Food and Agriculture Organization of the United Nations

SUSTAINABLE DEVELOPMENT GOALS

Guidelines on data disaggregation

for SDG Indicators using survey data

Guidelines on data disaggregation for SDG Indicators using survey data

Food and Agriculture Organization of the United Nations

Rome, 2021

Required citation:

FAO. 2021. *Guidelines on data disaggregation for SDG Indicators using survey data*. Rome. https://doi.org/10.4060/cb3253en

The designations employed and the presentation of material in this information product do not imply the expression of any opinion whatsoever on the part of the Food and Agriculture Organization of the United Nations (FAO) concerning the legal or development status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. The mention of specific companies or products of manufacturers, whether or not these have been patented, does not imply that these have been endorsed or recommended by FAO in preference to others of a similar nature that are not mentioned.

The views expressed in this information product are those of the author(s) and do not necessarily reflect the views or policies of FAO.

ISBN 978-92-5-133942-8

© FAO, 2021

Some rights reserved. This work is made available under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 IGO licence (CC BY-NC-SA 3.0 IGO; https://creativecommons.org/licenses/by-nc-sa/3.0/igo/legalcode).

Under the terms of this licence, this work may be copied, redistributed and adapted for non-commercial purposes, provided that the work is appropriately cited. In any use of this work, there should be no suggestion that FAO endorses any specific organization, products or services. The use of the FAO logo is not permitted. If the work is adapted, then it must be licensed under the same or equivalent Creative Commons licence. If a translation of this work is created, it must include the following disclaimer along with the required citation: "This translation was not created by the Food and Agriculture Organization of the United Nations (FAO). FAO is not responsible for the content or accuracy of this translation. The original [Language] edition shall be the authoritative edition."

Disputes arising under the licence that cannot be settled amicably will be resolved by mediation and arbitration as described in Article 8 of the licence except as otherwise provided herein. The applicable mediation rules will be the mediation rules of the World Intellectual Property Organization http://www.wipo.int/amc/en/mediation/rules and any arbitration will be conducted in accordance with the Arbitration Rules of the United Nations Commission on International Trade Law (UNCITRAL).

Third-party materials. Users wishing to reuse material from this work that is attributed to a third party, such as tables, figures or images, are responsible for determining whether permission is needed for that reuse and for obtaining permission from the copyright holder. The risk of claims resulting from infringement of any third-party-owned component in the work rests solely with the user.

Sales, rights and licensing. FAO information products are available on the FAO website (www.fao.org/publications) and can be purchased through publications-sales@fao.org. Requests for commercial use should be submitted via: www.fao.org/contact-us/licence-request. Queries regarding rights and licensing should be submitted to: copyright@fao.org

Contents

Figures

Tables

Boxes

Foreword

The overarching principle of the 2030 Agenda for Sustainable Development – "leave no one behind" – calls for more granular and disaggregated data than are currently available in most countries, in order to inform the Sustainable Development Goal (SDG) monitoring process.

Since its creation, the Inter-Agency and Expert Group on SDG Indicators (IAEG-SDG), which is tasked with developing and implementing the SDG Global Indicator Framework for the goals and targets of the Agenda 2030, has included work on data disaggregation in its annual activities. Indeed, at the core of the Framework there lies an overarching principle of data disaggregation, 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 Principle of Official Statistics".

Recognizing the fundamental role played by disaggregated data and information, the United Nations Statistical Commission (UNSC), at its Forty-seventh Session, requested the IAEG-SDG to form a working group on data disaggregation, with the objective of strengthening national capacities and developing the necessary statistical standards and tools to produce disaggregated data. The IAEG-SDG responded to this request by creating a dedicated work stream on data disaggregation that led, among other achievements, to the compilation of all categories and dimensions of data disaggregation currently in place or planned by custodian agencies, as well as policy priorities relating to the most vulnerable population groups. This data disaggregation matrix distinguishes between dimensions representing:

- **The minimum set of disaggregation,** including, for each indicator, the disaggregation dimensions specifically mentioned in the target or indicator name and information on the dimensions' categories. For these dimensions, reports are made as to whether data are currently available in the Global SDG Indicators Database and, if not, when data is expected to be produced.
- **Other current disaggregation,** which encompasses any additional data disaggregation beyond that covered in the minimum set for which data are currently available in the database.
- **Future additional disaggregation,** including data disaggregation dimensions and categories mentioned in the metadata for the indicator, but not currently available in the database.

As a member of the working group on data disaggregation, the Food and Agriculture Organization of the United Nations (FAO) has taken numerous steps towards supporting Member Countries in the production of disaggregated estimates. Within this framework, these Guidelines offer methodological and practical guidance for the production of direct and indirect disaggregated 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, including Small Area Estimation methods.

Pietro Gennari
Chief Statistician

Acknowledgements

The Guidelines on data disaggregation for SDG Indicators using survey data were prepared by the Office of Chief Statistician, Food and Agriculture Organization of the United Nations (FAO), under the general direction and encouragement of FAO Chief Statistician Pietro Gennari.

Piero Demetrio Falorsi is the principal author of these Guidelines and supervised the development of the entire publication. The individual chapters were drafted by the following authors:

Chapter 1: Piero Demetrio Falorsi and Ayça Dönmez

Chapter 2: Monica Scannapieco and Piero Demetrio Falorsi

Chapter 3: Piero Demetrio Falorsi

Chapter 4: Ayça Dönmez, Piero Demetrio Falorsi and Sara Viviani

Chapter 5: Ayça Dönmez and Piero Demetrio Falorsi

Chapter 6: Stefano Falorsi.

Additional technical support was generously provided by Carlo Cafiero, who gave valuable advice on setting up the experimental design on data integration presented in Chapters 4 and 5. Marcello D'Orazio made a significant contribution to the first review process and provided useful technical advice. Clara Aida Khalil substantially contributed to the second review process and finalization of the publication, providing valuable technical advice on making the Guidelines more tailored to specific country needs.

The document was edited by Sarah Pasetto.

Abbreviations and acronyms

AIC	Akaike information criterion
CV	Coefficient of variation
EU	European Union
FAO	Food and Agriculture Organization of the United Nations
FIES	Food Insecurity Experience Scale
GV	Global variance
GWP	Gallup World Poll
GWSM	Generalized Weight Share Method
GOF	Goodness of fit
HT	Horvitz-Thompson (estimator)
IAEG-SDGs	Inter-Agency and Expert Group on the Sustainable Development Goal Indicators
IHPS	Integrated Household Panel Survey
HIS	Integrated Household Survey
LSMS	Living Standard Measurement Survey
LSMS-ISA	Living Standards Measurement Study – Integrated Surveys on Agriculture
MOE	Margin of error
MSE	Mean squared error
MGRG	Modified GREG
NSI	National Institute of Statistics
NSO	National Statistical Office (Malawi)
NSS	National Statistical System
OECD	Organization for Economic Co-operation and Development (OECD)
POS	POSt-stratified estimator
PPS	Probability-proportional-to-size
PSU	Primary sampling unit

RSE	Relative standard error
SAE	Small area estimation
SE	Standard error
SDG	Sustainable Development Goal
SSRWOR	Stratified simple random sampling without replacement
TP	Tabulation plan
UN	United Nations
UNECE	United Nations Economic Commission for Europe
USU	Ultimate sampling units
VIF	Variance-inflation factor
WM	Working model

Chapter 1. Introduction

1.1. Background

Disaggregated data are key to reveal differences and disparities that are not usually reflected in broad aggregate figures in a comprehensive way. They are the prerequisite to further analysis that may be required to identify specific concerns, address issues and difficulties faced by various population subgroups, and understand the overall nature of disparities. Recognizing specific localities, households and individuals for priority actions can also guide policy interventions for efficient decision-making.

The 2030 Agenda objectives require more fragmented information to monitor the situation of the most vulnerable, especially in terms of poverty reduction and achieving food security. Accordingly, reliable disaggregated data are essential to monitor commitments under the 2030 Agenda for Sustainable Development and the overall goal of "Leaving no one behind." Tracking the Sustainable Development Goal (SDG) indicators is highly instrumental in assessing whether progress towards this core principle is being achieved.

During the Forty-seventh Session of the United Nations Statistical Commission, improving data disaggregation by developing necessary statistical standards and tools was agreed to be fundamental for the full implementation of the indicator framework (UNSC, 2016). In this regard, a working group on data disaggregation within the Inter-Agency and Expert Group on the Sustainable Development Goal Indicators (IAEG-SDGs) was established.[1] As part of its work, the Group compiled all data disaggregation dimensions and categories for the global SDG indicator framework, after consulting all custodian agencies on disaggregation dimensions (IAEG-SDG, 2019). The resulting disaggregation matrix classifies all of these dimensions into three categories (UNSD, 2019):

1. *minimum set of disaggregation*: disaggregation dimensions specifically mentioned in the target or indicator name and information on the categories, whether data are already available in the database and if not, when these disaggregated data are expected to be produced;
2. *other current disaggregation*: any additional data disaggregation dimensions beyond those included in the minimum set for which data are available in the database; and
3. *future additional disaggregation*: data disaggregation dimensions and categories mentioned in the metadata for the indicator, but not currently included in the database.

The Food and Agriculture Organization of the United Nations (FAO) is dedicated to collecting, analysing, interpreting, and disseminating sound and timely statistics, which are key to inform decisions, policies and investments that tackle issues related to food and agriculture. Such issues range from hunger and malnutrition to rural poverty, and from food systems productivity to the sustainable use of natural resources or to climate change. This is why developing and implementing methodologies and standards to assist countries in generating sound data and information is at the core of FAO's statistical work. The development of disaggregation techniques for food and agriculture statistics is an important driver of the success of FAO's statistical work as well as of FAO's overall mission.

In addition, FAO is the custodian United Nations (UN) agency for 21 SDG indicators and is a contributing agency for an additional five, which all lead to supporting countries' efforts in monitoring the 2030

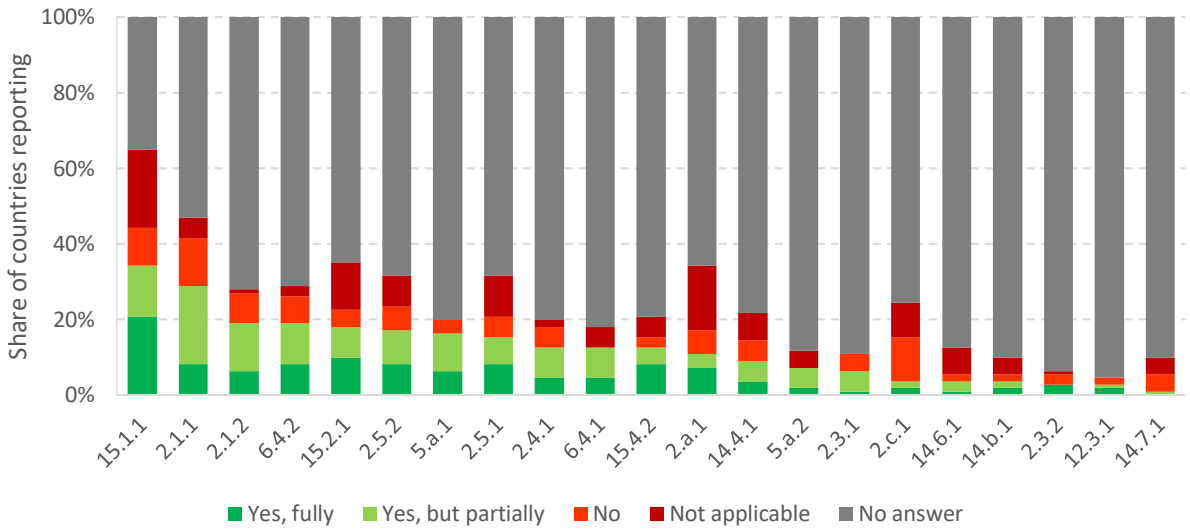
[1] For more details, visit the IAEG-SDGs Data Disaggregation for the SDG Indicators.

Agenda. Disaggregation of FAO-relevant SDG indicators is fundamental to make sure that required policies and plans are formulated and resources are spent on the areas and people where they are most needed and can have the greatest impact.

According to the results of the *Statistical Capacity Assessment for the FAO-Relevant SDG Indicators survey* conducted by FAO in 2018/2019 (FAO, 2019) (Figure 1), many countries do not publish the key SDG indicators for agriculture and food at the required level of disaggregation. Of the 111 countries that responded to the survey, less than 10 percent (mostly in Europe) are able to publish the majority of these indicators in a fully disaggregated manner.[2]

The use of traditional survey tools and sampling methods impose limitations on the production of disaggregated data and of relevant reliable estimates for small population groups or geographical areas. As a result, the data cannot drive the transformative changes required to achieve sustainable development, nor shed light on the situation of the most vulnerable groups and thus "leave no one behind". Innovative techniques that could address some of these issues are far from being mainstreamed.

Figure 1.1 Availability of disaggregated data by SDG indicator



Source: FAO, 2019.

[2] Only 6 percent reported that SDG Indicator 2.1.2 is available at the required level of disaggregation, while less than 20 percent reported having fully or partially disaggregated data.

1.2. Scope of these Guidelines

1.2.1. The main focus of these Guidelines

Ensuring the sustainability of the initiatives is one of the crucial factors in producing disaggregated data of good quality and that is regularly updated. In other words, National Statistical Systems (NSSs) and international organizations – FAO, the World Bank, Eurostat, the Organization for Economic Co-operation and Development (OECD), etc. – engaged in the production and dissemination of statistical information for monitoring commitments under the 2030 Agenda for Sustainable Development – should be capable of maintaining the effort required to produce disaggregated information regularly and continuously over time. This requisite makes it clear that various actions to tackle the limitations and obstacles on the regular production of disaggregated data must be implemented.

To prioritize these actions, which range from those engaging a higher strategic level to those having a more in-depth technical intensity, on the basis of their immediate efficacy, it is useful to focus on the large national surveys carried out by NSSs (for instance, agricultural surveys, multipurpose household surveys, and labour force surveys) as well as survey programs conducted by international organizations.

These surveys are generally carried out each year and their information can be immediately available for statistical production.

Although censuses are not the main focus of these Guidelines, they play a fundamental role as a source for data disaggregation. Indeed, censuses can directly provide a great deal of disaggregated information every ten years. Furthermore, they yield the auxiliary information essential for designing the surveys, computing the estimates and validating the results. Census data, properly processed, enables sample surveys to produce up-to-date and good-quality disaggregated information. Therefore, the integrated exploitation of survey and census data is a key aspect in ensuring the success of data disaggregation programmes.

The so-called “new data sources” – those having administrative nature or that have been obtained through electronic devices and different information-gathering channels – fall outside the scope of these Guidelines. Although this kind of data may overwhelmingly dictate which methodological and operational aspects are to be addressed and resolved by the official statisticians of various countries today, this publication focuses on traditional surveys, which allow for the regular production of disaggregated data to be achieved. New data sources merit separate and specific examination in a different technical document.

As the data sources for SDG indicators vary, in these Guidelines, priority is given to the SDGs having national surveys as their recommended data sources. Box 1 lists the FAO-relevant SDG indicators and their disaggregation dimensions reflected in the IAEG-SDGs disaggregation matrix. The other FAO-relevant SDG indicators have spatial relevance, and the indicators related to forestry and water especially can benefit, directly or indirectly, from geospatial information for disaggregation. Methods using only geospatial data and tools, including Geographical Information Systems (GIS), Remote Sensing (RS) and Global Positioning System (GPS), are not considered in these Guidelines. Nevertheless, geographical position, that can be collected by the above smart devices, can be considered a particular survey variable. Therefore, they have been included in this manual.

Box 1.1 Disaggregation matrix for FAO-relevant SDG 2 and SDG 5 indicators

IAEG-SDG data disaggregation matrix for FAO-relevant SDG Indicators 2.1.1, 2.1.2, 2.3.1 and 2.3.2 and 5.a.1:

M	Minimum set of disaggregation	The disaggregation dimensions specifically mentioned in the target or indicator name and information on the categories
O	Other current disaggregation	Any additional data disaggregation dimensions beyond those included in the minimum set for which data are available in the database.
F	Future additional disaggregation	Data disaggregation dimensions and categories mentioned in the metadata for the indicator, but not currently included in the database

Indicator	Gender	Age	Geographical location (urban/rural)	Other geographical location – Sub-national (e.g. province)	Income /economic status/poor and vulnerable	Ethnicity (indigenous)	Education level	Type of enterprise (farming /pastoral /forestry /fisheries)	Size of enterprise (small/ medium/ large)	Agroecological zone (climate variables / type of soil / geomorphology)	Type of products (crop / livestock / mixed)	Agricultural holding type (household / non household)	Water management	Type of tenure (customary / freehold / leasehold / other)	Type of legally recognized document
2.1.1	M	M	F	F	M										
2.1.2	M	M	F	F	M		F								
2.3.1	M	F		O		M		M	M	O					
2.3.2	M	F		O		M		M	M	O					
2.4.1				F							F	F	F		
5.a.1	M	F	F		F	F								M	F

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 food insecurity experience scale (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 2.4.1: Proportion of agricultural area under productive and sustainable agriculture

SDG Indicator 5.a.1: (a) Proportion of total agricultural population with ownership or secure rights over agricultural land, by sex; and (b) share of women among owners or rights-bearers of agricultural land, by type of tenure

1.2.2. Some useful concepts

This section introduces general definitions that are essential to clearly explain the structure of these Guidelines.

In a sample-survey context, the estimator of the parameter of interest for a given subpopulation is said to be a direct estimator when it is based only on sample information from the subpopulation itself. Unfortunately, for most surveys, the sample size is not large enough to guarantee reliable direct estimates for all subpopulations. A "small area" or "small domain" is any subpopulation for which a direct estimator with the required precision is not available. In the relevant literature, small area is intended as a general concept, and is used to indicate a general partition of the population according to geographical criteria or other structural characteristics (e.g. sociodemographic variables for household surveys or economic variables for business surveys).

When direct estimates cannot be disseminated because they are of unsatisfactory quality, an ad hoc class of methods, called small area estimation (SAE) methods, is available to overcome the problem (see Rao, 2003; Pfeffermann, 2002, 2013). These methods are usually referred to as indirect estimators as they cope with poor information for each domain, borrowing strength from the sample information belonging either to other domains or to previous survey occasions, resulting in an increase in the effective sample size for each small area.

Large-scale surveys are usually aimed at providing estimates of target parameters for the whole population, as well as for relevant subpopulations defined at the sampling stage. Design-consistent and design-unbiased direct estimates are produced for the parameters of interest. However, in most surveys, the sample size is not large enough to guarantee reliable estimates for all target subpopulations.

1.2.3. Summary of the Guidelines

Chapter 2 classifies the actions useful for data disaggregation into four main pillars, which range from those engaging a higher strategic level to those having a more in-depth technical intensity. Furthermore, the need to maintain a holistic view, implementing the actions within the context of a strategic framework taking together the different activities regarding which various institutional subjects (NSSs and international organizations) can fruitfully cooperate, is considered.

Chapter 3 illustrates the actions intended to define sample strategies for direct domain sampling estimates. The chapter further explores possible approaches to estimation that leverage various uses of auxiliary information. Furthermore, it proposes sampling designs that guarantee an observed set of sampling units for every subpopulation or domain for which disaggregated data must be produced. Thus, it would be possible to calculate the direct estimates. However, traditional sampling techniques present some issues when dealing with populations that are hard to reach (such as nomadic populations) or elusive.

The chapter first introduces the basic notation and describes the sampling approaches traditionally used to deal with disaggregated data. The construction of estimators is discussed. Then, the chapter briefly illustrates specific and innovative sampling methods that overcome the problems associated with traditional techniques and can be adopted to ensure the planned sample sizes, as well as to introduce a controlled measure of accuracy in disseminating disaggregated statistics for the relevant domains.

Chapter 4 details the methods of measuring sampling accuracy. This is a specific step that represents a prerequisite to implementing any action for data disaggregation. To derive, estimate and disseminate sampling errors of disaggregated data boosts confidence in and the transparency of NSSs. This allows data users to evaluate the fit and accuracy of the estimates for their use. Moreover, large sampling errors in direct estimates make evident the need to either initiate development of an estimation strategy based on small area techniques or to revisit the sampling design, in order to guarantee the desired level of error for the direct estimates.

Chapter 5 illustrates a useful approach to the integrated use of two independent surveys. This approach allows leveraging both a small survey, to measure a specific phenomenon precisely with a small measurement error, and from a more extensive survey, to produce cross-tabulation at a disaggregated level.

Chapter 6 provides information on the use of SAE techniques, strictly linked to forthcoming Guidelines currently being developed by a specific UN-level task force. The adoption of SAE methods is one of the possible approaches to deal with disaggregation when direct estimates cannot be computed at the required precision level. Since SAE methods are heavily based on models, and the validation of those models can be challenging, their implementation may not be straightforward.

The last section of each chapter summarizes the main findings and recommendations of the chapter. The information on the software and functions used are presented in Annex 1 at the end of these Guidelines.

1.2.4. Focusing on a specific application

To be effective, the Guidelines focus on a single case study, but the methods and approaches discussed are of much more general applicability. Focusing on a real problem helps to highlight the various issues that can arise in practical situations.

In particular, the focus is on SDG Indicator 2.1.2, on prevalence of moderate or severe food insecurity in the population, based on the Food Insecurity Experience Scale (FIES).

Disaggregated data is especially important in monitoring food insecurity and malnutrition. To ensure regular access to nutritious and sufficient food, and thus reductions in food insecurity, detailed and disaggregated information by age, gender and location and for disadvantaged population groups is required to target priority efforts. For this purpose, these Guidelines prioritize the study of food insecurity indicators and their disaggregation.

SDG Indicator 2.1.2 provides internationally comparable estimates of the proportion of the population facing moderate or severe difficulties in accessing food. The FIES produces a measure of the severity of food insecurity experienced by individuals or households, based on direct interviews. The FIES Survey Module (FIES-SM) is composed of eight questions with simple dichotomous responses (yes/no). Respondents are asked whether, at any time during a certain reference period, they have worried about their ability to obtain enough food, their household has run out of food, or if they have been forced to compromise on the quality or quantity of the food they ate due to limited availability of money or other resources to obtain food.

Food insecurity at moderate levels is typically associated with the inability to regularly eat a healthy, balanced diet. Severe levels of food insecurity, on the other hand, imply a high probability of reduced food intake and can therefore lead to more severe forms of undernutrition, including hunger. FAO (2020a)

reported that in 2019, 2 billion people, or 25.9 percent of the global population, experienced hunger or did not have regular access to nutritious and sufficient food.

The data source for most FAO Members is the Gallup World Poll (GWP). In addition, FIES-compatible data from official national surveys are already available for some countries (Brazil, Canada, Ecuador, Guatemala, Mexico, Seychelles and the United States of America). Moreover, since 2015, the FIES has been included in official surveys in Burkina Faso, Indonesia, Kenya, Pakistan and Saint Lucia. When food-insecurity prevalence estimates are based on FIES data collected in the GWP, the national sample size is usually of approximately 1 000 individuals.

As of 2018, data for SDG Indicator 2.1.2 are available for over 150 countries, from 2014 to 2018. In making the global assessment, preference is given to suitable and reliable FIES data available from large national surveys, whereas FAO data collected in the GWP are used to compile the estimates for countries for which there is no other data and/or to fill gaps in terms of time series (FAO, 2020a).

Table 1 gives a snapshot of the disaggregation matrix for SDG Indicator 2.1.2, on the prevalence of moderate or severe food insecurity in the population, based on FIES. Overall, although dimensions for disaggregation by age, sex and belonging to poor and vulnerable groups are listed in the minimum set of disaggregation, the disaggregated data are not currently available in the UN Global SDG Indicators Database, while for data collected by FAO through the GWP, indicator 2.1.2 could already be disaggregated by sex and, partially, by age, disaggregation for poor/vulnerable group is considered unfeasible with available data sources. The dimensions of geographical location and education level are reported to be available in the future. The FIES indicator does not cover any other dimensions classified as other current disaggregation.

To reduce the impact of year-to-year sampling variability, country-level estimates are presented as three-year averages, computed as the averages of all available years in the considered triennia. The data have been subject to a validation process, and only results validated by national statistical offices are published at the country level.

Table 1.1. Data disaggregation dimensions and categories for SDG Indicator 2.1.2

SDG Indicator 2.1.2 – Prevalence of moderate or severe food insecurity in the population, based on the food insecurity experience scale (FIES)		
Minimum set of disaggregation	**Minimum required disaggregation dimension**	1. Poor and vulnerable population 2. Age 3. Sex
	Minimum required disaggregation dimension available in Global SDG Database (Yes/No)	1. No 2. No 3. No
	Disaggregation category of minimum required disaggregation dimension	1. Income decile 2. Age 3. Female/Male
	If minimum required disaggregation dimension not currently produced, when will it be produced?	1. Not currently feasible 2+3. For data collected by FAO through GWP, the current indicator can already be disaggregated by sex and, partially, by age (only between classes of over and under 15 years of age), by computing the percentage of men and women, and of people in each of the two broad classes of living in households that are classified as moderately or severely food-insecure.
Future additional disaggregation	**Future additional disaggregation dimensions planned for indicator**	1. Geographical 2. Education
	Disaggregation categories of future additional disaggregation dimensions	1.1 Urban/rural 1.2. Subnational (e.g. province)
	When will these future additional disaggregation dimensions be produced?	Can already be produced for countries where FIES or compatible data is available from population surveys that are representative of the population in subnational areas and contain the relevant information at household/individual level.

Chapter 2. A strategic plan for data disaggregation

2.1. Why is a strategic plan for data disaggregation necessary?

The FAO results framework clearly specifies the importance of FAO's statistical work through its strategic objectives and cross-cutting technical objectives. In order to improve governance of the relationships between FAO and Member Countries in relation to the specific objective of data disaggregation, it is advisable to establish a dedicated strategic plan that is shared between FAO, the Member Countries and other international statistical organizations.

The ultimate objective of all actions carried out for data disaggregation is to enable NSSs to regularly produce and disseminate data (including on the SDG indicators) at a more detailed level and, eventually, to improve their decision-making processes. The main subjects in charge of the development of this objective are the NSSs, as well as the international organizations (FAO, the World Bank, Eurostat, OECD, etc.) engaged in the production and dissemination of the statistical information required to monitor commitments under the 2030 Agenda for Sustainable Development.

The sustainability of producing disaggregated data is crucial for the overall success of the initiative. In other words, the NSSs should be able to maintain the efforts required to produce new information regularly. This leads to the need to set up processes for various actions, ranging from those engaging a higher strategic level to those having a more in-depth technical intensity.

To achieve this objective, it is necessary to maintain a holistic view, implementing the required actions within the context of a strategic framework bringing together the different activities regarding which the various institutional subjects involved (NSSs, international organizations, etc.) can fruitfully cooperate. To this end, the multiple actors involved in the actions should share and agree on the same vision and accept to be responsible, in cooperation with others, for specific tasks. Thus, it is necessary to formally establish a strategic plan for data disaggregation, on which the various actors can fruitfully cooperate.

2.2. The four pillars of the strategic plan

The strategic plan for data disaggregation should flexibly leverage actions based on the integrated use of various approaches, statistical methodologies and tools that are useful for different phases of the statistical production chain.

These actions can be classified under the following four main pillars.

1. *Actions at the strategic level*, that establish the strategic choices for the data disaggregation. These choices are the drivers of activities conducted at the technical level. An example of strategic choice is the selection of dimensions (or domains) relevant to a specific indicator.
2. *Actions on sampling design*, aimed at defining the sample designs that guarantee the production of most of the disaggregated estimates with controlled quality, for relevant domains.
3. *Actions at direct estimation level*, that (i) measure sampling accuracy, and (ii) improve the quality of direct estimates, even defining auxiliary variables that can be used for both benchmarking sampling estimates and correcting sampling non-response.
4. *Actions at indirect estimation level*, to be implemented when direct estimates perform poorly. The SAE models fall under these actions.

2.3. Relationships among the pillars

The actions of the pillars are closely interrelated and can positively influence one another, as illustrated in Box 2.

Box 2.1 Examples of relationships among the pillars

- **Pillar 1**. An example of strategic action is defining the disaggregation domains for which a planned sample size is necessary. At a higher institutional level, another example could be the decision to use standard classifications and comparable definitions for surveys implemented by the NSS.
- **Pillar 2**. Based on the previous step, the sampling statisticians may define proper sampling designs, ensuring planned sample sizes for the disaggregation domains defined at the strategic level.
- **Pillar 3**. Thus, it would be possible to calculate direct estimates and assess their precision. Moreover, the indirect estimates (Pillar 4) could also benefit from having sampling units in each domain of interest. Indeed, in this way, the statisticians may decide to set up models at the domain level, enabling a substantial reduction of the model bias.
- **Pillar 3**. On the basis of the above actions, it is even possible to derive estimates and disseminate sampling errors of disaggregated data. This action boosts trust in and transparency of the NSSs. Moreover, the action allows data users to evaluate the fit and accuracy of the estimates for their specific use.
- **Pillar 4.** Large sampling errors affecting the direct estimates highlight the need to either initiate development of an estimation strategy based on SAE techniques or to revisit the sampling design (Pillar 2), to guarantee the desired level of error for the direct estimates.
- **Pillars 1, 2, 3 and 4**. The availability of a common set of definitions, metadata and classifications reinforces the overall coherence of the information system at country level, as all the disseminated statistics (computed with direct or indirect methods) must benchmark with the known totals.

The sequence of the pillars, from Pillar 1 to Pillar 4, is an approach whereby which statistical activity is a coherent and ordered streamline in which, first, action on a strategic level is determined. Then, the sampling designs are defined. The direct estimates are derived and, if need be, SAE techniques are leveraged. This approach is rational and, in the long term, it allows for obtaining maximum efficiency while at the same time ensuring the feasibility of the different activities.

However, a demand for detailed statistics for certain urgent and unforeseen need can often require the actions to be carried out in the reverse order to that given above. In these cases, disaggregated data can be produced using whatever information is available. Thus, it is possible to proceed either with a direct estimation (if all domains of the disaggregation have sample units) or with the SAE techniques (if some of the subpopulations have a null, or minimal, sample size). In these situations, too, after production of the

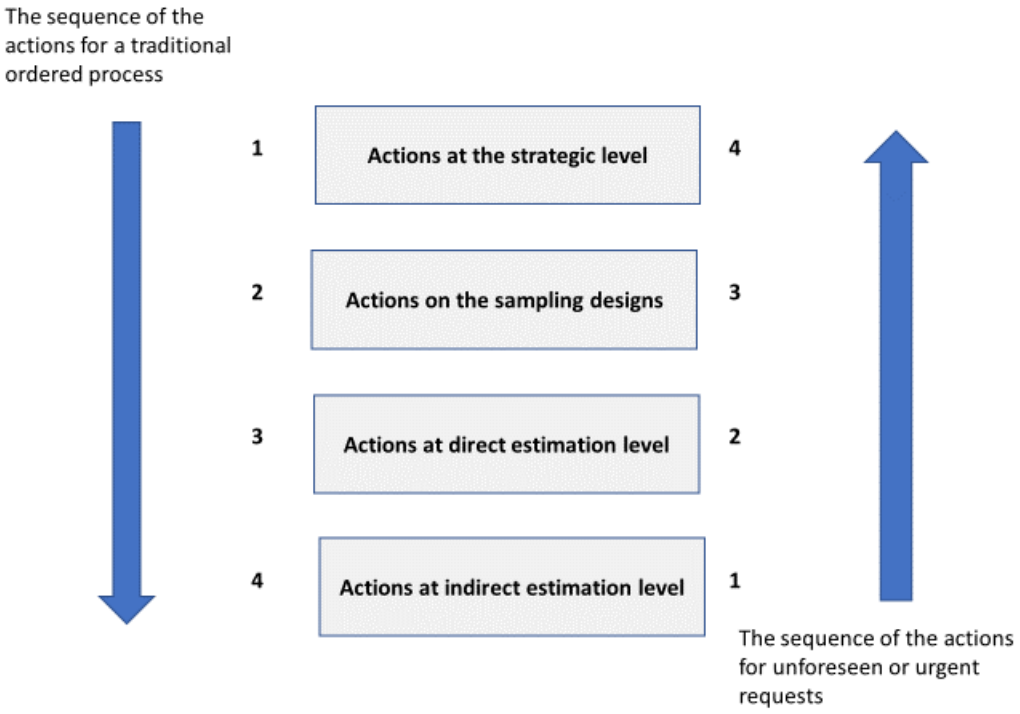
requested statistics, it is possible to proceed in the reverse order and consider intervening on the statistical plans and sampling designs because it is necessary to produce regular statistics. Briefly,

- the suggested list of pillars (or phases) should not be taken as a fixed set of steps to be followed in a rigid and specific order:
- certain practical situations, such as the implementation of the SDG monitoring agenda, may require omitting some of the abovementioned steps and producing estimates at a given disaggregation level using any data source available; and
- The production of direct and indirect estimates with the data sources available may reveal a need to return to the design stage and revisit sampling designs or to develop new data collection approaches.

This antinomy as to the sequence of the actions is reproduced in the picture below.

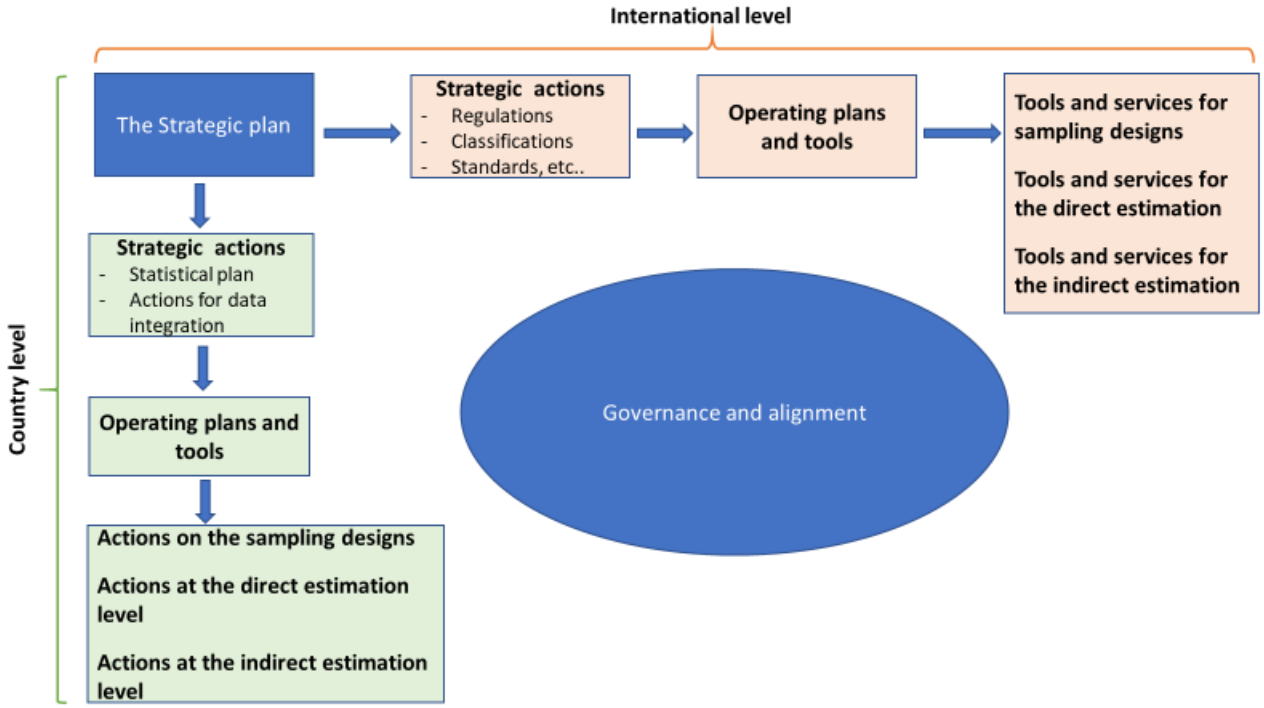
As there can be several different scenarios for data disaggregation, the various actions of the strategic plan must be developed at the same time.

