22 November 2019

Economic Commission for Europe

Conference of European Statisticians

Expert meeting on measuring poverty and inequality: SDGs 1 and 10

Geneva, Switzerland
5–6 December 2019
Item C of the provisional agenda
Improving response rate and sampling precision in surveys

Assessing and Improving Survey Methods

Note by the University of Siena [*]

1. Survey errors and quality

To establish trust in poverty measurement and prevent misguided policies, Statistical Offices have to regularly assess and continuously improve the quality of their processes and accuracy of their data. Quality reports which describe the quality criteria and explain any instances in which these criteria could not meet, or statistical concepts could not be correctly applied will not only assist the correct interpretation but can also provide the basis for future improvements. The World Bank's Report of the Commission on Global Poverty recommended that "The World Bank should make public the principles according to which household survey data are selected for use in the global poverty count; and there should be an assessment at national level of the availability and quality of the required household survey data…" The Commission suggested in particular to investigate potential survey underrepresentation and noncoverage. It also recommended that "…poverty estimates should be based on a 'total error' approach, evaluating the possible sources and magnitude of error…" (World Bank 2017 recommendations number 6, 3 and 5 on pages 33, 50 and 59). Certain errors are especially relevant for disaggregation.

1.1 A typology of survey errors

It is common to broadly separate sampling and non-sampling errors (Assael and Keon, 1982, Kalton, 1983; Kalton *et al*., 1986). For the assessment of poverty measurement however, Verma *et al.* (2010) applied a modified framework (Verma 1981, Hussmanns *et al.* 1990). Much of this section is drawn directly from Verma *et al.* (2010) with the permission of the authors.

[*] Prepared by Gianni Betti.

Three broad categories of errors should be distinguished:

(a) Errors in measurement

What is measured on the statistical units enumerated in the survey can be different from the actual (true) values for those units. These errors concern the accuracy of the *substantive content of the survey*: the definition of the survey objectives and questions; the ability and willingness of the respondent to provide the information sought; and the quality of data collection, recording and processing. A typical example for error in measurement would be underreporting of certain income components. This will not only increase uncertainties but can possibly also lead to significant bias in the estimates.

(b) Errors in estimation

The process of extrapolation from individual measurements to the entire study population adds further uncertainties. These result from *sample design and implementation*, notably coverage, sample selection and implementation, and also sampling errors and estimation bias.

(c) Item non-response

For poverty measurement, Verma *et al.* (2010) especially highlighted *item non-response* as special, mixed category that complements the common distinction between representation and measurement errors (Groves et al 2004). Item non-response is particularly important in surveys which collect detailed information on components of household and personal income. It is generated in the process of measurement but in its effect it adds to the existing non-response and thus also amounts to an error of estimation (that may be mitigated by estimation tools such as imputation).

Quality reports should describe these broad categories of error in sufficient detail.

1.1.1 Errors in measurement

It is useful to distinguish conceptual, response ('data collection') and processing errors. *Conceptual errors* concern the scope, concepts, definitions and classifications adopted in relation to the survey objectives. It is almost impossible to compensate conceptual errors. *Response errors* concern the process of data collection while *processing errors* concern the subsequent process of transforming the information into a micro database. They result from different survey operations but their effects are similar. Each type of error may further be decomposed into bias and variance components. These distinctions are useful in so far as the components differ in nature and in methods of assessment and control.

Measurement bias

Bias arises from shortcomings which affect the whole survey operation: basic conceptual errors in defining and implementing the survey content; incorrect instructions for interviewers; errors in the coding frame or programs for processing the data; etc. Some errors arise from inherent difficulties in collecting certain types of information given the general social situation and the type of respondents involved. The first step in identifying bias is through logical and substantive analysis of the internal consistency of the data. Beyond that, the assessment requires comparison with more accurate information: data from external sources or data collected with special, improved methods. When the same collection and processing tools are used for the whole population, most sources of measurement bias will be present across the whole population.

Measurement bias which is group-specific can systematically change disaggregated estimates. This will often be related to language and culture of the groups concerned. It matters for example if translations of a questionnaire are available and if the terms which are used have equal meaning. It is a good strategy to ensure equivalence by group translation and participation of representatives of the groups concerned in the questionnaire design (see Chapter 3). Other sources of measurement bias are directly related to the resource measures considered in poverty measurement. (See, for example, the discussion of different measures of cost of living in Chapter 5.)

Measurement variance

Different interviewers (coder etc.) often have a unique influence on measurements due to lack of uniformity and standardisation which can give rise to *correlated response variance.* By contrast, *simple response variance* is random, not correlated with any particular interviewer. Instability of particular items may indicate problems in the questionnaire's wording (e.g ambiguous terms). Its measurement requires comparisons between independent repetitions of the survey under the same general conditions. There is no way, in a single survey, to distinguish between variation among the true values of units (which contribute to the sampling error), and the additional variability arising from random factors affecting individual responses. Measurement variance contributes to the uncertainty of estimates and can therefore affect the robustness of disaggregated results.

1.1.2 Errors in estimation

Coverage and related errors

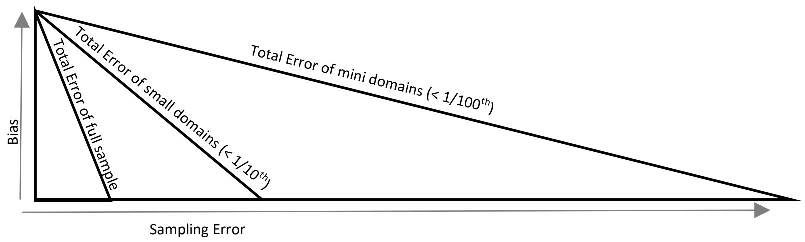
Coverage errors arise from discrepancies between the target and the frame populations, and also from errors in the way the sample is selected from the frame. Valid inference is only ensured by probability samples which meet all of these criteria: (a) the survey population is fully and correctly represented in the sampling frame; (b) units from the frame are randomly selected into the sample with known non-zero probabilities for all units; (c) all the units selected into the sample are successfully enumerated. Coverage error concerns primarily (a), but also (b); (c) concerns non-response. Poverty measurement must take any effort to ensure that all vulnerable groups are adequately covered (See also Chapter 3).

Sampling error

Sampling error is a measure of the variability that would be observed between estimates from different samples drawn using the same sample design as the survey, disregarding any variable errors and biases resulting from the process of measurement and sample implementation. Sampling error represents only one component of the total survey error. For estimates based on small samples, this component is often the dominant one. In other situations, non-sampling errors, in particular coverage, non-response and measurement biases, may be much more important. However, even in these cases, sampling error increases progressively as the estimates are produced for smaller and smaller subgroups of the population, such as for social classes or regions of a country. Therefore, when producing disaggregated poverty statistics, sampling error may well outweigh non-sampling errors.

The relative importance of sampling errors for disaggregation is portrayed by the right-angled triangles in Figure 1. Total error is often denoted as the root mean squared error and defined by taking the square root of variance and squared bias. It can thus be represented by the hypotenuses of a right-angled triangle (Kish, 1988). The smallest triangle of this figure depicts large survey samples where total error is dominated by the bias. In such situations the precision of poverty measures cannot be much increased by increases of sample size. Instead it is worthwhile to focus on reducing non-sampling errors. For subpopulations below the national level, the magnitude of bias is, however, often very similar, whereas the sampling error drastically gains in relative importance. Following the classification of Purcell and Kish (1980), groups which comprise less than 1/10th of the population may be considered as minor domains. The standard error for estimates for such domains is more than 3 times higher than for the full sample. For these domains, sampling error will overtake bias in many surveys. For mini domains, which Purcell and Kish categorised as groups which comprise between 1/100th and 1/10000th of the population, sampling error will often be the dominant factor and controlling the bias may not substantially improve total error.

Figure 1. Increasing Sampling Error/Bias Ratios for disaggregations


Total Error of mini domains (< 1/100th)
Total Error of small domains (< 1/10th)
Total Error of full sample
Bias
Sampling Error

1.1.3 Non-response errors

Non-response refers to the partial or complete failure to obtain a measurement on one or more study variables for one or more sample units. More specifically, unit non-response is a type of non-response occuring when no data are collected about a population unit designated for data collection. Item non-response means that a unit is included but information on some items for it is missed. In this typology, item non-response is in an intermediate category between errors in measurement and errors in estimation, whereas unit non-response is considered an error in estimation.

Non-response of both types causes an increase in variance due to decreased effective sample size and due to weighting and imputation introduced to control its impact. More importantly, it causes bias in so far as non-respondents are selective with respect to the characteristic being measured. For instance, one might expect persons with high incomes to be more reluctant to give information on their income; similarly, poorer, unemployed and socially excluded persons are more likely to be missed in surveys related to economic well-being.

Proposed solutions for item non-response will be presented in Section 2.3 on imputing missing values, while coverage and related errors and unit non-response will be treated in Section 2.4 with the introduction of weighting systems.

Framework to assess errors in poverty measurements (Verma et al. 2010)

Errors in measurement

A) conceptual errors; these include: i) errors in basic concepts, definitions and classifications; ii) errors in putting them into practice (questionnaire design, preparation of survey manuals, training and supervision of interviewers and other survey workers).

B) response (or 'data collection') errors; these include: i) response bias; ii) simple response variance; iii) correlated response variance.

C) processing errors; these include: i) recording, data entry and coding errors; ii) editing errors; iii) errors in constructing target variables; iv) other programming errors.

Mixed category

D) item non-response; this includes: i) only approximate or partial information sought in the survey; ii) respondents unable to provide the information sought ('don't knows'); iii) respondents not willing to provide the information ('refusals'); iv) information suppressed (for confidentiality or whatever reason).

Errors in estimation

E) coverage and related errors; these include: i) under-coverage; ii) over-coverage; iii) sample selection errors

F) unit non-response; this includes: i) unit not found or inaccessible; ii) not-at-home; iii) unable to respond; iv) refusal (potentially 'convertible'); v) 'hard core' refusal

G) sampling errors; these include: i) sampling variance; ii) estimation bias

Recalling the classical classification into sampling and non-sampling errors, the latter category is comprised of errors of types A) to F) above.

1.2 Multidimensional quality frameworks

Quality reports are essential tools to assess, improve and communicate the quality of poverty measurement. As a minimum, such reports should describe in sufficient detail all sources of error that limit the accuracy of poverty measures. Overall however, quality should be more broadly defined in terms of user needs, as “fitness to use” for the purpose for which the data were created (Juran and Gryna, 1970).

1.2.1 Questions which should be addressed to evaluate the utility of a survey

Table 1 reports an illustration of overlapping concepts and categories used by different organisations to identify dimensions of quality (taken from Lee and Shon, 2001) For the European Union, the legally required content of quality reports addresses all of these questions which are also reflected in Article 12 of the European Union Statistics Act (Regulation (EC) No 223/2009).

Table 1. Concepts and categories used by different organisations to identify dimensions of quality

Canada	Netherlands	R. of Korea	IMF	Eurostat
			Prerequisites of quality	
Relevance	Relevant	Relevance		Relevance
Accuracy	Accurate	Accuracy	Accuracy and reliability	Accuracy
Timeliness	Timely	Timeliness	Serviceability	Timeliness and Punctuality
Accessibility		Accessibility	Accessibility	Accessibility and clarity
Coherence				Coherence
		Comparability	Methodological soundness	Comparability
Interpretability				
			Integrity	Completeness
	Cost-effectively	Efficency		
	Without too much a burden			

Source: Lee and Shon (2001).

European Union: Content of Quality Reports as Required by the Law

In October 2019 the European Union established a new framework regulation ((EU) 2019/1700) which integrates all major social surveys in the European Statistical System including EU-SILC. Member States are thus legally required to meet specified quality criteria and produce regular quality reports. The content of these reports is specified in an implementing regulation which lists the required information as follows:

1. CONTACTS
2. STATISTICAL PRESENTATION
    - 2.1 Data description
    - 2.2 Classifications
    - 2.3 Sector coverage (main themes)
    - 2.4 Statistical concepts and definitions (including the reference period)
    - 2.5 Statistical units
    - 2.6 Statistical population
    - 2.7 Population(s) not covered
    - 2.8 Reference area
    - 2.9 Time coverage
3. STATISTICAL PROCESSING

3.1 Source data (e.g. interviews, administrative data)
3.2 Sampling frame
3.3 Sample design
3.4 Frequency of data collection
3.5 Data collection (mode such as CAPI, CAWI, CATI, etc., translated questionnaires)
3.6 Data validation (including explanation how it is reflected in the results).
3.7 Data compilation (e.g. data editing, imputation, weighting etc.)
4. QUALITY MANAGEMENT
4.1 Quality assurance (e.g. EFQM, ISO 9000)
4.2 Quality assessment (main strengths, trade-offs and deficiencies)
5. RELEVANCE
5.1 User needs
5.2 User satisfaction
5.3 Completeness (variables which are not transmitted)
6. ACCURACY AND RELIABILITY
6.1 Overall accuracy (esp. effect of random and systematic errors for key estimates).
6.2 Sampling error (methodology, national and regional standard errors for indicators)
6.3 Non-sampling error[1]
6.4 Seasonal adjustment (where applicable)
6.5 Data revision (policy and practice)
7. TIMELINESS AND PUNCTUALITY (dates of dissemination and end of fieldwork)
8. COHERENCE AND COMPARABILITY
8.1 Comparability – geographical
8.2 Comparability – over time
8.3 Coherence – cross domain
8.4 Coherence – National accounts
8.5 Coherence – internal
9. ACCESSIBILITY AND CLARITY (dissemination formats, documentation)
10. COST AND BURDEN (cost of collection and production, duration of interviews)
11. CONFIDENTIALITY (policy, data treatment)
12. COMMENT (Supplementary descriptive text that can be included in the quality report)

i) How relevant is the data?

