	14.1	0.24
Number of married	182 600	14.6	0.15
Number of divorced	14 200	14.6	0.70
Number of in same dwelling in 82	197 400	14.4	0.20
Number of employed	150 900	14.6	0.16
Number of retired	70 900	16.0	0.27
Total Number of persons	391 000	14.5	0.11

The next table shows the similar situation for Romans and for the urban area (within Drôme) of Romans, which comprises 1 large and 4 small communes; obviously, the annual sample cannot guarantee that a small commune will always be in sample.

Table 3
Romans

Variable	*Romans Urban Area, 1 large and 0.8 small*			*Romans, stratified sample of 8% of the dwellings*		
	1990 Census Value	Relative sampling error		1990 Census Value	Relative sampling error	
		1 Yr (CV%)	5 Yr (CV%)		1 Yr (CV%)	5 Yr (CV%)
Number of male	22 560	47.8	0.57	14 800	3.4	0.9
Number of persons 0 – 19	12 770	48.0	1.14	7 900	5.9	1.8
Number of persons 40 – 59	11 050	45.2	0.93	7 600	4.1	1.5
Number of married	21 090	48.3	0.62	13 400	3.3	1.0
Number of divorced	2 120	47.5	2.27	1 500	8.9	3.0
Number of persons in same dwelling in 82	22 100	46.3	0.85	15 500	4.1	1.3
Number of employed	17 950	48.3	0.67	11 600	3.5	1.0
Number of retired	8 780	50.5	1.08	5 900	5.2	1.6
Total Number of persons	44 230	48.4	0.41	31 200	2.9	0.6

These detailed estimates will make use of many years of census data, administrative data and modelling techniques. Detailed estimates will be available for any geographical level (within the bounds of statistical confidentiality) at the end of year *Y* for reference year *Y*-2. The idea behind the detailed estimates is to impute what could have been observed in the rotated-out areas. To this end, the relationship between administrative files and census observations for similar objects will be exploited.

For example, let's consider computing detailed estimates for year *Y*-2; there are four rotation groups to impute.

Table 4

Rotation	...	*Y*-5		*Y*-4		*Y*-3		*Y*-2		*Y*-1		*Y*	
I			A_I^{Y-5}	C_I^{Y-4}	A_I^{Y-4}		A_I^{Y-3}	?	A_I^{Y-2}		A_V^{Y-1}		A_I^{Y}
II			A_{II}^{Y-5}		A_{II}^{Y-4}	C_{II}^{Y-3}	A_{II}^{Y-3}	?	A_{II}^{Y-2}		A_{II}^{Y-1}		A_{II}^{Y}
III			A_{III}^{Y-5}		A_{III}^{Y-4}		A_{III}^{Y-3}	C_{III}^{Y-2}	A_{III}^{Y-2}		A_{III}^{Y-1}		A_{III}^{Y}
IV			A_{IV}^{Y-5}		A_{IV}^{Y-4}		A_{IV}^{Y-3}	?	A_{IV}^{Y-2}	C_{IV}^{Y-1}	A_{IV}^{Y-1}		A_{IV}^{Y}
V		C_V^{Y-5}	A_V^{Y-5}		A_V^{Y-4}		A_V^{Y-3}	?	A_V^{Y-2}		A_V^{Y-1}	C_V^{Y}	A_V^{Y}

One imputed value for Group I could be obtained by "forecasting" the census value observed in *Y*-4 using the trend in administrative data for the years of interest, thus:

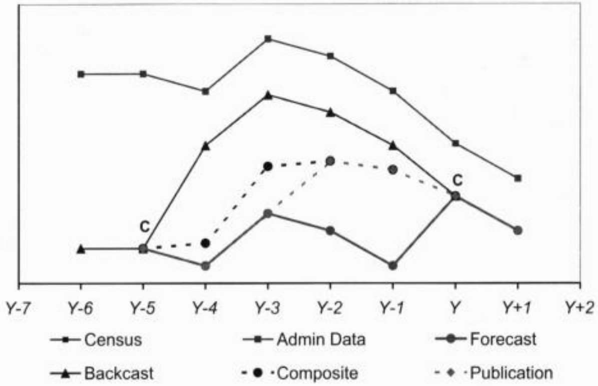
$$\tilde{C}_I^{Y-2} = C_I^{Y-4} \frac{A_I^{Y-2}}{A_I^{Y-4}}.$$

Such adjustments can be made for individuals, by sex – age groups, and for dwellings, depending on the variable of interest and the available administrative files, for all rotation groups. These adjustment factors will be stored on the data file, along with non-response adjustment factors, and estimation weights.

