RAF-AFCAS-21-E-54 – FAO E

November 2021

منظمة الأغذية والزراعة للأمم المتحدة | 联合国粮食及农业组织 | Food and Agriculture Organization of the United Nations | Organisation des Nations Unies pour l'alimentation et l'agriculture | Продовольственная и сельскохозяйственная организация Объединенных Наций | Organización de las Naciones Unidas para la Agricultura y la Alimentación

AFRICAN COMMISSION ON AGRICULTURAL STATISTICS

Twenty-Seventh Session

15 – 18 November 2021, Virtual Host – Dakar, Senegal

AGENDA ITEM 5

The FAO Guidelines on data disaggregation for SDG Indicators using survey data

NAME OF AUTHORS

Piero Demetrio Falorsi, Senior Statistician
Clara Aida Khalil, Statistician
Stefano Di Candia, Junior Statistician
Pietro Gennari, Chief Statistician

ORGANIZATION

FAO Office of the Chief Statistician

SUMMARY

"*Guidelines on data disaggregation for SDG indicators using survey data"* (FAO, 2021) present a comprehensive overview of survey methods and tools for the production of disaggregated estimates of SDG indicators having surveys as the main supporting data source. The publication represents one of the steps taken by the FAO towards supporting member countries in the computation of SDG indicators disaggregated by relevant population groups and territorial areas.

The guidelines address the main limitations posed by most sample surveys, having samples that are either not large enough to guarantee reliable direct estimates for all sub-populations, or that do not cover all possible disaggregation domains. Initially, the publication sets a framework to promote a holistic approach to data disaggregation, describing standard and innovative approaches to tackle these constraints at different stages of the statistical production process. At the sampling design stage, it describes a series of alternative sampling strategies (oversampling, deeper stratification, multiphase sampling with screening of respondents, and marginal stratification) allowing to ensure a "sufficient" number of sampling units for each disaggregation domain, but often resulting in increased cost and complexity of statistical operations. At the analysis stage, the guidelines discuss a series of indirect estimation approaches coping with the little information available for so-called small areas, by borrowing strength from other data sources or domains. In this respect, the publication introduces a model-assisted indirect estimation approach that allows integrating data from different surveys and censuses. The described estimator is operationalized for the production of disaggregated synthetic estimates of *SDG Indicator 2.1.2: Prevalence of Moderate and Severe Food Insecurity based on the Food Insecurity Experience Scale (FIES)*. Both for direct and indirect estimation approaches, methods and software to assess the accuracy of the disaggregated estimates are provided. Finally, the publication concludes with a general overview of small area estimation (SAE) methods, by presenting the key steps for their implementation.

1. INTRODUCTION

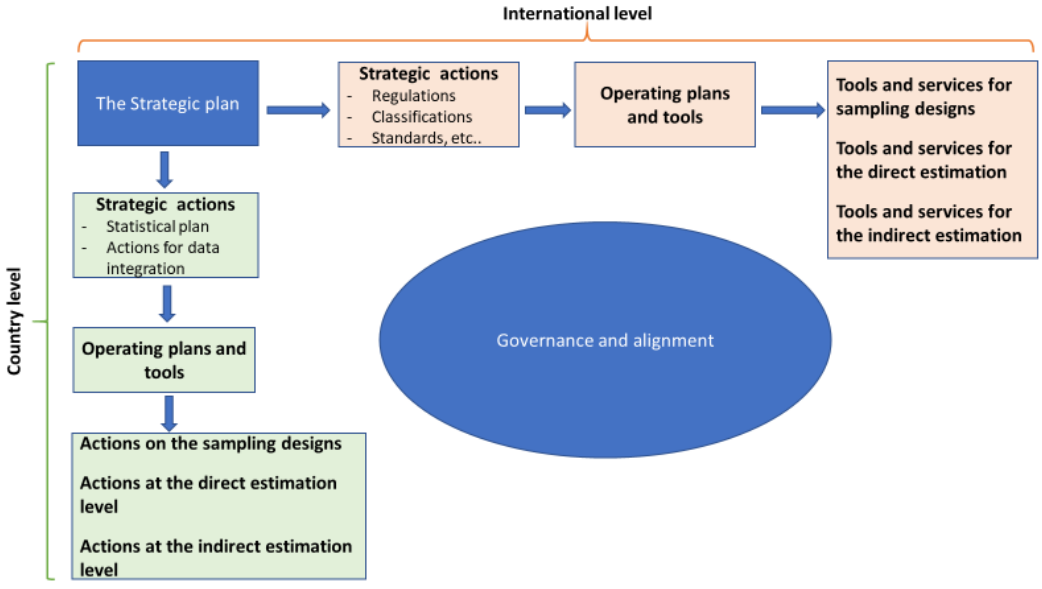
With the adoption of the Leave No one Behind principle (LNOB) as cross-cutting focus of the 2030 Agenda for Sustainable Development, the United Nations Member States have committed to eradicate poverty in all its forms, eliminate hunger, end discrimination, and reduce inequalities and vulnerabilities. In order to monitor all targets for all relevant population groups and geographical areas, detailed measures of progress and Sustainable Development Goal (SDG) indicators disaggregated by multiple dimensions are needed. To this end, the United Nations Statistical Commission (UNSC) – charged with developing the overall SDG measurement framework - embraced an overarching principle of data disaggregation in the development of the Global Indicator Framework to monitor the SDGs and their targets stating that: *"SDG Indicators should be disaggregated, where relevant, by income, sex, age, race, ethnicity, migratory status, disability and geographic location, or other characteristics, in accordance with the Fundamental Principles of Official Statistics".*

Nevertheless, producing high quality disaggregated estimates of SDG indicators imposes significant challenges to National Statistical Systems (NSSs), both in terms of data requirements and operational complexity. With this in mind, at its Forty-Seventh Session, the UNSC requested the Inter-agency and Expert Group on SDG Indicators (IAEG-SDG) to form a working group (WG) on data disaggregation, with the objective of strengthening national capacities and develop the necessary statistical standards and tools to produce disaggregated data. This led, among other outputs, to the compilation by custodian agencies of the main disaggregation categories of SDG indicators, as well as to the identification of a set of policy priorities targeting the most vulnerable population groups.