Relevance refers to the capacity of the data to meet users' needs. It implies the identification of users and their needs, and assessment of the extent to which their needs are actually met. The concept also covers the *potential* of the data in meeting the relevant needs. According to Statistics Canada's Survey Methods and Practice (2003): "… Assessing relevance is a subjective matter dependent upon the varying needs of users. The statistical agency's challenge is to weigh and balance the conflicting needs of current and potential users to produce a program that goes as far as possible in satisfying the most important needs within given resource constraints". Relevance also depends on the extent to which stakeholders and social groups which are considered for disaggregation were involved.

ii) How timely and punctual are results available?

As defined by Statistics Canada's Survey Methods and Practice (2003): "The *timeliness* of statistical information refers to the delay between the reference point (or the end of the reference period) to which

[1] Including: coverage of sub-populations; efforts made to limit measurement error in questionnaire design and testing; interviewer training; proxy interview rates; available characteristics of non-respondents; unit and item non-response rates; substitution rates; gross sample size (initial sample size); number of eligible units and net sample size, including substitution units (achieved sample size); checks; imputations.

the information pertains, and the date on which the information becomes available. It is typically involved in a trade-off against *accuracy (see below)*. The *timeliness* of information will influence its *relevance*."

It is important to note that the requirements of timeliness can conflict with those of accessibility and clarity, and above all with those of accuracy. At a minimum, the data must be checked and corrected to a high standard before their public release. Obviously, releasing data or results without adequate editing and correction can be misleading and wasteful. *It can also damage the credibility of the producer organisation*. For instance, Fellegi (2001) identifies credibility as a 'survival' issue for a statistical organisation.

Punctuality refers to adherence to a pre-established time schedule for the release of statistics. Timeliness is a more objective criterion, assessing how fresh are the data and whether they became available when most needed. Punctuality acquires increased importance in the EU-wide context. The so-called European semester is an annual coordination process for policies in EU Member States. The degree to which important fiscal and economic decisions may take social conditions into account, depends also on the timely availability of indicators for many countries simultaneously. The requirements of punctuality have been expressed very strongly in EU-SILC regulations.

iii) How precise are the results?

Data accuracy includes the assessment of survey errors which were discussed in the previous section. It is of such fundamental importance that it has been customary in survey practice to focus on accuracy, sometimes at the expense of – or even to the exclusion of – other dimensions of quality.

According to Statistics Canada (2003): "The *accuracy* of statistical information is the degree to which the information correctly describes the phenomena it was designed to measure. It is usually characterized in terms of error in statistical estimates and is traditionally decomposed into bias (systematic error) and variance (random error) components. It may also be described in terms of the major sources of error that potentially cause inaccuracy (e.g., sampling, coverage, measurement, nonresponse, processing)."

Ideally all indicators should be published with an indication of their accuracy, including whether there are conceptual differences with regard to international standards. In practice, sometimes sampling errors are presented for selected main indicators only. As an absolute minimum this should include an indication of the design effect, which is calculated as the variance of the main poverty estimate divided by the variance of the same indicator if simple random sampling had been used. For disaggregation of poverty measures it is especially important to identify clearly those results which provide only limited accuracy. Indicators with inacceptable inaccuracy should never be published. More information on best practices for dissemination are discussed in section 2.6.

iv) To what degree are findings comparable?

Comparability is increasingly considered a central requirement of data quality, especially for measures of poverty. Partnerships for development in the context of the 2030 Agenda for Sustainable Development require comparable measures for poverty. To improve the international comparability and availability of statistics on poverty and the related metadata, the Conference of European Statisticians (CES) established a Task Force in 2014, which worked through 2015 and 2016 to develop a Guide on Poverty Measurement (UNECE, 2017). This guide states that "… Many international organizations—the World Bank, OECD, UNDP, Eurostat, just to mention a few—produce poverty data. There have been continuous efforts to improve capacity in statistical offices to develop poverty measures in line with international standards. However, in most cases, these data are not comparable and often cover only a limited number of countries. A lack of comparable data across countries and time impedes effective policy actions. Data produced by countries are not always comparable internationally, largely for two main reasons: i) Country data primarily respond to national needs, which do not always correspond to international standards; and ii) Country data reflect national statistical capacities, which are not always able to meet international standards".

v) How coherent is the data with other statistics and over time?

According to Statistics Canada (2003): "The *coherence* of statistical information reflects the degree to which it can be successfully brought together with other statistical information within a broad analytic framework and over time. The use of standard concepts, classifications and target populations promotes coherence, as does the use of common methodology across surveys. Coherence does not necessarily imply full numerical consistency."

Coherence does not necessary mean identity: often there are genuine and inherent differences in the information coming from sources of different types. What it means is whether different sources together lead to a consistent picture, with each making a contribution towards the development of the picture. In the case of surveys on income and poverty, the most relevant sources for external comparison include national household budget and labour force surveys, national accounts, and various administrative and other sources depending on the country. In the European Union all countries conduct EU-SILC while some countries have also well-established official or academic surveys on the same topic. Quality reports should bring to the attention of users incoherences with external information. Although in practice it will often not be possible to say which source may be more accurate it is important for users to be informed about differences and their possible explanations.

In a panel survey and, in fact, in any continuing survey, coherence over time is also a fundamental requirement. Only under this condition can we study trends, aggregate data over time, or construct micro-level longitudinal measures.

vi) Are data accessible and clear?

These aspects refer to the extent to which the statistical data are available in the form and under conditions which meet users' requirements, and to how well the data are described and documented for the purpose. Conditions of availability include a whole range of factors such as restrictions on who can or cannot get access to the data, what items of information are suppressed, what restrictions apply on the conditions and purposes of data use, and also the difficulties, delays and the costs involved in gaining access to the micro data and timeliness of accompanying quality reports.

vii) What else would users need to know?

A variety of other aspects are also covered in the various data quality frameworks. Some of these overlap – different terms indicating more or less the same thing, perhaps from a somewhat different point of view or with a somewhat different emphasis. We have, for instance, 'completeness' in Eurostat terminology, 'integrity' in that of IMF and, along a different line of thought, 'interpretability' at Statistics Canada. Cost efficiency and minimisation of respondent burden are other aspects included as quality dimensions, especially in national frameworks (The Netherlands, South Korea in Table 1). Surprisingly, not all frameworks explicitly refer to 'comparability' as a dimension.

1.2.2 Relationship between different aspects of quality

It is safe to assume that no statistical agency is capable of meeting all of the above criteria to the same degree. To a certain extent, the *different dimensions of data quality compete against each other*, an obvious example being the common conflict between timeliness and data accuracy – 'quickly released but rough data, versus refined data but much delayed'. Different aspects of data quality can also *mutually support and reinforce each other*, one often forming a precondition for the other. For instance, it is hardly possible for two data sets to be comparable, when either or both lack statistical accuracy. Perhaps most critically, a survey loses its relevance if it is not timely and accurate enough.

Verma (1981) and later Verma *et al.* (2010) propose to think about reduction in data quality in any dimension as a loss in the utility of the information. The loss may be more or less steep depending on the particular context. Often the resources saved by reducing quality in one dimension can be used to improve quality in other dimensions; however, some dimensions can also be linked in such a way that a quality loss in one dimension necessarily implies a loss in the other as well. Beyond a certain point, there is likely to be a critical zone when further reduction in quality along a particular dimension would result in increasing drastically the loss in the overall utility of the data. A certain minimum degree of quality has to be present in *every* dimension for the statistical information to remain useful overall.

2. Improving Quality in each Survey Step

Once the quality profile of the survey is understood, action for methodological improvements can be taken. Improvements may be considered in each survey step, including design, data collection, edit and imputation, weighting, variance estimation and dissemination (Groves *et al.* 2004).

2.1 Survey design

Some survey designs lead to more accurate estimate or can produce more disaggregated statistics. However, global survey design decisions must balance these advantages with costs and other dimensions of quality such as timeliness and comparability.

Official measures of poverty that are based on sample surveys must use a probabilistic sample. Beyond this crucial requirement, the following design issues are particularly relevant for disaggregated poverty statistics. Issues specific to hard-to-reach groups are also discussed in Chapter 3.

2.1.1 Choice of data source

The form of data collection can have serious consequences for disaggregation. Income questions can be difficult to answer. When respondents provided income data directly, they will typically provide rounded numbers and may forget to include certain types of income. For instance, a respondent may remember their employment income but forget about or not be able to provide details about investment income.

As an alternative, in many countries, income data is taken from administrative sources, such as files created for the administration of income taxes or government programs, or from registers. This reduces the burden imposed on respondents and this data is often more accurate than respondent data as it is less prone to rounding and recall error. Moreover, administrative tax data may also be classified into more detailed income source categories, depending on the categories that are used for taxation, allowing for the production of statistics that are more disaggregated.

This is not to say that measurement errors are non-existent for administrative data. Conceptual errors are an important consideration since the categorization used on the administrative data may not align with the concepts that desired for the income survey. In particular, non-taxable or undeclared types of income may not be found on administrative sources but should be included in income statistics to give a complete picture of an individual's income. Processing errors must always be considered, including for data acquired from a source not controlled by the survey team. Also, the definition of household membership in registers may be quite different from what it is in reality.

Country Case. Use of Administrative Tax Data at Statistics Canada

At Statistics Canada, administrative tax data has been the primary source of income data for about 20 years. Using this source has numerous advantages. The Statistics Canada's surveys used to measure income, spending, and consumption all link to the same administrative tax data and process this data in a harmonized way. Additionally, the Canadian Census also uses the same administrative tax sources. This results increased coherence between these surveys. Using the administrative data led to more precise measurement of the various sources of income, while decreasing the burden place on respondents.

On the Canadian Income Survey, extra questions are asked to respondents to measure income concepts that are not included on tax form such as full amounts of spousal support (alimony) and other transfers between households. Collection is also required to obtain variables of interest for disaggregation of poverty statistics such as family composition.

Perhaps most importantly, the timeliness of administrative data can be a major drawback to their use for statistical purposes. Since tax data needs to be collected and processed by the tax agency before being provided the statistical agency, it can take quite a while before it is available. This is a classic example of the conflict between timeliness and accuracy.

The choice of data source will have a large impact on what is measurable by the survey. For instance, where administrative data includes details of government transfers such as credits given to families with children, the use of such data can be used to measure the impact of the programs on poverty which may then be different for different types of families. Details about these types of programs could be difficult or even impossible to obtain directly from survey participates directly as they may not be aware of the details of these transfer or even that a particular program affects them. Appropriate documentation and metadata should be made available and disseminated in order to make the users aware of the specificities of the data coming from countries making use of administrative and register data.

Country Case. Use of Registers in the European Union

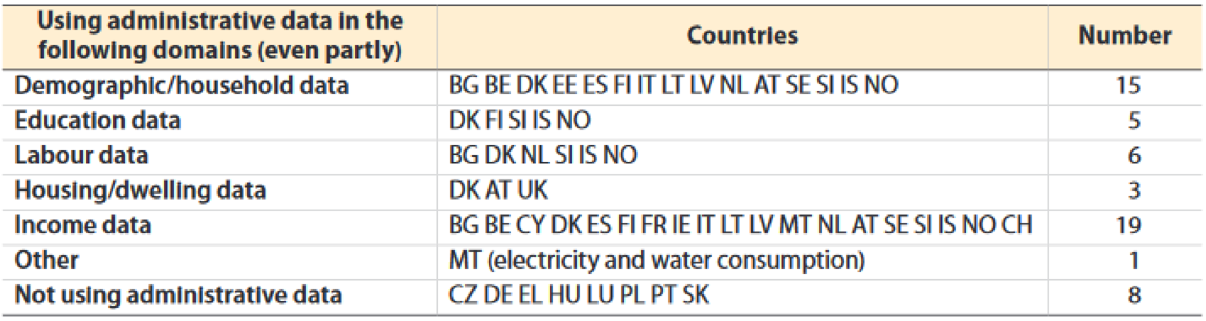
For the European Union the importance of registers for poverty measures is discussed in some detail in a volume by Jäntti, Törmälehto and Marlier (2013).[2] For the year 2012 it was found that 19 out of 28 EU Member States used registers to obtain income information in the year 2012 (Di Meglio and Montaigne, 2013). Apart from the traditional use of registers in the Nordic countries, several countries have seen transitions from interview based data collection towards use of registers or were planning to do so in 2012. The experience made by these countries appears particularly valuable for countries that intend to use more register information: FR, IT, LV, CH, IE, AT and ES.[3] Overall, was found good practice to assess carefully the impact and have at least one overlapping measurement from both interviews and registers. For example, in Austria register data was introduced in the year 2011. This lead to a decrease in poverty rates by about 2 percentage points for Austria. As register data was accessible also for previous years it was also possible to backcast to earlier waves of the survey including the year 2008.[4]

Table 2. Use of administrative data and registers for each domain covered by EU-SILC (2012)

[2] https://ec.europa.eu/eurostat/documents/3888793/5856365/KS-TC-13-004-EN.PDF

[3] https://ec.europa.eu/eurostat/documents/3888793/5856365/KS-TC-13-004-EN.PDF

[4] https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2017/CES_31-Poverty_measurement_lessons_learnt__Austria_.pdf



Using administrative data in the following domains (even partly)	Countries	Number
Demographic/household data	BG BE DK EE ES FI IT LT LV NL AT SE SI IS NO	15
Education data	DK FI SI IS NO	5
Labour data	BG DK NL SI IS NO	6
Housing/dwelling data	DK AT UK	3
Income data	BG BE CY DK ES FI FR IE IT LT LV MT NL AT SE SI IS NO CH	19
Other	MT (electricity and water consumption)	1
Not using administrative data	CZ DE EL HU LU PL PT SK	8

Source: Di Meglio and Montaigne (2013).

Register-based households are not always composed of the same persons as the survey-based household. In Finland for example, about 10 percent of households had a different composition according to the register definition than to the information than in the conventional housekeeping concept. This discrepancy was found to be only 5% among retired persons and 30% among student households (Jäntti and Törmälehto, 2013). Particular population groups may hence be inappropriately reflected in register data.