Figure 2.1. Sequence of the actions, with different scenarios



Source: FAO, 2020.

Figure 2.2 Factors that make the strategic plan effective



Source: FAO, 2020.

2.4. Making the strategic plan effective

To implement this strategic plan, certain crucial factors must be in place. These factors are summarized in Figure 2.2 (above).

The first factor is a **shared vision**. All actors involved should agree on the same vision. The vision is clearly stated in the 2030 Agenda for Sustainable Development and the overall objective of "Leaving no one behind", according to which reliable disaggregated data are essential for monitoring commitments and tracking progress under the SDG indicators.

The second aspect is related to the **levels of the actions**. We identify two main levels: the **international levels** and the specific **country levels**. Closely related to this aspect is that of the **actors** that can carry out the actions. The actors at the international level are international organizations (UN, FAO, etc.) that play a relevant role in the production of SDG indicators or monitor progress towards the SDGs (for instance, Eurostat in the European Union [EU] context), including by fostering international cooperation. The actors in each country are those of each country's NSS: the National Institutes of Statistics (NSIs) as well the ministries involved in the production of SDG indicators.

Each actor must carry out strategic actions that are the essential drivers of activities conducted at a more in-depth technical level. These are discussed in further detail in Section 2.5. below. Here, suffice it to mention that actions at the international level mainly concern regulations, classifications and standards. In contrast, at the country level, the actions focus on interventions on the country's statistical plan and specific activities for data integration.

The strategic plans and strategic actions activate the technical activities of Pillars 2, 3 and 4. To guarantee the ordered and harmonious conduction of such activities, the actors for each level must set up operating plans that can guide specific developments.

The final aspect that plays a crucial role is the governance and alignment of the strategic plan. This topic is beyond the scope of these Guidelines. It will only be mentioned that it is essential for governance to be clearly stated, and entail the involvement of the main actors.

Finally, it is noted that proper communication of the plans (strategic and operating) is critical to ensure alignment on the relevant changes within the technical staff involved in the actions for data disaggregation.

2.5. Focus on the strategic actions

These actions are twofold.

- The first-level actions aim to harmonize and standardize statistical processes at the international level and within the NSS, delivering new statistical services.
- The second level is more specific, concerning the statistical plan of the country, and seeks to identify the relationships among the different sources of information (among surveys, among surveys and censuses, etc.) to provide the disaggregated information requested.

The two levels are closely related to one another. Indeed, if the data are standardized and harmonized, many activities of the second level can be avoided, as they are the first level's implicit consequences. Even data manipulation can be simplified significantly. This topic will be discussed in further detail at the end of Section 2.5.2 below.

2.5.1. Harmonizing and standardizing the statistical processes

During the Forty-seventh Session of the United Nations Statistical Commission, it was agreed that improving data disaggregation by developing necessary statistical standards and tools is fundamental for the full implementation of the indicator framework (UNSC, 2016).

The ultimate aim of these actions is to streamline the statistical production process, making it more efficient, transparent and widely applicable, which in turn entails a substantial facilitation and an enabling prerequisite for many data disaggregation actions.

The impacts of these actions are twofold. On one hand, they make the statistical data production process similar to an industrial activity, defined by the sequencing of single standardized process steps (data collection, data editing, etc.) that are repeatable and of certified quality. This approach makes it much simpler to set up new processes that may be required to produce disaggregated data. Indeed, a new step can be introduced by simply assembling the elementary process steps. On the other hand, standardization enhances the use of shared definitions and metadata in different surveys, thus facilitating the integrated use of various sources of data (as shown in Chapter 5) to produce disaggregated statistics.

The main actors involved in carrying out these kinds of actions are international organizations, which can play a crucial role, including by fostering cooperation among NSSs.

The actions carried out by the United Nations Economic Commission for Europe (UNECE) and Eurostat on the modernization and standardization of statistical processes constitute best practices in this field.

A detailed and complete description of what is currently in place and what is in development in these fields is beyond the scope of these Guidelines. However, an overview of the various initiatives can be found in UNECE (2020) where, it can be seen that initiatives in this area include, among others, the topics mentioned below.

- *Human resources.* This area focuses on the strategic issues of enhancing NSS capabilities by improving the competencies of their staff.
- *Statistical production, methods and information technology*. The aim is to develop standards, guidelines, processes and tools to modernize and improve the efficiency of the statistical process. This topic covers, for instance, innovative services such as applications of machine learning and artificial intelligence to official statistics, and innovative statistical approaches such as the Enterprise Architecture. See, for example, the CSPA Service Catalogue, which provides information on shareable statistical services (CSPA, 2018); it is hosted by Eurostat and is publicly accessible.
- *Data collection and data sources.* These actions foster the integration of big data into statistical processes. (For an inventory of some ongoing projects, see: UNECE Big Data Inventory Home).
- *Standards and metadata*. The use of standards ensures that common definitions and processes are used within and between statistical organizations, helping to remove barriers to collaboration on technical projects, fostering the sharing of knowledge and experiences, and serving as the basis for streamlined statistical production. Projects in this area include, in particular, standards for metadata, such as statistical classifications. Efficient use and sharing of data rely on metadata to guarantee that everyone has the same understanding of the information and processes employed in producing official statistics.

In Eurostat, several current and future investments aimed at innovating the production of official statistics are being made, at both national and European level. These include the following.

- *New data sources*, which increase opportunities for the production of timely and/or more detailed statistics at the spatial or temporal level, as well as statistics on new topics. The Bucharest Memorandum on Official Statistics in a Datafied Society (Trusted Smart Statistics), adopted on 12 October 2018 by the presidents of the European NSIs, indicates the strategic direction that NSIs must take with reference to the use of big data sources and the production of smart statistics. Such a direction has a great impact on NSIs, as remarked under Point 3 of the Memorandum: “the variety of new data sources, computational paradigms and tools will require amendments to the statistical business architecture, processes, production models, IT infrastructures, methodological and quality frameworks, and the corresponding governance structures, and therefore invite the ESS to formally outline and assess such amendments.”
- *Consumer demand*. People increasingly expect data to be able to provide easy and accessible answers to their questions. They know they can access a variety of sources and official statistics must be able to convey quality and trustworthiness to them. Open data and citizen science demand specific innovation investments on the part of NSIs.
- *Technology advancement*. Innovation can be introduced in several fields of interest for NSIs. Examples of such innovation include cloud computing to optimize investments on IT capabilities, smart devices as supports to smart surveys, and artificial intelligence. In particular, artificial intelligence, recently recognized as “one of the most important applications of the data economy” (European Commission, 2020), is of utmost importance for NSIs. Building artificial intelligence capabilities can enable several innovative applications, including the use of new data source types (such as images and textual data), new interaction paradigms with respondents and users of official

statistics (for example, statistical chatbots supporting data collection and data dissemination), and the ability to process big data (such as massive sensor data).

FAO is particularly active in the field of statistical standards. In this regard, it is necessary to mention the work produced and disseminated for the Global Strategy to improve Agricultural and Rural Statistics (World Bank, FAO & UN, 2011), which is now in the phase of regional implementation. Particularly relevant to the topic of data disaggregation are the Technical Report and Guidelines on the Integrated Survey Framework (FAO, 2014 and 2015).

FAO has also played an indispensable role in the Census on Agriculture (FAO, 2020b and 2020c), which in many countries, is the main statistical source for data disaggregation.

2.5.2. Specific actions on the statistical plan at the country level

The aim of these actions is to intervene on the statistical plan of a country to introducing specific modifications that can be useful for data disaggregation.

The NSSs are the main actors carrying out these actions.

However, international organizations implementing specific survey programs can also be involved in the actions. Example of these programs are the GWP, the World Bank's Living Standard Measurement Study (LSMS), and the 50x2030 Initiative.

The two main strategic choices resulting from this activity are the following:

1. establishing that some surveys should regularly tabulate data at a more detailed level, for some domains; and
2. reviewing specific classifications to ensure that statistics can be computed considering standard categories of a disaggregation variable. For instance, if seeking produce urban/rural data disaggregation, the variable should be collected and recorded in the data sets with the survey data.

Both decisions above are the drivers of other activities conducted at the technical level. Various methodological solutions, all of which are reasonable, can be leveraged. For instance, let us consider point 1 of the above list. To ensure production of disaggregated data from a survey having a sample size that is too small for certain domains, it is possible, alternatively, to:

- maintain all characteristics of the sampling design, but oversize the overall sample size to ensure sufficient sample sizes at the domain level;
- ensure planned sample sizes (at the domain level) with a marginal stratification sampling design (see Chapter 4);
- maintain all characteristics of the sampling designs and achieve the objective with special SAE techniques (see Chapter 6); or
- adopt other solutions, based for instance, on the use of administrative records.

In addition to the actions listed above, interventions on the statistical plan may occur at a more complex level, fostering the integrated use of different probabilistic surveys. For instance, the decision might be to:

- gather information on a complex phenomenon with a small study characterized by a small measurement error;
- study, on the survey data, a regression model, linking the study variable with some auxiliary variables;
- predict the study variable on the units of a more extensive scale survey, that has the same auxiliary variables considered in the regression model; or
- produce the disaggregated tabulation, from the more extensive survey.

This strategy (illustrated in detail in Chapter 5) leverages the advantages of both the small survey (with a small measurement error) and the larger one (with the variables useful for the data disaggregation). It requires a strategic choice concerning the "integrated use of the two surveys." Moreover, the process outlined above requires the two surveys to share the same set of auxiliary variables. Specific technical work should follow the strategic decision (e.g. for fine-tuning the regression model).

None of the above solutions is cost-free. Oversize the sample entails all the costs associated with a more extensive data collection effort. However, the choice to adopt SAE techniques also involves expenses relating to the statisticians' need to set up proper small area models; moreover, additional data sources should be available, to fit effective models. All innovative solutions can affect the current organization of a survey, and this may encounter opposition by the persons in charge of specific statistical activities.

Thus, the relevance of strategic decisions to launching innovative technical activities (that can be expensive) becomes clear.

Furthermore, the strategic choices allow for overcoming some of the organizational barriers to the dissemination of disaggregated data. In this regard, the relevance of setting up proper communication activities (Barcaroli *et al.*, 2015), to ensure alignment on change throughout the organization, is noted.

Finally, both the strategic choice and the consequent technical activities can leverage the actions of harmonization and standardization illustrated in Section 2.5.1. Indeed, data manipulation can be simplified significantly if the surveys share the same metadata, classifications and definition and treatment of variables.

2.6 Chapter wrap-up and main recommendations

The main advice provided in this chapter is the following.

1. The data disaggregation must be sustainable as a regular statistical activity.
2. It is useful to establish a strategic plan for data disaggregation. This plan should consider the different activities regarding which various institutional subjects (at the international and national level) can fruitfully cooperate, sharing and agreeing on the same vision.
3. The actions of the strategic plan for data disaggregation range from those engaging a higher strategic level to those having a more in-depth technical intensity. The strategic activities and those of deeper technical intensity are closely related and can positively influence one another.
4. Together with the strategic plan, each country should intervene on its national statistical plan to define strategic choices that are essential for launching innovative technical activities and overcoming some of the organizational barriers that can hinder the production and dissemination of disaggregated data.
5. Proper communication of plans is essential to ensure alignment on change among all technical staff involved in the actions for data disaggregation.

Chapter 3. Direct sampling strategies for data disaggregation

3.1. Introduction

This chapter is twofold. On one hand, it illustrates the basic theory for sampling and direct estimation, focusing on disaggregated data. It describes the most common domain estimators, discussing their inferential properties related to the use of available domain information and sampling size. The pros and cons of each estimation option are extensively discussed, along with their applicability according with the specific context.

On the other hand, the chapter reviews the main actions that countries or international organizations can adopt regarding their surveys' sampling designs. These actions are intended to define sample designs that guarantee an observed set of sampling units for each subpopulation or domain for which disaggregated data must be produced.

Proper sampling designs for data disaggregation should ensure planned sample sizes for the domains of the disaggregation plan. This would allow for computing direct estimates. Furthermore, as illustrated in Chapter 6, having sampling units in each domain of interest would also benefit the computation of indirect estimators by enabling substantial reduction of 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 this case, the main issue is the extent of oversampling to employ when survey estimates are required for several domains and for the total population. 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, location sampling, panel surveys, and the use of multi-purpose surveys. 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). The relative size of the subpopulation is a key factor. Kish (1987) proposed a classification of major domains comprising approximately 10 percent or more of the total population, for which a general sample will usually produce reliable estimates; minor domains of 1 to 10 percent, for which the special sampling methods illustrated below in Sections 3.4 and 3.5 are required; mini-domains of 0.1 to 1 percent, estimates that mostly require the use of statistical models; and rare types comprising less than 0.01 percent of the population, that generally cannot be handled by survey sampling methods. Many surveys aim to produce estimates for some major domains, some minor domains and occasionally even some mini-domains.

Issues also arise with populations that are hard to reach (such as nomadic populations) or elusive**.** Verma (2013) gives a clear definition of the problem: "by elusive populations we mean populations for which – by virtue of their characteristics, or of the lack of suitable sampling frames, or difficulties in obtaining the required information – adequate samples cannot be defined, drawn or implemented using the normal procedures of general population sampling". Another issue that exacerbates the usual problems of coverage errors to which almost all sample surveys are subject is "a fundamental vagueness in the

definition of the population, beyond the usual problems relating to its precise demarcation in content, space" (Verma, 2013). The latter point would call for clearer regulations and better definitions. However, there is room for improvement, even leveraging on the sampling methodologies. The relevance of this problem as regards data disaggregation is highlighted by Indicators 2.3.1 (Volume of production per labour unit by classes of farming/pastoral/forestry enterprise size) and 2.3.2 (Average income of small-scale food producers, by sex and indigenous status), which should be disaggregated by, among other things, type of enterprise (Farming/Pastoral/Forestry/Fishery). Now, for these two indicators, some of the main problem 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, Righi and Lavallée 2019).

In this chapter, after a brief introduction of the basic sampling and theory in Section 3.2, the basic theory on sampling and estimation is expounded. Section 3.3 illustrates the direct estimation approaches to compute domain estimates. Section 3.4 sets out the traditional sampling approaches to ensure sizeable sample sizes for the subgroups of the population of interest, in the context of producing disaggregated indicators. Sections 3.5 and 3.6 illustrate new methods for sampling (marginal stratification sampling and indirect sampling) that allow some of the issues characterizing the traditional techniques to be partially overcome. Section 3.7 summarizes the chapter's main recommendations.

3.2. Basic theory on sampling and estimation

The content of this section is somewhat technical; however, it is necessary to introduce the main concepts. For non-technical readers, this section's main contribution is given in Equation 3.2, where it is stated that the direct sample estimate of a total for a particular domain is obtained by summing, over the domain units, the products of data (the y values) and the weights (the ω values). Section 3.3.1 and 3.3.2 will show how weights are derived under different assumptions. To simplify, the basic theory considering the estimation of a total is introduced here. However, Section 3.3.3 contains a brief illustration of how the basic theory can be easily extended to parameters different from the totals (e.g. mean values or ratios).

Basic notation and parameter of interest

Let us consider, as the parameter of interest, a domain total Y_d, where d indicates a generic domain for disaggregation ($d = 1, \dots, D$).

Let *U* be a target population of size *N* and let y_i indicate the value of the target variable $\mathcal{y}$ of the $i-th$ unit of *U*. In this text, $\mathcal{y}$ can denote either a real scaled variable or a dichotomous variable. In the latter case, which is relevant for frequencies, y_i assumes a value of 1 if the unit i has a specific characteristic of interest and 0 otherwise. In the case of a qualitative variable that assumes several modalities (e.g. employment with the modalities of Occupied, Not occupied and Not-labour force), a specific modality is considered: y_i assumes a value of 1 if the unit i presents that specific category and 0 otherwise.

Let U_d be a particular subpopulation of *U* (being $U_d \in U$) of size N_d, for which disaggregated data must be produced (e.g. U_d denotes a specific gender). In the text below, U_d can also be indicated as a disaggregation domain (or more simply a domain) or as a subpopulation.

Let Y_d be the total of the y_i values for the domain U_d:

$$Y_d = \sum_{i=1}^{N_d} y_i = \sum_{i=1}^{N} y_i \gamma_{di} \quad (3.1)$$

in which γ_{di} denotes the domain membership variable, being $\gamma_{di} = 1$ if $i \in U_d$ and $\gamma_{di} = 0$, otherwise.

Sampling

Let S be a specific sample that is a subset of units of U, and let $\mathbb{S}$ be the set of all possible samples that can be randomly selected from U. A sampling design, P, is a function that assigns a given probability, $P(S)$, to the sample S to be randomly selected, with

$$\sum\nolimits_{S \in \mathbb{S}} P(S) = 1.$$

Let π_i (for $i = 1, \dots, N$) be the inclusion probability of the unit i, which indicates the probability of the unit being included in a sample generated by sampling design *P.* The probability π_i is obtained as the expected value, defined over the sample space $\mathbb{S}$, of the sample-membership indicators, λ_i, with $\lambda_i = 1$ if $i \in S$ and $\lambda_i = 0$ otherwise. Thus, it is

$$\pi_i = E_P(\lambda_i) = \sum\nolimits_{S \in \mathbb{S}} P(S)\lambda_i = \sum\nolimits_{S \in \mathbb{S}} P(S \ni i),$$

where $E_P(\cdot)$ denotes the operator of the sampling expectation.

Let

$$n = \sum\nolimits_{i \in U} \lambda_i$$

be the realized size of sample S. Sample design *P* is of fixed size n, if in all sample selections, the sampling size is always equal to the same value n. Let

$$n_d = \sum\nolimits_{i \in U} \lambda_i \gamma_{di} = \sum\nolimits_{i \in S_d} \lambda_i$$

be the realized sample size in the $d-$th domain, where $S_d = S \cap U_d$ indicates the sample in the subpopulation U_d. In most of the empirical situations that characterize the use of sampling surveys for producing disaggregated domains, n_d is a random quantity that can vary from one sample selection to another. This is the case, for instance, when the disaggregation domains are demographic population subgroups (e.g. defined by gender and age class). The sampling expected value of n_d is given by

$$E_P(n_d) = \sum\nolimits_{i \in U} E_P(\lambda_i)\, \gamma_{di} = \sum\nolimits_{i \in U} \pi_i \gamma_{di}.$$

The domain sample size n_d may be fixed or not, depending on the sample design adopted. For instance, if a stratified simple random sampling without replacement (SSRWOR) design is adopted, in which the

domain U_d is obtained as an aggregation of entire strata, and then the domain sampling size is fixed. The domain U_d is said to be planned at the design stage if the sample size n_d is fixed in each sample selection.

That said, in the following paragraphs, a sample S of fixed size n is selected from the population U according to the sample design P, with inclusion probability π_i for $i = (1, \dots, N)$, where the domain U_d can be planned or not.

Sampling design involves a great deal of technicalities, such as stratification or two or more stages (or phases) of selection (Cochrane, 1976), that are beyond the scope of these guidelines. The following box, however, will provide insights on the stratified sampling designs having two or more stages, which are largely used for sampling on households and agricultural holdings. More specifically, according to Grosh and Munoz (1996), the most commonly used method for collecting household data in sub-Saharan Africa is the stratified two-stage sample. This was also confirmed by a more recent review of sampling designs used for agricultural surveys in developing countries performed by the Global Strategy to improve Agriculture and Rural Statistics initiative, according to which 90 percent of countries in Africa and 64 percent in Asia implement multi-stage sampling (Global Strategy, 2018).

Given its importance in concrete sampling applications, this case will be examined several times throughout these Guidelines, generally in specific in-depth boxes.

Box 3.1. Two-stage stratified sampling design

This box briefly introduces the stratified two-stage sampling design, which is largely used for sampling on populations and households. Let *h* $(h = 1, \dots, H)$ denote a generic stratum with M_h primary sampling units (PSUs). The PSU can coincide with, for example, a municipality or a census enumeration area. Let us select m_h PSUs in the stratum (out of the M_h) with varying probabilities and without replacement, $\pi_{1h\ell}$ being the first-stage inclusion of the $\ell - th$ $(\ell = 1, \dots, M_h)$ PSU in the stratum, where

$$\sum_{\ell=1}^{M_h} \pi_{1h\ell} = m_h.$$

Let us consider the $\ell - th$ PSU selected in the stratum having $N_{h\ell}$ ultimate sampling units (for instance, households). From this PSU, the second-stage sample $S_{(h\ell)}$ of $n_{h\ell}$ Ultimate sampling units (USUs) is selected out of the $N_{h\ell}$. These units are included in the sample without replacement and with second-stage equal inclusion probability $\pi_{2h\ell} = n_{h\ell}/N_{h\ell}$. The inclusion probability π_i of the $i - th$ USU $(i = 1, \dots, N_{h\ell})$ is determined by multiplying the *first-stage* inclusion probability of the PSU by the probability of selecting the $i - th$ USU in the second-stage sample of the PSU

$$\pi_i = \pi_{1h\ell}\, \pi_{2h\ell} \text{ for } i \in h\ell.$$

The extension to a three-stage (or more) sampling design is straightforward.

Estimation

In the interests of simplicity, this topic will be illustrated assuming full response to the survey or that the nonresponse is negligible. However, the theory illustrated here can easily be extended to consider nonresponse. Särndal and Lundström (2005) give a detailed overview of the subject.

That said, let $\hat{Y}_d$ be the direct estimate of Y_d. It is obtained by weighting the domain data observed in sample *S*

$$\hat{Y}_d = \sum_{i \in S_d} y_i \omega_i \,, \qquad (3.2)$$

where ω_i are the sampling weights, which allow for computing estimates that are unbiased due to the inferential approach adopted by the statistician. An alternative interesting expression of Equation 3.2 that will be used in some parts of this text, is

$$\hat{Y}_d = \sum_{i=1}^{n_d} y_i \omega_i = \sum_{i=1}^{n} y_i \gamma_{di} \omega_i = \sum_{i=1}^{N} y_i \gamma_{di} \lambda_i \omega_i \,. (3.2a)$$

The weights ω_i are computed differently if the inference is based on the properties of repeated sampling (see Section 3.2.1 below) or the model generating the data (Section 3.2.2). The difference is related to what is considered fixed and random in the inference. In repeated sampling, the y_i values are considered fixed, whereas the only random elements in Equations 3.2 and 3.*2a* are the sample membership indicators λ_i. Conversely, in the model-based approach, the observed sample S is considered fixed, and the randomness is concentrated only on the y_i values. In other words, the model-based approach assumes that the observed values of y_i are generated by a random mechanism formalized by a model. This mechanism is denoted as the model generating the data.

We see from Equation 3.2 that the direct estimate $\hat{Y}_d$ can be computed only if we have some units observed in the sample of the domain. Moreover, if too few units have been observed in the sample, it is not possible to compute quality direct estimates. In this case, the domain estimates can be computed with either the MDIFF, MGREG or MMPE estimators, illustrated in Section 3.2.1, or with special SAE techniques described in Chapter 6. However, these estimates are model-dependent, meaning that they are of good quality only if the model parameters estimated fit well those of the true model generating the data. The parameters can be estimated only by the observed sample data. Thus, if the true model is domain-dependent, the SAE techniques would induce substantial bias in the estimates.

3.3. Direct estimation for the data disaggregation

3.3.1. Repeated sampling

The standard approach: the Horvitz-Thompson (HT) Estimator

The basic estimator of the standard approach in the repeated sampling framework is the well-known Narain Horvitz-Thompson (HT) estimator (Narain, 1951; Horvitz and Thompson, 1952), in which the weights ω_i in (3.2) are given by:

$$\omega_i = a_i = \frac{1}{\pi_i}. \quad (3.3)$$

The resulting HT estimator

$$\hat{Y}_{HT,d} = \sum_{i=1}^{N} y_i \gamma_{di} \lambda_i a_i \qquad (3.3a)$$

is *P*-unbiased, which that means the expected value of the estimator $\hat{Y}_{HT,d}$ (averaging over the sampling space $\mathbb{S}$) equals the target parameter Y_d. Indeed,

$$E_P\left(\hat{Y}_{HT,d}\right) = \sum_{i=1}^{N} y_i \gamma_{di} E_P(\lambda_i) \frac{1}{\pi_i} = \sum_{i=1}^{N} y_i \gamma_{di} \pi_i \frac{1}{\pi_i} = Y_d.$$

Model-assisted survey sampling

The model-assisted survey sampling (Särndal, Swensson and Wretman, 1992) is an evolution of the standard approach in repeated sampling, which allows for leveraging known values of auxiliary variables. Today, it is the predominant approach used in the production of official statistics in developed countries. Here, the inference is based on the sampling design, but leverages the estimator with a working model (WM)[3] which helps make the estimates more efficient (see Comment 3.2 below) and improves the full consistency of the systems of estimates disseminated by the survey. Indeed, the main practical advantage of this approach is that the estimates of auxiliary variables benchmark the known totals, as, for instance, the number of male and females by age groups derived by demographic statistics or the census. This automatically ensures the full consistency of the various estimates with respect to the known totals.

The generalized DIFFerence (DIFF) estimator

To describe this estimator, consider the product variable y_{di}

$$y_{di} = y_i \gamma_{di} = \begin{cases} y_i & \text{if } i \in U_d \\ 0 & \text{if } i \in U_d \end{cases},$$

and suppose that it can be modelled with a WM *M*, according to which

$$y_{di} = m(x_i; \beta_d) + u_i, \quad (3.4)$$

where $m(x_i; \beta_d) = \tilde{y}_i$ is a known function applied on the column vector of auxiliary variables x_i (of the $i-$th unit) and u_i is a random residual, β_d being the unknown column vector of the model parameters that is domain-dependent. It is not necessary to know the full distribution function of the residuals, but only their model expected value $E_M(\cdot)$, variances $V_M(\cdot)$, and covariances $Cov_M(\cdot)$. For a general model $m(x_i; \beta_d)$, we suppose that the model is unbiased and the variances depend on the specific units, that is

$$E_M(u_i) = 0,\ V_M(u_i) \propto c_i\ ,\ Cov_M\left(u_i u_j\right) = 0. \quad (3.5)$$

[3] See Comment 3.2 for a discussion of the role of the WM.

In this chapter, it is supposed that $Cov_M(u_i u_j) = 0$. The more complex SAE models introduced in Chapter 6, relax this constraint, considering models with non-null covariance among different units.

The generalized form of DIFFerence estimator, (DIFF) (Breidt and Opsomer, 2017; Lehtonen and Veijanen, 1998) we propose below is computed in two steps.

1. First, $\hat{\beta}_d$ values are estimated by fitting the model $m(x_i; \beta_d)$ over all couples (x_i, y_{di}) observed in the sample and using the weights a_i given by Equation 3.3. To accomplish this step, it is necessary to have a non-null sample size in the domain.
2. Knowing the x_i values, we may compute the predicted values

$$\hat{y}_{di} = m(x_i; \hat{\beta}_d).$$

Then, we may obtain the DIFF estimator (Breidt and Opsomer, 2017), as

$$\hat{Y}_{DIFF,d} = \sum_{i \in U} \hat{y}_{di} - \sum_{i \in S} (y_{di} - \hat{y}_{di})\, a_i. \quad (3.6)$$

The first addendum on the right side of Equation 3.6,

$$\sum_{i \in U} \hat{y}_{di}$$

is the synthetic part of the estimator, and, for a general form of function m, the values of the auxiliary variables x_i for each unit in the population U must be known to calculate it. The second component on the right side of Equation 3.6,

$$\sum_{i \in S} (y_{di} - \hat{y}_{di})\, a_i,$$

is a weighted sum of the residuals and is computed only on the sample data.

The estimator at Equation 3.6 can be expressed in the weighted form, as given in Equation 3.2, by defining the weights ω_i as the solution of the calibration problem 3.7 below, which minimizes the chi-squared distance between the weights ω_i and a_i subject to the constraints that sampling estimates of the predicted values coincides with the sum of the predictions over the population U:

$$\begin{cases} \sum_{i \in S} \dfrac{c_i (\omega_i - a_i)^2}{a_i} & \text{function to be minimized} \\ \sum_{i \in S} \hat{y}_{di} \omega_i = \sum_{i \in U} \hat{y}_{di} & \text{calibration constraint} \end{cases}. \quad (3.7)$$

We may extend Problem 3.7 to consider the multivariate case. Suppose there are K different target variables $\mathcal{Y}_{(k)}$ $(k = 1, \dots, K)$. Let $y_{(k)i}$ denote the value of the variable $\mathcal{Y}_{(k)}$ of unit i of U and let $\hat{y}_{(k)di} = m(x_i; \hat{\beta}_{(k)d})$ indicate the prediction of the product variable $y_{(k)di} = y_{(k)i} \gamma_{di}$, with $\hat{\beta}_{(k)d}$ being the estimated vector of the parameter model $m(x_i; \beta_{(k)d})$ that links the auxiliary variables x_i to the target variable $y_{(k)di}$. The multivariate version of Problem 3.7 is

$$\begin{cases} \sum_{i\in S} \dfrac{c_i(\omega_i - a_i)^2}{a_i} & \text{function to be minimized} \\ \sum_{i\in S} \hat{y}_{(1)di}\omega_i = \sum_{i\in U} \hat{y}_{(1)di} & 1-th \text{ calibration constraint} \\ \dots & \\ \sum_{i\in S} \hat{y}_{(k)di}\omega_i = \sum_{i\in U} \hat{y}_{(k)di} & k-th \text{ calibration constraint} \\ \dots & \\ \sum_{i\in S} \hat{y}_{(k)di}\omega_i = \sum_{i\in U} \hat{y}_{(k)di} & K-th \text{ calibration constraint} \end{cases} . (3.7a)$$

The problems at Equations 3.7 and 3.7*a* can be solved with the software Regenesees (Istat, Regenessees).

Montanari and Ranalli (2002) show how the estimator can be applied with a general class of regression models, including non-parametric regression and estimators with non-null model covariances among different units.

Generalized REGression estimator

Consider now the Estimator 3.6 and suppose that the simple heteroscedastic linear model

$$m(x_i;\beta_d) = x_i'\beta_d, \qquad (3.8)$$

is adopted, where the model expectations are given by Equation 3.5 and ìx_i' is the transpose of x_i.

Plugging Equation 3.8 into Equation 3.6, we obtain the Generalized REGression (GREG) estimator

$$\hat{Y}_{GREG,d} = X'\hat{\beta}_d + \left(X - \hat{X}_{HT}\right)'\hat{\beta}_d \quad (3.9)$$

where

$$X = \sum_{i\in U} x_i \quad , \quad \hat{X}_{HT} = \sum_{i\in S} x_i a_i,$$

with

$$\hat{\beta}_d = \left(\sum_{j\in S} x_i x_j' \frac{1}{c_j} a_j\right)^{-1} \sum_{j\in S} x_j y_{dj} \frac{1}{c_j} a_j. \quad (3.10)$$

The weights ω_i of Equation 3.2 can be defined explicitly as

$$\omega_i = a_i g_{iS} \qquad (3.11)$$

where

$$g_{iS} = \left[1 + \left(X - \hat{X}_{HT}\right)' \left(\sum\nolimits_{j \in S} x_i x_j' \frac{1}{c_j} a_j\right)^{-1} x_i \frac{1}{c_i}\right] \quad (3.11a)$$

is the calibration-correction factor, with $g_{iS} \cong 1$ for large samples. To use the GREG estimator, it is not necessary to know all auxiliary variables for each unit in the population *U*, but only the x_i values for the sample data and the totals X.

A standard result of the GREG ((Särndal, Swensson and Wretman, 1992, Expression 6.5.16) estimator is that the estimator is calibrated on the totals of the x variables, ensuring that the sampling estimates of the totals of the auxiliary variables reproduce the known population totals X.

If the x_i vector includes the domain membership variable γ_{di}, then the estimates of the number of units in the domain is equal to the known total N_d.

Särndal, Swensson and Wretman (1992, Remark 10.6.2) suggest that one way to avoid the difficulty caused by sample domain counts that are too small is to consider the domain membership variables only for the larger domains. One can aggregate the smaller domains into macro-domains and include a macro-domain membership variable in the vector x_i, thus ensuring that the estimates of the number of units in the macro-domains are equal to the known totals.

Comment 3.1. Consistency of the estimates. The full consistency of the various estimates that can be produced from a given survey is realized if a unique weight ω_i is used (for the i −th unit) for the estimates in which it is applied, irrespective of the domain and the variable.

Comment 3.2. The role of the model. The role of the model can be examined considering (i) the bias and (ii) the variance. Särndal, Swensson and Wretman (1992) demonstrate that the GREG estimator is approximately P −unbiased irrespective of the shape of the finite population scatter. It follows that the estimator is P −unbiased irrespective of whether the assumptions of the model are true or false. On the other hand, the appropriateness of the model is crucial to achieving a small variance. As demonstrated in Särndal, Swensson and Wretman (1992), the more the population scatter conforms to the pattern induced by the predictions of the model, the smaller the population fit residuals $(y_i - \hat{y}_i)$, and the smaller the variance of the GREG estimator.

The POSt-stratified (POS) estimator

The POSt-stratified (POS) estimator is widely applied in social surveys. It ensures that the sampling estimates of the totals of demographic groups defined by age and sex reproduce population totals known from sources external to the survey (e.g. the census or demographic statistics). This characteristic is relevant in the context of these Guidelines because the disaggregation of SDG indicators is often requested for these demographic groups.

Suppose that the population U can be partitioned into B separate non-overlapping groups, U_b $(b = 1, \dots, B)$. Let N_b be the number of units of U_b known from a source external to the survey.