Within this framework, the FAO Guidelines on data disaggregation for SDG indicators using survey data, (FAO 2021), is one of the steps taken by the organization – as a member of the working group on data disaggregation - towards supporting member countries in the production of SDG indicators disaggregated by different population groups and territorial areas. As such, they offer methodological and practical guidance to produce direct and indirect disaggregated estimates of SDG indicators having surveys as their main or preferred supporting data source, and for the assessment of estimates accuracy.

The Guidelines promote a holistic approach to data disaggregation (Figure 1.1), which involves both national and international actors in the formulation of an agreed strategic plan to foster the integrated use of various approaches, statistical methodologies and tools at different stages of the statistical production chain. These strategic plans influence and guide all actions taken at a more technical level, such as those related to the sampling and estimation phases. National Statistical Offices (NSOs) are the key actors of any strategic plan concerning data disaggregation at national level. At the same time, international organizations can enhance coordination by promoting the adoption of statistical standards and common methodologies to ensure a better quality and comparability disaggregated statistics.

Figure 1.1. A holistic approach to data disaggregation



The Guidelines start with a discussion on the statistical challenges posed by data disaggregation in the context of the 2030 Agenda for Sustainable Development. Subsequently, technical solutions to define sampling strategies for direct domain estimation and methods relying on the use of auxiliary information are discussed. The guidelines also propose sampling designs that guarantee a sufficient number of sampling units for every subpopulation or domain for which disaggregated data must be produced, thus allowing the calculation of direct disaggregated estimates. Moreover, methods for measuring sampling accuracy are provided. The estimation and dissemination of quality indicators assessing estimates accuracy represents a fundamental step in the production of disaggregated estimates and has the potential of increasing the transparency of NSOs and, consequently, strengthening public confidence in official statistics. In addition, direct estimates presenting large sampling errors are an indication of the need to either resort to small-area techniques or revisit the adopted sampling design.

A large section of the guidelines is dedicated to present an indirect approach for producing disaggregated estimates relying on the integrated use of two independent surveys. This method allows integrating a small survey, measuring a target variable with a small measurement error, and a more extensive survey, collecting variables of general use, at least one of which is highly correlated with the target variable (proxy variable).

The guidelines end with an overview of small area estimation (SAE) techniques, as one of the possible approaches to produce indirect disaggregated estimates. Being heavily based on model assumptions, the validation and interpretation of results obtained with SAE approaches may be challenging.

Starting from this publication, the FAO Office of the Chief Statistician is continuing working on data disaggregation and indirect estimation approaches, by developing additional practical case studies on data disaggregation for SDG indicators under FAO custodianship. In particular, the technical report *"Using the projection estimator for data disaggregation of SDG indicators based on survey data"* (FAO, 2021b)[1] presents a case study – based on microdata from Guatemala, Malawi and South Africa, expanding the practical exercise presented in the Guidelines on data disaggregation, and providing a step-by-step guide for its replication. For each step, the software routines for its implementation are reported and explained. In addition, the technical report enriches the empirical application presented in the Guidelines, by providing the tools to assess the accuracy of indirect disaggregated estimates.

2. PLANNING FOR DATA DISAGGREGATION AT THE SURVEY DESIGN PHASE

In order to produce direct disaggregated estimates, the sampling design should ensure a planned sample size in each disaggregation domain. The presence of sampling units in all disaggregation domains also benefits the production of indirect estimates through a substantial reduction of the model bias. As stated in Kalton (2009), when membership of a rare subpopulation (or domain) can be determined from the sampling frame, selecting the required domain sample size is relatively straightforward. In such cases, the main issue is the extent of oversampling to employ to achieve the targeted level of estimates accuracy in each disaggregation domain. Sampling and oversampling rare domains whose members cannot be identified in advance present a major challenge. A variety of methods have been used in these situations. In addition to large-scale screening, these methods include disproportionate stratified sampling, two-phase sampling, the use of multiple frames, multiplicity sampling, and location sampling. Traditional sampling techniques address data disaggregation by oversampling or introducing a deeper stratification. More sophisticated techniques allow for improving sampling designs by geographically spreading the sample units (Gräfstorm, Lundström and Schelin, 2012) and diminishing the level of clustering. This would foster reaching segregated or rare subpopulations. Generally, traditional sampling techniques present certain issues when dealing with rare subpopulations (Kalton, 2009). Kish (1987) proposed a classification of disaggregation domains based on their relative size with respect to the total population. He identifies as "major domains" those comprising approximately 10 percent or more of the total population. For major domains, a traditional sampling design should normally produce reliable estimates. "Minor domains" are those containing from 1 to 10 percent of the total population. In these cases, special sampling approaches are needed to ensure a sufficient sampling size. "Mini-domains" include from the 0.1 to the 1 percent of the total population, and require the use of statistical models in order to get reliable estimates. Finally, "rare domains", comprising less than the 0.1 percent of total population, cannot be handled with survey sampling methods.

[1] The technical report is under review and will be published in the coming weeks.

Issues also arise with populations that are hard to reach (such as the homeless, the migrants, or nomadic populations) or elusive. The relevance of this problem as regards data disaggregation is highlighted by indicator 2.3.1 (Volume of production per labour unit by classes of farming/pastoral/forestry enterprise size), which should be disaggregated by, among other things, the type of enterprise (Farming/Pastoral/Forestry/Fishery). Now, some of the main problems lies in the fact that very often, agricultural surveys do not collect data for forestry, fishery and pastoral activities. Thus, it may be necessary to harmonize different data sources. If it is sought to design a survey for (or including) pastoral activities, in many developing countries, this would imply collecting data on nomadic populations – that can be very hard to locate (FAO, 2016). New approaches recently developed in the sampling literature allow some of the abovementioned problems to be overcome. These methods are, for instance, indirect or multisource sampling (FAO, 2014 and 2015; Lavallée, 2007; Singh and Mecatti, 2011) or marginal stratification sampling (Falorsi and Righi, 2008, and Falorsi Lavallée and Righi 2019). These approaches are extensively discussed in the guidelines (FAO, 2021).

Sections 2.1 and 2.2 below give a brief overview of some of the sampling techniques discussed in the guidelines.

2.1. Traditional sampling techniques to address data disaggregation

Oversampling

With oversampling, a larger size of the overall original sample is defined. This, in turns, results in a larger sample size at the domain level. If the initial sample size is augmented by a proportion Δ, this is expected to have an impact on the increase of the domain sample size equal to $n\Delta P_d$, where $P_d = N_d/N$ is the relative size of the domain d. Table 2.1 represents the increase in the domain sample size n_d due to a percentage increase Δ in the overall sample size of 10.000 households by different subpopulation proportions.