Out of the 19 countries that use income information from registers, 9 countries mentioned concerns about the coverage of all groups in the population. For example, in Austria the personal identification number required to link survey information with registers is missing for up to 5% of the sample persons. The information was more often missing for certain groups: Persons under age 40, persons living in the capital, persons with a foreign citizenship. Because register information is often generated in the context of employment, missing PINs were also more frequent among jobless persons and persons who are mainly fulfilling domestic tasks (Heuberger, Glaser and Kafka, 2013). This coverage error of units is potentially an important obstacle to using register data when poverty measures should be disaggregated. If this register under-coverage is known before the fieldwork, survey questions should be adjusted to collect specifically the information on groups which are otherwise missing. For example, for lone parents, transfers between households are often the main source of income. This component of income is usually not included in registers. To get appropriate measures for this group and their children, it is inevitable to put specific questions on alimonies.

When administrative data is used, survey data obtained directly from respondents is generally still required to obtain variables for disaggregation. In addition, valid information on certain sources of income such as transfers between households or self-employed income, capital income or hidden economy may only be collected through survey questions. As a result, record linkage is often required to combine the survey and administrative data.

Country Case. Record Linkage through Statistical Identifiers in Austria

In Austria, administrative data which are provided by various authorities are entered into registers with a unique personal identifier for statistical purposes. This unique identifier is provided also for each sample unit so that register information can be linked for almost every unit. For privacy concerns it is guaranteed that those identifiers cannot be matched to administrative data held by other authorities. For example, it would be impossible to report back respondent's answers to tax authorities. Also, identifiers have to be kept separate from names and addresses which are used only during fieldwork.

Register households are the sampling unit, but household membership is always verified by face to face interviewers. Individuals who are not living at the selected address are added and registered individuals not living at this address are not further considered. This sometimes involves cumbersome enquiries to obtain linkable person identification numbers (PIN). This practice also dictates the mode of data collection in the initial contact, leaving telephone or web interviews only as options in the case of follow up panel waves.

2.1.2 Level of disaggregation and sample design

The production of more disaggregated statistics generally requires larger samples. This objective has to be balanced against budget considerations and clearly communicated to all key stakeholders. For example, the European Union the Directorate General which is responsible for the allocation of regional funds had specified regional precision requirements such that EU-SILC. These requirements are partly reflected in Annex II of a newly established framework regulation which applies also to EU-SILC (EU 2019/1700).

Where disaggregation variables are available on the survey frame, the sample design may use them as stratification variables. This can help improving accuracy for disaggregations without considerable increases in total sample size. For example, to make inferences on differences in poverty between rural and urban areas, it is useful to consider such variables as strata. The stratification used for a survey will depend primarily on what is available on the frame, though geographic variables are often chosen. This is often done when surveys need to produce estimates for subnational regions. When these regions are of unequal size, it is often best to use different sampling rates by regions. For example, within the European Union, differences in sample size are much smaller than the differences in population size countries. This helps make comparisons between Member States.

It ought to be understood that different sampling rates may impact negatively on the precision of estimates for the population as a whole. The allocation of the sample to various strata is an important consideration. It will depend on many factors such as the variability of the variables of interest by strata, the size of the strata, the expected response rate and the cost of collection. Information about various allocation methods can be found in many of the classical survey sample texts (Kish 1965; Kish 1987; Cochran 1977; Särndal, Swensson and Wretman 1992; Lohr 1999).

While the theory is well-established, in practice there will not be one ideal allocation for a survey. On the one hand, this occurs because of the multivariate nature of surveys. The allocation that is ideal for one variable may not be for another. In a situation where multiple variables are of interest, it is important to remember to choose stratification variables that are related to many of them. Bethel (1989) discusses allocation in the context of multivariate surveys. Additionally, when the goal is to provide disaggregate statistics, the survey will generally be called upon to provide estimates for a whole hierarchy of domains. What is optimal for one class of domains will generally not be for another. In this case there will not be one formula that will give the best allocation. It is good practice to verify the impact of the allocation on a variety of domains that will be used for dissemination. It is also worth verifying that the allocation that is chosen is not too sensitive to small difference in the allocation since the final number of respondents in each stratum will be different from the number selected due to non-response.

In all cases, it is of utmost importance that sample design information be appropriately documented and stored as this is an essential condition for sampling variance to be appropriately assessed. If micro data are disseminated, information on stratification and primary sampling units (PSUs) should ideally be provided unless this information could compromise the confidentiality of respondents. In the European Union, the calculation of appropriate sampling errors has been substantially complicated because this sample design information is not accessible for all countries for Eurostat (Verma *et al.* 2010, Goedeme 2013, Trinidade and Goedeme 2016). [5]

[5] see also the important resources accessible here https://timgoedeme.com/eu-silc-standard-errors/

Country Case. Design of the American Community Survey in the United States

The American Community Survey (ACS)[6] is a relatively new survey conducted by the U.S. Census Bureau. It uses a series of monthly samples to produce annually updated estimates specifically for the small areas (census tracts and block groups). Formerly these areas were surveyed via the decennial census long-form sample. Initially, five years of samples were required to produce these small-area data. Once the Census Bureau, released its first 5-year estimates in December 2010; new small-area statistics now are produced annually. The ACS includes people living in both housing units (HUs) and group quarters (GQs). The ACS is conducted throughout the United States and in Puerto Rico, where it is called the Puerto Rico Community Survey (PRCS).

In total the ACS sample comprises about 3.54 million addresses per year (approximately 295,000 per month). These addresses are selected independently for each of the 3,143 counties and county equivalents in the U.S., including the District of Columbia, as well as for each of the 78 municipalities in Puerto Rico. Increased sampling rates were used for the smallest sampling entities.

The ACS complements, rather than replaces the monthly Current Population Survey which has an annual Social and Economic Supplement which is commonly used for poverty statistics.

2.1.3 Repeated surveys

Surveys on poverty and income are generally repeated in order to understand trends over time. It is therefore important to determine whether a cross-sectional or longitudinal design is desired. In a cross-sectional survey, the sample is used once and for the next repetition of the survey a new independent sample is chosen. In a longitudinal survey, data is collected from the selected sample on several occasions, often over many years. In this case, the sample is generally referred to as a panel.

For the measurement of income and poverty, longitudinal surveys can be of particular interest since they allow for the measurement of change at the individual level. As a result, issues such as the persistence of poverty can be measured by a longitudinal survey much better than by a cross-sectional survey. Measuring these types of issues using a cross-sectional survey relies on respondents being able to accurately report on their situation in the past which is known to be definitely less reliable than taking the measurement twice, at two points in time. Additionally, longitudinal surveys reduce the sampling variance for estimates of change ($\hat{Y}_1 – \hat{Y}_2$ where $\hat{Y}_1$ is the measure at time 1 and $\hat{Y}_2$ is the measure at time 2).

Longitudinal surveys are not without their challenges and disadvantages, however. Collection is complicated by the presence of movers that must be followed to their new address (see Iacovou and Lynn, 2013)[7]. The representativity of a longitudinal survey decreases the further you are from the time at which the panel was formed due to non-response, and to changes in the population such as birth, deaths, and immigration. This is a particular challenge for disaggregation. Without adequate sample refreshments the population of new migrants – which are often vulnerable to poverty – can not be represented which can lead to serious bias (Glaser *et al.* 2015).[8] Non-response due to respondent fatigue is a particular challenge for longitudinal surveys. Since non-response compounds over time, it is harder to measure and treat sufficiently in longitudinal surveys than in cross-sectional surveys. Finally, costs can be a concern for longitudinal surveys since they require budget to be guaranteed over an extended period of time.

There exist intermediate solutions between a cross-sectional and longitudinal survey and allow for objectives of both to be balances. In rotating panels designs part of the sample is replaced at each iteration of the

[6] https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html

[7] https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2013-17.pdf

[8] http://statistik.at/wcm/idc/idcplg?IdcService=GET_PDF_FILE&dDocName=081094

survey. This allows to more accurate measures of change while limiting the issues relating to representativeness and non-response attrition associated with longitudinal surveys. That being said, many of the collection and definitional issues carry over from a longitudinal design to a rotation panel design. The EU-SLIC uses this type of design.

2.1.4 Mandatory and voluntary surveys

Whether a survey is mandatory is another decision to be made. Mandatory surveys generally lead to higher response rates and lower collection effort being required which can lower costs. However, the decision as to whether a survey can or should be mandatory or voluntary depends on the legal framework under which a statistical agency operates and is often country-specific. Even when the legal framework provides for mandatory surveys, statistical agencies may decide not to make a survey mandatory in order to conform to social expectations and to maintain a good relationship with their constituents. In EU-SILC only few countries participation is mandatory. From the perspective of measurement errors, it is not always advisable to make survey participation mandatory (see Glaser *et al*., 2015).

Country Case. Canadian Income Survey Design

Statistics Canada has used a series of surveys to provide information on the income and income sources of individuals and families in Canada. Since 2012, this is done using the Canadian Income Survey (CIS), an annual cross-sectional household survey. The CIS is a supplement to the Canadian Labour Force Survey (LFS) which uses a probabilistic sample selected from an area frame using a multi-stage survey design. The LFS is composed of six independent samples as rotation groups. A sixth of the sample, corresponding to one rotation group, is replaced every month. The CIS inherits its sample from the LFS. From January to June every year, the LFS respondents in their last month of LFS collection are asked to answer the CIS questionnaire immediately following the LFS. Though the LFS is a mandatory survey, CIS is not.

All income data for the CIS is gathered from administrative tax files and record linkage is used to combine it with survey data. Though it has clear advantages from the point of view of accuracy, the principal disadvantage of using tax data is timeliness. The CIS is disseminated 14 months after the end of its reference year. The principal reason for this is that tax data only becomes available for processing with the survey data nine months after the end of the reference year.

By using administrative data as its income source and combining its collection with that of the LFS, during which personal and household characteristics have already been collected, the CIS can use a relatively short questionnaire that can generally be completed within 10 minutes. The CIS gathers additional data on labour market activity, school attendance, activity limitation, support payments, inter-household transfers, and characteristics and costs of housing.

Since its sample design and sample size is tied to that of the LFS, CIS can only be used to produce disaggregated statistics to the degree that the LFS design allows. Statistics Canada also builds complementary products produced directly from administrative tax data allowing further geographic disaggregation of income. Though administrative data can be used for producing statistics for small geographies, on its own, it cannot disaggregate along some other variables of interest such as household composition.

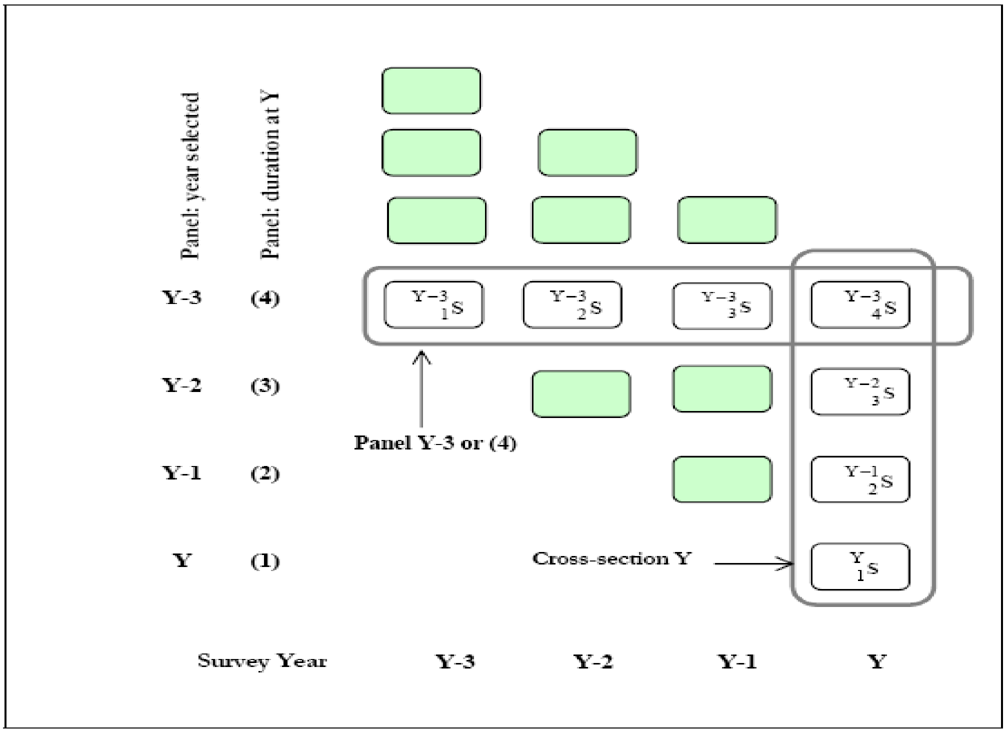
Country Case. EU-SILC Sample Design

The EU-SILC uses a rotating panel design that allows for the production of both cross-sectional and longitudinal statistics. In most EU-SILC, surveys a period of four years is taken as the duration for longitudinal follow-up at the micro level. A standard integrated design has been adopted by nearly all the participating countries.

This integrated design involves a rotational panel in which a new sample of households and persons is introduced each year to replace a part (normally one quarter) of the existing sample (see Figure 2 below). Persons enumerated in each new sample are followed-up in the survey normally for four years. A common rotational sample of this type yields each year a cross-sectional sample as well as longitudinal samples of various durations (Verma, 2001; Verma and Betti, 2006).