The POSt-stratified (POS) estimator is given by:

$$\hat{Y}_{POS,d} = \sum_{i=1}^{n}\sum_{b=1}^{B} y_i \delta_{bi}\, a_i \frac{N_b}{\hat{N}_{HT,b}}. \quad (3.12)$$

where $\delta_{bi}$is the group membership variable ($\delta_{bi} = 1$, if $i \in U_b$ and $\delta_{bi} = 0$, otherwise), and $\hat{N}_{HT,b}$ is the HT estimate of N_b, with

$$\hat{N}_{HT,b} = \sum_{i=1}^{n} \delta_{bi} a_i.$$

The POS estimator can be viewed as a particular case of the GREG estimator, by defining

$$x_i = (\delta_{1i}, \dots, \delta_{bi}, \dots, \delta_{Bi})' \text{ and } m(x_i; \beta_d) = (\delta_{1i}, \dots, \delta_{bi}, \dots, \delta_{Bi})'\beta_d$$

with

$$c_i = \sum_{b=1}^{B} \delta_{bi}\, \sigma_b^2,$$

with σ_b^2 being the homogenous model variance for the units of U_b.

The domain-specific auxiliary information DIFF estimator and the Domain-specific auxiliary information GREG estimator

In this case, the $\hat{\beta}_d$ values are estimated, using the weights a_i, by fitting the model $m(x_i; \beta_d)$ over all couples $(x_i\gamma_{di}, y_{di})$ observed in the sample S_d. To accomplish this step, it is necessary to have a sufficient sample size in S_d. Then, the domain-specific auxiliary information DIFF (DIFFD) estimator is given by:

$$\hat{Y}_{DDIFF,d} = \sum_{i\in U_d} \hat{y}_{di} + \sum_{i\in S_d} (y_{di} - \hat{y}_{di}) a_i. \quad (3.13)$$

The estimator, expressed in the weighted form, is:

$$\hat{Y}_{DDIFF,d} = \sum_{i\in S_d} y_i \omega_{di}$$

where the final weights, which are domain dependent, are obtained by solving the following calibration problem:

$$\begin{cases} \sum_{i\in S_d} \dfrac{c_i(\omega_{di} - a_i)^2}{a_i} & \text{function to be minimized} \\ \sum_{i\in S_d} \hat{y}_{di}\omega_{di} = \sum_{i\in U_d} \hat{y}_{di} & \text{calibration constraint} \end{cases} . \quad (3.14)$$

The weights ω_{di} now depend on the domain *d*. Therefore, if the domains U_d $(d = 1, \dots, D)$ represent a complete partition of the population U, the estimators $\hat{Y}_{DDIFF,d}$ do not add up to the DIFF estimator referred to the whole population, that is

$$\sum_{d=1}^{D} \hat{Y}_{DDIFF,d} \neq \sum_{i \in S} y_i \omega_i ,$$

where the weights ω_i are defined by Equation 3.7. This may hinder ensuring the consistency of the different estimates of the survey tabulation plan (see Comment 3.1 above). Also, $\hat{Y}_{DDIFF,d}$ is not approximately P −unbiased unless the domain sample size is large. However, the DDIFF estimator will be more efficient than the DIFF estimator if the expected domain-specific sample size is large (Rao, 2003; p. 19).

With a general function $m(x_i; \hat{\beta}_d)$, for the DDIFF estimator, it is necessary to know the domain-membership variables γ_{di} for all the units of U.

However, if we consider Linear WM 3.8 with the model expectations given by Equation 3.5, we have the domain-specific GREG (DGREG) estimator where only the domain totals of the auxiliary variables need to be known:

$$X_d = \sum_{i=1}^{N} x_i \gamma_{di}.$$

For instance, the total X_d may be the domain population counts by age and sex, determined by the census or by demographic statistics.

The sample estimate of β_d is given by:

$$\hat{\beta}_d = \left(\sum_{j \in S_d} x_j x_j' \frac{1}{c_j} a_j \right)^{-1} \sum_{j \in S_d} x_j y_j \frac{1}{c_j} a_j.$$

The DGREG estimator is given by:

$$\hat{Y}_{DGREG,d} = X_d' \hat{\beta}_d + \sum_{i \in S_d} \left[y_{di} - x_i' \hat{\beta}_d \right] a_i = \sum_{i=1}^{n} y_i \omega_{di}, \quad (3.15)$$

where the weights ω_{di} can be defined explicitly as

$$\omega_{di} = a_i \left[1 + \left(X_d - \hat{X}_{HT,d} \right)' \left(\sum_{j \in S_d} x_j x_j' \frac{1}{c_j} a_i \right)^{-1} x_i \gamma_{di} y_{di} \frac{1}{c_i} \right],$$

with

$$\hat{X}_{HT,d} = \sum_{i \in S_d} x_i \gamma_{di} a_i.$$

The Modified DIFF estimator and the Modified GREG estimator

The Modified DIFF (MDIFF) estimator uses y values from outside the domain. In particular, this estimator considers a model

$$y_i = m(x_i; \beta) + u_i$$

that is not domain-dependent. Having obtained an estimate, $\hat{\beta}$, of β, with the data of the full sample S, the MDIFF estimator is given by

$$\hat{Y}_{MDIFF,d} = \sum_{i\in U} m(x_i;\hat{\beta})\gamma_{di} - \sum_{i\in S}[y_{di} - m(x_i;\hat{\beta})\gamma_{di}]\, a_i. \quad (3.16)$$

The MDIFF estimator may be expressed in weighted form as

$$\hat{Y}_{MDIFF,d} = \sum_{i\in S} y_i \omega_{di}$$

where the weights ω_{di} are obtained as the solution of the following calibration problem (Rao, 2003; p. 20):

$$\begin{cases} \sum_{i\in S} \dfrac{c_i(\omega_{di} - a_i\gamma_{di})^2}{a_i} & \text{function to be minimized} \\ \sum_{i\in S} \hat{y}_i\omega_{di} = \sum_{i\in U} m(x_i;\hat{\beta})\gamma_{di} & \text{calibration constraints} \end{cases} . \quad (3.17)$$

$\hat{Y}_{MDIFF,d}$ allows for computing the domain estimate even with a null domain-sample size, since the synthetic part of the estimator, $\sum_{i\in U} m(x_i;\hat{\beta})\gamma_{di}$, is always greater than 0. Moreover, it is approximately P −unbiased as the overall sample size increases, even if the domain sample size is small or null. The main obstacle to its use in large-scale surveys is the fact that the weights are domain-dependent. Instead of a unique weight ω_i attached to the sample unit i for the computation of all the estimates, in the MDIFF estimator, the unit i has D different weights $(\omega_{1i}, \dots, \omega_{di}, \dots, \omega_{Di})$, each of which is used for a specific domain. Having different weights for each unit complicates the computation of the tabulation of the survey estimates greatly when the estimates refer to more than one domain. However, De Vitiis, Righi and Tuoto (2008) demonstrate that if the D domains represent a complete partition of the population U, then the $\hat{Y}_{MDIFF,d}$ estimators add to the total of the y obtained with the standard DIFF estimator, where the weights are obtained as the solution of Equation 3.7, that is

$$\sum_{d=1}^{D} \hat{Y}_{MDIFF,d} = \sum_{i\in S} y_i\omega_i.$$

With a general function $m(x_i;\beta)$, the MDIFF estimator requires knowledge of the domain membership variables γ_{di} for all the units of U.

With the linear regression model $m(x_i;\beta) = x_i'\beta$, with expectations given by Equation 3.5, we define the Modified GREG (MGREG) estimator where only the domain total, X_d, of the auxiliary variables needs to be known. The sample estimate of β is given by:

$$\hat{\beta} = \left(\sum_{j=1}^{n} x_j x_j' \frac{1}{c_j} a_i\right)^{-1} \sum_{j=1}^{n} x_j' y_j \frac{1}{c_j} a_j.$$

The MGREG estimator is given by:

$$\hat{Y}_{MGREG,d} = X_d'\hat{\beta} + \sum_{i\in S}[y_{di} - x_i'\hat{\beta}\gamma_{di}]\; a_i = \sum_{i=1}^{n} y_i\omega_{di},$$

where the weights can be defined explicitly as

$$\omega_{di} = a_i \left[\gamma_{di} + \left(X_d - \hat{X}_{HT,d} \right)' \left(\sum_{j \in S} x_j x_j' \frac{1}{c_j} a_i \right)^{-1} x_i \gamma_{di} y_{di} \frac{1}{c_i} \right].$$

3.3.2. The model-based approach

In this approach, the inference is based on the statistical WM *M*, which links the target variable $\mathcal{y}$ to the auxiliary x variables. As an example of applying the model-based approach to the SDG indicators relevant for data disaggregation, it is noted that, as illustrated in Chapters 4 and 5, this approach is used to estimate the individual probability unit of being food-insecure at a given level of severity of food insecurity.

In this approach, the model plays a much more crucial role than in the model-assisted approach. A wrong model here would determine unreliable estimates; thus, specifying a good model is crucial. In particular, a sufficiently reliable model and relevant good predictors x_i are necessary. However, it is noted that in actual empirical situations, it may only be possible to obtain a good approximation of the true model, because the task of identifying the true model might be complicated or even impossible: for instance, not all relevant auxiliary variables may be accessible. To quote from Box and Draper (1987; p. 74) "all models are wrong; the practical question is how wrong do they have to be to not be useful". This is why in this context, too, the statistical model used for the predictions is denoted as the WM.

In this approach, the inference is conditional on the observed sample values. In contrast, in the repeated sampling approach, the estimators' inferential properties (in terms of bias and variance) in the unconditional sample space $\mathbb{S}$ are evaluated. Furthermore, to estimate the unknown β values, the inclusion probabilities are not considered, since these would unnecessarily increase the variance, thereby making the estimators less efficient.

The basic estimator in this context (Chambers, 2015) is the Model-based Prediction Estimator (MPE), which is obtained as the sum of the sample-observed values and the model-predicted values over the non-sampled population units.

We consider three different cases.

1. The MPE estimator with no domain-specific auxiliary information. The domain membership indicators γ_{di} are not known for all population units but only for those observed in the sample.
2. The Domain-specific MPE (DMPE) estimator – the domain membership indicators γ_{di} are known for all the population units and there is a sizeable domain-specific sample size.
3. The Modified MPE (MMPE) estimator – the specific-domain sample size is null (or very small).

Case 1. **The *MPE estimator with no domain-specific auxiliary information***

The estimator may be expressed as

$$\hat{Y}_{MPE,d} = \sum_{i \in S} y_{di} + \sum_{i \in U \setminus S} m\left(x_i; \hat{\beta}_d\right) \qquad (3.18)$$

where β_d is estimated by fitting the model $m(x_i;\beta_d)$ over all couples (x_i, y_{di}) observed in the sample, and $U\backslash S$ is the subset of units of *U* that are not included in the sample.

The estimator is computed in two steps.

First, the $\hat{\beta}_d$ values are estimated, allowing for the construction of the predicted values

$$\hat{y}_{di} = m(x_i;\hat{\beta}_d).$$

Then, the predicted values are projected over all the $U\backslash S$ non-observed population units.

The estimator can be obtained in a weighted form (Equation 3.2) by defining the weights ω_i as the solution of the following minimum constrained problem:

$$\begin{cases} \sum_{i\in S} c_i(\omega_i - 1)^2 & \text{function to be minimized} \\ \sum_{i\in S} \hat{y}_{di}(\omega_i - 1) = \sum_{i\in U\backslash S} m(x_i;\hat{\beta}_d) & \text{calibration constraints} \end{cases} . \quad (3.19)$$

If we consider Linear WM 3.8, the weights ω_i are defined explicitly as

$$\omega_i = 1 + (X - X_S)' \left(\sum_{j\in S} x_j x_j' \frac{1}{c_j} \right)^{-1} \frac{1}{c_i} x_i,$$

with

$$X_S = \sum_{j\in S} x_j.$$

With WM 3.8, we do not need to know the auxiliary variables for each unit of the population, but only the totals of the auxiliary variables, X.

Case 2. The Domain-specific MPE estimator

In this case, β_d is estimated by fitting the model $m(x_i;\beta_d)$ over all couples $(x_i\gamma_{di}, y_{di})$ observed in the sample S_d. Moreover, the sum of the model-predicted values can be limited over the non-sampled population units in the domain. The DMPE is given by

$$\hat{Y}_{DMPE,d} = \sum_{i\in S} y_{di} + \sum_{i\in U_d} m(x_i;\hat{\beta}_d). \quad (3.20)$$

The weights ω_{di} are domain-specific and can be obtained as a solution of the following problem:

$$\begin{cases} \sum_{i\in S_d} c_i(\omega_{di} - 1)^2 & \text{function to be minimized} \\ \sum_{i\in S_d} \hat{y}_{di}(\omega_{di} - 1) = \sum_{i\in U_d\backslash S_d} m(x_i;\hat{\beta}_d) & \text{calibration constraints} \end{cases} . \quad (3.21)$$

If we consider Linear WM 3.8, with model expectations defined by Equation 3.5 the weights ω_{di} are given by

$$\omega_{di} = 1 + \left(X_d - X_{S_d}\right)' \left(\sum_{j \in S_d} x_j x_j' \frac{1}{c_j}\right)^{-1} x_i \gamma_{di} y_{di} \frac{1}{c_i},$$

being

$$X_{S_d} = \sum_{j \in S_d} x_j.$$

With WM 3.8, we do not need to know the auxiliary variable for each unit in the population, but only the specific domain totals of the auxiliary variables, X_d.

Case 3. Modified MPE estimator

If we consider Linear WM 3.12, which is not-domain dependent, and we know the domain membership indicators γ_{di}, we can define the Modified MPE estimator (MMPE), which is similar to the MDIFF (or the MGREG). With the MMPE, we can compute the domain estimates even with a null (or very small) domain sample size. The vector of unknown parameters β is estimated by fitting the model $m(x_i; \beta)$ over all couples (x_i, y_i) observed in the sample. The estimator MMPE is given by

$$\hat{Y}_{MMPE,d} = \sum_{i \in S} y_{di} + \sum_{i \in U_d \setminus S_d} m\left(x_i; \hat{\beta}\right). \qquad (3.22)$$

The MMPE estimator may be expressed in weighted form as:

$$\sum_{i=1}^{n} y_i \omega_{di}$$

where the weights ω_{di} are obtained as the solution of the following calibration problem:

$$\begin{cases} \sum_{i \in S} c_i (\omega_{di} - \gamma_{di})^2 & \text{function to be minimized} \\ \sum_{i \in S} \hat{y}_i \omega_{di} = \sum_{i \in U} m\left(x_i; \hat{\beta}\right) \gamma_{di} & \text{calibration constraints} \end{cases} . \qquad (3.23)$$

If we consider Linear WM 3.8, the weights ω_{di} are given by

$$\gamma_{di} + \left(X_d - X_{S_d}\right)' \left(\sum_{j \in S} x_j x_j' \frac{1}{c_j}\right)^{-1} x_i \frac{1}{c_i}.$$

With WM 3.8, we do not need to know the auxiliary variable for each unit in the population, but only the specific domain totals of the auxiliary variables, X_d.

This estimator is the basic form of SAE estimators, illustrated in Chapter 6.

Comment 3.3. Consistency of the estimates. The main obstacle to the full adoption of the model-based approach in large-scale surveys is the fact that the best model for one variable might not be the same as for another variable. Thus, the estimates for the different variables may not be congruous with one other. This may pose serious issues when producing contingency tables. However, the problem can be overcome by using more complex models for contingency tables. This is illustrated in Chapter 6.

Comment 3.4. Appropriateness of model for domain data. The appropriateness of the model for domain data is crucial. If the population scatter of the residuals in the domain is far from 0, this is a strong sign that the model is domain-dependent. Thus, in this case, it is essential to have sample data observed in the domain to achieve a good model fit for the domain predictions. A detailed discussion of the definition of models for domain estimation in the model-based approach can be found in Chambers and Clark (2015; Chapter 14).

3.3.3. Extensions to parameters different from the totals

This section considers the estimation of parameters of interest different from the totals. For the SDG indicators, the relevant target parameters are the mean of a quantitative variable, the proportion, expressed as relative frequencies of a categorical variable and the ratio among two totals or means.

Means and relative frequencies

Irrespective of whether it is a mean value or a relative frequency, the functional form of these parameters is expressed as the ratio of the domain total of the y variable and domain population size N_d:

$$\bar{Y}_d = \frac{1}{N_d}\sum_{i=1}^{N_d} y_i = \frac{Y_d}{N_d}.$$

In the repeated sampling approach, the direct estimate of parameter $\bar{Y}_d$ is obtained simply as the ratio of the estimates of the two terms of the ratio. Considering now the GREG estimator expressed by Equation 3.9, the GREG estimator of $\bar{Y}_d$ is given by

$$\hat{\bar{Y}}_{GREG,d} = \frac{\hat{Y}_{GREG,d}}{\hat{N}_{GREG,d}},$$

where

$$\hat{Y}_{GREG,d} = \sum_{i=1}^{n} y_i \gamma_{di}\, \omega_i \quad , \quad \hat{N}_{GREG,d} = \sum_{i=1}^{n} \gamma_{di}\, \omega_i,$$

where the weights ω_i are given by Equation 3.11. The form of the estimator given by estimator $\hat{\bar{Y}}_{GREG,d}$ is straightforward to implement and does not require knowledge of the population size, which may be estimated by directly summing the sample weights over the sample domain.

In the model-based approach, the same computational strategy defined for the estimator $\hat{\bar{Y}}_{GREG,d}$ can be adopted, and the estimate obtained by dividing the model-based estimate of the domain total $\hat{Y}_d$ by the sum over the domain sample data of the model-based weights ω_i defined for the y variable.

Ratios

The ratio of two specific-domain means (or totals) can be expressed as:

$$R_{d,YZ} = \frac{\bar{Y}_d}{\bar{Z}_d} = \frac{Y_d}{Z_d},$$

where $\bar{Z}_d$ and Z_d are the domain mean and total of the variable z, with

$$\bar{Z}_d = \frac{Z_d}{N_d} = \frac{1}{N_d}\sum_{i=1}^{N_d} z_i,$$

z_i being the value of z of the i −th unit.

In the repeated sampling approach, the direct estimate of the parameter $R_{d,YZ}$ is obtained simply as the ratio of the estimates of the two totals. Considering now the GREG estimator expressed by Equation 3.9, the GREG estimator of $R_{d,YZ}$ is given by

$$\hat{R}_{GREG,d,YZ} = \frac{\hat{Y}_{GREG,d}}{\hat{Z}_{GREG,d}},$$

where

$$\hat{Z}_{GREG,d} = \sum_{i=1}^{n} z_i \gamma_{di} \omega_i ,$$

where the weights ω_i are given by Equation 3.11.

In the model-based approach, the same computational strategy defined by the estimator $\hat{R}_{GREG,d,YZ}$ can be adopted, and the estimate obtained by dividing the model-based estimate of the domain total $\hat{Y}_d$ (with weights ω_i defined for the y variable), with the model-based estimate of the domain total $\hat{Z}_d$ (with weights ω_i defined for the z variable). To ensure consistency of the estimates at the numerator and denominator of the ratio, the auxiliary variables of the x_i vectors used for modelling the target variables (y and z) should both include the domain membership variables γ_{di}, thus guaranteeing that the sum of the model weights (for the numerator and the denominator of the ratio) over the domain units reproduce the same estimate of the domain size.

3.4. Traditional sampling techniques

The sampling size of the observed set for each subpopulation should be adequately large to produce sufficiently accurate directly disaggregated data.

The three main approaches to ensure an appropriate sampling size for every subpopulation (or domain) for which disaggregated data should be produced are:

- oversampling
- deeper stratification
- multiphase sampling with a screening of the respondents.

These solutions are not mutually exclusive and can be adopted jointly.

3.4.1. Oversampling

With the oversampling strategy, a larger size of the overall sample is defined. This affects obtaining a larger sample size at the domain level. If we augment the current sample size by a ratio of Δ, this may have an expected impact on the increase of the domain sample size of $n\Delta P_d$, where $P_d = N_d/N$ is the domain relative size.

To give an order of magnitude, Table 3.1 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.

We see that oversampling may be useful for major domains, that is, with 10 percent or more of the population (Kish, 1986). Still, it is ineffective for minor and mini-domains, with a proportion ranging from 1 to 10 percent, and less than 1 percent, respectively.

Table 3.1. Increase in the domain sample size n_d due to a percentage increase Δ in the overall sample size of 10 000 households by subpopulation relative size P_d

Percent relative increase (Δ) in the domain sample size (%)	$\%\boldsymbol{P_d}$			
	0.05%	**1%**	**5%**	**10%**
10	5	10	50	100
50	25	50	250	500
100	50	100	500	1 000

Furthermore, if the domain is not planned, oversampling can give uncertain results, as the domain sample size achieved may be different from the one expected. Suppose that only overall sample size *n* can be controlled, while n_d is a random outcome. Let $E_P(\cdot)$ and $V_P(\cdot)$ denote, respectively, the operators of expectation and variance under repeated sampling. Under as SRSWOR design, the expected sample size and the sampling variance of n_d are, respectively (see Appendix A.3.1),

$$E_P(n_d) = nP_d, \;\; V_P(n_d) \cong n\frac{N_d}{N}\left(1-\frac{n}{N}\right).$$

The lower bound Ln_d and the upper bound Un_d of the confidence interval (at the confidence level $1-\alpha$) of n_d are approximately

$$Ln_d \cong nP_d - t_\alpha\sqrt{V_P(n_d)} \;\;, \;\; Un_d \cong nP_d + t_\alpha\sqrt{V_P(n_d)},$$

where t_α is the percentile of the *t* distribution, with $t_\alpha \cong 2$ for a confidence level of 95 percent.

To ensure that estimate $\hat{Y}_d$ is adequately reliable, it would be useful for n_d to exceed a given threshold n_d^*, for instance $n_d^* = 50$ or $n_d^* = 30$. If only the overall sample size n can be controlled, the simplest way to guarantee that the realized domain-sample size n_d is larger than threshold n_d^* is to find the n value such that

$$E_P(n_d) - t_\alpha\sqrt{V_P(n_d)} \geq n_d^*. \quad (3.24)$$

Appendix A.3.1 shows that in the case of SRSWOR, the n value that guarantees respect of Equation 3.9 at a level of probability equal to 95 percent is

$$n = \frac{n_d^* + 1}{P_d}. \quad (3.25)$$

Table 3.2 illustrates the sample sizes n needed to guarantee the minimum threshold n_d^* by percentage values of the subpopulation proportion (P_d), for different values of the thresholds (30, 50, 100). It can be seen that for rare subpopulations (with $P_d \leq 1\%$), the overall sample size would be too large and substantially unfeasible for most surveys conducted at the country level. Thus, a better strategy would be to control the sampling sizes n_d directly at the sampling design phase, as proposed in Sections 3.3.2 and 3.3.3.

Table 3.2. Sample sizes n needed to guarantee the minimum threshold n_d^* by percentage values of the subpopulation proportion (P_d)

Threshold 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

Box 3.2. Example of a stratified two-stage probability-proportional-to-size sampling without replacement

Consider a population of 200 000 people grouped into four regions, as illustrated in Table 3.2 below. Suppose a stratified two-stage probability-proportional-to-size (PPS) sampling without replacement of 1 000 people. Each region is a stratum. The total sample size is allocated proportionally in each stratum: $n_h = n\ N_h/N$, with n_h being the sample size of the stratum h and N_h the total population in the stratum. Two municipalities are selected without replacement and with PPS in each stratum. In this sampling design, the municipality would represent the PSU. Municipality j in stratum h is selected with a first-stage inclusion probability equal to $\pi_{1hj} = 2\,N_{hj}/N_h$, where N_{hj} is the population in the municipality. From the selected municipality hj, $n_h/2$ final sample units are then extracted with an SRSWOR design. Every final unit i is selected with the second-stage inclusion probability $\pi_{2hji} = (n_h/2)/N_{hj}$. Then, the final inclusion probability of unit i of municipality j of stratum h is

$$\pi_i = \pi_{1hj} \times \pi_{2hji} = \left(2\,N_{hj}/N_h\right)\left[(n_h/2)/N_{hj}\right] = \frac{n}{N}.$$

The sampling design is summarized in Table 3.2 below.

Table 3.2. Example of stratification by region

Region (stratum)	**Population in the stratum (N_h)**	**Sample of people**	**First-stage sampling number of municipalities**	**Second-stage sampling number of sample units in each municipality**
Region 1	100 000	500	2	250
Region 2	50 000	250	2	125
Region 3	20 000	100	2	50
Region 4	30 000	150	2	75
Total	**200 000**	**1 000**	**8**	

3.4.2. Deeper stratification

Stratifying by domain is the traditional strategy adopted to control 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. To illustrate this case, consider the example of a survey stratified by region, illustrated in Box 3.2 above.

Continuing the above example, suppose that disaggregated data should also be disseminated for the three modalities of the living place, distinguishing among (1) rural areas; (2) urban non-metropolitan areas; and (3) metropolitan areas. Considering the sample design given in Box 3.1, it can be seen that the sample sizes of these disaggregation domains are unplanned and that the domain sample sizes n_d $(d = 1, \ldots, 3)$ are, thus, random outcomes. To ensure a fixed sample size for every disaggregation domain, stratification can be performed by cross-classifying the region and the living place. Thus, a deeper stratification would be achieved, into $12 = 4 \times 3$ strata. Maintaining the same characteristics of the sampling design illustrated in Box 3.1, 24 PSUs would be selected.

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 (since some cells can be empty, being structural zeroes) 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.

Box 3.3. Example of a deeper stratified two-stage PPS sampling without replacement

Table 3.3. Example of stratification by cross-classification of region and living place

Region	Living place								
	Rural			Urban, non-metropolitan			Metropolitan		
	Population in the stratum (N_h)	Sample PSU	Sample of people	Population in the stratum (N_h)	Sample PSU	Sample of people	People	Sample PSU	Sample people
Region 1	60 000	2	300	8 000	2	40	32 000	2	160
Region 2	20 000	2	100	10 000	2	50	20 000	2	100
Region 3	12 000	2	60	6 000	2	40	2 000	0	0
Region 4	9 000	2	45	15 000	2	75	6 000	2	30
TOTAL	**101 000**	**8**	**505**	**39 000**	**8**	**205**	**60 000**	**6**	**290**

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. A researcher may implement these strategies with the Mauss-R (Istat, Mauss-R) software, which enables the multivariate allocation of units in sampling surveys.

3.4.3. Multiphase sampling with a screening of respondents

The strategy based on a deeper stratification requires that the domain membership variables γ_{di} be available on the sampling frame. This can be the case with geographical variables, but, generally, not with many variables of the disaggregation plan, such as the income quantile or household characteristics.

A traditional sampling strategy to overcome this is to select a first-phase sample $S_{(1)}$ of size $n_{(1)}$. Then, the variables γ_{di} are collected from the units selected in S_1. The units in $S_{(1)}$ are stratified considering the variable collected in the first phase of the sampling. 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. An approximate indication of this size for SRSWOR designs is given in Table 3.1 above.

Several strategies can be employed to keep costs low (Kalton, 2009): (*i*) use an inexpensive mode of data collection, such as telephone interviewing or a mail questionnaire, for the screening; (*ii*) allow the collection of screening data from persons other than those sampled; and (iii) when screening is carried out by face-to-face interviewing in a multistage design, it is efficient to select a large sample size in each cluster. Costs are reduced and the precision of domain estimates is not seriously compromised if the average domain sample sizes in the clusters are relatively small.

A natural extension of the screening approach is to identify strata where the screening will be more productive. In the ideal scenario, a few strata covering all of the rare populations and no entities outside that population are identified. These circumstances allow for avoiding the screening process. Otherwise, samples must be selected from all the strata to complete coverage of the rare population. The use of disproportionate stratification, with higher sampling fractions in the strata where the prevalence of the rare population is higher, can reduce the amount of screening needed.

3.5. Marginal stratification designs

3.5.1. Motivating example

An excessively detailed stratification, which is useful to control the sampling sizes of the disaggregation domains, could be unsuitable, because it requires selection of at least as many sampling units as the product of the number of categories of the stratification variables.

To overcome some problems with the deeper stratification designs, a straightforward strategy is to drop one or more stratifying variables or to group some of the categories. Nevertheless, some planned domains can become unplanned and some of them can have a small or null sample size. The marginal stratification designs allow for this problem to be overcome. This topic will be illustrated starting with an example.

Consider the example developed in Boxes 3.2 and 3.3 and suppose that the country's NSI cannot conduct a survey involving more than eight municipalities. Therefore, the sampling design, based on a deeper stratification, as illustrated in Table 3.3, cannot be implemented, because it includes 24 PSUs. Moreover, even a sample design selecting only one PSU per stratum is not feasible, as there are 12 cross-classification strata.

Actually, if it is sought to produce the disaggregation estimate only for the marginal domains of the stratification variable, it is not necessary to select a sample from every cell of the cross-classified stratification. Instead, it is only necessary to have a sample for the marginals of the cross-classification defined by region and living place.

This serves to compute direct estimates for every modality of those two characteristics. Suppose that the aim is to define these marginal sample sizes (for the number of PSUs and individuals), as illustrated in Table 3.5. This example is introduced only for illustrative purposes; in fact, it is well known that in some countries, a "metropolitan" living place can exist in only one region.

Table 3.4. Example of marginal stratification design. Fixed sample of municipalities and individuals by region and living place

Region	Living place			Sample of municipalities	Sample of individuals
	Rural	Urban, non-metropolitan	Metropolitan		
Region 1				2	500
Region 2				2	250
Region 3				2	100
Region 4				2	150
Sample of municipalities	3	2	3	8	
Sample of individuals	505	195	300		1 000

With the above schema, for each disaggregation domain, at least two municipalities and a minimum of 100 sample individuals would be selected. In this way, direct estimates could be calculated separately by region and living place.

The marginal stratification designs allow for selecting a random sample such as that illustrated in Table 3.4, controlling only the marginals – and not each cell – of the cross-classified stratification. Table 3.5 provides an example of a marginal stratification design that ensures respect of the marginal sample sizes. It can be seen that not all cells of the cross-classified schema have a sample, although this design enables adequate sample sizes for all marginal categories of the stratification variables. Moreover, it can be seen that these techniques automatically solve the problems of structural zeroes in the cross-classified stratification.

We conclude this section by underscoring that we can use marginal stratification only if it is not necessary to produce estimates for the combination region/living place, but only for the two marginal distributions (as is often the case). However, if producing disaggregated data for each cell of the cross-classification is desired, it is somewhat obligatory to guarantee a sizeable sample in each cell (which increases survey costs).

Table 3.5. Example of marginal stratification design: selected municipalities and sample of individuals (in red brackets) in each cross-classification cell

Region	Living place			Total sample
	Rural	Urban, non-metropolitan	Metropolitan	
Region 1	1 (305)	0	1 (150)	**2 (500)**
Region 2	1 (75)	1 (175)	0	**2 (250)**
Region 3	0	1 (20)	1 (80)	**2 (100)**
Region 4	1 (80)	0	1 (70)	**2 (150)**
Total sample	**3 (505)**	**2 (195)**	**3 (300)**	**8 (1 000)**

3.5.2. General overview

Many methods have been proposed in the literature 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. Indeed, as described in Falorsi, Righi and Orsini (2006), it is not possible to implement the Latin Lattices schemes in many real survey contexts, for example if there are no population units in one or more of the cross-classification strata. The main weakness of the linear programming approach is its computational complexity. The sampling strategy proposed here, based on balanced sampling (see Section 3.5.3) does not suffer from the disadvantages of the abovementioned methods and grants control of the sample sizes for various domains of interest, defined by different partitions of the reference population. Furthermore, it guarantees that the sampling errors of domain estimates are lower than the given thresholds.

3.5.3. Balanced sampling for marginal stratification

Multi-way stratification designs can be treated in the context of balanced sampling.

Definition of a balanced sample depends on the assumed inferential framework. In the model-based approach, a sample is defined as balanced on a set of auxiliary variables if there is equality between the sample and the known population means of the auxiliary

variables (Valliant, Dorfmann and Royall, 2000). Following the design-based (or model-assisted approach) considered here, a sample is balanced when the HT estimates of the auxiliary variable totals are equal to their known population totals (Deville and Tillé, 2004).

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 set $\mathcal{S}$ of all subsets S of 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

$$\sum_{i \in S} \frac{x_i}{\pi_i} = \sum_{i \in U} x_i \quad (3.26)$$

for all $S \in \mathcal{S}$ 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, that is

$$\sum_{i \in U} \pi_i \gamma_{di} = n_d \ (d = 1, \dots, D), \quad (3.27)$$

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

$$x_i = \pi_i \gamma_i \ (3.28)$$

where γ_i is the *D* vector of domain membership variables, being

$$\gamma_i = (\gamma_{1i}, \dots, \gamma_{di}, \dots, \gamma_{Di})' .$$

If the $i - th$ unit belongs to five different disaggregation domains, Equation 3.13 defines the x_i vector with (*D* – 5) zeroes and with five entries equal to π_i in the places indicating the domains to which the unit *k* belongs.

When defining the x_i vector as in Equation 3.28, if the condition expressed in Equation 3.27 holds, the selection of samples satisfying the system of balancing equations 3.26 guarantees that the n_d values are non-random quantities.

The left-hand side of the balancing equation 3.26 is

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

The right-hand side is

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

Box 3.4. Examples of auxiliary variables for the balanced sampling illustrated in Table 3.4

Consider the marginal stratification design illustrated in Table 3.4. In this case, the vector of auxiliary variables chosen to select the municipalities has seven entries: the first four for the region, and the latter three for the living place. The vector γ_i has two entries equal to 1 and other five entries equal to 0. In the case of a municipality in Region 2 and in the rural area, vector γ_i is

$$\gamma_i = \left(\overbrace{0, 1, 0, 0}^{\text{Region}} , \overbrace{0, 1, 0}^{\text{Place}} \right)' .$$

Let π_{1i} be the first-stage inclusion of the $i -$ th municipality and let U be the population of *M* municipalities in the country. According to Equation 3.4, these probabilities must respect the following constraint:

$$\sum\nolimits_{i \in U} \pi_{1i} \gamma_{di} = (2, 2, 2, 2, 3, 2, 3)'$$

Deville and Tillé (2004) 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 (2004) show that with the specification at Equation 3.28 of the x_i vectors, the balancing equations 3.26 can be satisfied precisely. The cube method is implemented via an enhanced algorithm for large data sets (Chauvet and Tillé, 2006) available in a free software code.

Comment 3.5. Marginal stratification designs in the second phase of sampling. The marginal stratification design can be applied to select the second-phase sample $S_{(2)}$ (see Section 3.2), thus guaranteeing the planned final sample sizes n_d $(d = 1, \dots, D)$ and overcoming the problems associated with an excessively detailed stratification. To apply this design, the domain membership variables γ_{di} should be collected, through a screening on a subset of the Ultimate Stage Units (USUs) of the PSUs selected in the first phase. See Section 3.5.4 below.