Table 2.1. Increase in the domain sample size n_d due to an increase Δ of the initial sample size (10000 households) by domain relative size P_d

Δ	P_d%			
	0.05%	**1%**	**5%**	**10%**
10%	5	10	50	100
50%	25	50	250	500
100%	50	100	500	1.000

We see that oversampling may be useful when dealing with major domains (Kish, 1986). On the other hand, this approach is less sustainable when dealing with minor and mini-domains. In addition, when the disaggregation domain is not planned at the sampling design stage, the result of oversampling is uncertain, as the domain sample size achieved may be different from the expected one. Table 2.2 illustrates the overall sample size n needed to guarantee the minimum acceptable size n_d^*, for different values of n_d^* and the subpopulation proportion P_d. It can be seen that, in order to achieve the required n_d^* for rare subpopulations ($P_d \leq 1\%$), the overall sample size would need to be way too large and substantially unfeasible for most surveys conducted at national level.

Table 2.2. Sample size *n* needed to guarantee the minimum threshold n_d^* for different values of the subpopulation proportion P_d

n_d^*	%P_d			
	0.05%	**1%**	**5%**	**10%**
30	62.000	31.000	6.200	3.100
50	102.000	51.000	10.200	5.100
100	202.000	101.000	20.200	10.100

Deeper stratification

Stratifying by disaggregation domain is the traditional strategy adopted to control the sample size n_d at the sampling design stage. This implies including the domain-membership variables γ_{di} (with $\gamma_{di} = 1$ if $i \in U_d$ and $\gamma_{di} = 0$, otherwise) among those to be used for the stratification. In many practical situations, however, cross-classification of the stratification variables is unsuitable because it requires selection of a number of sampling units that is at least approximately as large as the product of the number of categories of the stratification variables. Moreover, to obtain unbiased estimates of the sampling variance, at least two units per stratum should be selected. Cochran (1977) illustrates this problem well, giving a clear example of an unfeasible cross-classification design. A combination of explicit and implicit stratification is often used in surveys to consider additional variables that cannot be considered in standard stratification. In the case of major non-planned domains, implicit stratification can facilitate estimation. Falorsi and Righi (2015) illustrate optimal sampling strategies with a priori (uncertain) information on the rare population rate in the strata. This strategy finds the least costly solution by oversampling only in the strata with an expected larger amount of the rare subpopulation. These strategies can be implemented with the Mauss-R[2] software, which enables the multivariate allocation of units in sampling surveys.

Multiphase sampling with a screening of respondents

The strategy based on a deeper stratification requires the availability of the domain membership variables γ_{di} in the sampling frame. This can be the case for geographical variables, but not for many other disaggregation variables such as the income quintile, the migratory or indigenous status, etc.

A traditional sampling strategy to overcome this is to select a first-phase sample $S_{(1)}$ of size $n_{(1)}$. Then, the membership variables γ_{di} are collected from the sampling units of S_1. Then, a stratified sample $S_{(2)}$ is selected to guarantee the planned final sample sizes n_d $(d = 1, \dots, D)$. Since a very large screening sample size is needed to generate an adequate domain sample size when one (or more) of the domains of interest is a rare population, the cost of screening becomes a major concern. Several strategies can be employed to keep costs low (Kalton, 2009): (*i*) use an inexpensive mode of data collection, such as telephone or web interviewing systems, for the screening; (*ii*) allow the collection of screening information from units not included in the screening sample; and (iii) when screening is carried out by face-to-face interviewing in a multistage design, select a large sample size in each cluster to increase efficiency.

2.2. Innovative sampling techniques to address data disaggregation

Marginal stratification designs

The literature on sampling designs provides various methods to keep under control the sample size in all categories of the stratifying variables without using a cross-classification design. These methods are generally referred to as multi-way stratification techniques and have been developed under two main approaches: (i) the Latin Squares or Latin Lattices schemes (Jessen 1978); and (ii) controlled rounding problems via linear programming (Lu and Sitter, 2002). Both approaches present drawbacks that have limited the use of multi-way stratification techniques as a standard solution when planning survey sampling designs in real survey contexts. The main weakness of the linear programming approach is its computational complexity. The sampling strategy proposed below, based on balanced sampling, does not suffer from the disadvantages of the abovementioned methods and grants control over the sample size of various disaggregation domains of interest, defined by different partitions of the reference population. Furthermore, it guarantees that the sampling errors of domain estimates are lower than a predefined threshold. To define the balanced sampling in the design or model-assisted approach, let us introduce the general definition of sampling design as a probability distribution $p(\cdot)$ on the set $\mathcal{S}$ of all samples S from population U. Let x_i be a vector of auxiliary variables x available for each population unit. Sampling design $p(S)$ with inclusion probabilities $\pi = \{\pi_i : i = 1, \dots, N\}$ is said to be balanced with respect to the auxiliary variables if and only if it satisfies the balancing equations

[2] https://www.istat.it/en/methods-and-tools/methods-and-it-tools/design/design-tools/mauss-r

$$\sum_{i\in S}\frac{x_i}{\pi_i}=\sum_{i\in U}x_i \quad (2.1)$$

for all $S \in S$ such that $p(S) > 0$. Let us suppose that a vector of inclusion probabilities π consistent with the marginal sampling distributions n_d $(d = 1, \dots, D)$ is available, i.e.

$$\sum_{i\in U}\pi_i\gamma_{di}=n_d\ (d=1,\dots,D), (2.2)$$

where D represents the total number of domains for which disaggregated data must be produced. Multi-way stratification designs are a special case of balanced designs, where for unit i, the auxiliary variable vector is given by

$$x_i=\pi_i\gamma_i\ (2.3)$$

where γ_i is the D vector of domain membership variables $\gamma_i=(\gamma_{1i},\dots,\gamma_{di},\dots,\gamma_{Di})'$.
When defining the vector x_i as in 2.3, if the condition expressed in 2.2 holds, the selection of samples satisfying the system of balancing equations 2.3 guarantees that the n_d values are non-random quantities.
The left-hand side of the balancing equation 2.1 is

$$\sum_{i\in S}\frac{x_i}{\pi_i}=\sum_{i\in U}\frac{\pi_i}{\pi_i}\gamma_{di}\lambda_i=\sum_{i\in U}\gamma_{di}\lambda_i=n_d.\ (d=1,\dots,D)$$