At any one time, the sample is made up of 4 subsamples or panels. Each year one new panel is added to stay in the survey for 4 years, and then dropped to be replaced by another new panel. Movers from the original sample are followed-up to their new location for up to the time their panel remains in the survey. This scheme provides both cross-sectional and longitudinal data from the same common set of units. The cross-sectional sample for year Y consists of four subsamples, 1-4, one introduced each year from (Y-3) to Y. A longitudinal sample consists of persons who have remained in the survey since they were first introduced into it. Three overlapping longitudinal samples of different durations are formed: of two-year duration from subsamples (2+3+4) of three-year duration from subsamples (3+4); and of four year duration from subsample (4).

Figure 2. EU-SILC standardised rotating scheme



Source: Verma and Betti (2006)

2.2 Data collection

Research on the Labour Force Survey in Austria (Glaser *et al*., 2015) has shown that even in official surveys where participation is compulsory and non-response may incur a penalty fine, systematic differences across groups can be observed which a may also lead to considerably biased results.[9] Moreover, non-response leads to increased variance by decreasing the size of the sample used at estimation. Data collection is the survey step at which these issues can be prevented, or at least minimized.

Lower response rates among people at both ends of the income distribution is a common occurrence and impacts the measurement of poverty. Factors potentially related to poverty, for instance lower levels of education or language barriers among recent immigrants, can also make responding to surveys more difficult and lead to higher non-response. Longitudinal surveys can be very useful for understanding longer-term trends in poverty but come with the extra challenge of minimizing attrition due to respondent fatigue.

In order to encourage high and balanced response rates, collection should be planned so that it minimizes the burden placed on respondents while maximizing their perceivable benefit. The use of administrative data sources to collect income information can greatly reduce the burden imposed on survey participants.

For the variables that cannot be obtained from administrative sources, there are many ways to simplify the process for respondents. A survey on income can be conducted at a time of year that is shortly after individuals prepare or review their income tax documents, thereby making it easier to answer questions on income. Letting respondents know ahead of time which type of information is required to answer the survey can help reduce item non-response. Surveys measuring consumption often do so by having respondents fill out a diary of their purchase over a period of time (often a week or two for example) and shorter periods may lead to higher rates of completion of the diary. Proxy interviews, that is, obtaining information for an absent respondent from another knowledgeable person, generally leads to higher response rate. Though the quality of proxy responses tends to be lower than if the response was obtained directly, depending on the question asked, it can often be adequate especially for members of the same household.

Communication strategies can be used to lower unit non-response. Letters sent to the sampled individual before they are contacted by interviewers can increase survey participation. Dillman, Smyth and Christian (2014) have demonstrated empirically that when attention is paid to the details of respondent communication, participation can be vastly improved substantially, notably when financial incentives are used to frame participation in the survey as social exchange.

Country Case. Collection Methodology of the American Community Survey

Because a high level of self-response is cost critical, the ACS employs multiple mailings to encourage respondents to complete the survey via the Internet or to return a paper questionnaire. ACS materials for U.S. addresses are printed in English, and Puerto Rico Community Survey (PRCS) materials sent to Puerto Rico are printed in Spanish. U.S. respondents can request Spanish mailing packages, and Puerto Rico respondents can request English mailing packages, via telephone questionnaire assistance (TQA).

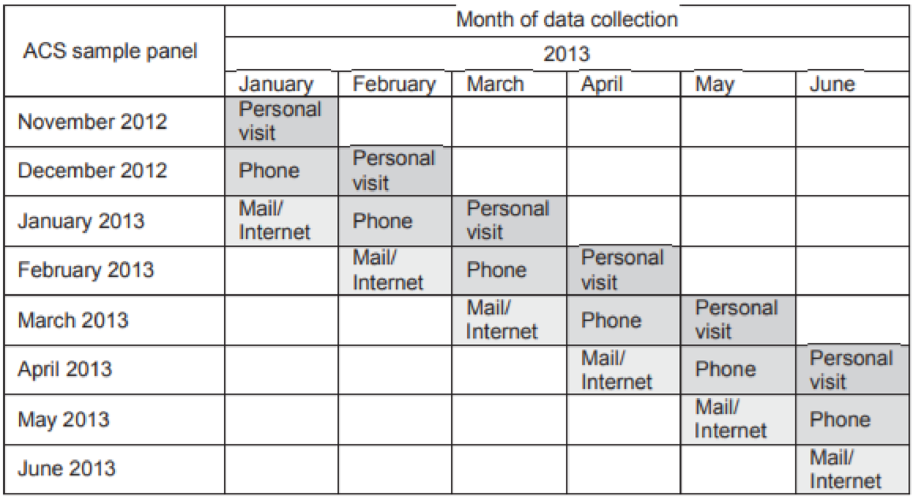
For most HUs, the first phase includes a mailed request to respond via Internet, followed later by an option to complete a paper questionnaire and return it by mail. If no response is received 5 weeks after the first mail, the Census Bureau follows up with computer-assisted telephone interviewing (CATI) when a telephone number is available. If the Census Bureau is unable to reach an occupant using CATI, or if the household refuses to participate, the address may be selected for computer-assisted personal interviewing (CAPI).

The ACS includes 12 monthly independent samples. Data collection for each sample lasts for three months, with mail and Internet returns accepted during this entire period. This three-phase process operates in

[9] https://www.ajs.or.at/index.php/ajs/article/view/doi10.17713ajs.v45i3.120

continuously overlapping cycles so that, during any given month, three samples are in the mail/Internet phase, one is in the CATI phase, and one is in the CAPI phase.

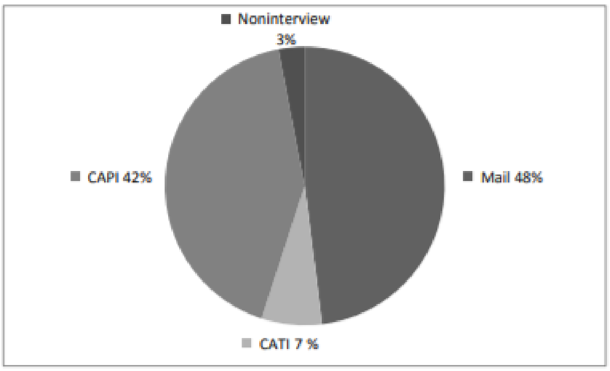
Figure 3: ACS Data Collection Consists of Three Overlapping Phases



ACS sample panel	Month of data collection					
	2013					
	January	February	March	April	May	June
November 2012	Personal visit					
December 2012	Phone	Personal visit				
January 2013	Mail/ Internet	Phone	Personal visit			
February 2013		Mail/ Internet	Phone	Personal visit		
March 2013			Mail/ Internet	Phone	Personal visit	
April 2013				Mail/ Internet	Phone	Personal visit
May 2013					Mail/ Internet	Phone
June 2013						Mail/ Internet

Figure 3 summarizes the distribution of interviews and non-interviews for the 2012 ACS. Among the ACS sample addresses eligible for interviewing in the United States, approximately 48 percent were interviewed by mail, seven percent by CATI, and 42 percent were represented by CAPI interviews. Three percent were non-interviews.

Figure 4: Distribution of ACS Interviews and Non-interviews



Source: 2012 ACS Sample

The Current Population Survey (CPS) uses CAPI/CATI interviewing with interviewers trained to probe respondents. It takes several months to get the ACS responses from a given month as nonresponse is followed up and there can be delays with the mail back. Increasingly internet responses are received which might alleviate some of these operational issues in the future. The ACS has a 3-month data collection window and not the one month or one-week collection period that CPS uses. This poses its own estimation challenges because each month's worth of interviews is a somewhat random collection of interviews from three different sample panels depending on the self-response rates and whether the housing units are interviewed in months 1, 2, or 3.

To reduce the potential bias due to non-response, response rates should be monitored during collection. Indicators of representativity which are based on the variance of response rates between groups should be calculated. A tool that can be used to evaluate non-response is the R-indicator.[10] Introduced by Schouten, Cobben and Bethlehem (2009), it measures the degree to which response propensities are constant over subpopulations. Originally, R-indicators had been limited to those variables which are available in the sampling frame only. More recently, the method has been extended so that information which is available only at the population level may be used (Bianchi *et al*., 2016). If certain areas or domains have lower response rates, collection staff can be reallocated to work on cases in these groups. If variables on the frame identify hard-to-reach groups or groups to which hard-to-reach groups are more likely to belong, then particular attention should be made to obtaining sufficient response rates in these groups. For longitudinal surveys, data from previous waves is a particularly rich source of data for monitoring collection, and if some groups are trailing in terms of response rates, tracing activities can be concentrated on the groups in question. The R-indicator was recently also used to evaluate the impact of panel attrition of representativeness of the EU-SILC (Luiten and Schouten, 2019).

Response rates among groups which are especially hard to contact can also be vastly improved by requesting a specified minimum number of documented visits or calls. Satisfactory contact rates require sufficient time. For groups which can be expected to be away from home such as young single adults contact modes (phone, text messaging or internet) may be adjusted. Different expected response rates may be considered when the order of interviews or even payment per interview are determined.

The practice of sample substitution is not an appropriate measure for managing non-response and should be avoided as it does not reduce the bias introduced by non-response. It can also encourage poor collection practices such as not making enough effort to obtain responses from households that do not respond at first contact, which can worsen non-response bias when present. Sample substitution makes it difficult to calculate appropriate response rates and assess the quality of the data which is obtained. In the European Union also, guidelines for EU-SILC are clear in this respect, stating that “As a rule, the units enumerated in the survey shall be exactly the same units as those selected for the purpose in accordance with the sampling design, i.e. not substituted for by other units.” (Eurostat, 2014)

The second type of error to be addressed during data collection is measurement error. Many of the strategies for reducing response burden mentioned above, also help reduce recall error and lessen measurement error. Yet again, one worth highlighting in the context of poverty measurement is the use of administrative or register data on income. Its use has been shown to reduce measurement error in many countries. Interviewer training is an important way to address measurement issues as well, especially for hard-to-interview populations.

When no statistical registers are available, software such as Italian RELAIS (Record Linkage At Istat) and Statistics Canada’s G-Link (Statistics Canada, 2017) is available to facilitate the process of data linkage to administrative sources. No matter which record linkage software is used, one of the most important steps of record linkage is the pre-processing of linkage variables such as names and addresses. The cleaning and standardization of linkage variables can substantially improve the quality of the subsequent linkage. The record linkage process can be thought of as simply another data collection strategy. Just as the response rate is an important indicator of quality that should be shared with data users, so too is the linkage rate.

For the measurement of poverty, it is still common to use personally assisted modes, rather than self-administered questionnaires. Mostly, computer assisted personal interviews (CAPI) have replaced the conventional paper and pencil mode (PAPI). This gives additional control over measurement errors due to routing mistakes in the questionnaire. The use of automatic checks can ensure that interviewers instantly

[10] https://www.cmi.manchester.ac.uk/research/projects/representative-indicators-for-survey-quality/

detect and can probe respondents on potential inconsistencies. For example, if a respondent state that someone in the household receives a pension, this should be aligned with the activity status of that person; gross salaries should always be larger than net salaries; social benefits will usually fall within certain limits; etc. This requires skills and sufficient time in programming the data collection software. With computer assisted data collection data can be processed without any further delay. Logistics with paper questionnaires becomes obsolete as well as provisions for scanning or manual data entry. On the other hand, issues such as data transmission, privacy, and performance of software solutions and reliability of communication technology become more pertinent when computers are used in data collection.

An increasingly relevant aspect of data collection comes with the – often simultaneous - use of different data collection modes such as telephone (CATI) and self-administered internet questionnaires (CAWI). In the European Union most countries are about to or have already turned to some kind of mixed mode design. Mixing survey modes is often seen as an opportunity to save survey cost. The potential switches between modes has however implications for the case management software. Furthermore, each mode is likely to exhibit specific selection bias and measurement errors. If data or a sub-population is accessible only by a certain mode, their measurements may reflect the specific methodological effects attached to that mode. When poverty measures are then disaggregated it may then be difficult to distinguish the effect of the different data collection modes from the true differences between groups. Many countries are using BLAISE or have developed their own data collection tools. The current state of mixed mode data collection methods has been documented in the MIMOD project, in which several EU Member States shared their views on survey organisation and software for mixed mode data collection.[11]

2.3 Data processing and imputation

Raw data that is received from collection is invariantly messy and difficult to use directly. Cleaning and processing of the raw data are crucial to transform it into a useable dataset. Some variables will have values that are missing. This can for many reasons, for instance, because of refusals to answer certain questions, because the respondent does not know the answer, or because the linkage did not find the relevant record. In addition, variables may take on invalid values or values that are incoherent when compared to other variables for the same record.

The first steps in the cleaning process is to determine which sampled units provided responses that are sufficiently complete to be considered as respondents. Units that have not provided responses to key questions should be considered to be non-respondents. The number of questions considered crucial for this purpose should be small. When a survey is carried out at the household level, the response status should be defined at the household level, so that either the entire household is determined to be respondent or non-respondent. On the Canadian Income Survey, the entire household is deemed to a respondent if at least one adult household member answers the key questions on the survey. Unit non-response is treated by weighting as described in the next section, while item non-response is treated through imputation which is described further in this section.

The second step, editing, involves the correction of values for which it is evident that there has been some measurement error. Generally, very few corrections of this type are made because is difficult to identify and correct measurement errors with certainty. One instance in which these types of errors can be corrected is when interviewers have left notes identifying problems. As another example, on monetary amounts, it may be possible to detect errors where a decimal point has been placed incorrectly. Extreme values and values that are inconsistent with other variables for the same sampling unit can also be treated at the editing step either by altering the values themselves or by deciding that the value should be imputed.

[11] https://www.istat.it/it/files//2011/07/MIMOD-project-Final-report-WP1-WP5.pdf

Though editing can be a helpful step and it can improve the overall data quality, it is strongly recommended to use it parsimoniously. On income variables whose distributions can be quite skewed, extreme values are to be expected and applying corrections should only be done rarely since it is likely to introduce bias. It is also important to remember that collected data will always contain surprising relationships between variables. The survey practitioner should not aim to produce dataset that is free of relationships that seem inconsistent at first view. Not only would he find the task near impossible, in trying to do so he would be likely to impose preconceived relationship on the data that may turn out to be incorrect.