Comment 3.6. Balanced sampling as a general design. It is emphasized that balanced sampling forms the basis for defining broad classes of sampling designs. For instance, stratified sampling designs require that

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

and each U_d is referred to as a stratum.

3.5.4. Marginal stratification design for two-stage or two-phase sampling designs

This section illustrates how to carry out marginal stratification designs for the two-stage or two-phase sampling designs commonly adopted in real survey contexts.

To introduce this topic, let *U* be a population of *M* PSUs (e.g. municipalities) and let U_i be the population of N_i USUs of the $i-\text{th}$ PSU (e.g. households).

Let m be the total number of PSUs to be selected in the first-stage sampling and let m_d $(d = 1, \dots, D)$ the number of PSUs to select in the $d-\text{th}$ disaggregation domain. Let n be the total number of USUs in the second-stage sampling and let n_d $(d = 1, \dots, D)$ be the number of USUs to select in the $d-\text{th}$ disaggregation domain. Suppose that the quantities, m_d $(d = 1, \dots, D)$, n and N_d are fixed and defined by constraints based on accuracy and budget. Falorsi and Righi (2015, 2018) explain how to define the latter quantities, minimizing the expected costs while ensuring predefined levels of accuracy of the sampling estimates.

Two-stage sampling

The m PSUs are selected without replacement and with first-stage inclusion probabilities π_{1i} $(i = 1, \dots, M)$.

In the $i-\text{th}$ selected PSU, n_i USUs are sampled with a SRSWOR design out of the N_i USUs of the PSUs, with

$$\pi_{2i} = \frac{n_i}{N_i} \ (i = 1, \dots, M)$$

being the second-stage inclusion probabilities.

To implement the sampling design, an initial value of the first-stage π_{1i}^{ini} and the second-stage π_{2i}^{ini} inclusion probabilities are defined, as

$$\pi_{1i}^{ini} = m\frac{N_i}{N} \ , \ \ \pi_{2i}^{ini} = \frac{n}{m} \ \ (i = 1, \dots, M).$$

Then, having defined the above elements, the operational steps for defining and implementing a balanced two-stage sampling that guarantees observing m_d PSUs and n_d USUs (as expected values over repeated sampling) are the following.

1. *Definition of the first-stage final inclusion probabilities* π_{1i} (i $= 1, \dots, M$). These are defined by the following calibration system, ensuring that the expected sample sizes are equal to the fixed ones:

$$\begin{cases} \sum_{i=1}^{M} D\left(\pi_{1i}, \pi_{1i}^{ini}\right) = \min \\ \sum_{i \in U} \pi_{1i}\gamma_{di} = m_d \ (d = 1, \dots, D) \\ \pi_{1i}^{ini} \le \pi_{1i} \le U\pi_{1i}^{ini} \ \text{ for } i = 1, \dots, M \end{cases} , \qquad (3.29)$$

   where $D\left(\pi_{1i}, \pi_{1i}^{ini}\right)$ is the *truncated logarithmic distance function* (Singh and Mohl, 1996) between π_{1i} and π_{1i}^{ini}; $0 \le L \le 1$; and $U \ge 1$. The truncated logarithmic distance function ensures that the final inclusion probabilities are bounded in the interval $(L\pi_{1i}^{ini}, U\pi_{1i}^{ini})$. The problem at Equation 3.29 can be solved with the software Regenesees (Istat, Regenessees).

2. *Definition of the second-stage final inclusion probabilities* π_{2i} (i $= 1, \dots, M$). These are defined by the following calibration system, ensuring that the expected sample sizes are equal to those fixed in advance:

$$\begin{cases} \sum_{i=1}^{M} D\left(\pi_{2i}, \pi_{2i}^{ini}\right) = \min \\ \sum_{i \in U} \pi_{1i}\pi_{2i}N_i \ \gamma_{di} = n_d \ (d = 1, \dots, D). \\ \pi_{2i}^{ini} \le \pi_{2i} \le U\pi_{2i}^{ini} \ \text{ for } i = 1, \dots, M \end{cases} \qquad (3.30)$$

   In the problem at Equation 3.20, the quantities π_{1i} are known, as they were defined in the first step.

3. *Selection of the first-stage balanced sampling* of PSUs, respecting the constraints

$$\sum_{i \in U} \gamma_{di}\lambda_i = m_d \ (d = 1, \dots, D),$$

   where the inclusion probabilities π_{1i} are those defined in the calibration system at Equation 3.29. The sample is selected with the Cube algorithm.

4. *Selection of the second stage sampling*. In the $i-$th selected PSU, n_i USUs are sampled with SRSWOR out of the N_i USUs of the PSUs, with

$$n_i = round(\pi_{2i}N_i) \quad (i = 1, \dots, m),$$

where the inclusion probabilities π_{2i} are those defined in the calibration system at Equation 3.30.

Note that with two-stage sampling design, while the sample sizes of the first-stage sampling m_d are equal to those established in advance, the sample sizes of the USUs may be not equal to those fixed in advance, even if the system provided by Equation 3.30 ensures that the constraint is respected only over repeated sampling. However, the domain realized and the fixed sample sizes of the USUs should be close to one other.

Two-phase sampling

To ensure that the domain realized and the fixed sample sizes of the USUs are strictly equal, the second step must be reformulated, defining the balancing constraints only over the selected sample PSUs. Thus, the second phase-inclusion probabilities $\pi_{2i|i\in S}$ are defined as the solution of the following calibration system:

$$\begin{cases} \sum_{i=1}^{m} D\left(\pi_{2i|i\in S}, \pi_{2i}^{ini}\right) = \min \\ \sum_{i\in S} \pi_{2i|i\in S} N_i \, \gamma_{di} = n_d \ (d = 1, \dots, D). \\ \pi_{2i}^{ini} \le \pi_{2i|i\in S} \le U\pi_{2i}^{ini} \ \text{ for } i = 1, \dots, M \end{cases} \tag{3.31}$$

Then, in the $i-\text{th}$ selected PSU, n_i USUs are sampled with an SRSWOR design out of the N_i USUs of the PSUs, with

$$n_i = round\left(\pi_{2i|i\in S} N_i\right) \quad (i = 1, \dots, m).$$

The above schema can be referred to as a two-phase sampling design, as the probabilities $\pi_{2i|i\in S}$ depend on the sample selected in the first stage.

3.6. Indirect sampling and multisource sampling designs

3.6.1. General background

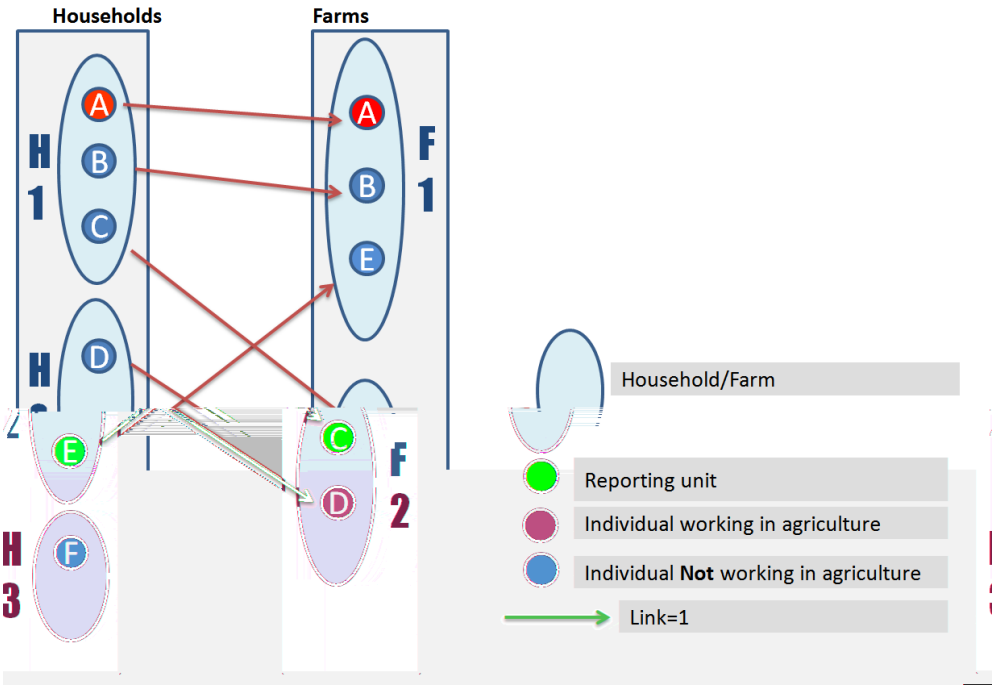
In any conventional survey, random selection of the sample requires an updated list that records all individuals eligible for the survey (and only these), each identified by a label. This perfect list, i.e. the sampling frame, is used to identify the elements of the target population. When the sampling frame is available, a crucial statistical issue is the assessment of the coverage actually 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 target population elements. However, in statistical practice, perfect frames seldom exist, and problems always arise to disrupt the ideal one-to-one mapping. For example, the sampling frame might suffer from either or both undercoverage and overcoverage. There is undercoverage when the available frame is incomplete, because it includes only part of the target population, and the missing elements cannot appear in any sample drawn for the survey. On the contrary, there is overcoverage when the sampling frame contains duplications of the same units or units that are not included in the target population. However, in statistical practice, there may also be frame imperfections of other types: for example, in certain circumstances, one may not possess the collection units desired, but rather another frame of units linked to the list of collection 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. The following strategy will be adopted: starting with the observation of one population, the units of the other populations are surveyed by reference to their links with the units of the first population. Thus, as would occur with an indirect sampling approach, the other populations can be considered sampled from an imperfect frame, i.e. the frame referring to the first population. Frame imperfections will also be considered in the observation of the first population.

Figure 3.1. below, taken from FAO (2015), illustrates the mechanism of the links in the case of farm surveys when only a list of households, derived from the last census, is available.

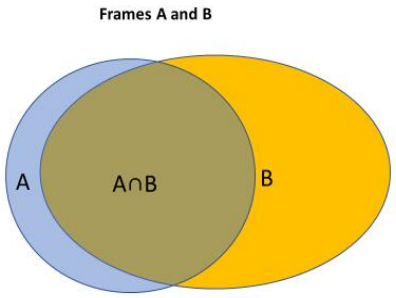
In practice, the links do not have to be known in advance; however, the enumerator obtains information on the links during the data collection phase. For instance, consider that in Figure 3.1, the enumerator who interviews the individuals A and B of Household H1 detects working in Farm F1. Thus, the enumerator identifies two links between Household H1 and Farm F1. Moreover, it may be seen that Farm F1 can be recognized by the enumerator who interviews Household H2. In total, Farm F1 may be identified by three links, each of which may be detected during the data collection. This is an example of the concept of *multiplicity* discussed below.

Figure 3.1. Example of links between a frame of households and the target population of agricultural holdings in the household sector



Source: FAO, 2015.

Figure 3.2. Example of multisource sampling: target population covered by the union of two sources



Source: FAO, 2020.

To identify the links, the survey questionnaires must be appropriately structured. FAO (2015; Chapter 3) illustrates the modules and operational rules for applying indirect sampling in agricultural surveys.

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 3.2. above, which displays two partially overlapping sources. A relevant example here is that of agricultural surveys with holdings in the household and non-household sector. In some circumstances, some of the holdings may fall under two different frames (of the families and the legal entities).

As can be seen, if a sample S^A is selected from Frame A and an independent sample S^B is selected from Frame B, the units in the intersection $\mathrm{A} \cap \mathrm{B}$ of the two sources can be observed 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. These two 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. Also known as multiplicity sampling or snowball sampling, this is a link-tracing sampling procedure in which a sample is obtained by following existing links from one respondent to another. This sampling methodology applies, for example, in estimating the country-prevalence of a rare disease, when a frame that fully represents the target population is not available. Selection units and target units may either coincide, be related or be unrelated, according to a one-to-many linkage rule. Thus, for each target unit, multiplicity is defined as the number of selection units to which it is linked, and a multiplicity adjusted estimator is suggested.

In indirect sampling, the notion of multiplicity is essentially the same, except that a many-to-many linkage pattern must be considered. To adjust for possible data duplication at the estimation stage, it is suggested to use the Generalized Weight Share Method (GWSM) to provide an estimation weight for each target unit in the selected sample; in fact, this is a multiplicity adjustment. On the other hand, in the context of multiple frames surveys, multiplicity is defined as the number of frames from which a unit can be selected.

Box 3.5. Examples of the concept of multiplicity

Multiplicity in indirect sampling

Consider Farm F1 in Figure 3.1 and suppose that a sample of households is selected. In this case, Farm F1 has a multiplicity equal to 3, since it can be reached from three links generated by Households H1 and H2.

Multiplicity in multisource sampling

With reference to Figure 3.1, suppose that target population *U* can be identified as the union of three disjoint subsets:

- A^* the set of units only in Frame A. These units have a multiplicity equal to 1.
- B^* the set of units only in Frame B. These units have a multiplicity equal to 1.
- the units in the intersection set $\mathrm{A} \cap \mathrm{B}$, which have a multiplicity equal to 2.

3.6.2. Indirect sampling: basic methodology

Indirect sampling is suitable when producing statistics of populations for which there is no sampling frame or for which the existing frame is imperfect. In such cases, the sampling procedure assumes that population U^A is related to population of interest U^B, but only the sampling frame of U^A is available. Then, a sample is selected from U^A, and using the links between the two populations, a sample of units of U^B is observed.

For instance, in the case of statistics on rural households, U^A is the population of farms and U^B the population of rural households. Let s^A be a sample selected from U^A without replacement and with fixed sample size m^A, where U^A contains N^A units. Let π_j^A represent the inclusion probability of the *j*-th unit in U^A with $\pi_j^A > 0$ and $\sum_{j \in U^A} \pi_j^A = m^A$ with $\pi^A = (\pi_1^A, \dots, \pi_j^A, \dots, \pi_{M^A}^A)'$.

Let M^B, N^B , U_i^B and M_i^B be the number of units in U^B, the number of clusters in U^B, the *i*-th cluster of U^B with $\bigcup_{i=1}^{N^B} U_i^B = U^B$ and the number of units in the *i*-th cluster U_i^B, respectively. Let us denote with y_{ik} the value of the variable of interest for the *k*-th unit of the *i*-th cluster of U^B and the population total of all y_{ik}'*s* by

$$Y_d = \sum_{i=1}^{N^B} \sum_{k=1}^{M_i^B} y_{ik}\, \gamma_{dik}.$$

Let $l_{j,ik}$ be an indicator variable of link existence: $l_{j,ik} = 1$ indicates that there is a link between the *j*-th unit in U^A and the *k*-th unit in U_i^B, while $l_{j,ik} = 0$ indicates otherwise.

Suppose that an indirect sampling process is performed: if unit $j \in U^A$ is included in s^A, then all clusters U_i^B for which $L_{j,i}^B = \sum_{k=1}^{M_i^B} l_{j,ik} > 0$ are observed (i.e. y_{ik}) in the indirect sample of population U^B. Let n^B be the size of the sample of clusters in population U^B obtained after the indirect sampling process. The variable Y is estimated according to the estimator based on the GWSM theory (Lavallée, 2007):

$$\hat{Y}_d = \sum_{i=1}^{n^B} y_{di}\, \omega_j^B\,, \qquad (3.32)$$

where $y_{di} = \sum_{k=1}^{M_i^B} y_{ik}\gamma_{dik}$ and $\omega_j^B = \sum_{j\in s^A} \omega_j^A\, \tilde{L}_{j,i}^B$ with $\tilde{L}_{j,i}^B = \frac{L_{j,i}^B}{L_i^B}$ and $L_i^B = \sum_{j=1}^{M^A} L_{j,i}^B$,

in which $\omega_j^A = 1/\pi_j^A$. Note that in the estimator at Equation 3.17, the multiplicity factor is L_i^B.

The theorem in Lavallée (2007; Section 3) states that Equation 3.32 provides an unbiased estimator for Y_r^B provided that all links $l_{j,ik}$ can be correctly identified and $L_i^B > 0$ for all $i \in U^B$. This is a key assumption, that is discussed in depth in Lavallée (2007). In practice, the links $l_{j,ik}$ are identified by the enumerators during the data collection phase, when interviewing the unit j of S^A, and L_i^B is collected by the enumerator interviewing unit i of S^B. For instance, consider the example of Figure 3.1; the quantity L_i^B for Farm F1 is given by the number of its workers (three). This information is easily captured by interviewing the farm.

By defining

$$z_{dj} = \sum_{i=1}^{N^B} \tilde{L}_{j,i}^B\, y_{di}\,, \qquad (3.33)$$

the estimator established under Equation 3.17 can be expressed as a usual HT estimator on the *z* values referring to the U^A population,

$$\hat{Y}_d = \sum_{s^A} z_{dj,}\omega_j^A\,. \qquad (3.34)$$

Therefore, the variance $V_B(\hat{Y}_d)$ of $\hat{Y}_r^B$ can be expressed as the variance of the HT estimator on the U^A population.

Indirect sampling for rare or hard-to-reach populations

In practice, rare populations are often difficult to target for surveying purposes. Often, no adequate sampling frames exist. Thus, these populations become unplanned domains. In these cases, it is often necessary to use a different but related sampling frame to reach the rare target population.

Indirect sampling is thus performed. For example, to target people having an infectious disease in a large city, it is possible to use lists of dwellings as sampling frames, which subsequently entails surveying the families of the selected dwellings. Other approaches developed in the literature for rare populations, such as adaptive cluster sampling (Thompson and Seber, 1992), network sampling (Sanders and Kalsbeek, 1990) and snowball sampling (Goodman, 1961) can be considered particular cases of indirect sampling, with a specific definition of the links. This is discussed in Lavallée (2007; Chapter 3). Fortunately for the statistician, it turns out that rare populations are often found in clusters. This is often the case, for example, with infectious diseases, including for Covid-19 (Alleva *et al.*, 2020). By surveying the complete clusters, considerable reductions in costs are possible because a large proportion of these are related to the identification of rare populations. Therefore, data for the clusters of surveyed units can be obtained

through indirect sampling. The problem for the statistician is to weight the survey data so that unbiased estimates for the characteristics of the rare target population can be produced. The GWSM provides a simple way to obtain this weighting.

Box 3.6. Examples of indirect sampling for hard-to-reach populations

Example 1. Indirect sampling can be used to sample nomadic populations as population U^B. In this case, list U^A can consist of the list of places where the nomadic population is known to stop, to obtain refreshment or stock up on water, such as oases.

Example 2. In the Italian National Institute of Statistics (Istat), indirect sampling has been adopted to sample the *homeless*, as population U^B. In this case, list U^A can consist of the list of places where the homeless went to receive some types of care, such as food or beds (De Vitiis, Falorsi and Inglese, 2014; Ardilly and Le Blanc, 2001).

3.6.3. Multisource sampling

The basic methodology is illustrated for cases involving two sources. The extension to cases involving three and more sources is straightforward. It is formally developed in FAO (2014) and Singh and Mecatti (2011).

To illustrate the value of this approach, consider a standard agriculture survey on farms. Two different sampling frames can be used: a frame of large farms and an area frame of the census enumeration areas to capture the small farms. In this context, a large farm can be captured from both frames, thereby identifying a multi-source situation. For another example, consider agricultural surveys where it is sought to cover holdings engaging in crop production, livestock rearing, fishery, forestry and aquaculture. In some cases, multiple frames may have to be used to reach all of these groups, and some holdings (engaging in more than one activity) may be included in more than one list.

To describe this approach, suppose that the target population U is expressed as the union of two subsets, U^A and U^B (having respectively N^A and N^B units), that partially overlap:

$$U = U^A \cup U^B. \qquad (3.35)$$

The total Y_d can be expressed as:

$$Y_d = Y_d^A + Y_d^B - Y_d^{AB}, \qquad (3.36)$$

where

$$Y_d^A = \sum_{i \in U^A} y_i \gamma_{di} \quad , \quad Y_d^B = \sum_{i \in U^B} y_i \gamma_{di} \quad \text{and} \quad Y_d^{AB} = \sum_{i \in U^A \cap U^B} y_i \gamma_{di}. \qquad (3.37)$$

are the domain totals of variable y in subpopulations U^A, U^B and in the intersection set $U^A \cap U^B$.

Two independent samples S^A and S^B, with fixed sample sizes n^A and n^B, are selected from U^A and U^B, respectively, without replacement and with vectors of inclusion probabilities $\pi^A = (\pi_1^A, \dots, \pi_i^A, \dots, \pi_{N^A}^A)'$ and $\pi^B = (\pi_1^B, \dots, \pi_i^B, \dots, \pi_{N^B}^B)'$.

A direct estimation of the total Y_d can be computed as:

$$\hat{Y}_d = \hat{Y}_d^A + \hat{Y}_d^B - \hat{Y}_d^{AB}, \quad (3.38)$$

with

$$\hat{Y}_d^{AB} = \alpha \hat{Y}_{d(A)}^{AB} + (1-\alpha)\hat{Y}_{d(A)}^{AB}, \quad (3.39)$$

where $\hat{Y}_d^A$ and $\hat{Y}_{d(A)}^{AB}$ are direct estimates of totals Y_d^A and Y_d^{AB} derived from sample S^A; $\hat{Y}_d^B$ and $\hat{Y}_{d(B)}^{AB}$ are direct estimates of totals Y_d^B and Y_d^{AB} derived from sample S^B and $\hat{Y}_d^{AB}$ is a convex combination of direct estimates $\hat{Y}_{d(A)}^{AB}$ and $\hat{Y}_{d(A)}^{AB}$, with $0 \leq \alpha \leq 1$,

where:

$$\hat{Y}_d^{A^*} = \sum\nolimits_{i \in S^A} y_i \omega_i^A \gamma_{di} \ , \hat{Y}_{d(A)}^{AB} = \sum\nolimits_{i \in S^A \cap U^B} y_i \omega_i^A \gamma_{di} \ , \quad (3.40a)$$

$$\hat{Y}_d^{B^*} = \sum\nolimits_{i \in S^B \cap B^*} y_i \omega_i^B \gamma_{di} \ , \hat{Y}_{d(B)}^{AB} = \sum\nolimits_{i \in S^B \cap U^A} y_i \omega_i^B \gamma_{di} \, , \quad (3.40b)$$

with ω_i^A and ω_i^B being the direct weights of samples S^A and S^B.

Information on the intersection of the samples with the intersection set $U^A \cap U^B$ can be collected either during the interview or by linking the two sampling frames A and B.

Singh and Mecatti (2011) give an in-depth illustration of the different approaches explored in literature to find the optimal value of α in the context of multiple frames surveys. Hartley (1974) proposed choosing α in Equation 3.39 to minimize the variance of $\hat{Y}_d$. Because the frames are sampled independently, the variance of $\hat{Y}_d$ is:

$$V(\hat{Y}_d) = V(\hat{Y}_d^A) + V(\hat{Y}_d^B) + \alpha^2 V(\hat{Y}_{d(A)}^{AB}) + (1-\alpha)^2 V(\hat{Y}_{d(B)}^{AB}) +$$

$$-2\alpha Cov\left(\hat{Y}_d^A, \hat{Y}_{d(A)}^{AB}\right) - 2(1-\alpha)\, Cov\left(\hat{Y}_d^B, \hat{Y}_{d(B)}^{AB}\right).$$

Thus, for general survey designs, the variance-minimizing value of α is:

$$\alpha^{opt} = \frac{V(\hat{Y}_d^B) + Cov\left(\hat{Y}_d^B, \hat{Y}_{d(B)}^{AB}\right) - Cov\left(\hat{Y}_d^A, \hat{Y}_{d(A)}^{AB}\right)}{V(\hat{Y}_d^A) + V(\hat{Y}_d^B)}. \quad (3.41)$$

Unfortunately, the above quantity depends on variable *y* and domain d. The coherence of the estimates across the variables and domains is not ensured. Note that if one of the covariances in Equation 3.41 is large, it is possible for α^{opt} to be smaller than 0 or greater than 1. Hartley (1974) suggests opting for the

alternative expression $\alpha^* = V(\hat{Y}_d^B)/[V(\hat{Y}_d^A) + V(\hat{Y}_d^B)]$. However, in this case too, the coherence of the estimates across the variables and domains is not ensured.

An alternative expression of estimator $\hat{Y}_d$ that guarantees the coherence necessary in official statistics is to fix the α value independently of the variable and domain, as $\alpha = n^B/(n^A + n^B)$. Another solution represented by the multiplicity estimator (Singh and Mecatti, 2011) is

$$\hat{Y}_d = \sum_{i \in S^A} y_i \frac{\omega_i^A}{m_i} \gamma_{di} + \sum_{i \in S^B} y_i \frac{\omega_i^A}{m_i} \gamma_{di}, \quad (3.42)$$

where m_i is the multiplicity of the $i-$ th unit given in Box 3.5 for multisource sampling.

3.7. Summary of the main recommendations

The main advice provided in this chapter is the following.

1. Weighting sample-domain data allows for computing the direct domain-sampling estimate.
2. Various estimators can be adopted if the available domain information is leveraged differently; the estimators can be domain-specific or referred to the entire population.

3. In choosing a particular estimator, it is necessary to consider: (i) its inferential properties; (ii) the level of detail at which the auxiliary information is made available (e.g. unit-level or aggregated); (iii) the consistency of the various estimates that can be produced from a given survey; and (iv) its computational feasibility.

4. Proper sampling designs for data disaggregation should ensure planned sample sizes for the domains of the disaggregation plan. Thus, it would be possible to calculate direct estimates. Indirect estimates could also benefit from having sampling units in each domain of interest.

5. Survey statisticians can improve sampling designs by geographically spreading the sample units and diminishing the level of clustering. This would improve the ability to reach segregated or rare subpopulations.

6. Traditional sampling techniques address this topic by leveraging oversampling, screening or a deeper stratification. However, these solutions may be costly and difficult to implement in some practical circumstances.

7. New sampling approaches (such as marginal stratification sampling, indirect sampling or multisource sampling) allow for some of the abovementioned problems to be overcome without excessively increasing survey costs. They also enable sampling of rare or hard-to-reach populations.

Appendix A3.1.

With a Poisson sampling, the variance $V_P(n_d)$ can be approximated by:

$$V_P(n_d) \cong \sum_{i=1}^{N} \gamma_{di}\, V_P(\lambda_i) = \sum_{i=1}^{N} \gamma_{di} \frac{n}{N}\left(1-\frac{n}{N}\right) = n\frac{N_d}{N}\left(1-\frac{n}{N}\right)$$

It is necessary to find the target overall sample size n^*, that guarantees that the lower bound of the confidence interval of n_d (at a probability of 95 percent) is bigger than the target n_d^*. The following inequality is defined:

$$n_d^* \leq n^* P_d - 2\sqrt{n^* P_d\left(1-\frac{n^*}{N}\right)} \cong n^* P_d - 2\sqrt{n^* P_d} \rightarrow (n_d^*)^2 = (n^*)^2 P_d^2 - 2n^* P_d.$$

The unknown value n^* can then be obtained, as the solution of the following second-degree equation

$$P_d^2 (n^*)^2 - 2P_d n^* - (n_d^*)^2 = 0.$$

Then, the following is obtained

$$n^* = \frac{2P_d \pm \sqrt{4P_d^2 + 4P_d^2 (n_d^*)^2}}{2P_d^2} = \frac{2P_d \pm 2P_d\sqrt{1+(n_d^*)^2}}{2P_d^2} \cong \frac{1 \pm n_d^*}{P_d} = \frac{n_d^* + 1}{P_d}.$$

Chapter 4. Computing the accuracy of disaggregated data

4.1 Introduction

This chapter discusses the importance of computing, and providing users with, the accuracy of disaggregated data and proposes a method with application. Its first section examines the need to compute accuracy and how to evaluate it. Section 4.2 introduces the basic theory, while Section 4.3 presents a case study based on the methodology explained using the available country-level data for SDG Indicator 2.1.2 – Prevalence of moderate or severe food insecurity in the population, based on the FIES.

4.1.1. Why must sampling errors be estimated?

Errors can and do occur at all stages of a survey or even a census (UNSD, 2020). The magnitude of these errors and, therefore, the quality of resulting estimates are crucial, given the power of data in shaping vital decisions defining a country's future. As presented in other parts of these Guidelines, the logical steps for releasing disaggregated data are the following:

1. compute direct estimates (when data are available);
2. compute the measure of accuracy;
3. evaluate the accuracy and decide whether to publish the disaggregated indicator; and
4. if needed and feasible, apply corrections and post-adjustments.

Thus, the estimation of errors is a preliminary and required step for every action carried out in data disaggregation. Before disseminating estimates, it is necessary to check their accuracy. If the estimates present high levels of inaccuracy, a decision should be taken as to whether to publish the data or not, while informing users of their level of reliability and thus alleviate the risk of their incorrect use. Alternatively, improvement actions could be launched (at the sample design or estimator level) to enable producing more reliable disaggregated data. At any rate, the estimation and dissemination of errors are essential to build public trust in data and their use, an element that eventually generates confidence in the NSS overall.

The dissemination of available data is generally tied to one leading strategic choice: whether to (a) limit the use of the data and allow only the dissemination of indicators of which the accuracy is certified;

or

(b) enable the dissemination of larger data sets with their accuracy profiles, to provide the data users with greater flexibility while ensuring that they are aware of the accuracy of the estimates.

Option (b) enhances the relevance of the information disseminated; it also reduces the risk of inappropriate use of the data.

4.1.2. The measure of accuracy

Given the need to compute the statistical errors in disaggregated data, it is necessary to determine the measures of accuracy to be calculated and communicated to users. There are multiple sources of error (ranging from sampling errors to coverage errors, etc.), and only some of the error components can be measured, under the condition that specific experimental designs are applied. Thus, providing users with full information on the statistical errors affecting the disseminated data tends to be unfeasible. Moreover,

a complete treatment of statistical errors would cover the entire field of statistics and exceed the scope of these Guidelines.

Here, it is supposed that the inferential processes for producing the disaggregated data rest on solid foundations. This necessary precondition assumes that the methodology embedded in the indicators is transparent and does not introduce bias in the estimates. Given this fundamental prerequisite, the main advice is to compute at least the leading components of the errors: sampling variance, model variance, or both.

Sampling variance is an adequate measure of accuracy when the construction of statistical indicators is based on the inferential properties of repeated sampling. Sampling variance measures the uncertainty deriving from the randomness of the observed set of data.

Model variance is a suitable measure of accuracy when the construction of statistical indicators is based on models using $x - \text{auxiliary}$ variables, generating the value of the target variable y for the units in the population.

Both sampling variance and model variance are well known in statistical literature (Cochran, 1977; Särndal, Swensson and Wretman, 1992; Chambers and Chandra, 2008; among others). However, some indicators of the data disaggregation plan can be obtained via statistical procedures that utilize model-based approaches jointly with inference based on the sampling design. For these cases, it is suggested to consider global variance (GV) (Wolter, 1985) as the measure of accuracy.

When disaggregated indicators are produced from the census or administrative records, it is also necessary to consider the bias in measuring accuracy. Bias generally derives from the measurement error (based on statistical models) and the coverage error, the latter deriving from erroneous inclusion, in the observation, elements extraneous to the population of interest (over-coverage) or from incorrectly excluding certain units from the target population. These types of error can be detected with special observational techniques (based on double and independent measurements), which may however be costly. In the case of official statistics, the techniques are implemented only in certain specific cases. Alleva *et al.* (2021) propose the Global Mean Squared Error (GMSE) as a more general measure of accuracy, that includes, as particular cases, the bias and the aspects discussed above (the GV, sampling and model variance). The GMSE may be useful when synthetizing the accuracy of disaggregate indicators deriving from census or administrative records.

4.1.3. Evaluating accuracy

It is challenging to identify concrete and comprehensive rules to assess the magnitude of errors. Indeed, as stated in Eurostat (2013), there are no general precision thresholds or sizes that apply to all surveys. The rules tend to be survey-specific and purpose-specific, depend on users' needs in terms of reliability, and are related to the resources available. However, Eurostat (2013) provides useful suggestions and presents examples of precision thresholds or sizes used by different institutions in specific cases. Statistics Canada (2010; pp. 30–31) applies the following guidelines concerning the reliability of data from labour force surveys (Statistics Canada, 2010; pp. 30–31): if the coefficient of variation (CV) < 16.5 percent, then there are no release restrictions; if 16.5 percent < CV < 33.3 percent, then the data should be accompanied by warnings (release with caveats); if the CV > 33.3 percent, then the data are not recommended for release. The British Office for National Statistics (ONS) dissemination policy (2004) established that ideally, the CV should be < 20 percent for a small area estimate to be considered publishable.

As illustrated in Section 4.2, the accuracy can be expressed in square terms; as the square of the expected difference between the estimator and the true (unknown) population value. A practitioner may find it difficult to relate the measurement of the accuracy with the metric used to measure a given phenomenon. Therefore, the derived measures of accuracy comprehensible to most users are:

- The Standard Error (SE), expressed as the square root of the square measure of accuracy. The SE is of the same order of magnitude as the disaggregated indicator, and is thus easily interpretable.
- The Coefficient of Variation (CV), given as the ratio of the SE to the estimate. It is very easily obtained and allows comparisons of accuracies to be drawn among the various SDG indicators.
- The Margin Of Error (MOE), which is the half of the confidence interval (based on the SE). It enables building well-founded inferential comparisons among different values of an indicator and hypothesis testing.

As stated in Eurostat (2013), it is recommended to use precision measures geared to the type of indicators of reference. The precision measures recommended are the following: (i) the CV for totals and means of continuous variables; and (ii) the SE for ratios and changes close to zero. The second recommendation aims to avoid situations in which precision requirements lead to a large increase in the sample size when the indicator approaches zero. Moreover, absolute precision measures for the percentages or proportions of any characteristic are symmetrical. However, for specific surveys, experts should decide whether or not to use the precision measure that is the most demanding, in the case of the proportion value that makes the study variable the most relevant. In other words, use of the SE may be preferred if the study variable becomes more relevant as the estimated proportions approach 0.5. Use of the CV may be preferred if the study variable becomes more relevant as the estimated proportions tend towards zero. However, use of either the SE or the CV is equally preferable if the study variable becomes more relevant as the estimated proportions approach one.

4.2. Basic theory: the measures of accuracy

This section briefly illustrates the basic theory considering a linear parameter of interest, such as that introduced in Chapter 3.

4.2.1. Sampling variance

The sampling variance of estimator $\hat{Y}_d$ is the expectation, under repeated sampling, $E_P(\cdot)$, of the squared differences

$$V_P\left(\hat{Y}_d\right) = E_P\left(\hat{Y}_d - Y_d\right)^2, \quad (4.1)$$

where the expectations and variance in Equation 4.1 operate on the random sample-membership indicators λ_i $(i = 1, \dots N)$, in which

$$E_P(\lambda_i) = \pi_i, \qquad V_P(\lambda_i) = \pi_i(1-\pi_i) \text{ and }, \ Cov_P\left(\lambda_i\lambda_j\right) = \pi_{ij} - \pi_i\pi_j, \qquad (4.2)$$

with $\pi_{ij} = E_P\left(\lambda_i\lambda_j\right)$.