For Group V, the "ageing" factors can be derived forecasting from C_V^{Y-5} or backcasting from the more recent C_V^{Y}; actually, both will be used and composite adjustment factors will be derived. The precision of said detailed estimates will be derived using Särndal's (1990) imputation for non-response model.

Obviously, the detailed estimates should agree with the global estimates that were published two years earlier. Such agreement will be forced upon the detailed series, as can be seen in the following chart.

Figure 4
Data Series for Small Commune



5 On-going developments

A bill describing the conditions under which the census will be undertaken is currently (Summer 2001) before the legislative assembly. Following possible amendments and adoption of the bill, some modifications to the sampling strategy in larger communes may prove necessary to reflect the spirit of the law.

As work progresses on data grooming, non-response corrections, data file management, etc., modifications to the estimation strategy may become necessary. As well, work must be done to foresee eventual shortcomings of administrative files. A way to mass impute rotated-out groups without complete coverage of the administrative source must be worked out.

More work is needed on efficient presentation and usage of measures of precision for both global and detailed estimates.

References

Borchsenius, L. (2000): From a Conventional to a Register-based Census of Population, Les Recensements après 2001, Séminaire Eurostat,-INSEE, Paris.

Deville, J. C.; Tillé, Y. (1999): Balanced Sampling by Means of the Cube Method, CREST-ENSAI, document interne, soumis pour publication.

Dumais, J. (2000): Quelques aspects méthodologiques du recensement rénové en France, 2è Colloque francophone sur les sondages, Bruxelles, juin 2000 (to appear).

Horvitz,D. G. (1986): Statement to the Subcommittee on Census and Population, Committee on Post Office and Civil Service, House of Representatives, Research Triangle Park, North Carolina.

Jacod, M.; Deville J. C. (1996): Replacing the Traditional French Census by a Large Scale Continuous Population Survey, in: Annual Research Conference Proceedings, USBC, Washington.

Kish, L. (1981): Population Counts from Cumulated Samples, Congressional Research Service, Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau, Prepared for the Subcommittee on Census and Population, Committee on Post Office and Civil Service, House of Representatives, Washington.

Kish, L. (1990): Recensement par étapes et échantillons avec renouvellement complet, in: Techniques d'enquêtes, Vol 16, N° 1, pp. 67 – 86, Statistique Canada, Ottawa.

Laihonen, A. (2000): 2001 Round Population Censuses in Europe, Les Recensements après 2001, Séminaire Eurostat,-INSEE, Paris.

UNO (1990): Principes et recommandations complémentaires concernant les recensements de la population et de l'habitat, Etudes statistiques, ST/ESA/STA/sérieM/67, New York.

Särndal, C. E. (1990): Méthodes pour estimer la précision des estimations d'enquête lorsqu'il y a eu imputation, Recueil du Symposium 90 de Statistique Canada : Mesure et amélioration de la qualité des données, Ottawa, octobre 1990, pp. 369 – 380.

Särndal, C. E.; Swensson, B.; Wretman, J. (1992): Model Assisted Survey Sampling, New York.