Having cleaned the data, the next step is imputation, that is, assigning values to replace missing data. There are many imputation methods available. Two examples of imputation for income variables are given below, the first is uses a regression-based approach and the second donor imputation. Regardless of the method chosen, the same principles apply. It is important to find a method that maintains coherence between the variables for the record that is being imputed. For example, on income surveys, the relationship between labour forces status and the types of income received should be respected. The process should be automated and objective, avoiding the application of preconceived models on the records to be imputed as manual intervention can. Additionally, imputation rates should be reported as part of the survey's quality report as an important complement to the response rates.

The most flexible imputation procedure for imputing missing data (unit non-response) in income variables in EU-SILC is based on the "sequential regression multivariate imputation" (SRMI). This approach was implemented by the University of Michigan in the imputation software (IVE-ware) which can be used with SAS, STATA, SPSS and R packages or as a standalone in Windows, Linux or Mac OS (except SAS) operating systems. The method proposed by the authors of the software (Raghunathan, Lepkowski, Van Hoewyk, and Solenberger, 2001) constructs the imputed values by fitting a sequence of regression models and drawing values from the corresponding predictive distribution, under the hypothesis of Missing at Random (MAR) mechanism, infinite sample size and simple random sampling.

The procedure is a variant of the estimation-maximisation (EM) algorithm and follows a Bayesian paradigm. The sequential multivariate model used made for more complete imputation of the variables, while at the same time safeguarding their variance and their inter-correlation. A brief outline of the approach may be described as follows:

- Initially, the variables are divided into two types: auxiliary variables used to impute the others, and target variables which are the subject of the imputation. In the initial stages the auxiliary variables are those relating to the demographic characteristics (sex, age) and to labour force characteristics.
- The auxiliary (exogenous) variables are supposed to be available for all cases. If not, some ad hoc procedures are used to perform the necessary imputations. The objective of this is not to impute 'final' values of these variable as such, but to provide a basis for their use in the imputation of the target (income) variables.
- The target variables are arranged in a sequence, starting with those with the smallest proportion of (or with no) missing values. (Alternatively, the ordering can be in terms of decreasing explanatory power of the variables.) Going down in sequence, each target variable is imputed using all the variables above it, for which all information is available (or has been previously imputed), as auxiliary variables in the multivariate regression.

  The model is as follows. With U as the matrix containing variables with no missing data (including as a result previous imputation), and Y_1, Y_2...Y_k are variables with increasing rates of missing data, the sequence of imputations is determined by the following factorisation:

  $[Y_1 \mid U]\ [Y_2 \mid U, Y_1]\ ...[Y_k \mid U, Y_1, ..., Y_{k-1}]$

  where $[Y \mid X]$ is the conditional joint distribution of Y where x is known.

  The form of regression depends on the nature of Y, such as a generalised linear regression for continuous variables (as in the case of income amounts), a logistical regression for binary variables, etc.

- Once a variable with missing values has been imputed, it is moved from the second set to the first, i.e. used as an auxiliary variable in imputation of the next variable in the list.
- After all variables in the list have been dealt with as above, the process is started again with the first variable in the target set, but this time using all the other variables as predictors, using for each the given or the most recently imputed value is used. The process is performed for each variable in turn, and is repeated iteratively.

At Statistics Canada, the linkage, processing and imputation of income variables is done in the same way for each of the three surveys measuring household spending, income, and wealth including the Canadian Income Survey. This is an important factor contributing to producing a more coherent picture of economic well-being by Statistics Canada.

On the Canadian Income Survey, around 12% of respondents need to have their income variables imputed. The imputation is carried in phases. In the first phase, the various components of individual market income, such as wages, self-employment income, investment income, and pensions are imputed. Imputing all of these income components is done together to maintain coherence between the various types of income. The next phases impute income variables that are related to family structure, such as alimony and transfers between spouses, and takes into account the vector of variables imputed at the first phase. The final phase imputes income tax amounts, again taking into account variables imputed at earlier phases.

At each phase, the income variables are imputed by nearest neighbour imputation using the Canadian Census Edit and Imputation System (CANCEIS), which is available freely to users outside of Statistics Canada. Nearest neighbour imputation consists of finding a record without any values needing to be imputed (donor) that resembles the record to be imputed (receiver) on range of auxiliary variables and, for the variables being imputed, using that donor's value for the receiver. The donor selected may be the record that is closest to the receiver or may be selected randomly from among a set of potential donor each of which is close to the receiver.

The auxiliary variables used to match donor and receivers include both categorical variables, such as age, sex, labour force status and family characteristics, and numerical values, such as income variables from the previous phases. It is well worth investing the time necessary to select the best auxiliary variables and to determine how much weight to give to each auxiliary variable when calculating the distance between the receiver and potential donors.

Additionally, special treatment is sometimes useful for income variables. For example, since income variables generally have a long tail of large values, they are often transformed so that matching between donor and receiver is done on the rank of units by the auxiliary variable rather than on the original values. As well, during the income tax imputation phase, the donor gives his tax rate rather than tax amount to the receiver and, based on this tax rate, the receiver's income tax to be imputed is then calculated.

2.4 Weighting

The calculation of weights is usual performed by a step-by-step procedure. The main steps common to the production of weights for most surveys include:

1. the calculation of design weights,
2. an adjustment for non-response,
3. calibration, and
4. the trimming of weight.

Implementing these steps can be more complicated for more complex designs. For instance, calculating weights for a cross-sectional survey or for the first wave of a longitudinal survey will generally be simpler than for subsequent waves of a longitudinal survey.

2.4.1 Design weights

In the first step, design weights are calculated on the basis of the sample design. These weights are of methodological interest since they are the starting point from which the weights will be created, but they are not meant to be used in substantive analysis. A design weight is assigned to all sampled units, not just responding units, and is defined when the survey sample is first selected. This weight will be based on the sampling unit.

For example, when a sample of households (or of addresses or other units containing households) is selected, the household design weight is computed as

$$\omega^{(HD)} = \frac{1}{\text{probability of selection of the household}}.$$

The probability of selection is based on the design of the survey and reflects design features such as stratification and multi-stage selection procedures.

2.4.2. Non-response adjustment

The next step of the weighting process in the non-response adjustment. At this step, the weight of non-responding units is redistributed to responding units. Non-response adjustment procedures aim to redistribute the weight of non-respondents to responding units that have a similar response propensity as this can help minimize the impact of non-response bias. For poverty surveys, the concern being addressed by the non-response adjustment is generally non-response at the household interview stage. Non-response by individuals in the household is often addressed by imputation as described in the previous section.

The problem of (unit) non-response can be particular problematic in some household survey in some countries; and it occurs for both cross-sectional surveys and longitudinal or panel surveys (in which case it is referred to as *attrition*). In a longitudinal survey, non-response compounds over the waves, with non-respondents at wave 1 being excluded from subsequent waves and so on. Good, efficient procedures to re-weight the responding cases is therefore a critical requirement at wave 1. However, the possibilities for non-response adjustment in cross-sectional surveys and at the first wave of longitudinal surveys are often constrained by lack of information since the non-response adjustment has to be based on characteristics which are known for both responding and non-responding households. For the later waves of a longitudinal survey, many variables are available for the non-response adjustment since the first wave data can be used.

The non-response adjustment procedure involves estimating response rates or propensities to response as functions of characteristics available for responding and non-responding households. This includes the use of characteristics of the areas where the households are located. This is also true when a sample of persons has been used. The main difference is that for samples of persons the characteristics of interest that can be used for the adjustment are not only those of households, but also (and perhaps more importantly), personal characteristics of the selected individuals.

Generally, it can be useful to apply the adjustment in two steps:

(i) for non-contact (of households and/or of selected individuals); and

(ii) for non-response, once a contact with the households or the person concerned has been made.

For both steps, especially for (i), area-level characteristics often provide a main part of the auxiliary variables explaining non-response. This is because they are the more easily available variables for both responding and non-responding units.

In dealing with the effect of non-response, it is of crucial importance to identify responding and non-responding units correctly. In this context, a "respondent" is not just a collection status, rather it is a unit whose interview is accepted after processing and will be used for estimation. In practice, determining which units are respondents and non-respondents can be complicated because the frames from which units are selected are generally not perfect. Continuing with the household sample example, an address frame will often contain units that do not correspond to a household. This can be because the address is non-existing, corresponds to an unoccupied structure, or is a business rather than a private dwelling. These selected units which turn out to be non-eligible or non-existent must be excluded and not counted as non-responding. Imputation has to be done for units with unknown status, i.e. when it is not clear whether they are non-eligible or non-respondents. Every unit has to be assigned uniquely to one category or the other.

In surveys where substitution has been allowed, non-responding original units for which successful substitutions have been made are to be considered as 'responding units' for the purpose of determining non-response weights.

Having done this preliminary step, there are two commonly used procedures for non-response weighting. The first is to modify the design weights by a factor inversely proportional to the response rate within each "weighting cells" (appropriately determined grouping of units). It is common to use sampling strata or other partitions, somethings geographical, as weighting cells. These classes can also be defined using classification tree, though this is more relevant if many variables are available as is the case after the first wave of a longitudinal survey. The non-response adjusted weight is:

$$\omega^{(HN)} = \frac{\omega^{(HD)}}{R_K}$$

where R_K is the response rate in weighting class k.

In this expression, the response rates should be computed with data weighted by the design weights:

$$R_K = \frac{sum \quad of \quad design \quad weights \quad of \quad responding \quad units \quad in \quad cell \quad k}{sum \quad of \quad design \quad weights \quad of \quad selected \quad units \quad in \quad cell \quad k}.$$

Numerous, very small weighting cells can result in a large variation in R_K values, and should be avoided. On the other hand, if only a few broad classes are used, little variation in the response rates across the sample may be captured, making the whole re-weighting process ineffective. On practical ground, cells of average size 100-300 units may be recommended. These cells must also include enough respondents so that the factor applied is not too large. The appropriate maximum factor will depend on the survey overall response rate. No absolute rule exists but it may be to useful use weighting cells for which the adjustment factor is no more than twice the average adjustment factor, for instance. In other words, if the overall weighted response rate for the survey is 80%, the average adjustment factor would be 1/0.80 = 1.25 and so the weighting cells used could be defined so that no adjustment over 2.5 would be applied in any of the cells.

The other alternative is to use a regression-based approach. Using an appropriate model such as logit regression, response propensities can be estimated as a function of auxiliary variables, which are available for both responding and non-responding cases. Each responding unit weight is adjusted by the inverse of the estimated response propensity, in the same way as by cell response rates in the previous method:

$$\omega_i^{(HN)} = \frac{\omega_i^{(HD)}}{R_K}.$$

When many auxiliary variables are available, this approach is often preferable to simply using sampling strata or a geographic partition.

A very important point when using the regression approach is to ensure that weights assigned are confined to be within reasonable limits. This is the case for all non-response adjustments, no matter the modelling approach used. In the case of a regression-based approach, the regression can predict zero or even negative values, which of course must be rejected. The problem is more general than that since extreme values should also not be permitted. To deal with this is a best practice to classify the units into response homogeneous groups (RHGs) based on the response propensity estimated using the regression. These are defined to be groups of units having similar response propensity. Once these classes are defined, the adjustment can then proceed as in the first method, within the cells. This is known as the score method (Little 1986, Eltinge and Yanseneh 1997).

Regardless of non-response adjustment method selected, the choice of variables is fundamental. In fact, the choice of variables will generally have more of an impact on the effectiveness of the non-response adjustment than the method used. The adjustment will only reduce non-response bias if the variables are related to both the response rate and the estimates being produced. Therefore, variables with a link to income, poverty, or to the variables that will be used to disaggregate the statistics should be prioritized, on the condition that they are also related to non-response of course. When too many variables are included in the model or when the variables are not related to both non-response and the statistics of interest, the non-response adjustment can increase the variance of the estimates.

2.4.2. Calibration

Calibration is a method that adjusts the weights assigned to sample units (individual or household) in order to satisfy (or approximately satisfy) some pre-determined constraints. These are typically based on Census data or other large surveys. The key idea is that estimates formed from the weighted sample should replicate the known values from other sources. The critical requirement in calibration is to ensure that the external control variables are strictly comparable to the corresponding survey variables, the distribution of which is being adjusted.

Calibration is used for multiple reasons. The first is to produce results that are coherent with other related surveys and with the Census data that is available. In addition to this, calibration can also improve the accuracy of estimates in two ways. Calibration can serve as a non-response adjustment and can stabilize estimates reducing the variance of the estimates. As a non-response adjustment, calibration is particular relevant if a control total is available for a variable that was not available for both respondents and non-respondents during the non-response adjustment step. For income statistics, one example of such a source can be administrative tax data that can give the distribution of some income components for the whole population. Calibrating to match these distributions more closely can be helpful as a way of adjusting for non-response.

In household surveys concerning poverty, income and social exclusion, where the household is the sampling unit and both the household and the individual are used as units of analysis, the so-called “integrated” calibration is recommended (Lemaître and Dufour, 1987). This is a calibration which retains the same weights for all members of the same household; characteristics of households and of the total population are controlled. When a separate personal interview sample exists, a further adjustment can be applied to the personal interview sample.

Mathematically, calibration is an optimization problem. The goal is to find weights as close as possible to the non-response adjusted weights that respect the chosen calibration constraints. Different choices of distance functions will result in different estimators. For example, post-stratification and raking ratio estimation can be expressed special cases of calibration. There are numerous software packages that have been built to implement calibration. Examples include CALMAR which is widely used in the EU and Statistics Canada’s G-Est, both of which are available at no cost.