Considering the model-assisted approach illustrated in Section 3.3.1 of these Guidelines, Särndal (1992; see result 6.1) demonstrates that a linear approximation for the sampling variance of $V_P\left(\hat{Y}_d\right)$ can be defined as:

$$V_P(\hat{Y}_d) = \sum_{i=1}^{N}\sum_{j=1}^{N} Cov_P(\lambda_i\lambda_j)\frac{u_i^o}{\pi_i}\frac{u_j^o}{\pi_j}\, y_{di}\gamma_{dj} \quad (4.3)$$

where u_i^o are the population fit residuals of the model in Equation 3.4:

$$u_i^o = y_i - m(x_i;\hat{\beta}_d). \quad (4.4)$$

An approximate sample-unbiased estimator of $V_P(\hat{Y}_d)$ is

$$\hat{V}_P(\hat{Y}_d) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{\pi_{ij}} Cov_P(\lambda_i\lambda_j)\, a_i g_{iS}\hat{u}_i a_j g_{jS}\hat{u}_j\gamma_{di}\gamma_{dj}, \qquad (4.5)$$

where $\hat{u}_i$ are the sample fit residuals of the regression set out at Equation 3.4, and

$$g_{jS} = \frac{\omega_i}{a_i} \quad (4.6)$$

are the multiplicative factors for correcting the direct weights a_i.

Comment 4.1. Conservative approximation of the variance. Simple conservative approximations of the estimate $\hat{V}_P(\hat{Y}_d)$ can be obtained by (i) substituting the y_i values for the sample residuals in Equation 4.3; and (ii) approximating the actual sampling design with sampling without replacement. An example of this for multi-stage stratified sampling is provided in Box 4.1 below.

Box 4.1. Estimate of the variance for stratified two-stage sampling designs

Consider the stratified two-stage sampling introduced in Box 3.1.

Approximation 1

An approximate (conservative) estimate of the variance can be obtained by using the equation for estimating sampling error in case of sampling with replacement in the first stage, and estimation through the regression estimator

$$\hat{V}_P(\hat{Y}_d) = \sum_{h=1}^{H} \frac{m_h}{m_h - 1} \sum_{\ell=1}^{m_h} \left(\hat{U}_{h\ell(d)} - \hat{\bar{U}}_{h(d)} \right)^2 \qquad (4.7)$$

$$\hat{U}_{h\ell(d)} = \sum_{i \in S_{(h\ell)}} \hat{u}_i \omega_i \gamma_{di} \quad , \quad \hat{\bar{U}}_{h(d)} = \frac{1}{m_h} \sum_{\ell=1}^{m_h} \hat{U}_{h\ell(d)}. \quad (4.8)$$

Approximation 2

A more conservative approximation can be obtained by applying the equation related to sampling with replacement in the first stage:

$$\hat{V}_P(\hat{Y}_d) = \sum_{h=1}^{H} \frac{m_h}{m_h - 1} \sum_{\ell=1}^{m_h} \left(\hat{Y}_{h\ell(d)} - \hat{\bar{Y}}_{h(d)} \right)^2 \qquad (4.7a)$$

$$\hat{Y}_{h\ell(d)} = \sum_{i \in S_{(h\ell)}} y_i \omega_i \gamma_{di} \quad , \quad \hat{\bar{Y}}_{h(d)} = \frac{1}{m_h} \sum_{\ell=1}^{m_h} \hat{Y}_{h\ell(d)}. \quad (4.8a)$$

The "survey package" in R calculates this estimator (Lumley, 2019).

4.2.1.1. Sampling variance of the balanced sampling

Consider a balanced sample design that respects the following balancing equations (see Section 3.5.3 of these Guidelines), on the x variables:

$$\sum_{i \in S} \frac{x_i}{\pi_i} = \sum_{i \in U} x_i. \quad (4.9)$$

Suppose that estimator $\hat{Y}_d$ (as given in Equation 3.2) is the HT estimator $\hat{Y}_{HT,d}$ given

$$\hat{Y}_{HT,d} = \sum_{i=1}^{N} y_i \gamma_{di} \lambda_i a_i. \quad (4.10)$$

Deville and Tillé (2005) proposed an approximate expression of the variance for $\hat{Y}_d$ (as expressed in Equation 4.10) for balanced sampling that respects Equation 4.9. This approximation, based on the Poisson sampling theory, is given as

$$V_P(\hat{Y}_d) \cong \sum_{i \in U} \left(\frac{1}{\pi_i} - 1\right) \eta_{di}^2, \quad (4.11)$$

where

$$\eta_{di} = y_{di} - \pi_i x_i' \beta_d \quad (4.12)$$

and

$$\beta_d = \left(\sum_{j=1}^{N} x_j x_j' \pi_j (1 - \pi_j)\right)^{-1} \sum_{j \in U} \pi_j \left(\frac{1}{\pi_j} - 1\right) x_j y_{dj} \quad (4.13)$$

Equation 4.11 uses a variance expression based on the Poisson sampling design that is not a fixed sample size design. Nevertheless, the use of the terms η_i in Equation 4.11, instead of the original values y_{dj}, introduces fixed sample sizes in variance computation. In practice, this is an application of the approximate variance of a Conditional Poisson design (Deville and Tillé, 2005). To clarify Expression 4.11, let us consider a stratified SRSWOR design. According to Equation 4.9 (Falorsi and Righi, 2015)

$$\hat{V}_P(\hat{Y}_d) = [N/(N-H)] \sum_{h=1}^{H} \sigma_{d,h}^2 N_h \left(\frac{N_h}{n_h} - 1\right) \quad (4.14)$$

where $\sigma_{d,h}^2$ is the variance of the $y_i \gamma_{di}$ values in stratum h. If $[N/(N-H)](1/N_h) \approx 1/(N_h - 1)$, the latter expression of the variance would approximate the variance of the HT estimate in the stratified SRSWOR design. The above approximation is proved true when the number of domains *H* remains small compared to overall population size N, and when domain sizes N_h are large.

Equation 4.11 can be estimated by

$$\hat{V}_P(\hat{Y}_d) \cong \sum_{i \in S} \frac{1}{\pi_i} \left(\frac{1}{\pi_i} - 1\right) \hat{\eta}_{di}^2 \quad \text{and}$$

$$\hat{\beta}_d = \left(\sum_{j=1}^{n} x_j x_j' (1 - \pi_j)\right)^{-1} \sum_{j \in S} \left(\frac{1}{\pi_j} - 1\right) x_j y_j \gamma_{dj}.$$

4.2.2. Model variance

The model variance of estimator $\hat{Y}_d$ is the model expectation of the squared differences (Chambers, 2015; p. 73)

$$V_M(\hat{Y}_d) = E_M[\hat{Y}_d - Y_d]^2, \quad (4.17)$$

where the model expectation and variance in Equation 4.17 operate on the y_i $(i = 1, \dots N)$, variables, keeping the λ_i indicators as fixed in which, considering the general model established at Equation 3.4,

$$y_i = m(x_i; \beta) + u_i,$$

which thus yields

$$E_M(y_i) = m(x_i; \beta),\ V_M(y_i) = \sigma_i^2\ ,\ Cov_M(y_i y_j) = \sigma_{ij}.$$

Using the linear WM in Equation 3.8, and the expectations given by Equation 3.5 results in (Chambers and Clark, 2015; p. 73):

$$V_M(\hat{Y}_d) = \sigma^2 \left\{ (N_d - n_d) + (X - X_S)' \left(\sum\nolimits_{j \in S} x_j x_j' \frac{1}{c_j} \right)^{-1} (X - X_S) \right\}, (4.17a)$$

where $V_M(u_i) = c_i \sigma^2$.

Since

$$\hat{\sigma}^2 = (n - p) \sum\nolimits_{i \in S} \left[y_i - x_i' \left(\sum\nolimits_{j \in S} x_j x_j' \frac{1}{c_j} \right)^{-1} \sum_{j=1}^{n} x_j y_{dj} \frac{1}{c_j} \right]^2 \quad (4.17b)$$

is an un unbiased estimator of σ^2 ,where p denoting the number of elements in x_j, the prediction variance $V_M(\hat{Y}_d)$ can be estimated as

$$\hat{V}_M(\hat{Y}_d) = \hat{\sigma}^2 \left\{ (N_d - n_d) + (X - X_S)' \left(\sum\nolimits_{j \in S} x_j x_j' \frac{1}{c_j} \right)^{-1} (X - X_S) \right\}. \quad (4.18)$$

4.2.3. Global variance

The GV proposed by Wolter (1985; see also Nedyalkova and Tillé, 2008) considers both elements of randomness: the model and the sampling design. The sampling design determines the observed set, and the model defines the random mechanism that generates the value of target variable y for each unit in the population.

GV is defined as the double expectation, for both the sampling design and the model, of the squared difference between the estimate and its overall expected value:

$$GV(\hat{Y}_d) = E_P E_M \left[\hat{Y}_d - E_P E_M (\hat{Y}_d) \right]^2 . \quad (4.19)$$

Using Kendall and Stuart (1976; p. 196), *GV* can be expressed as the sum of two terms:

$$GV(\hat{Y}_d) = V_P \left[E_M (\hat{Y}_d) \right] + E_P \left[V_M (\hat{Y}_d) \right]. \quad (4.20)$$

The first term,

$$V_P\left[E_M\left(\hat{Y}_d\right)\right] = V_P\left(\sum_{i=1}^{n} m(x_i;\beta)\ \omega_i\ \gamma_{di}\right), \quad (4.21)$$

represents the sampling variance of the theoretical values $m(x_i;\beta)$. In the following paragraphs, the square root of $V_P\left[E_M\left(\hat{Y}_d\right)\right]$ is denoted as the sampling error. Adopting a plug-in technique and substituting the unknown values of $m(x_i;\beta)$ with the estimates $\hat{y}_i = m\left(x_i;\hat{\beta}\right)$, the first term of the GV can be estimated with standard sampling variance estimation methods, as

$$\hat{V}_P\left[\hat{E}_M\left(\hat{Y}_d\right)\right] = \hat{V}_P\left(\sum_{i=1}^{n} m\left(x_i;\hat{\beta}\right)\omega_i\ \gamma_{di}\right). \quad (4.23)$$

The second term of the right-hand side of Equation 4.20 is the sampling expected value of the model variance. In the following equation, the square root of $E_P\left[V_M\left(\hat{Y}_d\right)\right]$ is denoted as the measurement error. As demonstrated in Appendix A.4.1 of these Guidelines, a plug-in roughly unbiased estimate of $E_P\left[V_M\left(\hat{Y}_d\right)\right]$ can be given by

$$\hat{E}_P\left[\hat{V}_M\left(\hat{Y}_d\right)\right] = \hat{V}_M\left(\hat{Y}_d\right), \quad (4.24)$$

where $\hat{V}_M\left(\hat{Y}_d\right)$ is the model-based estimate of the variance $V_M\left(\hat{Y}_d\right)$, which can be obtained bystandard statistical techniques.

Finally, the estimate of *GV* can be obtained as the sum of estimates of its terms, given respectively in Equations 4.23 and 4.24 and results in:

$$\widehat{GV}\left(\hat{Y}\right) = \hat{V}_P\left[\hat{E}_M\left(\hat{Y}_d\right)\right] + \hat{E}_P\left[\hat{V}_M\left(\hat{Y}_d\right)\right]. \quad (4.25)$$

Comment 4.2. If the estimator is expressed as in Equation 3.2 with weights given by Equations 3.6 and 3.7, there is no model expectation, and the GV collapses to the well-known sampling variance.

The GV can be particularly interesting when the countries produce the indicators of the data disaggregation plan with statistical procedures that utilize model-based approaches (to estimate the parameter of a specific statistical model, as a probability) jointly with the inference based on the sampling design, to define the total (or the mean) of the parameter over the whole population. A noteworthy example of this approach concerns the prevalence of moderate or severe food insecurity, as illustrated in Section 5.3 below. In this case the parameter of interest is expressed as

$$Y_d = \sum_{i=1}^{N} m(x_i;\beta_d)\,\gamma_{di}. \quad (4.25)$$

Equation 4.25 can be estimated as:

$$\hat{Y}_d = \sum_{i=1}^{n} m\left(x_i;\hat{\beta}_d\right)\omega_i\gamma_{di} = \sum_{i=1}^{n} \hat{y}_i\,\omega_i\gamma_{di}, \quad (4.26)$$

where $\hat{\beta}_d$ is the model-based sample estimate of the super-population parameterβ_d, ω_i are the sampling weights and $\hat{y}_i = m(x_i; \hat{\beta}_d)$.

It is noted that the GV constitutes a useful and straightforward tool to measure accuracy, as it enables a separate evaluation of the contribution of each component of randomness.

4.3. Case study: SDG Indicator 2.1.2 – Prevalence of moderate or severe food insecurity in the population, based on the FIES

This section illustrates how the GV approach can be adopted to estimate the accuracy of the prevalence of food insecurity by a disaggregation domain at a given level of severity (SDG Indicator 2.1.2). The methodology is based on the general theory presented in the previous section. In addition, specific methodological details on the inferential properties of the prevalence estimates are also presented.

4.3.1. Brief description of the methodology for SDG Indicator 2.1.2

This section summarizes the main equations used; all other methodological details are given in Appendix A.4.2.

Let $\tilde{y}_i = m(x_i; \beta)$ be the unknown probability for the $i-th$ unit of being food-insecure at a given level of severity of food insecurity. The parameter of interest, $\bar{Y}$, is the population mean of the individual $\tilde{y}_i$ values, which is actually the prevalence of food insecurity at a given level of severity:

$$\bar{Y} = \frac{1}{N}\sum_{i\epsilon U} m(x_i; \beta) = \sum_{i\epsilon U} \tilde{y}_i = \frac{Y}{N} \qquad (4.27)$$

where $Y = \sum_{k\epsilon U} \tilde{y}_i$ is the population total of probabilities $\tilde{y}_i$.

The prevalence of food insecurity for subpopulation U_d is:

$$\bar{Y}_d = \frac{1}{N_d}\sum_{i\epsilon U_d} \tilde{y}_i = \frac{Y_d}{N_d}, \qquad (4.28)$$

where $Y_d = \sum_{i\epsilon U_d} \tilde{y}_i$ is the total of the $\tilde{y}_i$ probabilities values in U_d.

Through a stratified two-stage random sampling without replacement, a sample *S* is selected, as illustrated in Box 3.1.The sample estimates of $\bar{Y}$ and $\bar{Y}_d$ are:

$$\hat{\bar{Y}} = \left(1/\sum_{i\epsilon S} a_{mod,i}\right)\sum_{i\epsilon S} \hat{y}_i\, a_{mod,i}\,, \qquad \hat{\bar{Y}}_d = \left(1/\sum_{i\epsilon S_d} a_{mod,i}\right)\sum_{i\epsilon S_d} \hat{y}_i\, a_{mod,i}, \qquad (4.29)$$

where $\hat{y}_i$ is the sample model-based estimate of $\tilde{y}_i$, $a_{mod,i}$ is the *modified* sampling weight. The sampling weights $a_{mod,i}$ are computed in such a way that their sum reproduces the overall sample size *n*:

$$\sum_{i\epsilon S} a_{mod,i} = \sum_{i\epsilon S} \omega_i n \Big/ \sum_{j=1}^{n} \omega_i \cong n. \qquad (4.30)$$

It is useful to note that the estimates $\hat{\bar{Y}}$ and $\hat{\bar{Y}}_d$ can be expressed in the form of ratio estimators:

$$\hat{\bar{Y}} = \frac{\hat{Y}}{\hat{N}} \quad , \quad \hat{\bar{Y}}_d = \frac{\hat{Y}_d}{\hat{N}_d}, \quad (4.31)$$

with

$$\hat{Y} = \sum_{i\epsilon S} \omega_i \ \hat{y}_i \ , \qquad \hat{N} = \sum_{i\epsilon S} \omega_i, \ \ \hat{Y}_d = \sum_{i\epsilon S_d} \omega_i \, \hat{y}_i, \qquad \hat{N}_d = \sum_{i\epsilon S_d} \omega_i \, \hat{y}_i. \ (4.32)$$

Therefore, it can be seen that the estimates $\hat{\bar{Y}}$ and $\hat{\bar{Y}}_d$ are non-linear estimators of the corresponding population parameters.

For the sake of brevity, this section only discusses the estimate $\widehat{GV}\left(\hat{\bar{Y}}_d\right)$ of $GV\left(\hat{\bar{Y}}_d\right)$. The expression of $GV\left(\hat{\bar{Y}}_d\right)$ is provided, as well as the equations for $GV\left(\hat{\bar{Y}}\right)$ and $\widehat{GV}\left(\hat{\bar{Y}}\right)$ (see Appendix A4.1). The estimate of the *GV* for the subpopulation is obtained as the sum of the estimates of its simple components

$$\widehat{GV}\left(\hat{\bar{Y}}_d\right) = \hat{V}_P\left[\hat{E}_M\left(\hat{\bar{Y}}_d\right)\right] + \hat{E}_P\left[\hat{V}_M\left(\hat{\bar{Y}}_d\right)\right], \qquad (4.33)$$

where

$$\hat{V}_P\left[\hat{E}_M\left(\hat{\bar{Y}}_d\right)\right] = \left(\frac{\hat{N}}{n\,\hat{N}_d}\right)^2 \sum_{h=1}^{H} \frac{m_h}{m_h - 1} \sum_{\ell=1}^{H} \left(\hat{Z}_{h\ell(d)} - \hat{\bar{Z}}_{h(d)}\right)^2 \qquad (4.34)$$

and

$$\hat{E}_P\left[\hat{V}_M\left(\hat{\bar{Y}}_d\right)\right] = \hat{V}_M\left(\hat{\bar{Y}}_d\right), \qquad (4.35)$$

in which

$$\hat{Z}_{h\ell(d)} = \sum_{i\in S_{h\ell(d)}} \hat{z}_{di} a_{mod,i} \quad , \quad \hat{\bar{Z}}_{h(d)} = \frac{1}{m_h} \sum_{\ell=1}^{m_h} \hat{Z}_{h\ell(d)},$$

with $\hat{z}_{di} = \dfrac{\hat{N}}{n\hat{N}_d}\left(\hat{y}_i - \hat{\bar{Y}}_d\right) a_{mod,i}$ and $S_{h\ell(d)} = S_{(h\ell)} \cap U_d$.

The above results are obtained by using the plug-in estimates $\hat{y}_i$ and $\hat{\bar{Y}}_d$ of the unknown $\tilde{y}_i$ and $\bar{Y}_d$ terms. The resulting expression (Equation 4.35) derives from Equation A1.2. Furthermore, the linear approximations of the estimators are considered. Finally, the estimation of $\hat{V}_P\left[\hat{E}_M\left(\hat{\bar{Y}}_d\right)\right]$ can be implemented with the help of the *Survey* function in R, considering the variance of the subpopulation mean of variable $\hat{y}_i$. As illustrated in Section 4.3.2., this component constitutes the leading part of the GV.

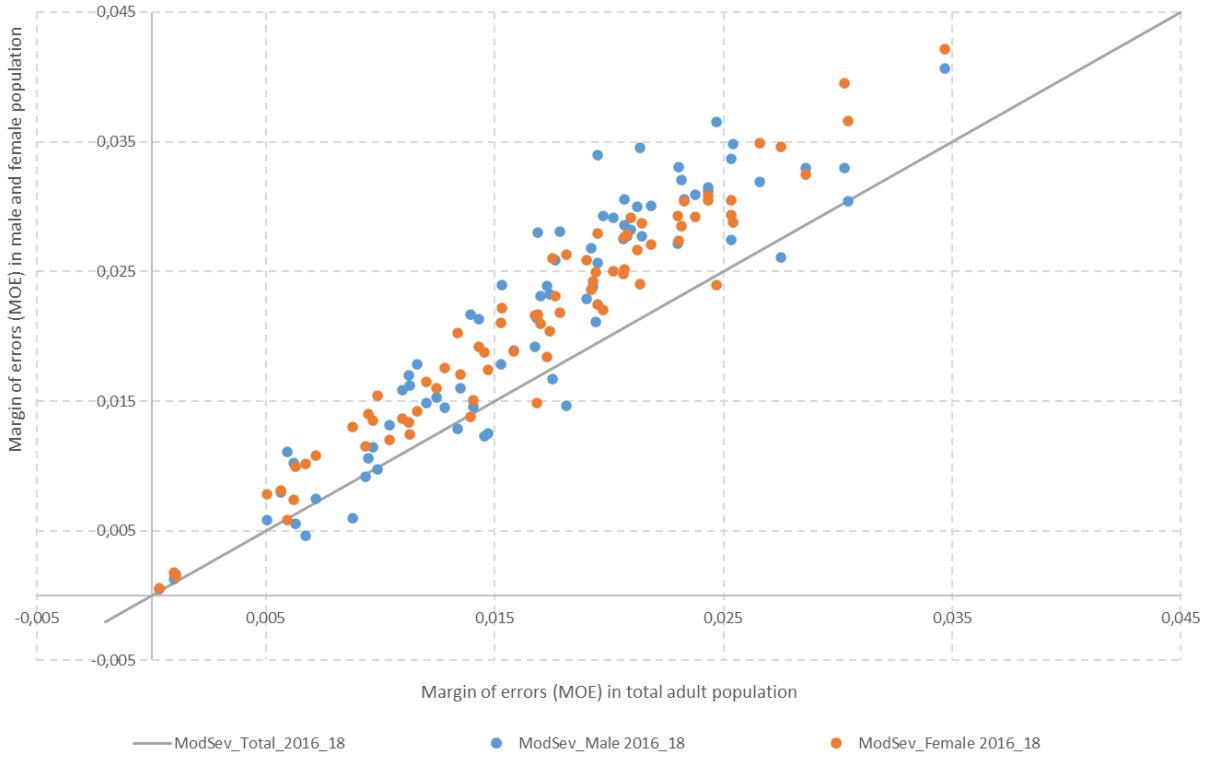
In the following sections, the square root of $\hat{V}_P\left[\hat{E}_M\left(\hat{\bar{Y}}_d\right)\right]$ is denoted as the sampling error and the square root of $\hat{E}_P\left[\hat{V}_M\left(\hat{\bar{Y}}_d\right)\right]$ as the measurement error.

4.3.2. Results by subpopulation: gender

For moderate or severe prevalence of food insecurity in the adult population overall and in the population disaggregated by gender, MOE at a confidence level of 90 percent is produced. The methodology explained in the previous section is applied to the full list of countries (77 in total) surveyed in the GWP, which gave its consent to the publication of its data on SDGs. The data reflected as three-year averages between the years 2016 and 2018 are used to derive GV estimates by gender.

Figure 4.1 shows how MOE by gender compares with those in the total adult population. It can be observed that, in most countries, MOE by gender is larger than that in the total adult population. Table 4.1 shows that, on average (unweighted), the MOE in the male adult population is 1.27 times larger than in the total adult population, while the MOE in the female adult population is 1.25 times larger. This magnitude of difference is expected, as the populations of males and females are smaller, and a larger variability is therefore expected. Table 4.1 also presents average ratios between the errors (due to measurement or sampling) and prevalence rates at moderate or severe food insecurity, for the total adult population and by gender. MOE by gender is, on average, approximately 1.3 percent larger than MOE in the prevalence referred to the whole population.

Figure 4.1. Margins of error for moderate or severe food insecurity prevalence, in male and female populations



Source: FAO, 2020.

Table 4.1. Average margins of error for moderate or severe food insecurity prevalence, in male and female populations

Prevalence at moderate or severe level	Average Margin of Errors	Gender
	2016–2018	2016–2018
Gallup World Poll (GWP) – Total	0.017	
Gallup World Poll (GWP) – Female	0.021	1.25
Gallup World Poll (GWP) – Male	0.021	1.27

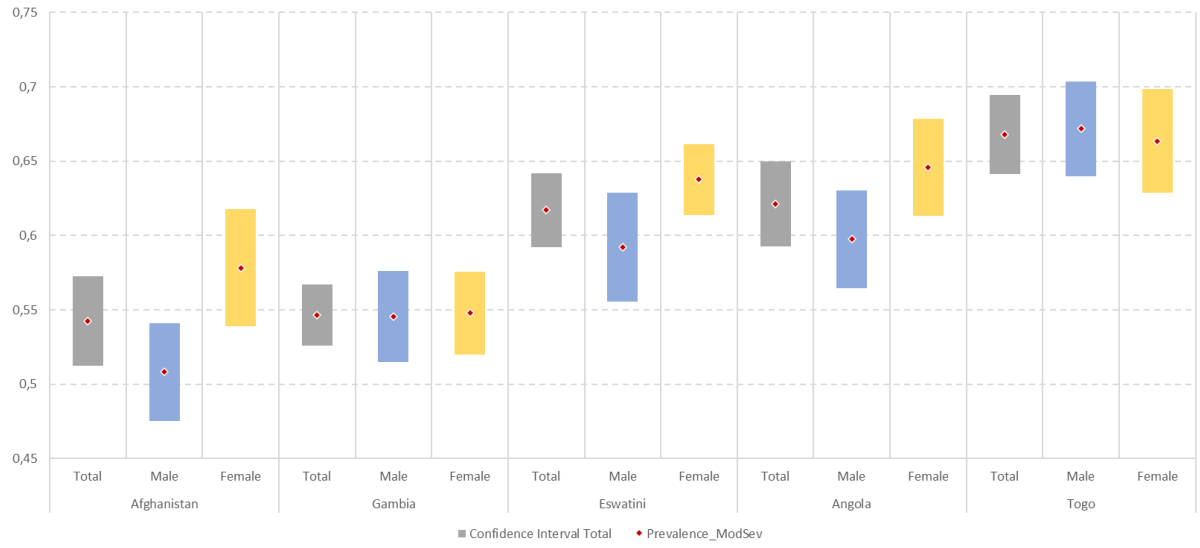
Table 4.2. Relative standard error for the prevalence of moderate or severe food insecurity (SDG Indicator 2.1.2), total and by gender, 2016–2018

Prevalence at moderate or severe level 2016–2018:	Average Relative Standard Error (RSE) = Average of SE/Prevalence*100	
	Sampling	Measurement
Gallup World Poll (GWP) – Total	7.07	0.59
Gallup World Poll (GWP) – Female	8.62	0.83
Gallup World Poll (GWP) – Male	9.00	0.89

Figure 4.2 illustrates confidence interval estimates of total, male and female populations for countries with different population sizes and with similar prevalence levels (between 0.5 and 0.7). It can be seen that the results do not change based on the size of the population.

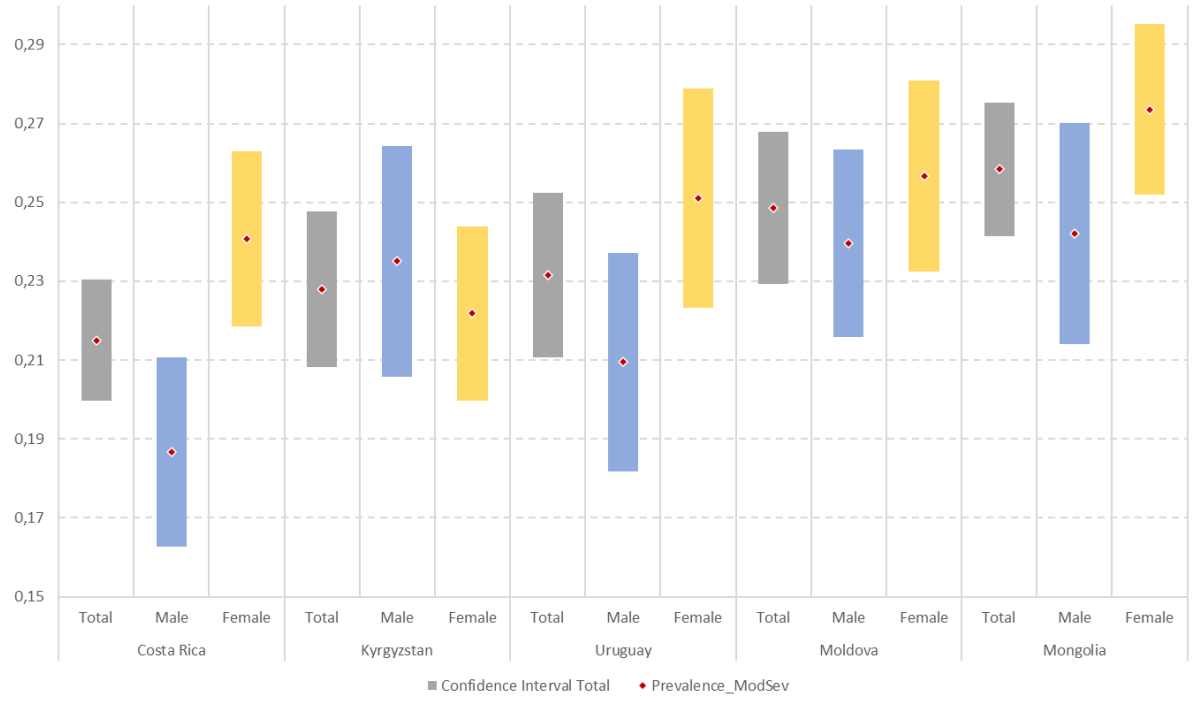
Figure 4.3 shows confidence interval estimates of total, male and female populations for countries with different population sizes and with similar prevalence levels (between 0.2 and 0.3).

Figure 4.2. Confidence interval for the prevalence of moderate or severe food insecurity (SDG Indicator 2.1.2) in Afghanistan, Gambia, Eswatini, Angola and Togo, total and by gender, 2016–2018



Source: FAO, 2020.

Figure 4.3. Confidence interval for the prevalence of moderate or severe food insecurity (SDG Indicator 2.1.2) in Costa Rica, Kyrgyzstan, Uruguay, Moldova and Mongolia, total and by gender, 2016–2018



Source: FAO, 2020.

More results for various countries are presented in Annex A.4.4 of these Guidelines, to enable comparison of countries with different characteristics, i.e. region, size of the population.

4.4. Summary of main recommendations

The main advice provided in this chapter are the following.

1. Calculating errors is a preliminary and required step for all actions carried out for data integration.

2. If the calculation of errors flags an excessively high inaccuracy, it must be decided whether to publish the data all the same, informing users of their level of reliability. Alternatively, improvement actions to ameliorate the accuracy of the data could be undertaken.

3. It is essential to measure and communicate the accuracy of disaggregated estimates. Users should have a say in determining the fitness for use of an estimate. Moreover, this supplies greater flexibility for data users, while ensuring that they are aware of the accuracy of the results.

4. The measure of accuracy can be based on the uncertainty resulting from the model and the sampling design. The sampling determines the observed set, and the model defines the random mechanism that generates the value of the target variable for each unit in the population.

5. A useful measure of accuracy is the GV, which explicitly considers the two sources of randomness above.

6. As particular cases, the GV includes the traditional standard indicators of accuracy: sampling variance and model variance.

7. In the experiment illustrated in this chapter, it can be seen that sampling variance constitutes the leading part of the GV. It can be estimated with R functions (see the *Survey* function and Annex 1).

Appendix A4.1. Estimate of the global variance component $E_P\left[V_M\left(\hat{Y}_d\right)\right]$

Let $V_M\left(\hat{Y}_d\right)$ be the model variance of $\hat{Y}_d$ and let $\hat{V}_M\left(\hat{Y}_d\right)$ be its model-based estimate. Let us suppose that $V_M\left(\hat{Y}_d\right)$ can be either directly expressed, or linearly approximated, as the model variance of a linear combination:

$$V_M\left(\hat{Y}_d\right) \cong V_M\left(\sum_{i=1}^{n} y_i \omega_i \gamma_{di}\right) = \sum_{i=1}^{n} \sigma_i^2 \omega_i^2 \gamma_{di} + \sum_{i=1}^{n}\sum_{j\neq i}^{n} \sigma_{ij}\omega_i\omega_j \gamma_{di}\gamma_{dj}. \quad (A1.1)$$

Thus, we can formulate $\hat{V}_M\left(\hat{Y}_d\right)$ as

$$\hat{V}_M\left(\hat{Y}_d\right) \cong \sum_{i=1}^{n} \hat{\sigma}_i^2 \omega_i^2 \gamma_{di} + \sum_{i=1}^{n}\sum_{j\neq i}^{n} \hat{\sigma}_{ij}\omega_i\, \omega_j \gamma_{di}\gamma_{dj},$$

where $\hat{\sigma}_i^2$ and $\hat{\sigma}_{ij}$ are the model-based estimate of the unknown terms σ_i^2 and σ_{ij}.

Starting from Equation $A1.1$, an approximate estimate of $E_P\left[V_M\left(\hat{Y}_d\right)\right]$ can be obtained as

$$E_P\left[V_M\left(\hat{Y}_d\right)\right] \cong \sum_{i=1}^{N} \sigma_i^2 \omega_i^2\, E_P(\lambda_i)\gamma_{di} + \sum_{i=1}^{N}\sum_{j\neq i}^{N} \sigma_{ij}\omega_i\, \omega_j\, E_P\left(\lambda_i\lambda_j\right)\gamma_{di}\gamma_{dj}$$

$$= \sum_{i=1}^{N} \sigma_i^2 \omega_i^2\, \pi_i\gamma_{di} + \sum_{i=1}^{N}\sum_{j\neq i}^{N} \sigma_{ij}\omega_i\, \omega_j\, \pi_{ij}\, \gamma_{di}\gamma_{dj}.$$

Thus, a sampling-based unbiased estimate of $E_P\left[V_M\left(\hat{Y}_d\right)\right]$ is

$$\hat{E}_P\left[V_M\left(\hat{Y}_d\right)\right] = \sum_{i=1}^{n} \frac{1}{\pi_i} \sigma_i^2 \omega_i^2\, \pi_i\gamma_{di} + \sum_{i=1}^{n}\sum_{j\neq i}^{n} \frac{1}{\pi_{ij}} \sigma_{ij}\omega_i\omega_j\, \pi_{ij}\, \gamma_{di}\gamma_{dj}.$$

$$= \sum_{i=1}^{n} \sigma_i^2 \omega_i^2\, \gamma_{di} + \sum_{i=1}^{n}\sum_{j\neq i}^{n} \sigma_{ij}\, \omega_i\omega_j\gamma_{di}\gamma_{dj} = V_M\left(\hat{Y}_d\right). \quad (A1.2)$$

Therefore, a plug-in roughly unbiased estimate of $E_P\left[V_M\left(\hat{Y}_d\right)\right]$ can be obtained by substituting, in the above expression, the estimates $\hat{\sigma}_i^2$ and $\hat{\sigma}_{ij}$ in place of the unknown terms σ_i^2 and σ_{ij}

$$\hat{E}_P\left[\hat{V}_M\left(\hat{Y}_d\right)\right] = \sum_{i=1}^{n} \hat{\sigma}_i^2 \omega_i^2\, \gamma_{di} + \sum_{i=1}^{n}\sum_{j\neq i}^{n} \hat{\sigma}_{ij}\omega_i\omega_j\, \gamma_{di}\gamma_{dj} = \hat{V}_M\left(\hat{Y}_d\right).$$

The above approximation holds if it can be assumed that for large sampling, the weights ω_i approximate the weights that would be obtained if the whole population were observed.