The right-hand side is

$$\sum_{i\in U}x_i=\sum_{i\in U}\pi_i\gamma_{di}=n_d\ \ (d=1,\dots,D).$$

Tillé (2020) proposed the cube method that allows for selection of balanced (or approximately balanced) samples for a large set of auxiliary variables and with respect to different vectors of inclusion probabilities. In particular, Deville and Tillé (2005) show that, with x_i vectors satisfying expression 2.3, the balancing equation in 2.1 can be satisfied precisely. The cube method is implemented via an enhanced algorithm for large datasets (Chauvet and Tillé, 2006) available in a free software code.[3]
It is important to notice that balanced sampling forms the basis to define broad classes of sampling designs. For example, stratified sampling designs require that:

$$\sum_{d=1}^{D}\gamma_{di}=1,$$

and each U_d is referred to as a stratum. Section 3.5.4 of the Guidelines (FAO,2021) illustrates how to carry out marginal stratification designs for the two-stage or two-phase sampling designs, which are the commonly adopted strategies in real survey contexts.

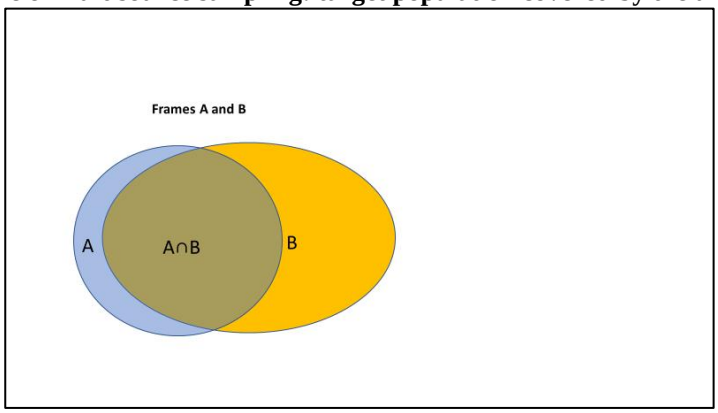
Indirect sampling

In any conventional survey, the random selection of a sample requires the availability of an updated sampling frame recording all units of the target population eligible for the survey, each identified by a label. When the sampling frame is available, a crucial statistical issue is the assessment of the actual coverage provided by this list of the target population. A sampling frame is perfect when there is a one-to-one mapping of frame elements to the target population elements. However, in the statistical practice, perfect frames seldom exist, and problems always arise to disrupt this ideal one-to-one correspondence. For example, the sampling frame might suffer from either or both under-coverage and over-coverage. Under-coverage occurs when the available frame is incomplete, i.e. it includes only part of the target population. As a consequence, the missing elements cannot appear in any sample drawn from the sampling frame. On the contrary, there is over-coverage when the sampling frame contains duplications of the same unit or units that are not included in the target population. In practical situations there may also be frame imperfections of other kinds: for example, in certain circumstances, one may not possess the collection of desired units, but rather a frame of units somehow linked to the list of target units. Also, although a frame may be available, in a dynamic environment it quickly becomes outdated, thus representing a situation that might be rather different from reality. In all these circumstances, the following strategy can be adopted: starting with the observation of one population, the units of the linked and target population are surveyed by reference to their links with the units of the first population. Thus, as would occur with an indirect sampling approach, the target populations can be considered as sampled from an imperfect frame, i.e. the frame referring to the first population. To identify these links, the survey questionnaires must be appropriately structured. FAO (2015) illustrates the modules and operational rules for applying indirect sampling in agricultural surveys.

[3] Available at http://www.insee.fr/fr/nom_df_met/outils_stat/cube/accueil_cube.htm

Multiple-source sampling is another useful approach when dealing with imperfect frames, in particular, when the target population is covered by the union of two or more frames. The case is illustrated by Figure 2.1 below, which displays two partially overlapping frames. A relevant example here is that of agricultural surveys covering holdings in the household and non-household sector. In some circumstances, some of the holdings may fall under two different frames, that of households and that of legal entities.

Figure 2.1. Example of multisource sampling: target population covered by the union of two sources



As it can be seen from Figure 2.1, if a sample S^A is selected from Frame A and an independent sample S^B is selected from Frame B, the units falling in the intersection A ∩ B of the two frames could be included in both samples. FAO (2014) proposes a methodological approach that extends the use of indirect sampling (Lavallée, 2007) to the production of integrated estimates on more than one target population, in the context of multiple frame surveys (Hartley, 1974; Singh and Mecatti, 2011). The techniques proposed are relatively flexible. Furthermore, under rather general conditions, they enable the production of unbiased statistics, thus overcoming most of the problems caused by imperfect sampling frames. The Guidelines (FAO, 2021) show how these approaches can be combined through the concept of multiplicity, first introduced by Birnbaum and Sirken (1965) in their presentation of network sampling as a strategy for surveying rare or elusive populations.

3. ADDRESSING DATA DISAGGREGATION AT THE ANALYSIS STAGE

3.1. The Projection Estimator

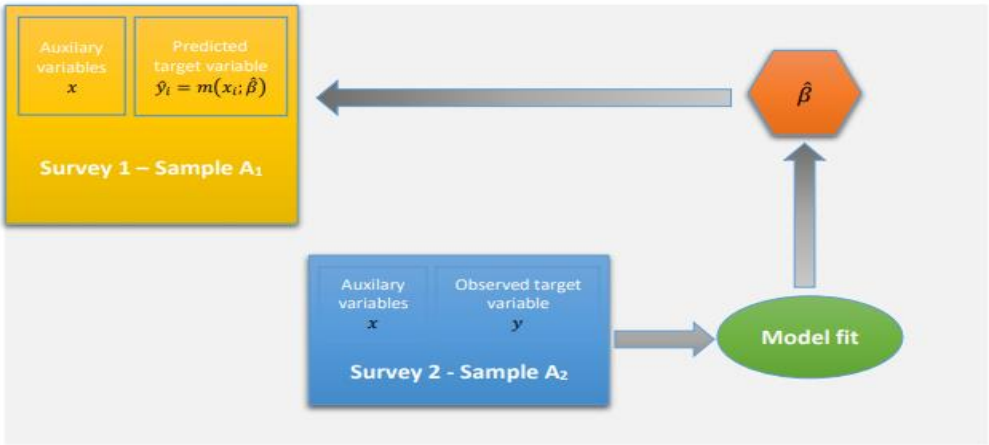
At the analysis stage, data disaggregation can be addressed adopting indirect estimation approaches coping with the little information available for so-called small areas by borrowing strength from additional domains. In particular, the integrated use of different data sources offer a powerful approach for achieving the desired level of disaggregation by preserving estimates accuracy. Typical data sources that could be integrated with data from a particular household and/or agriculture survey are: 1) other surveys; 2) censuses; 3) administrative registers; 4) geospatial information; and 5) big data. Indirect estimation approaches range from model-based to model-assisted approaches.