There are important conditions to be respected when implementing calibration. First of all, it is important to be selective in terms of the calibration constraints. If the constraints are not related to the statistics to be produced by the survey or to the domains for which they are produced, the calibration can increase the variance of the estimates without a gain in accuracy. Along the same lines, when too many calibration constraints are used, this can also increase the variance rather than stabilize the estimates.

The main symptoms of excessive calibration include non-convergence (i.e. no solution being found that satisfies all constraints), the presence of negative weights, and weights that are very close to 0 or that are very large. The calibration factors (ratio of the calibrated weight divided by the pre-calibration/non-response adjusted weight, often referred to as g-factors) should be neither too small nor too large. It therefore a good practice to apply bounds to the calibration totals. While there is no specific rule for which range of calibration totals is acceptable. Trying to keep these factors between 0.3 and 3 (or between 0.3 and 3 times the average calibration factor) can be preferable. In this range, it is most important to respect the upper bound for the calibration factors.

Country Case. Calibration in the Canadian Income Survey

This case study briefly describes the calibration totals used for the Canadian Income Survey. This strategy uses two main sources of control totals: demographic projections derived from Census estimates and administrative tax files providing wage and salary information for all paid employees in Canada. The demography totals used include:

Counts of individuals in 15 sex × age groups	$0-6\text{ (both sexes)}$ $\left.\begin{matrix}\text{male}\\\text{female}\end{matrix}\right\} \times \begin{cases}7-17\\18-24\\25-34\\35-44\\45-54\\55-64\\65+\end{cases}$
Number of households by household size	sizes 1, 2 and 3+
Number of economic families by family size	sizes 1 and 2
Counts of individuals in 6 Canadian cities	Montreal Toronto Winnipeg Calgary Edmonton Vancouver

The administrative tax data on wages and salaries is used in the following way:

- An administrative tax file provides wage and salary amounts for all paid employees in Canada, not just those who file their taxes. Using this file, the 10th, 25th, 50th, 65th, and 75th percentiles of the wages and salaries amount is calculated, as is the number of employees in each of the six classes defined by these cut-off points (0th – 10th percentile, 10th – 25th percentile, 25th – 50th percentile, 50th – 65th percentile, 65th – 75th percentile, and 75th – 100th percentile).
- On the survey data, a new variable is derived indicating in which of these six classes each survey respondent who has received wages and salaries the respondent belongs. If the respondent did not receive wages and salaries, he is put in a seventh class of non-wage earners.
- Six calibration totals are used corresponding to the number of employees in each of the six wage and salary classes.

Using these control totals, calibration is carried out separately for each of the ten Canadian provinces.

The wage and salary counts are a particularly important part of the Canadian Income Survey calibration strategy. They provide a way to ensure that the distribution of wages and salaries after calibration matches the distribution coming from administrative tax data. In practice, their effect is to compensate for higher non-response rates at the top and bottom of the income distribution that cannot be sufficiently corrected for during the non-response adjustment.

Along with using a common imputation strategy for income variables, Statistics Canada's surveys on household spending, income, and wealth all use a similar calibration strategy. This has helped make them much more coherent with each other.

2.4.3 Treatment of extreme or influential weights

Trimming or winsorisation refers to recoding of extreme weights to more acceptable values. The objective of trimming is to avoid excessive increase in variance due to weighting (the so-called Kish effect). It is important to realise that the process will introduce some bias. Even so, the aim is to seek a procedure which reduces the mean squared error. Though treatment of extreme or overly influential weights introduces some bias, the overall error may still be reduced.

At each step of the weighting procedure, the distribution of the weight adjustments and of the weights should be checked. In principle, the results of every step can be subject to the trimming procedure. This applies to weight adjustments for non-response and to calibration as well but, of course, if the adjustment factors are already limited by the non-response or calibrations strategy, this step may not need to be repeated separately.

It can also be useful to reduce the weights of unit that are influential for certain important variables, such as key income components. Even if the weight of a unit on its own is not too large, the product of the weight and the value of the variable together may make it influential. This can be dealt with by adjusting the value of the variable in question or the value of the weight. When the value of the variable is reasonable but large or if there are relationships between the variables that must be maintained, it can be more practical adjust the weight.

There is no rigorous or absolute procedure for general use for determining the limits for trimming or windsorising and it is very important to use it parsimoniously because it does introduce bias. While sophisticated approaches are possible, it is generally desirable to have a simple and practical approach.

The following approach, given as an example for the non-response adjustment, may be quite adequate for the purpose if the permitted limits are wide enough. Where

$\omega_i^{(HD)}$ is the household design weight,

$\omega_i^{(HN)}$ the weight determined after non-response adjustment, and

$\overline{\omega}^{(HD)}, \overline{\omega}^{(HN)}$ their respective mean values,

any computed non-response weights outside the following limits are recoded to the boundary of these limits: $1/C \le \dfrac{\omega_i^{(HN)} / \overline{\omega}^{(HN)}}{\omega_i^{(HD)} / \overline{\omega}^{(HD)}} \le C$.

A reasonable value for the parameter is C=3.

As a second example, a unit can be determined to be influential for a statistic in a cell if removing that unit changes the estimate of the statistics in the cell by more than a predetermined percentage. For example, if removing one unit changes the average of wages in a particular age group by more than 10%, it could be deemed influential. The appropriate threshold depends on how common the variable is in the population. As a weight adjustment, units that are influential can have their weight reduced to the point of no longer being influential and have the amount by which their weight was reduced redistributed to other units in the same domain.

Since trimming alters the mean and total value of the weights, these types of adjustments may need to be applied iteratively, with the mean re-determined after each cycle. It may also be necessary to iteratively repeat this step with the calibration. In both cases, a very small number of cycles normally suffices.

The most important factor to remember when reducing the weight of certain units is to do so very parsimoniously since it will introduce bias. It should only be done for a small number of units that are particularly extreme.

Country Case. Calibration in Household Living Condition Survey 2009 in the Ukraine

This case study briefly describes the calibration process performed by State Statistics Service of Ukraine (SSSU or Ukrstat) for the 2009 round of the Household Living Condition Survey (HLCS). The sample design for this survey consists of a stratified multistage probability sample design with a three-stage sampling procedure for urban area and a two-stage sampling procedure for rural area. The procedures for calculation of final weights for the 11,182 interviewed households were implemented under the generally established steps: i) calculation of design weights; ii) adjustment of design weights for unit non-response; and iii) calibration of weights to external sources. The population characteristic variables used for calibration were quite numerous and disaggregated over 27 geographical regions (25 Oblast plus cities of Kiev and Sevastopol): the household size; the presence of children in the household; and the number of men and women in the household, classified into four age groups each. The software used at the SSSU, the SPSS g-calib (Statistics Belgium, 2002, Vanderhoeft, 2002), was not therefore not able to reach convergence without producing some negative weights and a semi-automatic procedure was used to make these negative weights positive. This procedure leads to a weighting system with the following statistical characteristics:

- Mean weight: 1528.96 (for a total population of 17,096,871 households).
- Standard deviation: 790.23
- Coefficient of variation: 0.5168
- 5 minimum value weights: 1.44196, 9.06557, 13.48944, 14.48621, 18.19603
- 5 maximum value weights: 7733.00, 7799.99, 7864.44, 8807.92, 9175.46

Such weights have clearly been calibrated too much, with too many constraints imposed. Although they permit the sample statistics to be unbiased, they introduce extra variability (instability) in such statistics. According to Kish (1992), the increase of variance of a generic statistics y, is given by 1+ cv^2, where cv is the coefficient of variation of weights. In this case the Kish weight effect is equal to 1.267, which means an increase of the variance of about 27% compared to a sample with equal weights. Verma and Betti (2011) show that the Kish effect of weights may depends to the statistic under observation, and that such an effect is multiplicative with the overall design effect; they recommend an optimal ratio of about 10 between the highest and the lowest weights in the sample (here we can observe a ratio of 6363!!).

2.5 Variance estimation

Among the various types of survey errors, sampling error is unique in that it does not need an external source serving as 'gold standard' in order to be measured; it can be estimated based on the sample design. Moreover, as noted in the first part of this chapter, the sampling variance is often the largest component of error for disaggregated statistics for a domain having a sample size. Being able to estimate the sampling variance is the basis on which inference can be made in a design-based approaches to sample surveys. It is therefore very important to have practical procedures for estimating sampling variance. This section will outline options that are often used for social surveys, including those on income and poverty, while the next section discusses how to provide these estimates to the data users.

Practical procedures for estimating sampling errors for such a survey: (i) must take into account the actual, complex structure of the design; (ii) should be flexible enough to be applicable to diverse designs; (iii) should be suitable and convenient for large-scale application, producing results routinely for diverse statistics and subclasses; (iv) should be robust against departure of the actual sample design from the ideal model assumed in the computation method; (v) should have desirable statistical properties such as small mean-square error of the variance estimator; (vi) should be economical in terms of effort and cost; and (vii) suitable computer software should be available for application of the method (Verma, 1991).

Linearization methods and replication methods are two broad practical approaches to the computation of sampling errors. A major advantage of replication methods is that they do not require an explicit expression for the variance of each particular statistic, and hence can more easily handle complex statistics and designs, including multi-wave and longitudinal situations. As a result, replication methods are more commonly used on social survey that generally use complex designs. Under these methods, the variance is estimated by

- taking repeated subsamples, or replicates, from the data, each of which reflect the structure of the full sample;
- re-computing the weighted survey estimates for each replicate and for the full sample; and
- estimating the variance as a function of the resulting estimates.

Examples of replication methods include the bootstrap, Jackknife, and Balanced Repeated Replication (BRR).

The variance estimates should also take into account the effect on variance of aspects of the estimation process by repeating these steps on each of the replicates. In principle, this can include complex effects such as those of imputation and various steps of weighting, though often full repetition of these procedures for each replication is not feasible.

Step by step, this means first creating the replicates, taking into account the sample design. The way in which this is done depends on the replication method used. Subsequently, each step of weighting (and imputation if feasible) is redone on each replicate. At the non-response adjustment stage, idealing the entire modeling of non-response is redone for each replicate but, when this in not practical, the same non-response adjustment factor that was applied to the original sample can be applied to each replicate instead. Similarly, for influential values, while it is ideal to re-identify which values are excessively influential on each replicate individually, it is generally not feasible to do so. In this case, the records whose weights were modified because of influential values on the original sample can have the same modification applied to them in each of the replicates. For calibration on the other hand, it is very important to repeat the calibration on each replicate individually since it has such a large impact on the final variability of the survey estimates. In other words, each individual replicate should be calibrated individually to the same control totals as are used for the whole sample.

Once the replicates are created and have each gone through the weight adjustment process, estimating the variance can be done in most statistical software packages, such as SAS or Stata. Though the bootstrap is not always supported explicitly, the bootstrap variance can be calculated using the BRR functionality when it exists. More information about replication methods can be found in Wolter (2007), Rust and Rao (1996) or Lohr (1999).

Country Case. Calculating Variance for EU-SILC in Austria

Statistics Austria has developed its own tool for variance calculation. A package called surveysd[12] is freely available a for the open source software R. It was developed specifically to take into account the EU-SILC overlapping sample structure which affects the variance properties of estimators when data is pooled over several years (see section 3). The package has three basic elements. A typical workflow with this package consists of three steps which are described in a hands on manner[13] in the accompanying documentation:

- Bootstrap samples are drawn with rescaled bootstrapping in the function `draw.bootstrap()`.
- These samples can then be calibrated with an iterative proportional updating algorithm using `recalib()`.
- Finally, estimation functions can be applied over all bootstrap replicates with `calc.stError()`.

The first function creates any desired number of bootstrap replicates which consider the sampling design. Each sampling unit receives a slightly altered selection weight. If sampling units are followed over time this can be specifically requested so that the longitudinal structure is preserved. Individual sampling units carry their weights as long as they are part of the sample.

The second function specifies controls for calibration and adjusts the original replicate weights accordingly. These controls should be the same as used in the actual survey. They ensure that each sample replicate does indeed represent the same population. This ensures also that the variance estimation will consider the impact of calibration on variance. Depending on the characteristics and the controls which are used this effect will usually imply a gain in precision over the uncalibrated estimates. This point was considered particularly important for Statistics Austria to be able to demonstrate how regional precision requirements can be met (as stipulated in Annex II of regulation (EU) 2019/1700).

The third element implemented in surveysd repeats estimations over the previously defined and calibrated replicates. At statistics Austria this function is used in combination with a tool which automatically supresses tables according to customizable filtering and flagging rules to ensure that only reliable estimates will be published.

2.6 Dissemination

Disseminating disaggregated poverty statistics from survey data can be a challenge. Domains of interest from the analytical and policy perspective may be small, especially when disaggregation variables are considered together. Particular attention must therefore be paid to the accuracy of estimates before they are disseminated. Estimates will have lower precision when the sample contains only a small number of units in the domain of interest. Though the relevance of poverty statistics clearly is increased by producing disaggregated indicators of economic well-being, these statistics are only useful if they are sufficiently accurate for their intended use.

As outlined in the first part of this chapter, the guiding principle in evaluating the quality of statistical estimates is fitness-for-use. It is, of course, impossible to anticipate all eventual uses of a survey's data before its publication. Moreover, the accuracy of survey estimates that is required varies by use. It is particularly important to inform users of the quality of estimates and the elements that affect the quality of the survey or surveys from which disaggregated poverty statistics are produced, so that they may determine whether the data is fit for their intended use. Earlier chapters give recommendations on which statistics are most useful for making international comparisons and by which variables the statistics should be disaggregated. In this section, we outline best practices for what complementary information should be provided to the data

[12] https://github.com/statistikat/surveysd

[13] https://statistikat.github.io/surveysd/articles/surveysd.html

users. The principal recommendation is that users should be informed of the quality of the estimates and of the quality of the survey more generally.