Appendix A4.2. Indicator of the prevalence of food insecurity

The estimator and its properties

The weights ω_i are computed, ensuring that the sample estimates of some known distributions reproduce the benchmark totals:

$$\omega_i = (1/\pi_i) g_{is}$$

with $g_{iS} \cong 1$ for large samples. Therefore, the following relation approximately holds

$$\sum_{j=1}^{n} \omega_i = N.$$

The estimate $\hat{\bar{Y}}$ can be expressed as:

$$\hat{\bar{Y}} = \frac{\hat{Y}\frac{n}{N}}{\hat{N}\frac{n}{N}} = \frac{\hat{Y}^*}{\hat{N}^*} = \frac{\hat{Y}}{\hat{N}},$$

where $\hat{Y}^* = \sum_{i\epsilon S} a_{mod,i}\, \hat{y}_i = \hat{Y}\frac{n}{N}$ and $\hat{N}^* = \sum_{i\epsilon S} a_{mod,i} = \hat{N}\frac{n}{N}$.

Similarly, the expression of $\hat{\bar{Y}}_d$ can be reformulated as

$$\hat{\bar{Y}}_d = \frac{\hat{Y}_d\frac{n}{N}}{\hat{N}_d\frac{n}{N}} = \frac{\hat{Y}_d^*}{\hat{N}_d^*} = \frac{\hat{Y}_d}{\hat{N}_d},$$

where $\hat{Y}_d^* = \sum_{i\epsilon S_d} a_{mod,i}\, \hat{y}_i = \hat{Y}_d\frac{n}{N}$ and $\hat{N}_d^* = \sum_{i\epsilon S_d} a_{mod,i} = \hat{N}_d\frac{n}{N}$.

To derive the model and sampling expectations of the estimates $\hat{\bar{Y}}$ and $\hat{\bar{Y}}_d$, their linear approximations are considered:

$$\hat{\bar{Y}} \cong \hat{Y} + \frac{1}{N}\left[(\hat{Y} - Y) - \hat{Y}(\hat{N} - N)\right]$$

$$= \hat{Y} + \frac{1}{N\frac{n}{N}}\left[(\hat{Y} - Y)\frac{n}{N} - \hat{Y}(\hat{N} - N)\frac{n}{N}\right], \quad \text{(A. 2.1)}$$

$$\hat{\bar{Y}}_d \cong \bar{Y}_d + \frac{1}{N_d}\left[(\hat{Y}_d - Y_d) - \bar{Y}_d(\hat{N}_d - N_d)\right],$$

$$= \bar{Y}_d + \frac{1}{N_d\frac{n}{N}}\left[(\hat{Y}_d - Y_d)\frac{n}{N} - F_d(\hat{N}_d - Y_d)\frac{n}{N}\right]. \quad \text{(A. 2.2)}$$

Considering the above linear approximations, and because for large samples, the calibration factors g_{is} converge towards 1, the (approximate) unbiasedness of the estimates can be proved as below:

$$E_P E_M(\hat{Y}) = E_P\left[\sum_{i\epsilon U} \omega_i\, \lambda_i\, \tilde{y}_i\right] \cong E_P\left[\sum_{i\epsilon U} \frac{1}{\pi_i} \lambda_i\, \tilde{y}_i\right] = \sum_{i\epsilon U} \frac{1}{\pi_i} E_P\,(\lambda_i)\tilde{y}_i = Y,$$

$$E_P E_M(\widehat{N}) = E_P\left[\sum_{i\epsilon U} \omega_i\, \lambda_i\ \right] \cong E_P\left[\sum_{i\epsilon U} \frac{1}{\pi_i} \lambda_i\ \right] = \sum_{i\epsilon U} \frac{1}{\pi_i} E_P\,(\lambda_i) = N,$$

$$E_P E_M(\hat{Y}_d) = E_P\left[\sum_{i\epsilon U_d} \omega_i\, \lambda_i\, \tilde{y}_i\right] \cong E_P\left[\sum_{i\epsilon U_d} \frac{1}{\pi_i} \lambda_i\, \tilde{y}_i\right] = E_P \sum_{i\epsilon U_d} \frac{1}{\pi_i} E_P\,(\lambda_i)\tilde{y}_i = Y_d,$$

$$E_P E_M(\widehat{N}_d) = E_P\left[\sum_{i\epsilon U_d} w_i\, \lambda_i\right] \cong E_P\left[\sum_{i\epsilon U_d} \frac{1}{\pi_i} \lambda_{ki}\right] = E_P \sum_{i\epsilon U_d} \frac{1}{\pi_k} E_P\,(\lambda_i) = N_d.$$

Inserting the above results into the linear approximations at Equations A.2.2 and A.2.3, the unbiasedness of the estimator can be derived.

The GV and its estimate

The GV of the estimator $\hat{\bar{Y}}$, is

$$GV\left(\hat{\bar{Y}}\right) = V_P\left[E_M\left(\hat{\bar{Y}}\right)\right] + E_P\left[V_M\left(\hat{\bar{Y}}\right)\right].$$

We have: $E_M\left(\hat{\bar{Y}}\right) = \frac{1}{\widehat{N}} \sum_{i\epsilon S} E_M(\hat{y}_i)\, \omega_i = \frac{1}{\widehat{N}} \sum_{i\epsilon S} \tilde{y}_i\, \omega_i = \frac{\hat{Y}}{\widehat{N}} = \frac{N}{\widehat{N} n} \sum_{i\epsilon S} \tilde{y}_i\, a_{mod,i} = \frac{\hat{\bar{Y}}^*}{\widehat{N}^*},$

with $\hat{\bar{Y}}^* = \sum_{i\epsilon S} \tilde{y}_i\, a_{mod,i}$.

Applying a linear approximation on the above ratio,

$$V_P\left[E_M\left(\hat{\bar{Y}}\right)\right] = V_P\left(\frac{\hat{\bar{Y}}^*}{\widehat{N}^*}\right) \cong V_P\left[\frac{1}{N^*}\left(\hat{\bar{Y}}^* - \bar{Y}\widehat{N}^*\right)\right]$$

where $N^* = E_P(\widehat{N}^*) = E_P(\widehat{N})\frac{n}{N} = n.$

Adopting the Woodruff (1971) method, $V_P\left[E_M\left(\hat{\bar{Y}}\right)\right]$ can be expressed as the standard expression of the variance of the totals (with sampling weights $a_{mod,i}$) of the transformed variables z_i:

$$V_P\left[E_M\left(\hat{\bar{Y}}\right)\right] = V_P\left[\sum_{i\in S} z_i\, a_{mod,i}\right],$$

where

$$z_i = \frac{1}{N^*}(m(x_i;\beta) - \bar{Y}) = \frac{1}{n}(m(x_i;\beta) - \bar{Y}),$$

with $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} m(x_i;\beta)$.

Assuming that there is no model covariance among the $\hat{y}_i$ values, the model variance $V_M\left(\hat{\bar{Y}}\right)$ can be expressed as

$$V_M\left(\hat{\bar{Y}}\right) = \left(\sum_{i\in U} a_i\,\lambda_{ki}\right)^{-2} \sum_{i\in U} V_M(\hat{y}_i)\, a_{mod,i}^2\,\lambda_k.$$

To derive the sampling expected value of the above model variance, the linear approximation of the model variance is considered. Thus,

$$E_P\left[V_M\left(\hat{\bar{Y}}\right)\right] = E_p\left[\sum_{i\in S} V_M(\hat{y}_i)\, a_{mod,i}^2\right] \Big/ E_p\left[\left(\sum_{i\in S} a_{mod,i}\right)^2\right]$$

Since $\sum_{i\in S} a_{mod,i} = n$, then $E_p\left[\left(\sum_{i\in S} a_{mod,i}\right)^2\right] = n^2$. Thus,

$$E_P\left[V_M\left(\hat{\bar{Y}}\right)\right] \cong \sum_{i\in U} V_M(\hat{y}_i)\frac{1}{\pi_i^2}\frac{n^2}{N^2}E_p(\lambda_i)/n^2 = \sum_{i\in U} V_M(\hat{y}_i)\frac{1}{\pi_i}\frac{n^2}{N^2}/n^2$$

$$= \frac{1}{N^2}\sum_{i\in U} V_M(\hat{y}_i)\frac{1}{\pi_i}. (A.2.3)$$

The estimate of *GV* is obtained as the sum of the estimates of its simple components:

$$\widehat{GV}\left(\hat{\bar{Y}}\right) = \hat{V}_P\left[\hat{E}_M\left(\hat{\bar{Y}}\right)\right] + \hat{E}_P\left[\hat{V}_M\left(\hat{\bar{Y}}\right)\right] \quad .$$

The above is obtained using the plug-in estimates of the unknown terms.

The estimate of $V_P\left[E_M\left(\hat{\bar{Y}}\right)\right]$ can be approximated by considering the variance of stratified two-stage sampling with replacement (see Boxes 4.1 and 4.2):

$$\hat{V}_P\left[\hat{E}_M\left(\hat{\bar{Y}}\right)\right] = \sum_{h=1}^{H} \frac{m_h}{m_h - 1} \sum_{\ell=1}^{m_h} \left(\hat{Z}_{h\ell} - \hat{\bar{Z}}_h\right)^2,$$

where

$$\hat{Z}_{h\ell} = \sum_{i \in S_{(\ell h)}} \hat{z}_i a_{mod,i} \quad , \quad \hat{\bar{Z}}_h = \frac{1}{m_h} \sum_{\ell=1}^{m_h} \hat{Z}_{h\ell},$$

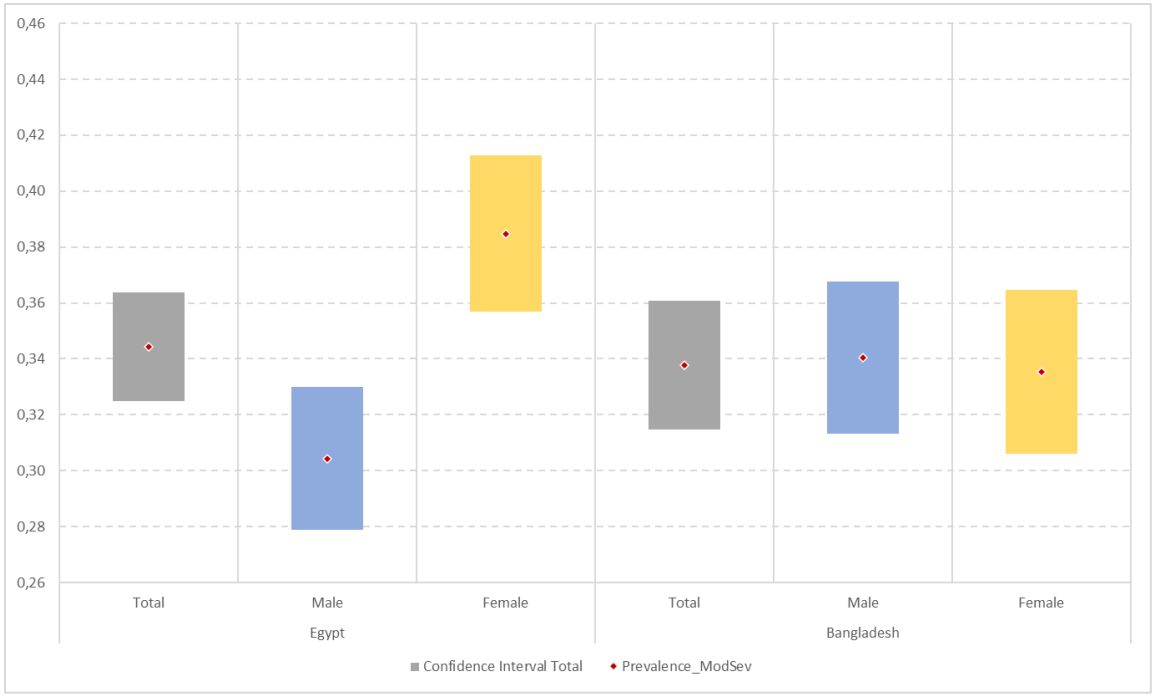
with $\hat{z}_i = \frac{1}{n}\left(\hat{y}_i - \hat{\bar{Y}}\right)$.

Then, an approximated unbiased estimate of $E_P\left[V_M\left(\hat{\bar{Y}}\right)\right]$ is

$$\hat{E}_P\left[\hat{V}_M\left(\hat{\bar{Y}}\right)\right] \cong \frac{1}{n^2} \sum_{i \in S} \hat{V}_M(\hat{y}_i).$$

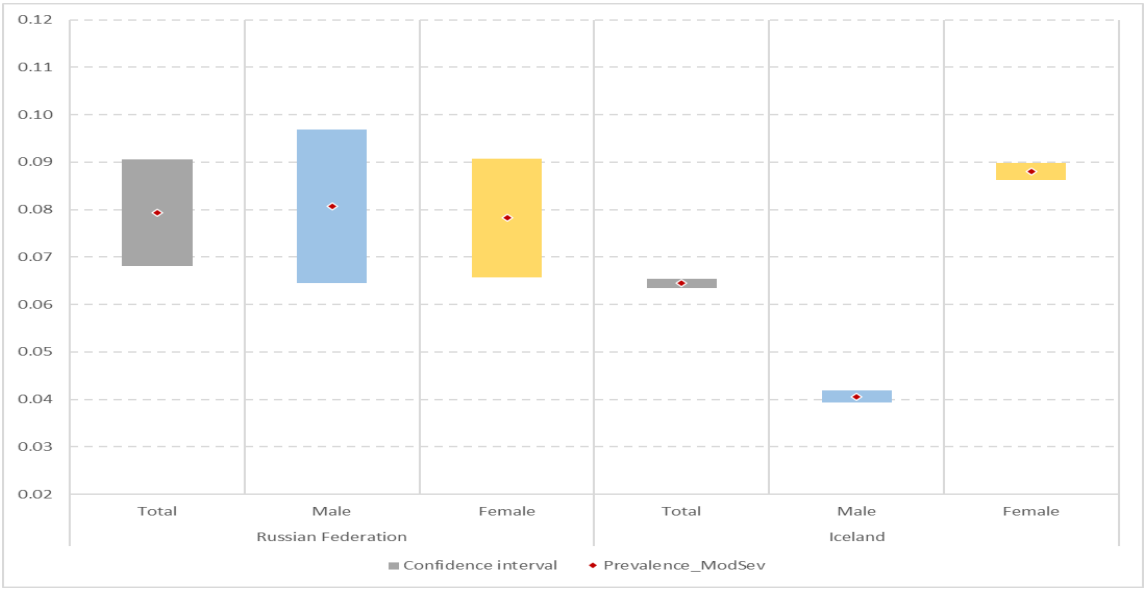
Appendix A4.3. Estimates of confidence intervals for different countries

Figure A.4.1. Confidence interval for the prevalence of moderate or severe food insecurity (SDG Indicator 2.1.2), total and by gender, 2016–2018



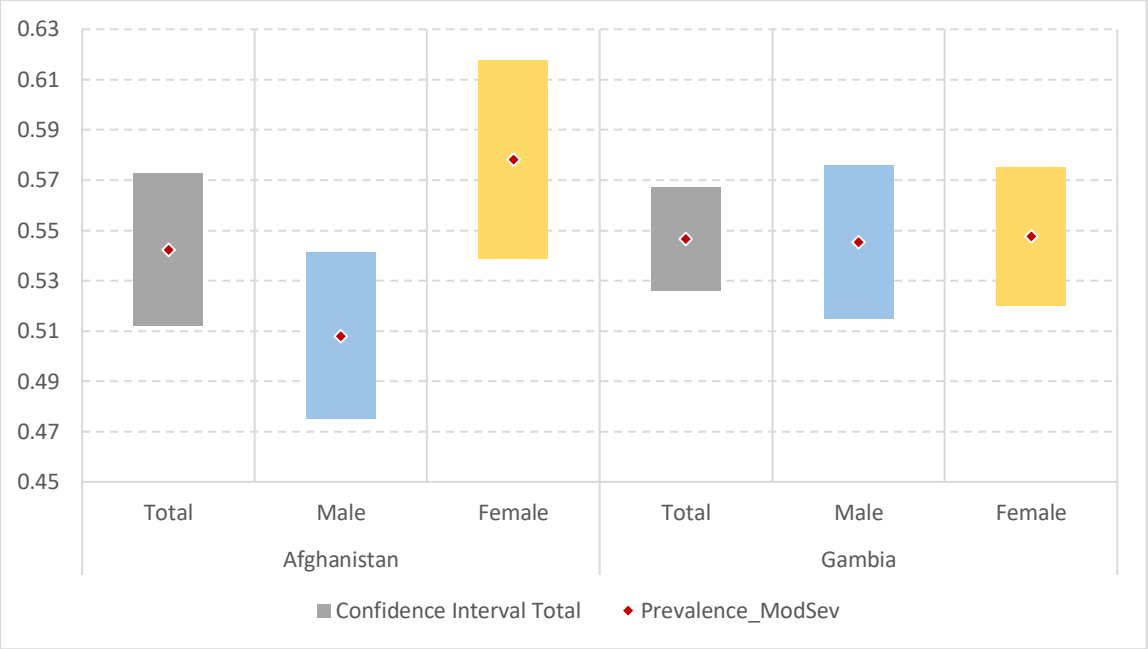
Source: FAO, 2020.

Figure A.4.2. Confidence interval for the prevalence of moderate or severe food insecurity (SDG Indicator 2.1.2) in the Russian Federation and Iceland, total versus by gender, 2016–2018



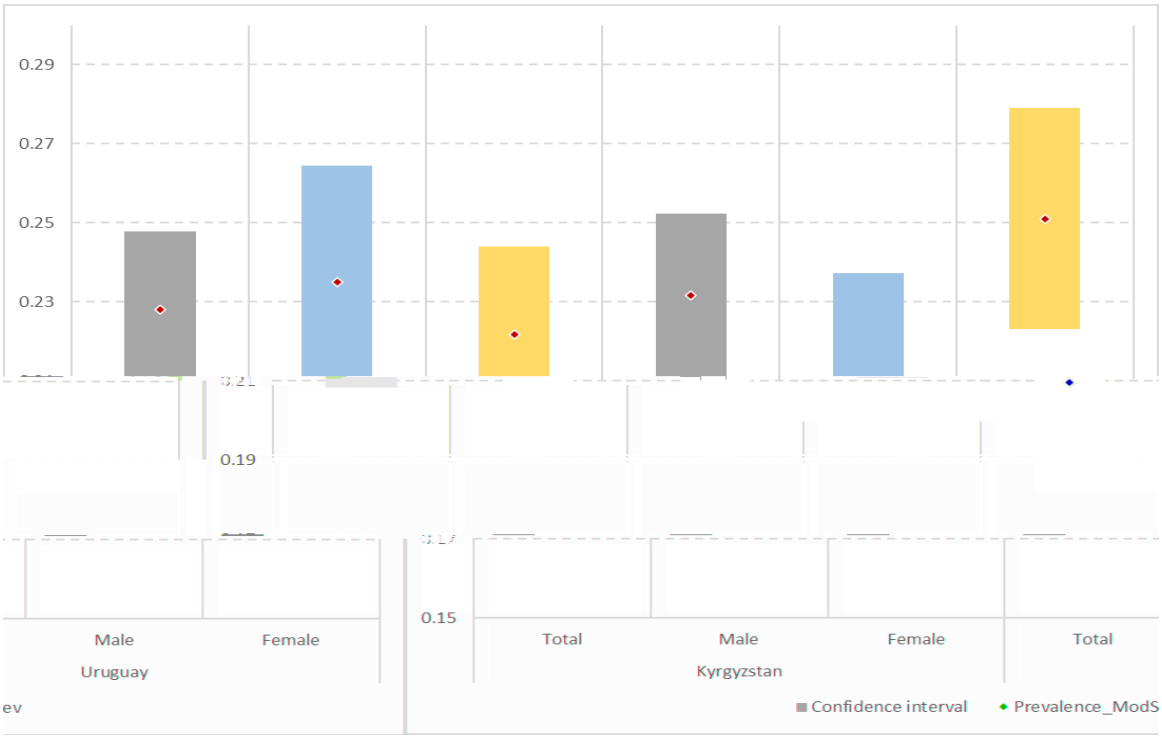
Source: FAO, 2020.

Figure A.4.3. Confidence interval for the prevalence of moderate or severe food insecurity (SDG Indicator 2.1.2) in Afghanistan and Gambia, total and by gender, 2016–2018



Source: FAO, 2020.

Figure A.4.4. Confidence interval for the prevalence of moderate or severe food insecurity (SDG Indicator 2.1.2) in Kyrgyzstan and Uruguay, total and by gender, 2016–2018

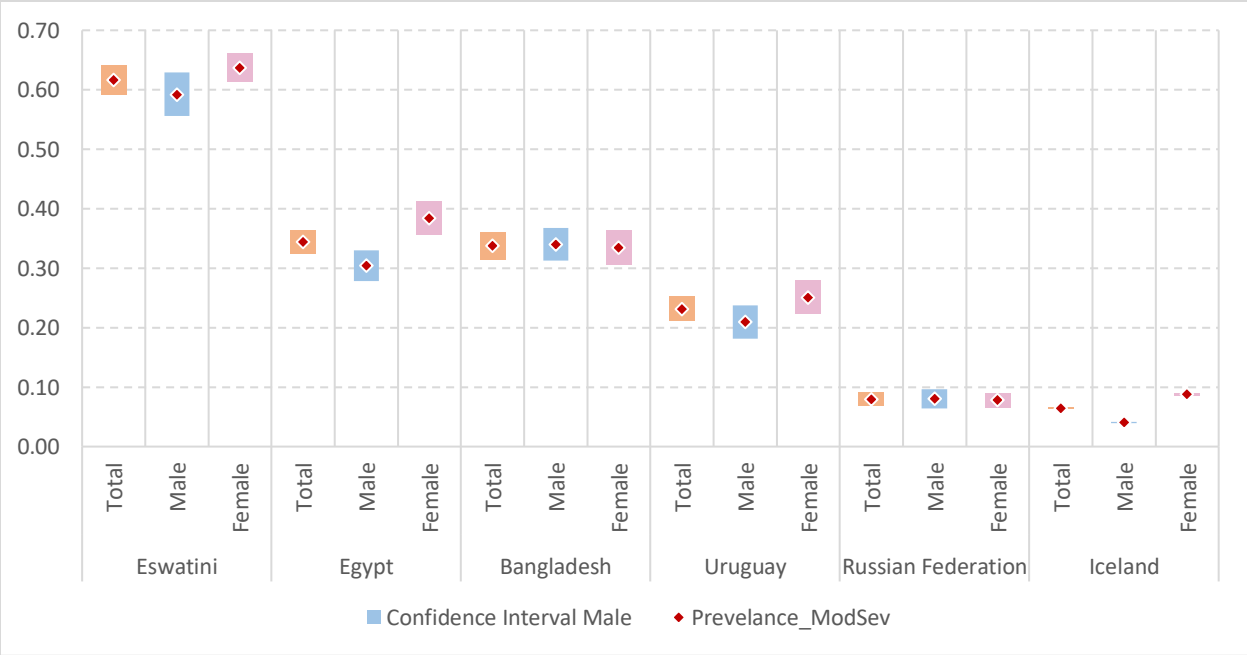


Source: FAO, 2020.

Figure A.4.5. Confidence interval for the prevalence of moderate or severe food insecurity (SDG Indicator 2.1.2) in Moldova, Mongolia, Mexico, Tajikistan and Argentina, total and by gender, 2016–2018: Countries with different population sizes, with similar prevalence levels

Source: FAO, 2020.

Figure A.4.6. Confidence interval for the prevalence of moderate or severe food insecurity (SDG Indicator 2.1.2) in Eswatini, Egypt, Bangladesh, Uruguay, the Russian Federation and Iceland, total and by gender, 2016–2018: Countries with different population sizes, with similar prevalence levels



Source: FAO, 2020.

Chapter 5. Integrated use of two surveys

5.1. Introduction

This chapter focuses on how to leverage the integrated use of different surveys to achieve data disaggregation. In particular, the combined use of a small survey that focuses on a target phenomenon– is considered, together with a more extensive survey or census that does not measure the target phenomenon, but rather gathers data on general-use variables (such as demographic variables or geographical variables) .

The case study discussed covers a great deal of interesting empirical contexts that can occur when producing disaggregated data. In particular, most countries have at least one large-scale survey, which collects general-use variables; examples are the census, administrative data, and large-scale household surveys. Some of the target phenomena required for data disaggregation are difficult (or too costly) to measure with a large-scale survey, such as the gender gap, the risk of malnutrition or food waste. Therefore, a wise choice would be to measure the target phenomenon with a small-scale survey; the costs and measurement errors could thus be limited and controlled. On the other hand, small surveys often do not collect the variables required for data disaggregation.

The strategy proposed here overcomes these problems. In particular, small surveys are used to estimate the parameters of a regression-type statistical model that links the target variable to some explanatory (or auxiliary) variables. Then, the target values for the units of the large-scale survey are predicted applying the regression parameters to the units' auxiliary variables. The two surveys must share the same set of auxiliary variables selected as regressors in the statistical model.

Thus, it is possible to take advantage of both the small survey, to measure a specific phenomenon precisely and with low costs, and of a more extensive study, to produce cross-tabulation at the disaggregated level.

This chapter first illustrates the basic methodology (Section 5.2) and then provides details on a case study focused on food insecurity in Malawi (Section 5.3). This application aims to show the different steps of the approach and how to overcome the possible difficulties that can characterize the application of the method in real empirical contexts. Section 5.4 gives advice based on the findings of the practical exercise.

5.2. Methodology

5.2.1. The projection estimator

The basic methodology is illustrated in the pioneering work of Kim and Rao (2012), that considered two independent surveys and a model-assisted approach for inference. The main points of Kim and Rao (2012) are summarized below.

Let us consider the case of two surveys with two independent samples. The first survey is characterized by a large sample A_1 that collects only auxiliary information or general-use variables (demographic characteristics, employment, household composition, etc.). The second survey, with a much smaller sample A_2, provides information on both the variable of interest and the auxiliary variables.

Kim and Rao (2012) propose a model-assisted projection method of estimation based on a WM that results in asymptotically unbiased projection estimators. With this approach, synthetic or proxy values of a variable of interest are generated by, first, fitting the WM, linking the variable of interest to the auxiliary

variables, to the data from survey 2, A_2. Then, the variable of interest associated with the auxiliary variables observed in survey 1is predicted. The projection estimators are obtained from sample A_1 (based on survey 1) and associated synthetic values. Furthermore, as can be seen in Kim and Rao (2012), the synthetic data obtained through the model-assisted projection method can provide a useful tool for efficient domain estimation when the size of the sample in survey 1 is much larger than the size of sample in survey 2 (Figure 5.1).

Let

$$Y = \sum_{i=1}^{N} y_i$$

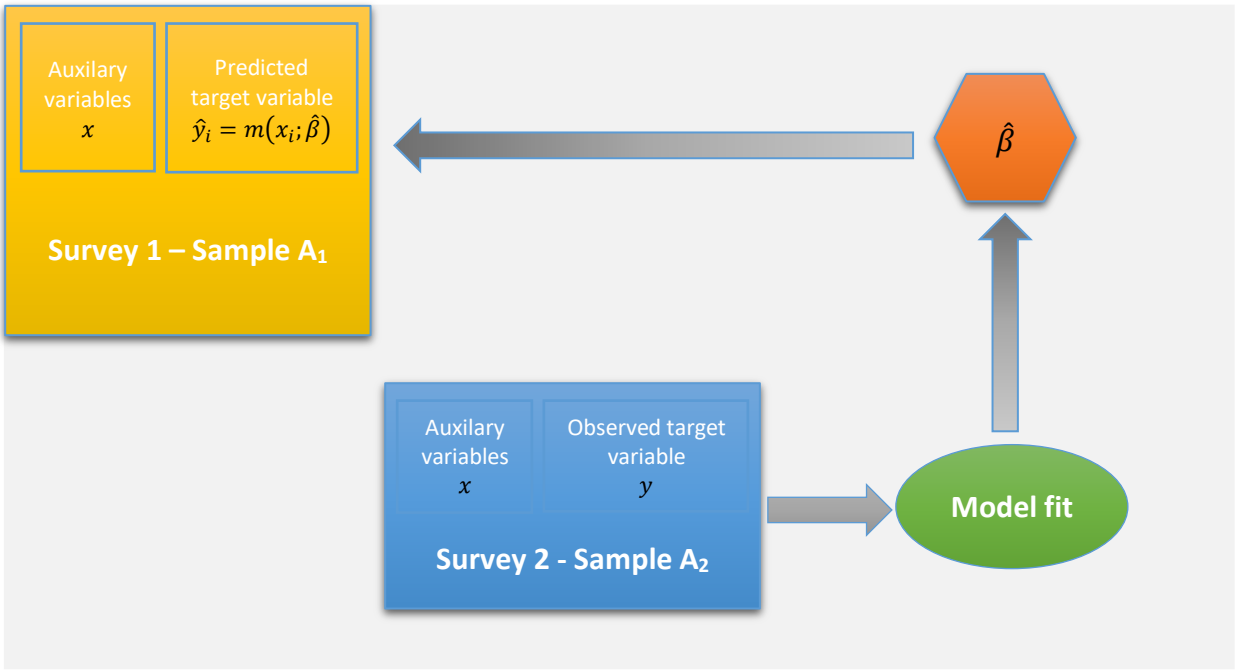
be the target population total, where y_i is the value of the variable of interest of the unit i , with *N* being the population size.

Suppose that the WM M is introduced, according to which the y_i values can be modelled as

$$y_i = m(x_i; \beta) + u_i,$$

where u_i is a random residual and $m(x_i; \beta)$ is a known function applied on the column vector of auxiliary variables x_i (of the $i-$ th unit), with β as the column vector of the model parameters. In the case of a simple regression model, $m(x_i; \beta) = x_i'\beta$, where x_i' is the transpose of x_i. The model expectation of u_i equals zero; this is denoted here as $E_M(u_i) = 0$. The two vectors x_i and β are congruent in terms of dimensions and the x_i values are available on both samples $A_1$and A_2.

Figure 5.1. Projection estimator



Source: FAO, 2020.

Let $\hat{\beta}$ be the estimator of β obtained from the second survey, using the data $\{(y_i, x_i): i \in A_2\}$, and let

$$\hat{y}_i = m(x_i; \hat{\beta})$$

indicate the predicted value of y_i, with $E_M(\hat{\beta}) = \beta$.

Let E_P denote the expectation under repeated sampling and let ω_{i1} be the sampling weights of sample A_1 that allow for computing sample-unbiased estimates. If y_i is available in sample A_1,

$$\hat{Y}_1 = \sum_{i \epsilon A_1} \omega_{i1}\, y_i$$

would be a sample-unbiased estimator Y, where the sample-unbiasedness implies that, under repeated sampling, the expected value of $\hat{Y}_1$ is equal to the unknown *Y*:

$$E_P(\hat{Y}_1 - Y) = 0.$$

However, estimator $\hat{Y}_1$ cannot be implemented from sample A_1, unlike the following estimator of *Y*, which is based on the synthetic values $\hat{y}_i = m(x_i; \hat{\beta})$ reported in the first survey data file:

$$\hat{Y}_{PR} = \sum_{i \epsilon A_1} \omega_{i1}\, \hat{y}_i = \sum_{i \epsilon A_1} \omega_{i1}\; m(x_i; \hat{\beta}). \qquad (5.1)$$

Estimator $\hat{Y}_p$ is called a PRojection estimator (PR) or a synthetic estimator, because $\hat{y}_i = m(x_i; \hat{\beta})$ can be viewed as a projection of y_i using the auxiliary variable x_i, or as a synthetic value of y_i. The PR estimator in Equation 5.1 is derived from the WM $E_M(y_i | x_i) = m(x_i; \beta)$; however, the results do not depend on the validity of the WM, although this affects the efficiency of the estimators.

5.2.1.a. Bias and variance

The estimator $\hat{Y}_{PR}$ is unbiased with respect to both the model and the sampling design, as follows:

$$E_P E_M[\hat{Y}_{PR} - E_M(Y)] = 0.$$

The asymptotic sample bias of $\hat{Y}_p$ is

$$Bias(\hat{Y}_{PR}) = E_P(\hat{Y}_{PR}) - Y \cong \sum_{i=1}^{N} [y_i - m(x_i; \beta_0)],$$

with β_0 denoting the estimate of β when observing the entire population.

The asymptotic sample bias from the second survey can be estimated as

$$\hat{B}ias(\hat{Y}_{PR}) = \sum_{i \epsilon A_2} \omega_{i2}\, [y_i - m(x_i; \hat{\beta})],$$

where ω_{i2} are the sampling weights of A_2, which allow for computing sample-unbiased estimates for the second survey.

Thus, $\hat{Y}_{PR}$ is not sample-unbiased, except when

$$\sum_{i \epsilon A_2} \omega_{i2} \left[y_i - m(x_i; \hat{\beta}) \right] = 0 \qquad (5.2)$$

Therefore, to guarantee sample-unbiasedness, estimate $\hat{\beta}$ should be obtained by respecting the condition established in Equation 5.2. For generalized linear models (such as heteroscedastic linear regression models or logistic models) to satisfy Equation 5.2, it is assumed that the first element of x_i is equal to unity, which means that the model has an intercept.

Kim and Rao (2012) demonstrate that the sample variance of $\hat{Y}_p$ is given by

$$Var\left(\hat{Y}_{PR}\right) = V_p\left(\sum_{i \epsilon A_1} \omega_{i1}\, m(x_i; \beta_0)\right) + V_p\left(\sum_{i \epsilon A_2} \omega_{i2} \left[y_i - m(x_i; \beta_0) \right]\right), \quad (5.3)$$

where the first term on the right-hand side is the variance due to sampling, in survey 1, of the population predictions (with the β_0 value), and the second term is the variance due to sampling, in survey 2, of the population residuals (for the predictions with the β_0 value). The latter term tends to be small if the residuals are small, because model *m* is sufficiently predictive. Kim and Rao (2012) present a pseudo-replication method for correct variance estimation without requiring access to the data $\{(y_i, x_i): i \in A_2\}$ from survey 2.

5.2.1.b. Domain estimation

Let *d* denote a particular domain for which disaggregated data must be produced.