Among the various methods available to produce indirect estimates, the Guidelines (FAO, 2021) present and apply the so-called "*Projection estimator"* (Kim and Rao, 2012). This model-assisted approach (Figure 3.1) allows integrating data from two independent sample surveys – or a sample survey and a census – where the first survey, is characterized by a large sample A_1, but only collects auxiliary information or variables of general use (e.g. socio-economic variables); while the second survey has a smaller sample A_2 but collects information on the target variable y, along with the same set of auxiliary variables available in A_1. In this statistical setting, the total of variable y in the disaggregation domain d can be obtained as

$$\hat{Y}_{PR,d} = \sum_{i \in A_1} w_{i1} m(x_i;\, \hat{\beta}) \gamma_{id}\,, \quad (3.1)$$

where w_{i1} is the sampling weight of unit i in survey A_1, $m(x_i;\, \hat{\beta})$ is the predicted value of the y variable with the regression parameter $\hat{\beta}$ estimated from survey A_2, and γ_{id} is the domain membership variable, i.e. a dummy variable taking value 1 if unit i belongs to the *d-th* domain.

Figure 3.1: Implementation of the projection estimator



This case covers a great deal of possible empirical situations relevant to data disaggregation. As a matter of fact, most countries have at least one large-scale survey collecting general-use variables, such as censuses, household surveys, but also administrative registers. On the other hand, some of the target variables to be disaggregated in the context of the SDGs are too costly to be measured with a large-scale survey. In these circumstances, a possible solution could be to measure the phenomenon of interest using a small-scale survey and then improve estimates accuracy by relying on auxiliary information collected through a larger-scale survey. The only requisite to be satisfied for the implementation of this approach is that the two surveys must share the same set of auxiliary variables used to fit the regression model.

In many cases, SDG indicators based on survey data present the following functional form:

$$R_d = \frac{Y_d}{Z_d}$$

where

$$Z_d = \sum_{i=1}^{N} z_i \gamma_{di},$$

z_i being the value of the variable z on unit i, where the variable z is observed in the survey A_1. In all these cases, the projection estimator can also be expressed in the form of the ratio:

$$\hat{R}_{PR,d} = \frac{\hat{Y}_{PR,d}}{\hat{Z}_d} \quad (3.2)$$

where $\hat{Y}_{PR,d}$ is defined in Formula 3.1 and

$$\hat{Z}_d = \sum_{i \epsilon A_1} \omega_{i1} z_i \gamma_{di}$$

is the direct estimate of the total Z_d from the survey A_1. When $z_i = 1$, Expression 3.2 provides the projection estimator of a proportion. In order to study the asymptotic properties of estimator (3.2), we consider its linear approximation, given by the first order terms of Taylor's series approximation:

$$\hat{R}_{PR,d} = R_d + \frac{1}{Z_d}\left[\left(\hat{Y}_{p,d} - Y_{p,d}\right) - R_d\left(\hat{Z}_d - Z_d\right)\right] + o_i \quad (3.3)$$

where o_i is a rest of minor order. Starting from Expression 3.3. and considering the variance formulation in Kim and Rao (2012) it can be shown that the sample variance of $\hat{Y}_p$ can be expressed as

$$Var\left(\hat{R}_{PR,d}\right) = Var\left(\sum_{i \epsilon A_1} w_{i1}\, t_{di}\right) + Var\left(\sum_{i \epsilon A_2} w_{i2}\, [y_i - m(x_i;\beta_0)]\right) \quad (3.4)$$

with β_0 denoting the estimate of β when observing the entire population, i.e. the estimation that we would get using census data, and t_{di} is the *Woodruff* (1971) transformation:

$$t_{di} = \frac{1}{Z_d}\gamma_{di}[m(x_i;\beta_0) - R_d z_i].$$

We can derive a plug-in asymptotically unbiased estimator of $Var\left(\hat{Y}_p\right)$ by substituting the super-population value β_0 with the estimate $\hat{\beta}$, as reported below:

$$\hat{V}ar\left(\hat{R}_{PR,d}\right) = \hat{V}ar\left(\sum_{i\epsilon A_1} w_{i1}\,\hat{t}_{di}\right) + \hat{V}ar\left(\sum_{i\epsilon A_2} w_{i2}\left[y_i - m(x_i;\,\hat{\beta})\right]\right)$$

where $\hat{V}ar(\cdot)$ denotes the sampling estimate of $Var(\cdot)$, and

$$\hat{t}_{di} = \frac{1}{\hat{Z}_d}\gamma_{di}\left[m(x_i;\hat{\beta}) - \hat{R}_{p,d}z_i\right].$$

This extension of the basic approach presented in Kim and Rao (2012) allows adopting the projection estimator for many FAO-relevant SDG Indicators, such as:

- **SDG Indicator 2.1.1:** Prevalence of Undernourishment;
- **SDG Indicator 2.1.2:** Prevalence of moderate or severe food insecurity in the population based on the FIES;
- **SDG Indicator 2.3.1:** Volume of production per labour unit by classes of farming/pastoral/forestry enterprise size;
- **SDG Indicator 2.3.2:** Average income of small-scale food producers, by sex and indigenous status;
- **SDG Indicator 5.a.1.a** (Percentage of people with ownership or secure rights over agricultural land (out of total agricultural population), by sex) **and 5.a.1.b.** (share of women among owners or rights-bearers of agricultural land, by type of tenure).

The approach based on the projection estimator allows producing cross-tabulations of the variable of interest y also for disaggregation domains not originally included in the data collection instrument used to get A_2 (sample providing information on y). For example, let's suppose to be interested in estimating a parameter related to y, disaggregated by indigenous status. Let us also assume that the information on the indigenous status of respondents is not available in A_2, but only in A_1. By projecting the values of y on A_1, it is possible to use the auxiliary information on the indigenous status to estimate the parameter of interest considering this disaggregation dimension. Finally, it is important to stress that the projection estimator is a very flexible tool. Practitioners in National Statistical Offices (NSOs) and international organizations can adopt this approach for the integration of survey data with different data sources such as censuses, administrative records, and/or geospatial information. In addition, predicting a variable of interest on the sample of a more extensive survey from which most national official statistics are produced, allows improving estimates' consistency.