Teilnehmerverzeichnis

A

Aasness, Dr. Jørgen; *Statistics Norway, Oslo*

Arnež, Marta; *Statistical Office of the Republic of Slovenia, Ljubljana*

B

Becker, Dr. Irene; *Johann Wolfgang Goethe-University, Frankfurt*

Berke, Paul; *State Office for Statistics and Data Processing North Rhine - Westphalia, Düsseldorf*

Bihler, Wolf; *Federal Statistical Office of Germany, Wiesbaden*

Brinner, Dr. Karin; *University of Rostock*

Buck, Ute; *Federal Ministry of Labour and Social Affaires, Bonn*

C

Clemenceau, Anne; *Statistical Office of the European Communities (SOEC), Luxembourg*

Cornali, Anne; *Swiss Federal Statistical Office, Neuchâtel*

Correia, Dr. Luis Paulo; *National Statistical Institute of Portugal, Lisboa*

D

Deckl, Silvia; *Federal Statistical Office of Germany, Bonn*

Dumais, Jean; *National Institute for Statistics and Economic Studies (INSEE), Lyon*

E

Ehling, Dr. Manfred; *Federal Statistical Office of Germany, Wiesbaden*

Ernst, Nicole; *University of Trier*

Essig, Lothar; *University of Mannheim*

F

Fuhrmann, Anja; *University of Trier*

G

Girndt, Tanja; *Federal Ministry of Labour and Social Affaires, Bonn*

Göbel, Jan; *German Institute for Economic Research (DIW), Berlin*

Graf, Dr. Monique; *Swiss Federal Statistical Office, Neuchatel*

H

Haslinger, Alois; *Statistics Austria, Wien*

Hawliczek, Ingo; *State Office for Statistics of Rheinland-Pfalz, Bad Ems*

Holz, Erlend; *Federal Statistical Office of Germany, Wiesbaden*

Horneffer, Birgit; *Federal Statistical Office of Germany, Wiesbaden*

J

Jacob, Dr. Rüdiger; *University of Trier*

K

Kordos, Prof. Dr. Jan; *Central Statistical Office, Warszawa*

Kowalczyk, Barbara; *Warsaw School of Economics, Warszawa*

Krug, Prof. Dr. Walter; *University of Trier*

L

Langers, Jean; *National Statistical Institute of Luxembourg, Luxembourg*

Lindqvist, Markku; *Statistics Finland, Helsinki*

Linz, Stefan; *Federal Statistical Office of Germany, Wiesbaden*

M

MacFeely, Steve; *Central Statistics Office of Ireland, Cork*

Mejer, Lene; *Statistical Office of the European Communities (SOEC), Luxembourg*

Missong, Dr. Martin; *University of Kiel*

Møller, Bo; *Statistics Denmark, Kopenhagen*

P

Pfister, Dr. Martin; *Infratest Burke, Munich*

R

Reeh, Dr. Klaus; *Statistical Office of the European Communities (SOEC), Luxembourg*

Rehm, Dr. Norbert; *University of Trier*

Reich, Prof. Dr. Utz-Peter; *University of Applied Sciences Mainz*

Rendtel, Prof. Dr. Ulrich; *Johann Wolfgang Goethe-University, Frankfurt*

Ritz, Daniel; *Federal Statistical Office of Germany, Wiesbaden*

Rompel, Heinz-Kurt; *Statistical Office of the State of Hesse, Wiesbaden*

S

Schadendorf, Dr. Felix; *Federal Ministry of Labour and Social Affairs, Bonn*

Schmerbach, Dr. Sibylle; *Humboldt-University, Berlin*

Schmidt, RD Jürgen; *Federal Statistical Office of Germany, Wiesbaden*

Schupp, Dr. Jürgen; *German Institute for Economic Research (DIW), Berlin*

Schwill, Rilana; *University of Trier*

Selén, Dr. Jan; *Statistics Sweden, Stockholm*

Sonnemann, Prof. Dr. Eckart; *University of Trier*

T

Thill, Jean; *Statistical Office of the European Communities (SOEC), Luxembourg*

V

Verma, Prof. Vijay; *ORC Macro International Social Research, London*

von der Heyde, Christian; *Infratest Burke, München*

W

Wagenhals, Prof. Dr. Gerhard; *University of Hohenheim, Stuttgart*

Walter, ORR Mario; *Bavarian State Office for Statistics and Data Processing, Munich*

Weins, Cornelia; *University of Trier*

Wiedenbeck, Michael; *Center for Survey, Research and Methodology (ZUMA), Mannheim*

Wiegert, Dr. Rolf; *University of Tübingen / IAW Tübingen, Mössingen*

Wirtz, Christine; *Statistical Office of the European Communities (SOEC), Luxembourg*