Let

$$Y_d = \sum_{i=1}^{N} y_i \gamma_{di}$$

be the total of the target variable for the *d*-th domain, where γ_{di} is the domain membership variable. The projection estimator of the total Y_d is given simply by:

$$\hat{Y}_{PR,d} = \sum_{i \epsilon A_1} \omega_{i1}\, m(x_i; \hat{\beta}) \gamma_{di} \qquad (5.4)$$

The condition for sample-unbiasedness is

$$\sum_{i \epsilon A_2} \omega_{i2} \left[y_i - m(x_i; \hat{\beta}) \right] \gamma_{di} = 0. \qquad (5.5)$$

To satisfy Equation 5.5, vector x_i must include the γ_{di} values, which means that the model has a domain intercept. The condition in Equation 5.5 is fulfilled if, in sample A_2, domain *d* has a sufficient sample size. The latter condition is satisfied for larger domains or by planning the sample size (for sample A_2) of domain *d* when defining the sample design (as illustrated in Chapter 3).

However, in general, the condition established under Equation 5.5 cannot be ensured in the sampling design phase for minimal geographical domains. Therefore, it is preferable to focus on the model conditions that provide negligible bias. In Kim and Rao (2012; Expression 13), it can be seen that the relative $E_P\left(\hat{Y}_{PR,d} - Y_d\right)/Y_d$ bias can be expressed as

$$\frac{E_P\left(\hat{Y}_{PR,d} - Y_d\right)}{Y_d} = -\frac{Cov\left[\gamma_{di}, \left(y_i - m(x_i;\beta)\right)\right]}{\bar{N}_d\, \bar{Y}_d}, \quad (5.6)$$

where $Cov\left[\gamma_{di}, \left(y_i - m(x_i;\beta)\right)\right]$ is the population covariance between the domain membership indicators, γ_{di}, the model residuals $y_i - m(x_i;\beta)$, $\bar{N}_d$ is the population mean of the domain membership indicators, and $\bar{Y}_d$ is the population mean of the product variable $\gamma_{di}\, y_i$.

Therefore, to make sure that the relative bias is close to zero, the model should be specified to ensure that the model residuals depend slightly on the domain membership variables:

$$Cov\left[\gamma_{di}, \left(y_i - m(x_i;\beta)\right)\right] \cong 0. \quad (5.7)$$

This constraint will be satisfied if the WM is correctly specified.

From the relationship in Equation 5.6, it can also be seen that in large domains, for which $\bar{N}_d\, \bar{Y}_d$ is large, the relative bias becomes negligible.

Finally, the variance can be obtained easily from Equation 5.3 as

$$Var\left(\hat{Y}_{PR,d}\right) = Var\left(\sum_{i\epsilon A_1} \omega_{i1}\, m(x_i;\beta_0)\, \gamma_{di}\right) + \left(\sum_{i\epsilon A_2} \omega_{i2}\, [y_i - m(x_i;\beta_0)]\right). \quad (5.8)$$

Here too, a pseudo-replication method can be used to compute correct variance estimation without requiring access to the data $\{(y_i, x_i): i \in A_2\}$ from survey 2.

5.2.1.c. Extensions

The Kim and Rao (2012) approach can be modified to leverage some enhancements that can yield interesting results in some empirical applications. This section shows some of the possible extensions.

Small area estimation

Consider a well-known model with domain random effects

$$y_i = m(x_i;\beta) + u_i + z_d \text{ for } i \in U_d,$$

in which u_i is random noise and z_d a domain effect, such as $E_M(u_i) = 0$, $E_M(u_i^2) = c_i\sigma_u^2$, $E_M(\varepsilon_i\varepsilon_j) = 0$ for $i \neq j$ and $E_M(z_d) = 0$, $E_M(z_d^2) = \sigma_d^2$, $E_M(z_d z_{d'}) = 0$ for $d \neq d'$.

In this case, the estimates $\hat{\beta}$ and $\hat{z}_d$ and the parameters β and z_d are obtained from sample A_2, with standard SAE techniques. The projection estimator at Equation 5.4 is reformulated as:

$$\hat{Y}_{PR(SAE),d} = \sum_{i\epsilon A_1} \omega_{i1}\left[\, m(x_i;\hat{\beta}) + \hat{z}_d\right]\gamma_{di}.$$

FAO (2014) gives a detailed description of this case, defining the theoretical conditions for unbiasedness and variance; it also reports simulations on real data that demonstrate the effectiveness of the estimator. Chapter 6 of these Guidelines explores the projection estimator with SAE techniques in further depth.

Projecting on the census

A useful extension is when the first survey is the census on the population *U*. In this case, we can obtain the projection estimator by summing the predictions computed over the census records:

$$\hat{Y}_{PR} = \sum_{i\epsilon U} m(x_i;\hat{\beta}).$$

Subsampling from the first survey

Another useful extension that can be developed quickly is the case where A_2 is a subsample of A_1. In this case, illustrated in Figure 5.1a, the conditions at Equations 5.5 and 5.6 for sample-unbiasedness still hold.

This should be defined when designing the overall sampling strategy.

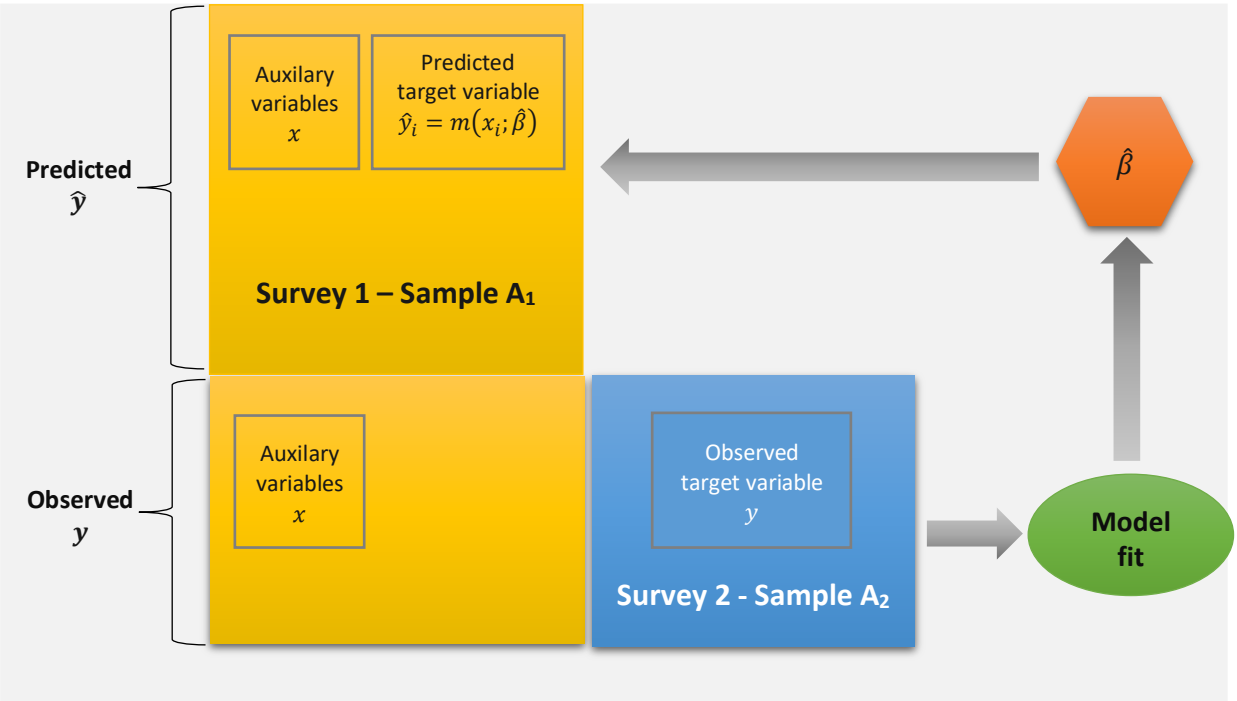
The practical application of the estimator is strongly enhanced, as the two surveys share the same metadata and definitions.

On the other hand, the sampling variance is larger than in the case of two independent surveys, because there is covariance between the two surveys:

$$Var\left(\hat{Y}_{PR}\right) = V_p\left(\sum_{i\epsilon A_1}\omega_{i1}\, m(x_i;\beta_0)\right) + V_p\left(\sum_{i\epsilon A_2}\omega_{i2}\,[y_i - m(x_i;\beta_0)]\right) +$$

$$+ Cov_P\left[\left(\sum_{i\epsilon A_1}\omega_{i1}\, m(x_i;\beta_0)\right)\left(\sum_{i\epsilon A_2}\omega_{i2}\,[y_i - m(x_i;\beta_0)]\right)\right].$$

Figure 5.1.a. Projection estimator for subsampling



Source: FAO, 2020.

5.2.2. Selection of auxiliary variables

Proper identification of the predictors x_i is a crucial step to ensure the quality of the projection estimator. The use of variable selection methods can be helpful when there are many potential regressors, although it may be challenging to select regressors when there is multicollinearity.

The literature on this topic is ample (Ryan, 2008), and a detailed description of the various possible approaches is beyond the scope of these Guidelines. Harrell (2015) provides a comprehensive summary on the common methods of variable selection in regression methods. The tools used could also facilitate the selection of these methods, as various available statistical packages/functions are available in different types of statistical software; examples are the statistical packages to run lasso regression and the Boruta or random forest regression packages that are available in R and Python. Boruta is the auxiliary variables selection method formulated in Kursa and Rudnicki (2010), and it was used to select the auxiliary variables in the empirical exercise illustrated below in Section 5.3.

5.2.3. Model assumptions and performance

Another relevant issue is that of verifying model assumptions and performance based on the selected model type – ordinary regression, generalized linear regression, *etc*. Since the model selection and the model assumptions can be highly volatile, these Guidelines cannot fully cover all possible options. Nevertheless, to provide some material for reflection on the methods available to assess model

performance and assumptions, as well as to inform users as to the steps to follow, the following section presents some methods for generalized linear models based on the empirical exercise illustrated.

Those common methods include Pearson's chi-square test, and Hosmer and Lemeshow's goodness of fit (GOF).[4] Pearson's chi-square test basically checks whether the model with predictors fits significantly better than a model with only an intercept (i.e. a null model). An associated *p*-value of less than 0.001 shows that the predicted model as a whole fits significantly better than an empty model. Hosmer and Lemeshow's GOF test deals with binary data. The model fits well when there is no significant difference between the model and the observed data (i.e. when the *p*-value is above 0.05). Besides, the Akaike information criterion (AIC) can also be used when assessing the quality of a model through comparison of related models. For instance, the performance of the model after the selection of variables, and all variables in the model, can be compared through AIC. In all cases, it is observed that the model with the smallest AIC is the model with a variable selection based on Boruta. However, it is important to consider that most general methods to assess inference in case of independent and identically distributed (iid) variables (simple random sampling) can be misleading when applied to a sample obtained with stratified two-stage selection and unequal weighting of the units. Archer *et al.* (2007) demonstrate that standard goodness-of-fit tests are not always suitable for complex sample survey data, and propose alternative tests that account for complex design features.

5.3. Case study: food insecurity in Malawi

5.3.1 Background

Malawi was chosen to test the method proposed in Kim and Rao (2012) for various reasons:

1. **Availability of microdata from the World Bank's LSMS:**
   The Integrated Household Survey (IHS) is implemented by the Government of Malawi through the country's National Statistical Office (NSO) every five years to monitor and evaluate the changing conditions of Malawian households. The IHS is an important source of information on the country's socio-economic indicators, which are fundamental to the evidence-based policy formulation process and monitoring progress towards achieving the SDGs.

   The Fourth Integrated Household Survey (IHS4) is the fourth full survey conducted under the umbrella of the World Bank's Living Standards Measurement Study – Integrated Surveys on Agriculture (LSMS-ISA), and was fielded from April 2016 to April 2017. It was also the third round of the Integrated Household Panel Survey (IHPS) 2016, which ran concurrently with the IHS4 main cross-section fieldwork. The IHS4 cross-section collected information from a sample of 12 480 households statistically designed to be representative at both national, district, urban and rural levels. The IHPS 2016 collected information from a sample of all families and a subsampling of 102 individuals out of the 204 in the original baseline enumeration areas deemed representative at the national and urban/rural levels.

   The IHS4 consists of five core questionnaire instruments: the Household Questionnaire, the Agriculture Questionnaire, the Fishery Questionnaire, the Community Questionnaire, and the

[4] Rather than goodness of fit, measuring explained variation could also be considered to assess the model's performance. For logistic regression, pseudo R-squares – especially McFadden (1974) and Cox and Snell (1989) – could be considered as alternatives. The pR2() function in R produces pseudo R-square estimates and more.

Individual Questionnaire. Details on the structure and scope of the questionnaire instruments are provided in NSO (2017).[5]

The IHS4 Individual Questionnaire includes a module on the food insecurity scale, Module L: Subjective Assessment of Well-being. Therefore, it enables full control over the quality check for the resulting estimators. That is, the results of the proposed method (or projected estimates) can be compared easily with the estimates already derived from the IHS4 data collected on the FIES.

2. **Availability of microdata from the Gallup World Poll**
   Voices of the Hungry, a FAO initiative launched in 2013 in collaboration with Gallup, Inc., has been measuring food insecurity worldwide by using an experience-based tool. An eight-question Global FIES, that can be applied easily in many different contexts, has been developed and included in the annual GWP to generate estimates of the prevalence of moderate or severe food insecurity (SDG Indicator 2.1.2). The FIES survey module collects information on the experience of people (individuals over the age of 15) with food insecurity, through annual nationally representative samples (of a size of approximately 1 000 individuals) in more than 150 countries. This enables a global standard of reference to compare the measures obtained in different parts of the world and in different contexts.

   In making the global assessment, preference is given to suitable and reliable FIES data available from large national surveys, whereas FAO data collected in the GWP are used to compile the estimates for countries for which there are no other data and/or to fill gaps in terms of time series (FAO, 2020a).

   To identify the best wording and phrasing (in the local Chichewa and Chitumbuka languages) to express the intended meaning of each question, a FIES linguistic adaptation exercise took place in Malawi in July 2013 (Manyamba, 2013).

   In tandem with the release of *The State of Food Security and Nutrition in the World 2020* (FAO, 2020a), access to the FIES microdata for all countries where FAO has collected data through the GWP and for which national statistical authorities have consented to their use (in total, 77 countries) is granted through the Food and Agriculture Microdata Catalogue (FAM).[6]

3. **The FIES estimates from both data sources, the IHS4 2016 and the GWP 2016, are similar:**

   The IHS4 and GWP microdata for 2016 were used to estimate SDG Indicator 2.1.2 by the *Voice of the Hungry* team and the estimates were reported to be close to one other.

5.3.2. Available auxiliary information

Smith, Rabbitt and Coleman-Jensen (2017) studied determinants of food insecurity in the world and considered a list of variables collected in the GWP that include individual, household, and socio-economic characteristics:

[5] All other relevant documents can be accessed in the World Bank Microdata Library (Integrated Household Panel Survey 2010-2013-2016 (Long-Term Panel, 102 EAs).
[6] FAM microdata website.

1. demographic characteristics – gender, age, number of adults and children in the household, marital status, education level, residential information (e.g. rural/urban);

2. social capital characteristics – social capital, social network;

3. economic characteristics – household income, employment status; and

4. country characteristics – unemployment rate, GDP per capita.

Smith, Rabbitt and Coleman-Jensen (2017) concluded that low levels of education, less social capital, weak social networks, low household income, and being unemployed are strongly associated with the likelihood of experiencing food insecurity. The authors stated that causality cannot be inferred from their results, because they did not attempt to correct for the potential endogeneity of the determinants of food insecurity. However, strong correlations between those variables and food insecurity are observed and can be used to understand the global determinants of food insecurity.

Smith, Rabbitt and Coleman-Jensen (2017) and other similar studies in the literature guided the selection of auxiliary variables in this section. However, the availability of data also posed constraints on the selection, as in any economic and statistical analysis. For Malawi, IHS4 provides access to a vast range of information thanks to its five different questionnaires (see Section 5.3.1, Point 1, above). The available information is mostly at household level; individual-level information is provided for a more restricted set of variables. In contrast, the GWP provides individual-level information on food insecurity in addition to a limited number of demographic and economic variables.

Overall, information on age, sex, income, education level, employment status and size of households are available in both surveys. Since the model proposed by Kim and Rao (2012) requires common auxiliary variables, that auxiliary information is used for the projection model here. An important issue to consider when selecting those variables is that their definitions should ideally be the same in both surveys, in order to obtain a consistent estimation model. The selected variables also satisfy this condition in general. New categorical variables were created, or available information was used to calculate the new variables. See Annex 5 for details.

5.3.3. Projection model

As explained in Section 5.2.1, the synthetic values $\hat{y}_i = m(x_i; \hat{\beta})$ are constructed through a known function $m(x_i; \hat{\beta})$ of $\hat{\beta}$, where the estimator $\hat{\beta}$ is obtained from the survey with the smaller sample.

The selection of the functional form for m relies heavily on the type of variables considered. Our variables of interest are:

1. the probability of being moderately or severely food-insecure (prob.ms), and

2. the probability of being severely food insecurity (prob.s).

These are required to estimate the prevalence of severe or moderate, and severe, food insecurity among adults,[7] as well as the number of food-insecure adults in the national population, that can be obtained easily with the help of population estimates. Based on their responses, the individuals or households

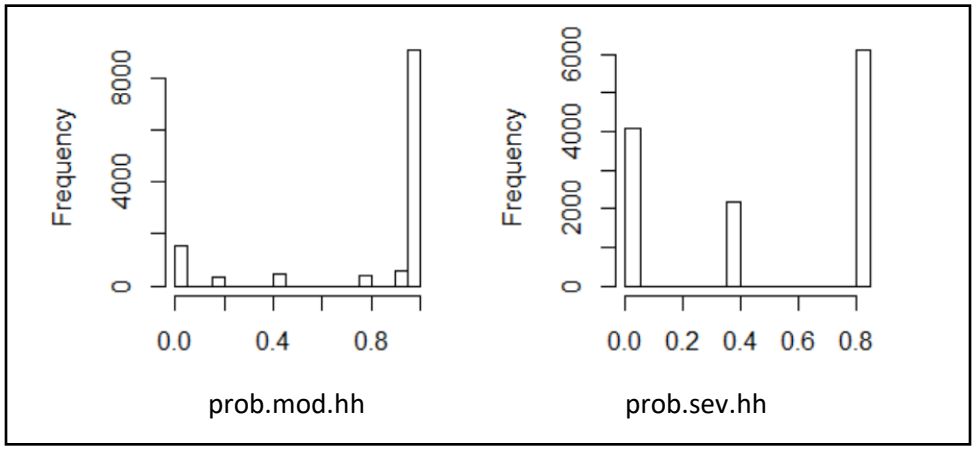
[7] Adults, defined as individuals older than 15 years of age, compose the reference population of the GWP.

interviewed in a nationally representative survey are assigned a probability of being in one of two classes: moderately food-insecure and severely food-insecure, as defined by two globally set thresholds[8].

The probability values are expressed as percentages/rates, thus ranging from 0 to 1. Figure 5.2 shows the distribution of the probability estimates for moderate or severe (prob.mod.hh) and severe (prob.sev.hh) food insecurity (see the IHS4 2016 data). Although the variable of interest is a continuous one, the values are accumulated around numbers that are very close to one another. In fact, the difference is very small and is observed only after the third or fourth decimal numbers.

Having seen that the probability values are very close to each other, it was decided to create categorical dependent variables, which led to using logistic or ordinal logistic regression models for projection. Initially, the dependent variables were grouped into three categories and the ordinal regression model was applied. However, it was observed that using ordinal regression models brought more complexity into the overall estimation process, without a significant contribution towards improving the estimates. Besides, the results were not easy to interpret the numbers of categories varied based on the data sets available. As the objective is to develop a flexible method for disaggregation that is also easy to implement, it was decided to use logistic regression with binary response variables y_ℓ : with $y_{\ell i} = 1$ if the probability of being ℓ – level food-insecure is higher than 0.5 for individual i and, $y_{\ell i} = 0$ otherwise, where $\ell = \{s, m\}$ for severe food insecurity and moderate or severe food insecurity, respectively. In other words, Category 1 is associated with a high probability of being food-insecure.

Figure 5.2. Histogram of the prevalence of moderate and severe food insecurity



Source: FAO, 2020.

[8] See FAO (2020a) for more details on the methodology.

Given a set of discrete and categorical auxiliary variables $x_i' = (1, x_{i1}, x_{i2}, \dots, x_{ik})$, it is assumed that $\mathcal{y}_\ell$ follows a Bernoulli distribution with a parameter $\mathcal{p}_{\ell i}$:

$$y_{\ell i} = \begin{cases} 1 & \text{with } P(y_{\ell i} = 1|x_i) = \mathcal{p}_{\ell i} \\ 0 & \text{with } P(y_{\ell i} = 0|\, x_i) = 1 - \mathcal{p}_{\ell i} \end{cases}$$

Therefore, $\mathcal{p}_\ell$ actually represents the odds of being food-insecure. The natural log (*ln*) of the odds, also known as the logit, is as follows:

$$ln\left(\frac{\mathcal{p}_{\ell i}}{1 - \mathcal{p}_{\ell i}}\right) = logit(\mathcal{p}_{\ell i}) = logit[Prob(\mathcal{y}_{\ell i} = 1)]$$

As a result, the variable of interest is modelled as a multinomial logistic regression:

$$\begin{aligned} logit(\mathcal{p}_{\ell i}) &= \beta_{\ell 0} + \beta_{\ell 1} x_{1i} + \beta_{\ell 2} x_{2i} + \cdots + \beta_{\ell k} x_{ki} + \varepsilon_{\ell i} \qquad (5.9) \\ &= x_i' \beta_\ell + \varepsilon_{\ell i}, \end{aligned}$$

where $x_i' = (1, x_{i1}, x_{i2}, \dots, x_{ik})$ and $\beta_\ell' = (\beta_{\ell 0},\ \beta_{\ell 1},\ \beta_{\ell 2}, \dots, \beta_{\ell k})$ at ℓ – level food insecurity for individual i. Recall that $\varepsilon_{\ell i}$ takes only two values, $\varepsilon_{\ell i} = -x_i' \beta_\ell$ when $y_{\ell i} = 0$ and $\varepsilon_{\ell i} = 1 - x_i' \beta_\ell$ when $y_{\ell i} = 1$. Therefore, $\varepsilon_{\ell i}$ cannot be assumed to follow a normal distribution.

The model's coefficients may not be straightforward to interpret because they are scaled in terms of logs. Another way to interpret logistic regression models is to convert the coefficients into odds ratios. Since the estimated logistic regression model is

$$logit(\hat{\mathcal{p}}_{\ell i}) = x_i' \hat{\beta}_\ell = \hat{\beta}_{\ell 0} + \sum_{j=1}^{k} \hat{\beta}_{\ell j} x_{ji}, \qquad (5.10)$$

a backward logit transformation on the odds would result in:

$$\hat{\mathcal{p}}_{\ell i} = \frac{\exp(x_i' \hat{\beta}_\ell)}{\exp(1 - x_i' \hat{\beta}_\ell)}. \qquad (5.11)$$

5.3.4. Variable selection

The case study only illustrates the Boruta feature selection method (Kursa and Rudnicki, 2010), that was found useful and was applied in selecting the auxiliary variables in the empirical exercise here illustrated. The method uses a wrapper approach built around a random forest (Breiman, 2001) classifier (in Slavic mythology, Boruta is a forest divinity). The wrapper approach allows for ranking and classification of features that cannot be produced via a random forest algorithm, because the statistical significance of the features in question cannot be fully estimated (Kursa and Rudnicki, 2010). The method is wholly relevant, and thus aims to classify all features connected through decision rather than a minimal optimal class that deals only with the non-redundant auxiliary variables.

The random forest algorithm proposed by Breiman (2001) was successfully applied to reduce high dimensional and multi-source data for both classification and regression problems. The random forest algorithm is a collection of Classification and Regression Trees (CART) trained on data sets that are the same size as the training set, while creating samples (or bootstraps) by bootstrapping (or random resampling) on the training set itself. Once a tree is constructed, a set of bootstraps that does not include any particular record from the original data set (out-of-bag samples) is used as the test set.

Using a random forest as a classifier, the Boruta method proceeds according to the steps below.

1. It duplicates the data set and shuffles the values in each column, resulting in shadow features.

2. It trains the random forest classifier on the data set, which ensures that measures or scores of importance (Mean Decrease Accuracy or Mean Decrease Impurity) are produced for each feature in the data set. The higher the importance score produced, the more important the feature.

3. The algorithm compares real versus shadow features by checking whether the original feature has a higher score than the maximum score of its shadow features. If yes, the algorithm marks the feature as important and records it as a hit in a vector. It then moves on to the next iteration.

4. It repeats this process up to a predefined number of iterations, which will result in a table of hits.

5. By comparing the scores of the random shuffled copies of the features iteratively, the algorithm manages to compare the number of times a feature did better than the shadow features (binomial distribution). This boosts the robustness of the selection, because the importance of the feature is validated.

5.3.5. Results

To develop the projection model, the steps below were taken.

1. The relevant auxiliary variables were listed and their definitions checked in both samples.

2. Boruta was used to list relevant auxiliary variables for each level of prevalence, in the small survey.

3. Assumptions – linearity, multicollinearity – were checked for in the small survey.

4. The resulting variables were used to estimate the projection model parameters in the small survey.

5. The estimated projection model was used to estimate the variable of interest in the large survey.

The estimates from the projection model were used in the comparison with the probability values already available for severe and moderate or severe levels of food insecurity. The comparison was performed for different disaggregation domains, such as age, sex and income level, which are all listed in the IAEG-SDG disaggregation matrix. It should be noted that dummy variables were created for the auxiliary variables that were deemed categorical.

5.3.4.1. Results of the prevalence of severe food insecurity

As the first step, all available and relevant auxiliary variables[9] were plugged into the Boruta algorithm to assess their relevance, with their order of importance as auxiliary variables to explain the prevalence of severe food insecurity. In Figures 5.3 and 5.4, the boxplots of different colours represent various Boruta outputs: the red, yellow and green boxplots represent the scores of the rejected, tentative and confirmed variables, respectively, while blue was assigned to shadow features. Tentative variables are included in the projection models here: Boruta could not indicate a clear decision concerning those variables with the desired confidence because their importance levels are very close to their best shadow features. The *y*-axis in Figure 5.3 displays the importance of the variables. The most important variable is income (inccat), followed by location (rural), education (educat) and size of the household (sizeHH), which are all significantly important.

Various models are estimated, and it has been observed that the selection of the reference levels of categorical variables to be set aside with respect to the model affects the estimation. Therefore, the Boruta algorithm is applied on various levels of the auxiliary information to assess their importance.

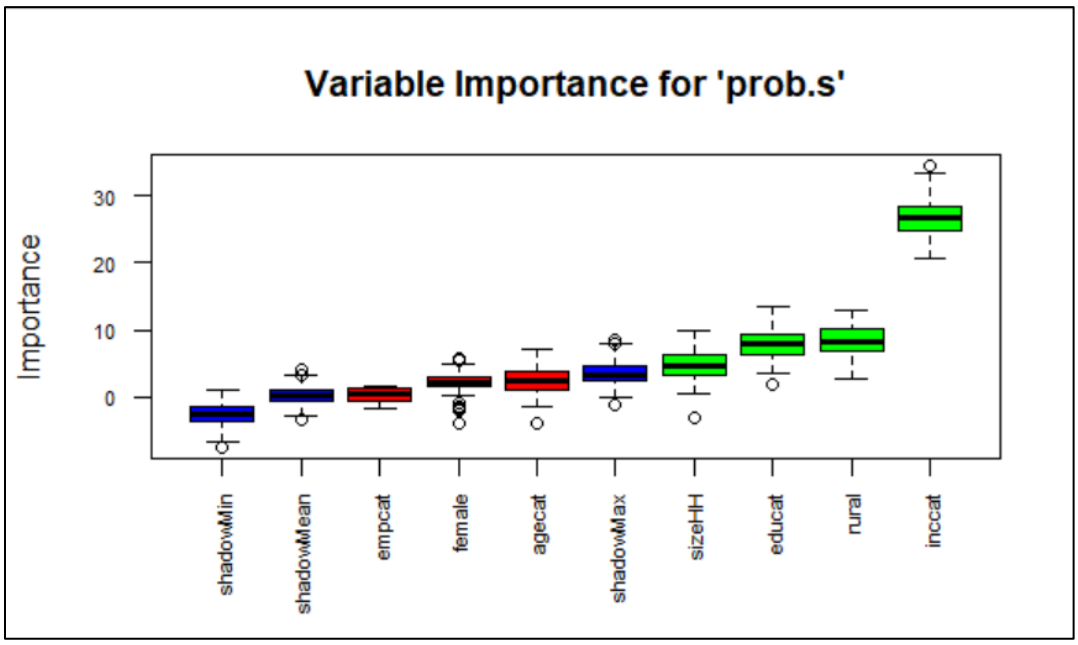
Figure 5.4 presents the order and importance of the different levels of auxiliary variables. The results in Figures 5.3 and 5.4 are consistent, as Boruta chooses the same set of auxiliary variables at the end of two independent runs.

The significantly important variables (including both “Confirmed” and “Tentative”) are as follows, in order of importance:

1. inccat_5 – the first quintile, representing the 20 percent of the population with the highest income
2. inccat_1 – the last quintile, representing the 20 percent of the population with the lowest income
3. inccat_2 – the second income quintile, i.e. between 21 percent and 40 percent
4. rural – being in rural area (1: rural)
5. educat_1 – completed elementary education or less (up to eight years of basic education)
6. inccat_3 – the middle-income quintile, i.e. between 41 percent and 60 percent
7. educat_3 – completed four years of education beyond high school and/or received a four-year college degree.
8. educat_2 – completed secondary education and some education beyond secondary education (9 to 15 years of education)
9. sizeHH – size of household

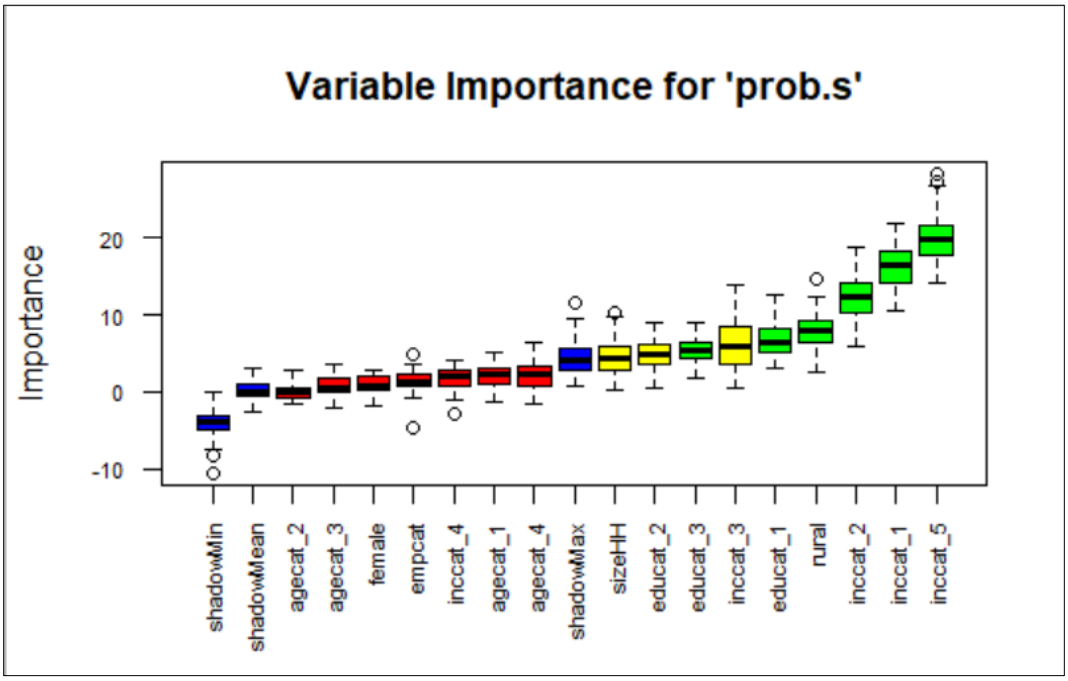
[9] The list of variable names and details on the levels of categorical variables is available in Annex 5, Section A5.1.

Figure 5.3. Level of importance of the auxiliary variables for severe food insecurity



Source: FAO, 2020.

Figure 5.4. Level of importance of the various levels of auxiliary variables for severe food insecurity



Source: FAO, 2020.

Generalized linear models for logistic regression (the *glm* function in R) is used to develop the projection model based on the significantly important variables listed after Boruta. The resulting model (see Annex 5, A5.2.1 for the overall estimation results) with significant variables is used as the projection model for various disaggregation domains (see the above list):

$$\widehat{Y_s} = 1.285 * \text{inccat}_1 + 0.807 * \text{inccat}_2 - 0.669 * \text{inccat}_5 - 0.081 * sizeHH \qquad (5.12)$$

To assess whether different variables have similar predictive relationships with the variable of interest, i.e. if there is multicollinearity, the Fox and Monette (1992) generalized variance-inflation factors (VIF) are computed.[10] The variables with high VIF value above (as a rule of thumb) 5 or 10 will be removed from the mode, which will lead to a simpler model without compromising the model accuracy. The VIF scores for Model 5.12 showed that there is no multicollinearity problem for the variables in the projection model, i.e. that all variables have a VIF value well below five.

Table 5.1. Variance-inflation factors for prevalence of severe food insecurity

Variables:	inccat_1	inccat_2	inccat_5	sizeHH
VIF score:	1.315	1.390	1.457	1.153

In this case study, some well-known GOF tests were considered – such as Pearson's chi-square test and the Hosmer and Lemeshow GOF – to evaluate the performance of the prediction model, while warning users that those tests perform best under the assumption of independent observations rather than observations from stratified and/or cluster sampling (Archer, Lemeshow and Hosner, 2007). The results indicate that Pearson's chi-square p-value is 9.880476e-21 and the GOF test p-value is 0.1757, both of which support the fitted model's accuracy.

Although it is not presented here, the AIC could also be considered when assessing the quality of a model through comparison of related models. For instance, the performance of the model after selection of variables with the help of Boruta and all variables in the model are compared through AIC. In all cases, the model that has the smallest AIC is observed to be the model with variable selection based on Boruta[11].

The reference category defined for the projection model is crucial for understanding and reading the results properly. In our case, the reference category is defined as "0" where the probability of being severely food-insecure is less than 0.5, and the non-reference category is "1" for a high probability/greater likelihood of being severely food-insecure.

For individuals in the poorest quintile (inccat_1=1), the odds of being more likely to be severely food-insecure (i.e. very likely as opposed to less likely) is 3.616 (=exp(1.285)) times that of individuals in higher-

[10] The "vif()" function in R can be used to calculate VIF scores (see Rdocumentation).
[11] See Annex 2 for further details on the estimation results.

income levels, holding all other variables constant. On the contrary, the individuals in the richest quintile (inccat_5=1) are less likely to be severely food-insecure, as shown by the negative sign. Moreover, for every unit increase in the number of household members, the odds of being more likely to be severely food-insecure are expected to decrease by approximately 7.8 percent (=100*(1- exp(-0.081))), holding all other variables constant.