3.2. Some empirical results

In FAO (2021b), the projection estimator was adopted to produce disaggregated estimates of SDG Indicator 2.1.2 on the *Prevalence of Moderate and Severe Food Insecurity based on the Food Insecurity Experience Scale (FIES)* using the following data sources:

- The Malawi's Fourth Integrated Household Survey (IHS4) 2016-17 ;
- The Malawi FIES survey module collected through the Gallup World Poll (GWP) – 2016.

FAO (2021b) expands the practical exercise presented in the Guidelines on data disaggregation, and provides a step-by-step guide for its possible replication. For each step, the R scripts for its implementation are reported and explained; in addition, this work enriches the empirical application presented in the Guidelines, by providing the tools to assess the accuracy of disaggregated estimates.

Steps for the implementation of the Projection estimator

One of the most relevant outputs of the FAO technical report on data disaggregation (FAO, 2021b) is a clear identification of all the main operational steps to be carried out in order to implement the projection estimator, which can be summarized as follows:

1. **Identifying and recoding auxiliary variables.** The implementation of the projection estimator requires the availability of the same set of auxiliary variables in the two surveys to be integrated. These variables also need to share common structure and definitions.
1. **Definition of the function $m()$ and estimation of projection parameters in the small sample.** The selection of the functional form for the link function $m()$ to estimate the projection parameters heavily relies on the type of variable y considered (e.g. scale, nominal, dichotomous).
2. **Computation of synthetic values.** Using the estimated projection parameter, the synthetic values of the variable of interest are computed in the large dataset. This, in turn, allows producing indirect disaggregated estimates of the indicator of interest.
3. **Assessment of estimates accuracy.** After producing synthetic estimates, their accuracy can be assessed estimating their variance, coefficient of variation and confidence intervals.

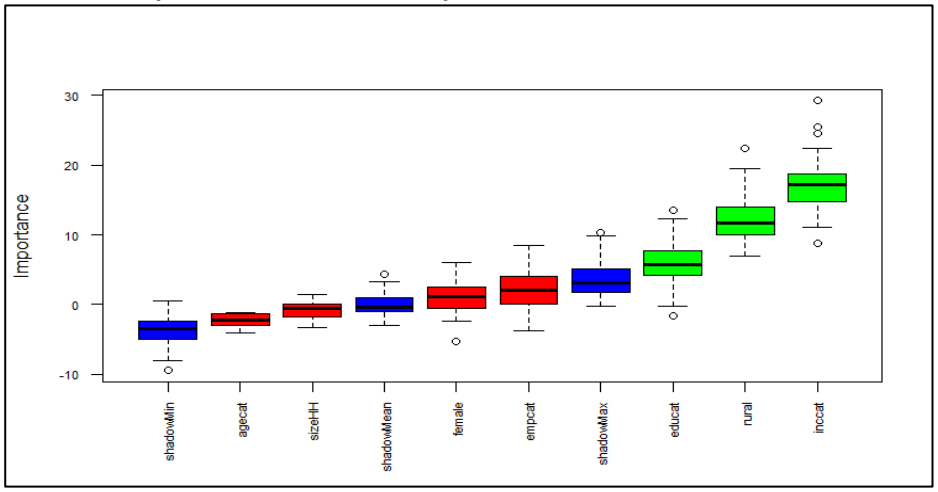
Step 1: Identifying and recoding auxiliary variables

Proper identification of the auxiliary variables x_i in the small survey is a crucial step to ensure the quality of the projection estimator. In this context, the use of variable selection methods can be helpful when there are many potential auxiliary variables, although in some cases problems of multicollinearity could increase the complexity of this task. The literature on variable selection approaches is very ample. For example, Ryan (2008) or Harrel (2015) provide a comprehensive summary of the common methods used for the selection of auxiliary variables in regression models.

Despite the availability of a relatively small number of auxiliary variables common to the two datasets, FAO (2021b) illustrates the use of the Boruta feature selection method, proposed in Kursa and Rudnicki (2010), using a wrapper approach built around a random forest classifier (Breiman, 2001).

Figure 3.2 reports the output of Boruta, represented with a series of boxplots of different colors: red, yellow and green boxplots represent the scores of the rejected (unimportant), tentative and confirmed (important) auxiliary variables respectively, while blue boxplots represent the shadow features identified by the algorithm. Tentative variables are those for which Boruta could not indicate a clear decision concerning their relevance, as their importance level was not significantly different from their best shadow feature.

Figure 3.2. Level of importance of the auxiliary variables for moderate or severe food insecurity



Source: FAO, 2021b

All the levels of auxiliary variables identified as tentative or important by Boruta have been used to fit a logistic regression on the probability of being moderately or severely food insecure. In addition, all the relevant dimensions for data disaggregation (sex, age class, income, rural/urban location) have been included in the regression model, in order to increase the sample unbiasedness of the projection domain estimator.

It is important to stress that, one of the conditions to be satisfied by auxiliary variables before applying the projection method, is for these to share similar definitions and structure in the two samples to be integrated. Hence, before implementing this indirect estimation approach, all the selected auxiliary variables have been recoded and harmonized across the two surveys, as detailed in FAO (2021b).

Step 2: Definition of the function $m()$ and estimation of projection parameters in the small sample

In the case study, a weighted multivariate logistic regression has been implemented in the small sample to estimate the projection parameters $\hat{\beta}$ to be used to predict the value of the variable of interest in the large survey. Let us indicate with $\hat{p}_{ms,i}$ the probability of being moderately or severely food insecure for the $i-$th individual in the small sample. This probability was estimated using GWP data collected with the FIES individual module.