The published tables should contain not only the point estimates but also an indication of the accuracy of the estimates. Most often a measure of the sampling variability is used as a measure of the exactitude of the estimates. There are many ways in which to present sampling variability. The standard error, the coefficient of variation (CV) or relative standard error, and confidence intervals are all possibilities. No matter which indicator is used, it is a best practice to make the indicator available in the same table as the point estimate to which it corresponds, in order to make it easily accessible to users.

The most commonly used measure used in dissemination tables is generally the CV which, by definition, is the ratio between standard error and the mean of the variable of interest. The CV is particularly useful for comparing the precision of two estimates that are on different scales, as can be different types of income or income from different countries. One common communication strategy for dissemination tables is to classify the quality of the estimates based on CV scale. For example, estimates from the Australian Bureau of Statistics' Survey of Income and Housing are annotated by an asterisk (*) when the CV is between 25% and 50% to indicate that the estimate should be used with caution. Estimates with a CV greater than 50% are annotated with a double asterisk (**) to indicate that the estimates are considered too unreliable for general use and should only be used to aggregate with other estimates to provide derived estimates with RSEs of 50% or less (Australian Bureau of Statistics, 2019). Statistics Austria has developed a special R-package (surveysd) for obtaining sampling errors using a bootstrap algorithm that is suitable also for longitudinal data collections such as EU-SILC. An extension of this package is currently under development which should automatically flag cells to be suppressed because of their large standard errors.

Unfortunately, the CV is not as useful for estimates of proportions, of change or differences, and of statistics that can take on negative values and these are all common type of statistics when analysing poverty. Measures of poverty (such as the AROP for example) often take the form of the proportion of a group whose income is below a threshold. In general CVs tend to understate the quality of estimates of small proportions and overstate the quality of large proportions.

Confidence intervals, on the other hand, are appropriate for all types of estimates and have the advantage of being easier to interpret than the CV. Including confidence intervals in the same table as the point estimates is an excellent practice. When the sample is sufficiently large for a central limit theorem to apply, a symmetric confidence interval around the mean of a variable whose width is based on the sampling error, that is $[\bar{y} - z_\alpha SE(\bar{y}), \bar{y} + z_\alpha SE(\bar{y})]$, may be appropriate. However, this may not be the case for small proportions especially if they are based on a small sample, which is often the type of variables that is desired when disaggregating poverty measures. In this case, alternate methods such as a bootstrap confidence interval or Wilson's method could be used.

As mentioned above, some estimates may not be reliable enough to be published. Criteria that can be used to determine which estimates to suppress are always subjective. In addition to release criteria based on the CV as mentioned above, it can be a good idea to suppress cells of a table that are based on too few records to be reliable. For estimates based on a sample that is too small, not only may the estimate be less precise than desired but the estimate of the variance may also be imprecise making it difficult to reliably inform users of the quality of the estimate. As an example on Statistics Canada's Canadian Income Survey, estimates that are derived from fewer than 25 records are suppressed. Here the number of records is either the number of individuals or the number of families depending on the statistic.

EU-regulations on EU-SILC require that the European Commission shall not publish an estimate if it is based on fewer than 20 sample observations, or if non-response for the item concerned exceeds 50 %. The data shall be published by the Commission with a flag if the estimate is based on 20 to 49 sample observations, or if non-response for the item concerned exceeds 20 % and is lower than or equal to 50 %. The data shall be published by the Commission in the normal way when based on 50 or more sample observations and the item non-response does not exceed 20 %.(European Commission Regulation 1982/2003)[14] Following these guidelines, Statistics Austria puts numbers which are based on less than 50 observations in brackets in its publications from the EU-SILC survey and uses a hyphen ("-") for cells in tables which contain fewer than 20 observations.

Country Case. United States Suppression Rules

In the United States, one-year estimates from the American Community Survey (ACS) are published for an extensive set of tables for any geography or group with a population of 65,000 or more. Until the series was discontinued due to budgetary reasons also three-year estimates had been produced for any group/geography of 20,000 or more units. Instead a streamlined set of "supplementary" tables using one-year data for these smaller geographies (>20,000 but < 65,000) is produced regularly. Every other geography/group gets 5-year ACS estimates. The smallest geography published in tables are Census Block Groups which typically have a population of 600 to 3,000 people. Public use micro data show only PUMAs - public use microdata areas - which tend to have populations of approximately 100,000 or more. In addition, tables are limited to those that pass a number of data quality filtering rules. Firstly, if more than half of the estimates in the table are not statistically different from 0 (at a 90 percent confidence level), then the table fails to meet the rule's requirements and is restricted from publication. Secondly, if the median CV value for the table is less than or equal to 61 percent, the table passes for that geographic area and is published; if it is greater than 61 percent, the table fails and is not published. (If the estimate is 0, a CV of 100 percent is assigned). CVs are calculated for each table's estimates, and the median CV value is determined. Whenever a table fails these rules, a simpler table that collapses some of the detailed lines together can be substituted for the original. If the simpler table passes, it is released. If it fails, none of the estimates for that table and geographic area are released. These release rules are applied to single year estimates, but are not applied to the 5- year estimates. Tables with 5-year estimates are subject to some disclosure avoidance suppressions when cell sizes are small.

Depending on the sample design, it can be useful to consider the design effect to determine what is too small of a sample size for dissemination. Though suppression for confidentiality reasons is not addressed here, suppressing cells based on fewer than a predetermined number of records can be part of a strategy of disclosure control for confidentiality reasons as well. Since the sampling variability is not the only factor influencing the accuracy of survey estimates, additional information of the survey should also be made readily available to survey users. This additional information should include information on the survey methodology with a focus on aspects that affect the accuracy of estimates. For example, the Australian Bureau of Statistics presents this material in the User Guide for its Survey of Income and Housing (Australian Bureau of Statistics, 2019).

Even if poverty statistics that are sufficiently disaggregated are disseminated, these statistics may not address all questions. Making microdata sources available to users can be a complementary way of increasing the usefulness of surveys that can be used to measure poverty. The confidentiality of survey respondents is the most important concern to be addressed for microdata sources. Beyond the necessary first step of removing all personal identifiers, such as name, address and other contact information, there are two main ways of protecting the confidentiality of respondents. The first is to restrict access to the microdata set to certain individuals in a controlled setting. The second is to prepare a confidentialised microdata file that has been treated to protect privacy and confidentiality using a variety of techniques while preserving the variables of interest to the degree possible.

[14] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32003R1982

Statistics Canada uses both of these strategies. The Canadian Income Survey data is available to researchers at Research Data Centres across the country. At these centres, researchers must have their research projects reviewed and take an oath not to disclose individual information. These researchers are asked to follow the same suppression rule for small sample sizes as are used on Statistics Canada tables and their final tables are reviewed to ensure that there is no breach of confidentiality. Additionally, a Public-Use Microdata File (PUMF) is prepared based on the CIS. This file can be shared more broadly and used outside of the Research Data Centers. (Statistics Canada, 2019)

Outreach activities may support the relevance and quality of the data. For example, Statistics Austria has successfully held several user conferences to stimulate research and obtain feedback on EU-SILC micro data. On a European level, GESIS has established regular user meetings for EU-SILC microdata and other official data sets which bring together expertise from leading academic researchers and data producers.[15]

3. Other methodological issues related to measuring poverty

This section treats some further aspects that are not always part of the survey process, such as pooling, small area estimation (SAE), and rapid estimates.

3.1. Pooling

According to Verma *et al.* (2013), two types of measures can be constructed at the regional level by aggregating information on individual elementary units:

(a) Average measures, i.e. ordinary measures such as totals, means, rates and proportions constructed by aggregating or averaging individual values (such as population proportions in the area having certain characteristics relating to welfare).

(b) Distributional measures, such as measures of variation or dispersion among households and persons in the region. Such measures may depend on the distribution of characteristics in each region, or on the overall distribution in the whole national population.

The patterns of variation and relationship for the two types of measures can differ from each other, and hence involve separate statistical considerations. Average measures are often more easily constructed or are available from alternative sources. Distributional measures tend to be more complex and are less readily available from sources other than complex surveys; at the same time, such measures are more pertinent to the analysis of poverty, social exclusion and other aspects of well-being.

Where there is not enough data to produce reliable estimates from a single iteration of a survey, pooling of data may be used. For instance, using three-years of data pooled could produce useful estimates in this context. Both measures of types (a) and (b) could be performed using pooled data.

An important point to note is that, more than at the national level, many measures of averages can also serve as indicators of disparity and deprivation when seen in the regional context: the dispersion of regional means is of direct relevance in the identification of geographical disparity. In particular, the interest is in pooling over survey waves in a national survey in order to increase the precision of regional estimates.

A difficulty in pooling samples is that, in the presence of complex sampling designs, proper variance estimation may not be possible, at least in an almost exact way as introduced in Section 2.5, because the structure of the resulting pooled sample can become too complex or even be unknown. In any case, different waves of a survey (such as, for instance, EU-SILC) do not necessarily correspond to exactly the same population. The problem is akin to that of combining samples selected from multiple frames.

[15] https://www.gesis.org/en/services/events/gesis-conferences/european-user-conference-6

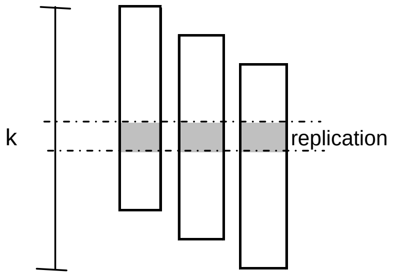
Pooling of estimates from each wave rather than of micro data sets is generally the appropriate approach to aggregation over time from such surveys. Under this approach, for each wave, a person's poverty status (poor or non-poor) is determined based on the income distribution of that wave separately, and the proportion of poor for each wave is computed. These proportions are then averaged over a number of consecutive waves.[16]

Now, when we consider the variance, the issue is to quantify the gain in sampling precision from such pooling, given that data from different waves of a rotational panel are highly correlated.

A large proportion of the individuals are common in the different cross-sections of the panel. However, a certain proportion of individuals are different from one wave to the other. The cross-sectional samples are not independent, resulting in correlation between measures from different waves. Apart from correlations at the individual level, we have to deal also with additional correlation that arises because of the same structure (stratification and clustering) of the waves of a panel. Such correlation would exist, for instance, in samples coming from the same clusters even if there is no overlap in terms of individual households.

The estimation of measures such as differences or averages of poverty measures can be straightforward when a replication method is used for the variance estimation. For this purpose, a replication variance estimation method, as introduced in Section 2.5, can be easily extended by building a coordinated set of replicates for the whole pooled sample along the following lines.

The total sample of interest is formed by the union of all the cross-sectional samples being compared or aggregated. Using as basis the common structure of this total sample, a set of replicates is defined in the usual way. Each replicate is formed such that when a unit is to be excluded in its construction, it is excluded simultaneously from every wave where the unit appears. For each replicate, the required measure is constructed for each of the cross-sectional samples involved, and these measures are used to obtain the required averaged measure for the replicate, from which variance is then estimated in the usual way. So, as example, if we have a dataset with three consecutive years and we want to estimate the average of the three years, we proceed as follows.



It is suggested to construct a common structure of strata and PSUs from the union of the three datasets and assign to this common structure new weights equal to the average of the weights of the three years:

$$w_t^{(Common)} = (w_t)^{Average} = (w_1 + w_2 + w_3) / 3$$

For each year (t) and for each replication (k), we can estimate $y_k^{(t)}$ where t=1,2,3 and from this, the required statistic $y_k^{Average} = \sum_t a_t y_k^{(t)}$; that in our case is just $y_k^{Average} = (y_k^1 + y_k^2 + y_k^3) / 3$.

[16] It may be clarified that when averaging over waves is normally done at the macro level (e.g. poverty rates), and not at micro level (e.g. poverty status of individual persons). Nevertheless, estimation of the measures and their sampling precision requires access to micro data.

Verma *et al.* (2013), have implemented such methodology for Austria and Spain for 2011 EU-SILC data. In this case single year estimates for 2011 are compared with the averages of three years (2009-2010-2011). Also in this case, averaging the poverty rate over three waves leads to a variance of this averaged estimator that is 30% less than the variance of the ARPR estimated from just a single wave for Austria and 35% less for Spain. Similar reduction is also found at regional level (NUTS2).

While the pooling of data can be quite easy to implement and can be used for both average and distributional measures. It is important to remember that access to micro data is required to properly estimate variance of pooled samples and that variance gains are not as large when then the yearly estimates are highly correlated as is the case with rotating panel design which are common for surveys on income and poverty.

3.2. Small Area Estimation

There is a wide variety of small area estimation (SAE) techniques available, and the field is rapidly expanding. SAE has been used successfully on poverty rates and, in fact has its roots in the field, as demonstrated by the now classical example of Fay and Herriot (1979). In general, the suitability and efficiency of a particular technique depends on the specific situation and on the nature of the statistical data available for the purpose. A standard reference on small area estimation methodology is Rao (2003). See also, among others, Gosh and Rao (1994) and Henderson (1950).

It is, of course, not possible within the framework of this project to develop and evaluate SAE models for diverse poverty and related indicators in the specific situation of individual countries. Nor would it be appropriate to make such an attempt, given that the applicability of SAE methods is generally very country-specific since it depends greatly on the data available in each country. The knowledge and experience of national statisticians and other researchers about the specific possibilities and limitations in their own country can be expected to be superior to the approach from the single EU-SILC example mentioned below.

There are some serious limitations to the application of SAE methodology in the context of regional estimation in EU-SILC. But let us first note some potential merits of the procedure.

(a) SAE methods such as EBLUP make use of external data aggregated to NUTS2 (area level) only. R-codes are available under projects funded by the EU 7th Framework program (such as SAMPLE, AMELIE, etc…); estimates could be performed every year, given that such external sources are available.

(b) Poverty mapping permit to obtain estimation at NUTS2 level with high precision (very low standard errors).

There are four types of limitation to be faced.