Since $\hat{p}_{ms,i}$ was concentrated around few discrete values in the [0,1] interval, it was recoded into a dummy variable $y_{ms,i}$ such that: $y_{ms,i} = 1$ if $\hat{p}_{ms,i} \geq 0{,}5$, and $y_{ms,i} = 0$ otherwise.
Then, the $y_{ms,i}$ values were modeled with a multivariate logistic function of the set of discrete categorical auxiliary variables $x_i' = (1, x_{i1}, x_{i2}, \dots, x_{ik})$:

$$P(y_{ms,i} = 1|x_i) = m(x_i;\beta) = \frac{\exp(\beta_{ms,0} + \beta_{ms,1}x_{i1} + \beta_{ms,2}x_{i2} + \ \dots + \ \beta_{ms,k}x_{ik})}{1 + \exp(\beta_{ms,0} + \beta_{ms,1}x_{i1} + \beta_{ms,2}x_{i2} + \ \dots + \ \beta_{ms,k}x_{ik})},$$

with $\beta = (\beta_{ms,0}, \beta_{ms,1}, \beta_{ms,2}, \dots, \beta_{ms,k})$.

Step 3: Computing the synthetic values in the large sample

Having obtained the estimates $\hat{\beta} = (\hat{\beta}_{ms,0}, \hat{\beta}_{ms,1}, \hat{\beta}_{ms,2}, \dots, \hat{\beta}_{ms,k})$ of the parameters β with standard statistical tools the predicted probabilities are given by

$$\hat{P}(\hat{y}_{ms,i} = 1|x_i) = \frac{\exp(\hat{\beta}_{ms,0} + \hat{\beta}_{ms,1}x_{i1} + \hat{\beta}_{ms,2}x_{i2} + \ \dots + \ \hat{\beta}_{ms,k}x_{ik})}{1 + \exp(\hat{\beta}_{ms,0} + \hat{\beta}_{ms,1}x_{i1} + \hat{\beta}_{ms,2}x_{i2} + \ \dots + \ \hat{\beta}_{ms,k}x_{ik})}$$

Using the $\hat{P}(\hat{y}_{ms,i} = 1|x_i)$, values we can obtain the projection estimator:

$$\hat{Y}_{PR,ms,d} = \sum_{i \in A_1} w_{i1}\hat{P}(\hat{y}_{ms,i} = 1|x_i)\, y_{di}$$

for the total in the target population, and

$$\hat{R}_{PR,ms,d} = \frac{\sum_{i \in A_1} w_{i1} P(\hat{y}_{ms,i} = 1|x_i) y_{di}}{\sum_{i \in A_1} w_{i1}\, y_{di}}$$

for the proportion in the target population.

Step 4: Disaggregated estimates and the assessment of their accuracy

Estimates, standard errors and confidence intervals have been calculated for the relevant disaggregation dimensions (e.g by sex, age_class, income quintile and urban/rural location). The main empirical results are presented in Table 3.1 below. The comparison of projected versus direct estimates in terms of their coefficient of variation (CV) and confidence intervals shows that the former have a better (or at least equal) accuracy than the latter in almost all cases.

Table 3.1 Projected versus direct estimates of the probability of being moderately or severely food insecure (prob.ms)

		Moderate or severe food insecurity			
		prob.ms	**CV (%)**	**Lower_CI**	**Upper_CI**
IHS4*	**Total**	0.91	1.2	0.89	0.93
GWP**		0.91	1.3	0.89	0.93
IHS4	**Female**	0.91	1.4	0.88	0.93
GWP		0.90	1.5	0.89	0.94
IHS4	**Male**	0.91	1.9	0.87	0.94
GWP		0.91	2.0	0.87	0.94
IHS4	**Rural**	0.93	1.2	0.90	0.95
GWP		0.92	1.3	0.90	0.94
IHS4	**Urban**	0.81	5.7	073	0.92
GWP		0.82	5.9	0.74	0.93
IHS4	**15-24**	0.91	2.0	0.87	0.94
GWP		0.89	2.1	0.85	0.93
IHS4	**25-49**	0.91	1.6	0.88	0.93
GWP		0.92	1.6	0.89	0.95
IHS4	**50-64**	0.87	3.6	0.82	0.94
GWP		0.90	3.5	0.84	0.96
IHS4	**65+**	0.97	1.6	0.94	1.0
GWP		0.98	1.7	0.95	1.0
IHS4	**Inc_1**	0.96	1.5	0.94	0.99
GWP		0.97	1.5	0.94	1.0
IHS4	**Inc_2**	0.96	1.5	0.93	0.99
GWP		0.96	1.6	0.93	0.99
IHS4	**Inc_3**	0.97	1.1	0.95	0.99
GWP		0.97	1.1	0.95	0.99
IHS4	**Inc_4**	0.89	3.6	0.82	0.95
GWP		0.88	3.7	0.82	0.94
IHS4	**Inc_5**	0.74	3.8	0.68	0.80
GWP		0.76	3.8	0.71	0.82

* IHS4: Malawi Fourth Integrated Household Survey – 2016/17

** GWP: Malawi FIES survey module collected through the Gallup World Poll (GWP) – 2016

4. CONCLUSIONS AND RECOMMENDATIONS

The necessity of producing disaggregated estimates of indicators in the SDG Monitoring Framework imposes significant challenges to NSSs. In this framework, the Food and Agriculture Organization of the United Nations (FAO) - as a member of the WG on data disaggregation - is well positioned to support countries who lack the capacity to report SDG indicators at the required disaggregation level. To this end, the FAO Office of the Chief Statistician (OCS) has developed guidelines on data disaggregation for SDG Indicators using survey data (FAO, 2021), which offer methodological and practical guidance for the production of direct and indirect estimates of SDG indicators having surveys as their main or preferred data source. Furthermore, the publication provides tools to assess the accuracy of these estimates and presents strategies for the improvement of output quality through indirect estimation, including SAE methods.

When planning data disaggregation at the sampling design stage, the guidelines illustrate how the sampling design should ensure a planned sample size for the all disaggregation domains. Traditional sampling techniques address data disaggregation by oversampling, deeper stratification, or by introducing multiphase designs with screening of respondents. However, for small domains, or segregated and hard-to-reach populations, standard techniques are generally unfeasible, as they tend to produce an exponential increase of survey costs. Other sophisticated techniques allow for improving sampling designs by geographically spreading the sample units and diminishing the level of clustering. More recent approaches – such as marginal stratification techniques, indirect sampling, multisource and balanced sampling - allow overcoming some of the abovementioned limitations. However, their main drawback is the fact that these techniques are far from being mainstreamed in countries' NSOs, and their adoption would require the implementation of technical assistance and capacity development programs. Strengths and weaknesses of all the presented methods are extensively discussed in the Guidelines (FAO, 2021; Chapter 3). In addition, the publication provides a useful appendix with software packages to be used in empirical applications. Finally, methods and tools to estimate the accuracy of direct disaggregated estimates are discussed (Chapter 4).

At the analysis stage, the guidelines present and apply a model-assisted indirect estimation approach that allows the generation of disaggregated estimates of SDG indicators by leveraging on the integrated use of two independent surveys. In particular, the application of the projection estimator allows combining a big survey or a census, collecting a set of auxiliary information, with a smaller survey, collecting data on a variable of interest along with the same set of auxiliary variables. The discussed indirect estimation approach covers a great deal of interesting and relevant empirical applications for the production of disaggregated data for SDG (and other) indicators. In particular, most countries can normally rely on auxiliary variables provided by large-scale surveys, censuses, administrative records, or geospatial information. In this context, some of the target phenomena for SDG monitoring and data disaggregation are often too costly or complex to be incorporated in large-scale data collection campaigns. The presented approach allows measuring the variable of interest with a small-scale survey, on the sample of which the parameters of a regression-type statistical model can be estimated by linking this variable to a set of auxiliary variables. Based on these parameters, the values of the target variable can be predicted on a larger-scale data source collecting the auxiliary information used to fit the model. Relying on a larger sample allows increasing the accuracy of disaggregated estimates and consider disaggregation domains that are not available in the small survey. In addition, predicting a variable of interest on the sample of a more extensive survey from which most national official statistics are produced, allows improving estimates' consistency. In conclusion, it is important to highlight that the proposed strategy could be easily extended to other empirical contexts where, instead of integrating two independent surveys, a small survey could be integrated with auxiliary information coming from other types of data sources, such as censuses, administrative registers, and/or earth observation data.

5. REFERENCES

Birnbaum, Z.W. & Sirken, M.G. 1965. Design of sample surveys to estimate the prevalence of rare diseases: three unbiased estimates. *Vital Health Statistics*, 2(11): 1–8.
Breiman, L. 2001. Random forests. Machine Learning, 45(1): 5–32.
Cochran W. G. 1977. *Sampling Techniques*. Wiley. New-York.
Chauvet, G., Tillé, Y. 2006. A fast algorithm for balanced sampling. *Computational Statistics* 21, 53–62 (2006). https://doi.org/10.1007/s00180-006-0250-2.
Deville, J.-C. & Tillé, Y. 2005. Variance approximation under balanced sampling, *Journal of Statistical Planning and Inference*, 128: 569–591.
FAO 2014. *The Global Strategy to Improve Agricultural and Rural Statistics. Technical Report on the Integrated Survey Framework*, Technical Report Series GO-02-2014, http://gsars.org/wp-content/uploads/2014/07/Technical_report_on-ISF-Final.pdf. Accessed on 1 December 2014.
FAO. 2015. *Integrated Survey Framework*, Guidelines. Rome, FAO. (also available at http://www.gsars.org/wp-content/uploads/2015/05/ISF-Guidelines_12_05_2015-WEB.pdf).
FAO. 2016. *Guidelines for enumeration of nomadic and semi-nomadic livestock*. http://www.fao.org/3/ca6397en/ca6397en.pdf.
FAO. 2021. *Guidelines on data disaggregation for SDG Indicators using survey data*. http://www.fao.org/documents/card/en/c/cb3253en.

FAO, 2021b. *Using the projection estimator for data disaggregation of SDG indicators based on survey data*. Technical Report. To be published.
Falorsi, P.D. & Righi, P. 2015. Generalized framework for defining the optimal inclusion probabilities of one-stage sampling designs for multivariate and multi-domain surveys. *Survey Methodology*, 41(1): 215–236.
Falorsi P. D., Righi P., Lavallée P. 2019. Optimal Sampling for the Integrated Observation of Different Populations. *Survey methodology*, Vol. 45, No. 3, pp. 485-511. Statistics Canada, Catalogue No. 12-001-X.
Grafström, A., Lundström, N.L.P. & Schelin, L. 2012, Spatially Balanced Sampling through the Pivotal Method, *Biometrics*, 68: 514, 520.
Jessen, R, J, 1978. *Statistical Survey Techniques.* New York City, John Wiley & Sons.
Harrell J., **F.E**. 2015. Describing, Resampling, Validating, and Simplifying the Model. In: Harrell Jr., F.E., *Regression Modeling Strategies*, *Springer Series in Statistics*. Switzerland, Springer International Publishing, pp. 103–126.
Kalton, G. 2009. Methods for oversampling rare subpopulations in social surveys. *Survey methodology*, 35(2): 125–142.
Kursa, M. & Rudnicki, W., 2010. Feature Selection with the Boruta Package. Journal of Statistical Software. September 2010. Volume 36, Issue 11. https://www.jstatsoft.org/article/view/v036i11
Kim, J.K. & Rao, J.N.K. 2012. Combining data from two independent surveys: a model-assisted approach. Biometrika, 99(1): 85–100.
Lu, W. and Sitter, R. R. 2002. Multi-way Stratification by Linear Programming Made Practical, *Survey Methodology*, 2, 199-207.
Rao, J.N.K. 2003). Small Area Estimation. New York City, USA, John Wiley & Sons
Särndal, C.-E., Swensson, B. & Wretman, J. 1992 Model Assisted Survey Sampling. New York City, USA, Springer-Verlag.
Singh, A.C. & Mecatti, F. 2011. Generalized Multiplicity-Adjusted Horvitz-Thompson Estimation as a Unified Approach to Multiple Frame Surveys. *Journal of Official Statistics*, 27(4): 633–650.
Tillé, Y. (2020). Sampling and estimation from finite populations. *John Wiley & Sons*.
Valliant, R., Dorfmann, A.H & Royall, R.M. (2000). Finite Population Sampling and Inference: A Prediction Approach. New York City, USA, John Wiley & Sons.
UNSD 2020. Report on the results of the UNSD survey on 2020 round population and housing censuses. Background document presented at the Fifty-first session of the United Nations Statistical Commission, 3–6 March 2020, New York City, USA (https://unstats.un.org/unsd/statcom/51stsession/documents/BG-Item3j-Survey-E.pdf).
Woodruff, R.S. 1971. A Simple Method for Approximating the Variance of a Complicated Estimate. Journal of the American Statistical Association. 66(334): 411–414.
Verma, V. 2013. *Sampling for household-based surveys of child labour*. Geneva, Switzerland, International Labour Office. (also available at https://www.ilo.org/ipec/ChildlabourstatisticsSIMPOC/Manuals/WCMS_304559/lang--en/index.htm).