(a) The first concern is the lack of external data for the purpose of making SAEs. The methodology needs information from census data, which are usually available every ten years in many countries. The poverty mapping model, for instance, is mainly used for consumption data, although some applications with income data have been successful. In any case, often such external sources are not correlated sufficiently highly to the poverty measures under investigation. Also, most of the models assume the external data to be error-free, which is certainly not the case when the data come from other large-scale field studies and surveys.

(b) The methodology tends to be complex and require specialised knowledge and software.

(c) The major concern in application to a multi-country undertaking such as EU-SILC is that the results may lack comparability. Generally, the procedures and application would have to be country-specific, and ensuring the application of common standards required for EU-SILC may be very difficult.

(d) A most important merit of EU-SILC is the provision of public-use microdata files. A major limitation of using SAE methodology in this context is that the results cannot be replicated by researchers since the microdata files do not include the auxiliary information (nor the software tools) used in constructing the original small area estimates.

3.3. Rapid estimates

Surveys used to estimate income and poverty are often not as timely as users would like. This is especially the case when they are based on administrative data that is generally not available to the statistical office until many months after the end of the reference year. With this in mind, many countries are looking to develop rapid estimates or nowcasts of key income statistics.

In 2017, Eurostat, the statistical office of the European Union, and the United Nations Statistics Division jointly published a document called the “Handbook on Rapid Estimates” (Eurostat, 2017). It presents four options for producing rapid estimates:

- *Extrapolation:* This method consists of using a historical data series to produce future estimates. It is characterized by good performance under normal conditions but is unable to predict turning points or the effect of changes to government programs.
- *Nowcasts:* This is an increasingly popular method that uses available data to provide early estimates shortly after the end of the reference period. Data that is available from the reference period is used along with modelling to provide an estimate of the relevant statistics.
- *Flash estimates:* Unlike nowcasts, these estimates use the usual statistical process, but with incomplete survey or administrative data. Not waiting for the complete versions of datasets saves time in production, but the quality of the estimates may be lower when incomplete data is used.
- *Leading indicators:* These are indicators linked to the variable of interest that are characterized by better timeliness. Indicators can be based on a variable that is highly correlated with the variable of interest or created through modelling.

A number of national statistical offices have started to evaluate rapid estimates so they can produce new statistical products that are timelier than traditional published statistics. Eurostat and the national statistical offices of the United Kingdom (ONS) and France (INSEE), are three organizations that have developed and started to publish nowcast estimates of income measures over the past few years (Eurostat 2017, Fontaine and Fourcot 2015, INSEE 2015, INSEE 2018, ONS 2015, ONS 2018, Stoyanova and Tonkin 2016).

With respect to disaggregation objectives, it is important to note that rapid estimation cannot replace the publication of traditional statistics. Instead, rapid estimates aim to complement traditional statistics and are published while awaiting the official figures. They are generally only produced for a few important but high-level statistics. The traditional survey remains essential for addressing questions that require disaggregate statistics.

References

Assael, H. and Keon, J. (1982), Nonsampling vs. Sampling Errors in Survey Research, *Journal of Marketing*, 46(2), pp. 114-123.

Australian Bureau of Statistics (2019), *Survey of Income and Housing, User Guide, Australia, 2017-18*, https://www.abs.gov.au/ausstats/abs@.nsf/PrimaryMainFeatures/6553.0?OpenDocument

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology*, 15, 1, 47-57.

Bianchi A., Biffignandi S. and Lynn P. (2016), Web-Face-to-Face Mixed-Mode Design in a Longitudinal Survey: Effects on Participation Rates, Sample Composition, and Costs. *Journal of Official Statistics*, Vol. 33, No. 2, 2017, pp. 385–408,

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.

Dillman, D. A., Smyth, J. D., and Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. Hoboken, NJ: John Wiley & Sons

Di Meglio E., and Montaigne F. (2013), Registers, timeliness and comparability: Experiences from EU-SILC, in Markus Jäntti, Veli-Matti Törmälehto and Eric Marlier (2013, eds.), *The use of registers in the context of EU–SILC: challenges and opportunities*, Luxembourg: Publications Office of the European Union, 2013.

Eltinge, J. L. and Yansaneh, I. S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey. *Survey Methodology*, 23, 33-40.

European Union (2019), Regulation (EU) 2019/1700 of the European Parliament and of the Council of 10 October 2019. Official Journal of the European Union.

Eurostat (2014), METHODOLOGICAL GUIDELINES AND DESCRIPTION OF EU-SILC TARGET VARIABLES, https://circabc.europa.eu/sd/a/2aa6257f-0e3c-4f1c-947f-76ae7b275cfe/DOCSILC065%20operation%202014%20VERSION%20reconciliated%20and%20early%20transmission%20October%202014.pdf

Eurostat (2017), *Handbook on Rapid Estimates.* Luxembourg: Publications Office of the European Union. https://ec.europa.eu/eurostat/documents/3859598/8555708/KS-GQ-17-008-EN-N.pdf

Eurostat (2018), *Flash estimates of income inequalities and poverty indicators for 2017 (FE 2017) - Experimental results.* https://ec.europa.eu/eurostat/documents/7894008/8256843/Flash-estimates-of-income-inequalities-and-poverty-indicators-experimental-results-2017.pdf

Fay, R.E. and Herriot, R.A. (1979), Estimation of Income from Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, 74, 269-277.

Fellegi, I.P. (2001), “Comment”, *Journal of Official Statistics*, 17, pp. 43-50.

Fontaine M. and Fourcot J. (2015), *Nowcasting du taux de pauvreté par la micro-simulation.* Institut National de la Statistique et des Études Économiques, Série des documents de travail de la Direction des Statistiques Démographiques et Sociales. https://www.insee.fr/fr/statistiques/1304142

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: an appraisal (with discussion). *Statistical Science*, 9, 65-93.

Glaser, T.; Kafka, E.; Lamei, N.; Lyberg, L.; Till, M. (2015): European Comparability and National Best Practices of EU-SILC: A Review of Data Collection and Coherence of the Longitudinal Component.

Goedemé, T. (2013). ‘How much Confidence can we have in EU-SILC? Complex Sample Designs and the Standard Error of the Europe 2020 Poverty Indicators‘, *Social Indicators Research*, 110(1): 89-110.

Groves R. M., Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, Roger Tourangeau (2004), *Survey Methodology*, John Wiley & Sons.

Henderson, C.R. (1950) Estimation of Genetic Parameters. *Annals of Mathematical Statistics*, 21, 309-310.

Heuberger R., Glaser T. and Kafka E. (2013), The use of register data in the Austrian SILC survey, in Markus Jäntti, Veli-Matti Törmälehto and Eric Marlier (2013, eds.), *The use of registers in the context of EU–SILC: challenges and opportunities*, Luxembourg: Publications Office of the European Union, 2013.

Hussmanns, R., Mehran, F. and Verma, V. (1990), *Surveys of the Economically Active Population, Employment, Unemployment and Underemployment*, International Labour Organisation, Geneva.

Iacovou, M. and Lynn, P. (2013), "*Implications of the EU-SILC following rules, and their implementation, for longitudinal analysis*", ISER Working Paper Series 2013-17, Institute for Social and Economic Research.

INSEE - Institut National de la Statistique et des Études Économiques (2015), *Des indicateurs précoces de pauvreté et d'inégalités - Résultats expérimentaux pour 2014.* https://www.insee.fr/fr/statistiques/1304063

INSEE - Institut National de la Statistique et des Études Économiques (2018), *Estimation avancée du taux de pauvreté et des indicateurs d'inégalités.* https://www.insee.fr/fr/statistiques/3623841

Jäntti M., Törmälehto V.M. and Marlier E. (2013, eds.), *The use of registers in the context of EU–SILC: challenges and opportunities*, Luxembourg: Publications Office of the European Union, 2013. https://ec.europa.eu/eurostat/documents/3888793/5856365/KS-TC-13-004-EN.PDF

Jäntti M., Törmälehto V.M. (2013), Survey- and register-based estimates of income distribution and poverty, in Jäntti M., Törmälehto V.M. and Marlier E. (2013, eds.), *The use of registers in the context of EU–SILC: challenges and opportunities*, Luxembourg: Publications Office of the European Union, 2013.

Juran J.M. and Gryna F.M. (1970), *Quality Planning and Analysis*. McGraw Hill.

Kalton, G. (1983), Introduction to survey sampling, 35. Sage.

Kalton, G., McMillen, D. and Kazprzyk, D. (1986), *Non-sampling error issues in SIPP*. In Proceedings of the Bureau of the Census Second Annual Research Conference. Washington, D.C., pp.147-164.

Kish, L. (1965), Survey Sampling. Wiley.

Kish, L. (1987a). *Statistical Research Design*, New York: Wiley.

Kish, L. (1988). Multipurpose sample designs. *Survey Methodology*, 14, 19-32.

Kish, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics*, 8, pp. 183-200.

Lee D. and Shon A. (2001). Korea's experiences in statistical quality assessment. Proceedings, Statistics Canada Symposium "Achieving data quality in a statistical agency: a methodological perspective".

Lemaître, G. and Dufour, J. (1987), An integrated method for weighting persons and families, *Survey Methodology*, vol.13, pp.199-207

Little, R.J.A. (1986): Survey Nonresponse Adjustments for Estimates of Means, *International Statistical Review*, 54, 2, 139-157.

Lohr, S.L. (1999) *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove, CA.

Luiten A. and Schouten B. (2019), Impact of panel attrition on representativeness of EU-SILC, paper presented at Net-SILC3 International Best Practice Workshop "Unit non response and weighting" and "Item, non-response and imputation" University of Essex, 20-22 February 2019.

ONS - Office for National Statistics (2015), *Nowcasting household income in the UK: initial methodology.* https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/articles/nowcastinghouseholdincomeintheukinitialmethodology/2015-07-01

ONS - Office for National Statistics (2018), *Effects of taxes and benefits on UK household income – flash estimate: financial year ending 2018.*https://www.ons.gov.uk/releases/theeffectsoftaxesandbenefitsonhouseholdincomeflashestimatefinancialyearending2018

Purcell, N. J. and Kish, L. (1980). "Postcensal Estimates for Local Areas (Small Domains)", *International Statistical Review*, 48, pp. 3-18.

Raghunathan, T. E., Lepkowski, J. L., Van Hoewyk, J. H., Solenberger, P. W. (2001). A multivariate technique for imputing the missing values using a sequence regression models, *Survey Methodology*, 27, 85-95.

Rao J.N.K. (2003), *Small Area Estimation*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Relais 1.0. User's guide, Istat, http://www.istat.it/strumenti/metodi/software/analisi_dati/relais/

Rust, and J.N.K. Rao (1996), Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*. 5. pp. 283-310.

Särndal, C.-E., B. Swensson and J. Wretman (1992), *Model Assisted Survey Sampling*. Springer. New-York.

Schouten B., Cobben F. and Bethlehem J. (2009), Indicators for the representativeness of survey response, *Survey Methodology*, June 2009 Vol. 35 No. 1.

Statistics Belgium (2002), g-CALIB; Generalised Calibration under SPSS®; Statistics Belgium, release 1.0, April 2002.

Statistics Canada (2003), Survey Methods and Practice, Ottawa: Ministry of Industry, Statistics Canada, Social Survey Division.

Statistics Canada (2017), G-Link A Probabilistic Record Linkage System https://pdfs.semanticscholar.org/abd4/47d16fa837714d582ccd43b18d2285980a8f.pdf

Statistics Canada (2019), Canadian Income Survey (CIS), https://crdcn.org/datasets/cis-canadian-income-survey

Stoyanova, S. and Tonkin, R. (2016), *Nowcasting Household Income in the UK: Financial Year Ending 2015.* Presented at the 34th International Association for Research on Income and Wealth General Conference. http://www.iariw.org/dresden/stoyanova.pdf

Trindade Z. L. and Goedemé, T. (2016), Notes on updating the EU-SILC UDB sample design variables 2012-2014, CSB Working Paper 16/02, Antwerp: Herman Deleeck Centre for Social Policy, University of Antwerp.

UNECE (2017), Guide on Poverty Measurement. United Nations publication issued by the Economic Commission for Europe. https://www.unece.org/stats/publications/guidepm.html

Vanderhoeft, C. (2002), g-Calib Generalised Calibration Under SPSS. Statistics Belgium.

Verma, V. (1981), 'Assessment of errors in household surveys', *Bulletin of the International Statistical Institute*, 49(2), pp. 905-919.

Verma, V. (2001), LESLIE KISH: DEVELOPMENT OF STATISTICS INTERNATIONALLY, Proceedings of Statistics Canada Symposium 2001 Achieving Data Quality in a Statistical Agency: A Methodological Perspective. https://www150.statcan.gc.ca/n1/pub/11-522-x/2001001/session4/6242-eng.pdf

Verma V., Betti G. (2006), EU Statistics on Income and Living Conditions (EU-SILC): Choosing the survey structure and sample design, *Statistics in Transition*, 7(5), pp. 935-970.

Verma V., Betti G. (2011), Taylor linearization sampling errors and design effects for poverty measures and other complex statistics, *Journal of Applied Statistics*, 38(8), pp. 1549-1576.

Verma V., Betti G. and Gagliardi F. (2010), *An assessment of survey errors in EU-SILC*, Eurostat Methodologies and Working papers, Luxembourg: Publications Office of the European Union.

Verma V., Gagliardi F., Ferretti C. (2013), *Cumulation of poverty measures to meet new policy needs*, in Advances in Theoretical and Applied Statistics. Torelli, Nicola; Pesarin, Fortunato; Bar-Hen, Avner (Eds.) 2013, XIX, Springer

Wolter (2007), An Introduction to Variance Estimation, Springer.

World Bank (2017), Monitoring Global Poverty, Report of the Commission on Global Poverty, International Bank for Reconstruction and Development / The World Bank 1818 H Street NW. https://openknowledge.worldbank.org/bitstream/handle/10986/25141/9781464809613.pdf