The projection model is used to estimate the likelihood (more likely, as opposed to less likely) of being severely food-insecure for various disaggregation domains, such as sex, location, age and income. The resulting estimates (Predicted) are presented in Table 5.2 and compared with the FIES estimates available for IHS4 (Actual). It is observed that overall, the predicted values are very close to the actual values for all domains. The reasons for the lower performance of the projection estimator for the lowest income level are being explored. This result could be observed because income is defined differently in the two surveys, with an impact on the first category of the variable.

Table 5.2. Estimates for prevalence of severe food insecurity

	ALL	
prob.s	Actual	Predicted
0	15 602	15 397.23
1	14 131	14 335.77

	FEMALE		MALE	
prob.s	Actual	Predicted	Actual	Predicted
0	8 099	8 136.892	7 503	7 260.341
1	7 675	7 637.108	6 456	6 698.659

	RURAL		URBAN	
prob.s	Actual	Predicted	Actual	Predicted
0	11 139	12 068.71	4 463	3 328.517
1	12 739	11 809.29	1 392	2 526.483

	AGE<25		AGE>=65	
prob.s	Actual	Predicted	Actual	Predicted
0	5 554	5 350.249	999	1 097.997
1	5 070	5 273.751	1 289	1 190.003

	INCOME – Poorest (1st Q)		INCOME – Richest (5th Q)	
prob.s	Actual	Predicted	Actual	Predicted
0	3 215	1 869.109	4 094	3 946.894
1	3 103	4 448.891	1 261	1 408.106

5.3.4.2. Results for prevalence of moderate or severe food insecurity

As in the previous section, the first step in the projection of the likelihood of moderate or severe food insecurity was feature selection. All available and relevant auxiliary variables listed in Annex 5, Section A5.1 are evaluated by the Boruta algorithm. Figure 5.5 shows the level of importance of those variables and whether they are redundant or non-redundant. It is observed that income (inccat), location (rural) and education (educat) are all relevant (green), in order of importance. To check whether there are specific levels of those variables that are more relevant when modelling the likelihood of moderate or severe food insecurity, the Boruta algorithm is also applied on different levels of those relevant variables and the results are presented in Figure 5.6. It can be observed that Figures 5.5 and 5.6 are consistent because they highlight the same variables as relevant. Based on Figure 5.6, the variables that are significantly important (including both "Confirmed" and "Tentative") are listed below, in order of importance:

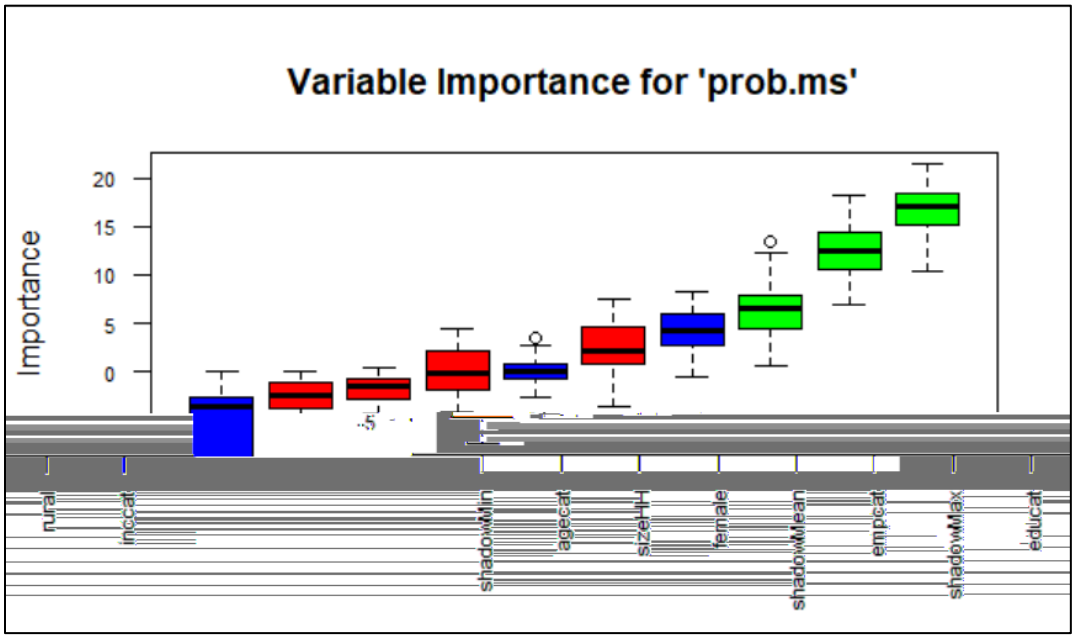
1. educat_2: completed secondary education and some education beyond secondary education (9–15 years of education)

2. educat_1: completed elementary education or less (up to eight years of basic education)

3. rural: being in rural area (1: rural)

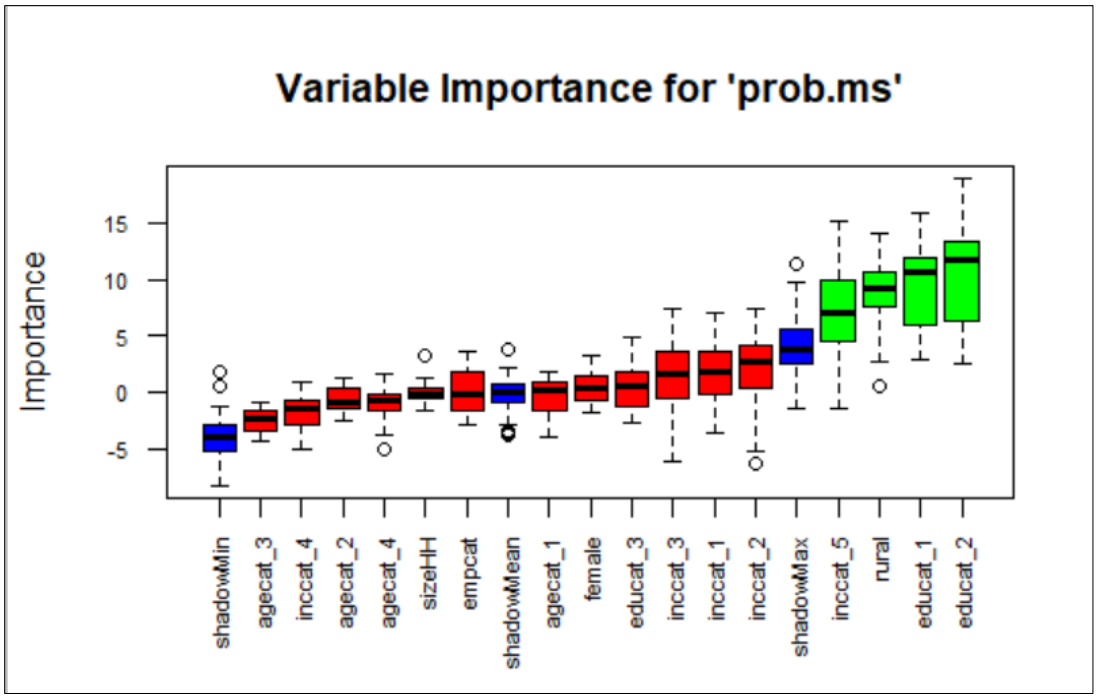
4. inccat_5: the first quintile, representing the 20 percent of the population with the highest income

Figure 5.5. Level of importance of auxiliary variables for moderate food insecurity



Source: FAO, 2020.

Figure 5.6. Level of importance of various levels of auxiliary variables for moderate food insecurity



Source: FAO, 2020.

The generalized linear models for logistic regression in R (*glm()* function) produced the projection model below (see Annex 5, Section A5.2.2 for the overall estimation results):

$$\hat{Y}_{ms} = 1.818 - 1.414 * inccat_5$$

The Pearson's chi-square (p_value = 1.459197e-15) and GOF tests (p-value = 0.9079) do not yield any evidence such as to reject the null hypothesis, which assumes that the fitted model is correct. As noted in the previous section, other GOF tests considering complex sampling designs could be preferred for this specific exercise.

The resulting projection model produces an intercept term, that can be interpreted as the log odds of predicting the non-reference category. Here, the reference category is "0", where the probability of being moderately or severely food-insecure is less than 0.5, and the non-reference category for a high probability of being moderately or severely food-insecure is "1". Overall, the intercept coefficient is the log odds of someone who is not in the richest quintile of the population (inccat_5 = 0) having a high level of moderate or severe food insecurity. We can translate log odds into just odds (of the high probability of being moderately or severely food-insecure):

exp(1.818)= 6.158

6.158/(1+6.158) = 0.860

It can be observed that being at the highest income level (inccat_5=1) has a negative impact on the probability of being moderately or severely food-insecure. In fact, being in the richest income quintile

entails a probability of 0.20 (exp(-1.414)=0.243 and 0.243/(1+0.243)=0.195) of displaying a value of “1” (high probability of severe or moderate food insecurity) and a probability of 0.80 of displaying a value of "0" on the outcome variable.

The likelihood (more likely, as opposed to less likely) of being moderately or severely food-insecure is predicted by the resulting projection model above. The projections (predicted) for various disaggregation domains presented in Table 5.3 allow for comparisons with the FIES estimates available for IHS4 (Actual). Overall, it is observed that the projected values are very close to actual values for all domains in general, which supports the conclusion that the proposed method performs well for this data set. The reasons for the subpar performance of the projection estimator for the lowest income level are still being explored. This result could be observed because income is defined differently in the two surveys, with an impact on the first category of the variable.

Table 5.3. Estimates for prevalence of moderate or severe food insecurity

	ALL	
prob.ms	Actual	Predicted
0	6 037	5 548.857
1	23 696	24 184.143

	FEMALE		MALE	
prob.ms	Actual	Predicted	Actual	Predicted
0	3 123	2 927.641	2 914	2 621.215
1	12 651	12 846.359	11 045	11 337.785

	RURAL		URBAN	
prob.ms	Actual	Predicted	Actual	Predicted
0	3 382	4 128.118	2 655	1 420.739
1	20 496	19 749.882	3 200	4 434.261

	AGE<25		AGE>=65	
prob.ms	Actual	Predicted	Actual	Predicted
0	2 134	1 875.174	332	394.91
1	8 490	8 748.826	1 956	1 893.09

	INCOME – Poorest (1st Q)		INCOME – Richest (5th Q)	
prob.ms	Actual	Predicted	Actual	Predicted
0	1 224	882.4605	2 382	2 143.883
1	5 094	5 435.5395	2 973	3 211.117

5.4. Lessons learned

The method and the resulting projection models performed very well for both levels of food insecurity indicators. The most important lessons learnt are listed below:

1. The definition/methodology of the auxiliary variables in both small and large surveys may not be fully consistent. For example, the definitions of potential auxiliary variables, such as income or economic activity, can be substantially different in national surveys and other smaller surveys such as the GWP. It might be necessary to reorganize the available data to create similar definitions, when possible. For instance, in one survey, more disaggregated data may be available by education level, while in another, the education level might be more aggregated. This could be improved by aggregating the more disaggregated variable.

2. National surveys are based on households rather than individuals (as in the GWP); this is another source of variability in the available auxiliary information. Strong assumptions might be necessary to improve the situation. For instance, to estimate individual-level income in surveys for which only household-level income is available, it was decided to divide the household-level income by the number of adults in the household. A similar problem may be that the reference respondent is the head of the household, while individual-level data is needed for all members of the household. The assumption is that the relationship between the auxiliary variables selected and the variable of interest is stable over time.

3. Using model-assisted projection for another domain for which data collection is expensive could be a good investment. An example of such a domain is food consumption.

The main steps to build a projection model based on two surveys are listed below.

1. List relevant auxiliary variables and check their definitions in both samples.

2. Use Boruta to list relevant auxiliary variables for each level of prevalence – small survey.

3. Check for assumptions – linearity, multicollinearity; small survey.

4. Use the resulting variables to estimate the projection model parameters – small survey.

5. Check the model's GOF.

6. Use the estimated projection model to estimate the variable of interest – large survey.

Appendix A5.1. List of auxiliary variables

Age: Converted into categorical variables following UN[12] definitions and classification of age groups:

agecat_1: 15-24 (youth)[13]
agecat_2: 25-49
agecat_3: 50-64
agecat_4: 65 and over (older persons)
agecat_99999: NA

Education: Converted into categorical variables:

GWP

educat_1: Completed elementary education or less (up to eight years of basic education)

educat_2: Secondary – Three-year tertiary secondary education and some education beyond secondary education (9–15 years of education)

educat_3: Completed four years of education beyond high school and/or received a four-year college degree

educat_99999. This covers "DK: Don't know" and "RF: Refused" in GWP and "NA" in LSMS

IHS4 lists various answer options for education applicable for Malawi (Startfishers Malawi):

A. NONE

B. PSLC: Primary School Leaving Certificate – Primary School Leaving Exam assesses academic achievement at the Primary School level (ages 13–14)

C. JCE: Junior Certificate of Education is a school-based junior schooling qualification awarded to eligible students at the end of Year 9 on completion of the junior phase of learning (ages 15–16)

D. MSCE: The Malawi School Certificate of Education exam, taken during the last year of secondary school (ages 17–18)

E. NON-UNIV.DIPLOMA

F. UNIVER.DIPLOMA,DEGREE

G. POST-GRAD.DEGREE

[12] For statistical purposes, the UN – without prejudice to any other definitions applying in Member States – defines "youth" as persons between the ages of 15 and 24 years.
[13] The reference population of the GWP consists of adults, i.e. individuals older than 15 years of age.

These are merged to create a consistent categorical variable with the GWP:

1. A+B
2. C+D
3. D+E+F

Income: The income estimates for IHS4, as extracted from the FAO Rural Livelihoods Information System (FAO, RuLIS). Income at individual level is estimated by the income of the household divided by the number of adults in the household. GWP provides income data in quintiles at the individual level. As a result, RuLIS estimates for IHS4 are converted into categorical variables following the GWP definitions:

Income Category
inccat_1: Poorest 20%
inccat_2: 21% - 40%: Second 20%
inccat_3: 41% - 60%: Middle 20%
inccat_4: 61% - 80%: Fourth 20%
inccat_5: Richest 20%

Employment: New categorical variables created:

Original GWP

A. Employed full-time for an employer
B. Employed full-time for self
C. Employed part-time, wants full-time
D. Employed part-time, does not want full-time
E. Unemployed
F. Out of workforce

Original IHS4 – Extracted from RuLIS

a. not employed (inactive or unemployed)……..1 (0 in RuLIS)
b. employed ……..2 (1 in RuLIS)
c. in the age group (>=5 years old) but no data on employment ……missing (99 in RuLIS)
d. not in applicable age range, i.e. <5 years old……. NA

Employment Category:
empcat_1: Unemployed + Out of workforce
empcat_2: Employed part-time, wants full-time + Employed part-time, does not want-full time
empcat_3: Employed full-time for an employer+ Employed full-time for self

Appendix A5.2. Results – R output for *glm ()* function

A5.2.1. Severe food insecurity

```
Deviance residuals:
   Min      1Q  Median      3Q     Max
-2.2562  -1.1139   0.6118   0.8137   1.7673
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.1416     0.8627   1.323 0.185725
df$educat_1  -0.1481     0.8436  -0.176 0.860634
df$educat_2  -0.5653     0.8461  -0.668 0.504101
df$educat_3  -1.7402     1.1694  -1.488 0.136740
df$inccat_1   1.2852     0.2823   4.553 5.28e-06 ***
df$inccat_2   0.8074     0.2438   3.311 0.000929 ***
df$inccat_3   0.3988     0.2290   1.742 0.081586 .
df$inccat_5  -0.6695     0.2029  -3.299 0.000970 ***
df$rural1     0.2659     0.2196   1.211 0.226063
df$sizeHH    -0.0810     0.0370  -2.189 0.028607 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1219.8 on 994 degrees of freedom
Residual deviance: 1104.1 on 985 degrees of freedom
AIC: 1124.1

Number of Fisher Scoring iterations: 4
```

A5.2.2. Moderate or severe food insecurity

```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.4636   0.3140   0.3140   0.4222   1.0114
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.81781    0.62297   2.918  0.00352 **
df$educat_1  0.65427    0.63208   1.035  0.30062
df$educat_2  0.04162    0.62352   0.067  0.94678
df$inccat_5 -1.41395    0.22213  -6.365 1.95e-10 ***
df$rural1    0.51321    0.27067   1.896  0.05795
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 691.86 on 994 degrees of freedom
Residual deviance: 616.22 on 990 degrees of freedom
AIC: 626.22
Number of Fisher Scoring iterations: 5
```

Chapter 6. Small area estimation techniques

6.1. Introduction

An estimator of the parameter of interest for a given subpopulation is said to be a direct estimator when it is based only on sample information from the subpopulation itself. Unfortunately, for most surveys, the sample size is not large enough to guarantee reliable direct estimates for all subpopulations. A "small area" or "small domain" is any subpopulation for which a direct estimator with the required precision is not available. In the literature, "small area" is intended as a general concept and is used to indicate a general partition of the population according to geographical criteria or other structural characteristics (sociodemographic variables for household surveys or economic variables for business surveys). This chapter will use the broad definition of small area. When direct estimates cannot be disseminated because they are of unsatisfactory quality, an ad hoc class of methods – the SAE methods – is available to overcome the problem (see Rao, 2003; Pfeffermann, 2002, 2013). These methods are usually referred to as indirect estimators because they cope with poor information for each domain, borrowing strength from the sample information belonging to other domains; this results in an increase in the effective sample size for each small area.

Large-scale surveys are usually aimed at providing estimates of target parameters for the whole population, as well as for relevant subpopulations defined at the sampling stage. Design-consistent and design-unbiased direct estimates are produced for the parameters of interest. However, in most surveys, the sample size is not large enough to guarantee reliable estimates for all target subpopulations.

Over the last few years, the paradigm underlying the statistical process has been gradually changing the production of official statistical data by the major information statistical centres, both nationally and internationally. In fact, alongside the data collected using traditional "statistical surveys", the growing availability of data from so-called "new data sources" – both those of an administrative nature and those obtained through new electronic devices and information-gathering channels on the Internet – overwhelmingly dictates the agenda of the methodological and operational aspects to be addressed and resolved by official statisticians in each country. As far as the more strictly statistical-methodological aspects are concerned, the following aspects must be addressed:

1. the need to estimate multiple contingency tables, which arises from the fact that the sampling surveys produce hypercubes obtained from the intersection of numerous variables;

2. territorial and structural classification;

3. the use of estimators of projection type allows for producing the predicted values at the level of the single unit;

4. post-stratified ratio estimators – traditional estimators that use totals obtained as special cases of regression estimators – can be extended to linear models with fixed effects, as well as to those with mixed ones, by defining the estimators as a function of matrices of totals. This can produce significant computational efficiencies.

6.2. Process flow for computing small area estimates

This section summarizes the main findings of the ESSnet research project (2012) on SAE techniques. The project proposes a standardized process for SAE, in which: (i) there is a sample survey from which statistics at the national level, or by major subnational domains, are available by means of design-based direct estimation; and (ii) the aim is to investigate whether it is acceptable to produce statistics at lower aggregation levels, possibly using SAE techniques. The standardized process flow is divided into the following phases.

(1) Clarification for the identification and prioritization of the needs and uses of small area estimates; identification of the survey of the relevant data available and choice of criteria to evaluate the small area estimates obtained.

(2) Calculation of direct estimates together with basic design smoothing techniques, i.e. synthetic and composite estimators calculated under a design-based approach. At this stage, no change of the inferential framework or additional data is needed, compared to the existing regular survey.

(3) Enhancement of the basic design smoothing techniques. This step is needed if the results of direct estimates or basic design smoothing estimators are not acceptable. The quality assessment of the design-based small area estimates should identify the weaknesses to be improved. The general idea is to utilize additional information, though modelling, targeted at the perceived shortcomings in the basic smoothing results.

6.2.1. Clarification for the identification and prioritization of needs

Referring to the first phase of the process flow, the following types of user needs are often been mentioned in the literature: policy and programme formulation and evaluation; allocation of funds; local government and business planning. In any case, from a methodological point of view, the following distinctions regarding the nature of small area statistics are important:

- cross-sectional totals, or means, and their changes over time;

- area-specific best prediction and ensemble small area characteristics. Examples of ensemble characteristics include, but are not limited to, the difference between the maximum and minimum small area parameter values, the distribution of parameter values across the small areas and the rank of the small areas according to their respective parameter values.

In this context, important factors to be considered are the following.

- Practitioners who are only familiar with producing national-level estimates often find it difficult to prioritize the different objectives of small area estimates and, therefore, set preferences and a balance among their statistical properties. It is nevertheless important to be aware of potential bias in the indirect small area estimates.

- The most important metadata to be clarified consists in the hierarchy of aggregation levels: from the national level to the small areas of interest, and possibly domains within the small areas. Moreover, there may be an interaction towards the clarification/prioritization of the needs or uses of the small area estimates.

- It is good practice to make the repository of data as thorough as possible.

- It is important to consider the auxiliary variables used in the existing estimation method for national, or major subnational, statistics. In particular, it is important to clarify whether these auxiliary variables are available at the small area level, such that the existing estimation method can be applied within each small area to produce the corresponding direct estimates. Furthermore, additional covariates (from previous censuses and surveys, administrative sources, etc.) that may help to explain statistical variations in the target variables must be considered.

- It is necessary to consider the available data, including proxies of target variables – either exact for the target population or in proximity – that can be used to set up realistic Monte Carlo simulation studies.

- It is also helpful to obtain an idea of the accuracy that can be expected of the small area estimates, in relation to that of the existing statistics. It is generally unrealistic to expect small area estimates to have the level of accuracy that one may be accustomed to in the case of national-level estimates. For example, Statistics Canada applies the following guidelines on the reliability of data from labour force surveys (Statistics Canada, 2010; pp. 30–31): if the coefficient of variation (CV) < 16.5 percent, then there are no release restrictions; if 16.5 percent < CV < 33.3 percent, then the data should be accompanied by warnings (release with caveats); if the CV > 33.3 percent, then the data are not recommended for release. The British Office for National Statistics (ONS) dissemination policy (2004) established that ideally, the CV < 20 percent for a small area estimate to be considered publishable.

- The bias of the indirect small area estimates implies that it is insufficient to consider the CV on its own. In practice, the area-specific mean squared error (MSE) of the small area estimator is the most common measure of uncertainty; less often, it is the use of confidence intervals, or prediction intervals from the model-based point of view. It is also rare to find uncertainty measures of the ensemble characteristics of the small area estimates. Apart from choosing the summary measure of uncertainty, it is important to consider and select a set of diagnostics and checks that may help to: better understand the data used; assess the estimation methods or the underlying model assumptions; and form a more complete picture of the quality (strengths as well as weaknesses) of the small area estimates obtained.

    - One should consider the feasibility of a realistic Monte Carlo simulation study based on an artificial “target population”, i.e. from the design-based point of view. The artificial population need not faithfully reflect every aspect of reality. However, it should be realistic with respect to the key uses and needs of the small area estimates, as well as their corresponding desirable statistical properties. Such Monte Carlo simulation studies, when feasible, often provide the most reliable evidence when choosing from among the alternative methods or models.

    - The various diagnostics and measures of uncertainty are helpful when comparing alternative estimators to identify the best available method. However, it is difficult to provide a set of explicit and absolute generic conditions that the estimation method and its results must meet in order to be considered acceptable for dissemination, just as in the case of official statistics at the national level. The decision is more likely to be made on a case-to-case basis with regard to a fit-for-purpose assessment, and the “acceptance” margin is unlikely to be uniform across all small areas of interest.

6.2.2. Calculation of direct estimates together with basic design smoothing techniques

Direct estimates are derived separately within each small area, based only on the data from the given area or domain. In theory, the auxiliary information provided is used for regular national-level estimates and is available at the small-area level of interest, and it is possible to apply the existing estimation approach within each small area. In practice, however, a lack of data can cause problems. First, if there are no sample data at all from a particular small area, then no direct estimate can be produced, with or without auxiliary information. However, the estimation method could also fail because there are too few direct sample data. For instance, the post-stratified estimator may break down due to empty within-area sample post-strata; similarly, the calibration estimator may be unfeasible because of "empty" within-area sample margins.

The next option is indirect synthetic estimation (Rao, 2003; Section 4.2). Essentially, this amounts to replacing the direct estimates with regression estimates, whereby the regression coefficients are estimated based on the data in the larger area (or domain) to which the small area of concern belongs. The simple examples below illustrate this notion. Indirect synthetic estimates can have much smaller variances compared to direct estimates. Nevertheless, they rarely provide satisfactory solutions to the problems affecting SAE, because they tend to reduce the between-area variation of interest – hat is, they might have unacceptably large biases. Moreover, from a smoothing perspective and at least in theory, it is always possible to further improve the bias-variance trade-off by means of composite estimation. A small area composite estimate is given as a weighted average of the corresponding direct estimate and a synthetic estimate of choice. The idea is to give more weight to the direct estimate if it is reliable, and less weight otherwise. Indeed, the composite weights can be derived to minimize the area-specific MSE. As a result, the composite estimator can be assumed to have a reduced variance but an increased bias compared to the direct estimator, whereas it has a reduced bias but increased variance compared to the synthetic estimator. In practice, however, one must allow for extra uncertainty associated with the estimation of the optimal "weights".

Composite estimates are sometimes known as shrinkage estimates, because by construction all direct estimates are pulled towards the corresponding synthetic estimate of a broader area. A consequence is that, together, the composite estimates derived by minimizing the area-specific MSE generally display less between-area variation than they should. This is referred to as the over-shrinkage problem.

Composite estimation is generally easier to implement compared to explicit model-based estimation, especially as the model grows increasingly complicated. Moreover, when the composite weights depend only on subsample sizes, composite estimates can be derived for a large number of target variables at the same time. In contrast, a model applies only to one variable, or perhaps very few variables, at a time. It is usually impractical to build models for all statistical variables collected in the sample, both at the national level and at the small-area level.

6.2.3. Enhancement of the basic design smoothing techniques

Basic smoothing may not be satisfactory for one or several reasons. For example, there may be too many empty sample domains, where the synthetic estimator causes unacceptable over-smoothing. Even the best composite estimator may have an excessively large bias and/or MSE in general. The optimal composite weights are numerically too unstable, as explained above, while the alternative practical choices of composite estimator are inadequate. The small area parameters may have a skewed non-

normal distribution across the areas, such that the composite estimates may fail for areas at either of the two ends of the distribution. An example is binary means close to 0 or 1.

An explicit model-based approach may help address the perceived shortcomings; for example, additional covariates or correlations can be included in a model. For instance, the latest census population totals (or means) can be used as the explanatory variables for the current population parameters, while the information cannot be naturally utilized in design-based estimation unless the census values of the current sample units can be identified. The incorporation of spatial and temporal correlations is yet another example. Further reductions of both bias and variance may therefore be possible with model-based estimates. However, additional model assumptions are necessary. The main numerical instability of the composite estimator is associated with the direct (unbiased or consistent) estimation of the variance of the direct small area estimator. Models specified for area-level sample statistics face a similar issue, if the sampling design effect must be taken into account. Being completely model-based, in these cases, the unit-level models can provide an alternative.

Large classes of generalized linear and non-linear models exist that can be useful in handling non-normally distributed small area parameters.

The first, or most basic, choice is whether to specify the model at the unit level or at the area level. The choice depends on the nature and the availability of data at the two levels. However, sampling design can also play a part. There may be a strong design effect for reasons such as stratification, multistage sample selection or clustering. An area-level model must take this into account, in order to appropriately describe the variation in the area-level sample statistics. Meanwhile, a strong design effect may nevertheless be considered non-informative from a model-based perspective, given appropriate auxiliary information. For example, the design effects due to stratification by age and sex can be handled in a unit-level model by including age and sex as explanatory variables. Also, the household clustering effects for the labour force survey's employment status may be ignored, provided that the model makes use of good administrative employment data at the individual level, even if the administrative data are not used in the sampling design at all.

The next choice is the one between linear and generalized linear (or non-linear) models. In theory, generalized linear models are preferable for categorical data. In practice, however, linear models are computationally much easier and often yield similar results. For example, employment status can be modelled as binary data using a logistic regression model; however, it is often viable to apply a linear model at unit or perhaps area level, just like implicit models underlying design-based survey weighting are usually linear, regardless of the type of data. However, the linear probability model presents some drawbacks (prediction beyond the admissible range; heteroscedastic errors; non-meaningful R-square).

The conceptual relationship between (fixed effects) regression models and (random effects) mixed models is analogous to that between synthetic and composite estimation. In SAE, mixed models are always more appropriate than the corresponding regression model, as the latter simply excludes all the random effects of the former because it allows for heterogeneity across the small areas. In terms of smoothing, however, the questions are empirical in nature: (i) is the regression model good enough? (ii) does the extra computational effort of the mixed model pay off? In particular, going from a generalized linear model to a generalized linear mixed model can be much more technically complicated than going from a linear regression model to a linear mixed model.

Finally, questions of multivariate modelling can sometimes arise. Multivariate modelling is more efficient (or appropriate) when there are multiple target variables from each small area, and these are either correlated with (or mutually restrictive of) each other. For instance, the average household incomes of

different household types from the same small area may be positively correlated. Also, a set of counts may sum up to a known total in each small area, such as the number of persons in different household types or the number of persons with the three different labour market statuses. Again, in terms of smoothing, the question is practically of an empirical nature, as the choice between regression and mixed models.

In summary, it is not necessary to start the SAE process with modelling. Rather, explicit modelling can be considered after the basic smoothing results have been obtained and examined. Modelling should be targeted to address the perceived shortcomings of the basic smoothing results in the given situation. This is because, in terms of smoothing in the context of SAE, the goal of modelling is not to build the most plausible theoretical construct to explain the data, but to find a more powerful and, hopefully, acceptable tool to predict the statistical variables of interest from the existing sample survey.

6.3. Parameters of interest and the working model

6.3.1. Notation

This section refers to the vectorial rewriting of the multivariate regression model with multiple random effects proposed in Datta *et al.* (1999), which enables extending the usual formulae of the univariate model for small areas to the multivariate case.

Let us consider the target population *U* and a set of subpopulations U_d $(d = 1, \dots, D)$, related to *D* domains of interest. Let us indicate with $y_i = \{y_{c,i};\ c = 1, \dots, C\}$ the vector of the *C* dummy elementary variables referred in the i –th $(i = 1, \dots, N_d;\ d = 1, \dots, D)$ unit of the target population, with $y_{c,i}$ = 1 if the $i-$ th unit has the $c-$ th characteristic of interest while $y_{c,i} = 0$ otherwise; $y_{dc,i} = y_{c,i}\gamma_{di}$ $(i = 1, \dots, N_d;\ d = 1, \dots, D;\ c = 1, \dots, C)$ the $c-$ th element of the generic elementary vector $y_{d,i}$ equal to 1 if the $i-$ th unit belonging to the $d-$ th domain (subpopulation U_d) has the $c-$ th characteristic of interest while $y_{dc,i} = 0$ otherwise; $y_d = \{y_{d,i}\,; i = 1, \dots, N_d\}$ the vector for all units of subpopulation U_d $(d = 1, \dots, D)$ and with $y = \{y_d\,; d = 1, \dots, D\}$ the vector for all units of target population U.

To introduce the related regression model, we denote, for each target variable and for each unit of the target population, as follows: *X* the design covariate matrix and *β* the corresponding fixed effect (i.e. regression coefficients) vector; *Z* the design matrix of the random effects that denotes the belonging of each unit to the different domains and $u = \{u_d\ ;\ d = 1, \dots, D\}$, is the vector of the random effects $u_d = \{u_{d,c}\ ;\ c = 1, \dots, C\}$, which are part of vector u ; Σ_e , is the diagonal matrix of the *C* variances, σ^2_{ec} $(c = 1, \dots, C)$, of the vector of random errors $e_{d,i} = \{e_{c,i}\gamma_{di}\,; i = 1, \dots, N;\ c = 1, \dots, C\}$; Σ_u is the matrix of the *C* variances, σ^2_{uc} for *c = 1,. . . , C* and of the $C \times (C-1)$ covariances, $\sigma_{ucc'}$ $(c \neq c = 1, \dots, C)$ of the vector of random effects $u_d = \{u_{d,c}\ ;\ c = 1, \dots, C\}$**.** Note that in the simplest case of the models in which an absence of correlation between the *C* characteristics of interest is assumed, $\Sigma_u = diag^C_{c=1}\{\sigma^2_{uc}\}$.. Finally, $\omega = col^C_{c=1}\{\omega_c\}$, with $\omega_c = [\sigma^2_{ec}, \sigma_{ucc'}]^T$ indicates the vector of the so-called variance components whose elements coincide with the unknown parameters of the model's variance and covariance matrices, Σ_e and Σ_u.

On the other hand, it is also useful to add the corresponding notation that serves to define the direct estimators referred to each domain of interest. Sample S, of size n, is selected from the population U according to sample design P, from which to extract the $i-th$ unit in the sampling with inclusion probability $\pi = \{\pi_i\,.1_C\ ;\ i = 1, \dots, N\}$. Let *n* be the overall sample size and n_d the realized sample size for

the domain U_d, with $N_r = N - n$ and $N_{rd} = N_d - n_d$. Let $R = U\backslash S,$ and $R_d = U_d\backslash S_d$, indicating the set of the residual units that are not included in the sample in population *U* and subpopulation U_d, respectively.

Accordingly, for our purposes, it is useful to divide the formal structures involved in the estimation process into those referring to the units of the sample, denoted with the subscript *S*, and those referring to the remaining units of the population, denoted with the subscript *R*. Thus, for, example, the vectors y, Z and *e*, referring to all units of the population, are divided into the submatrices y_S, Z_S and e_s, and y_R, Z_R referring to the units belonging to the sets *S* and *R*, respectively.

6.3.2. The General Working model

To construct the various estimators that will be discussed in the following sections, in addition to the quantities presented above, it is useful to introduce, concerning the i − th unit of the domain U_d $(i = 1, \dots, N_d;\ d = 1, \dots, D)$, both the vector of *C* synthetic values $\hat{y}_{d,i} = col_{c=1}^{C}\{\hat{y}_{dc,i}\}$ and the relative vector of the estimated residuals $\hat{e}_{d,i} = col_{c=1}^{C}\{\hat{e}_{dc,i}\}$. In particular, for the aforementioned quantities, general expressions valid under the different formulations of the WMs adopted for the direct and indirect estimators examined are considered below. To this end, we introduce, for $y_{d,i}$ $(i = 1, \dots, N_d;\ d = 1, \dots, D)$, the following general WM, given by

$$y_{d,i} = \mu_{d,i} + e_{d,i}, \quad (6.1)$$

where $e_{d,i} = \left[row_{c=1}^{C}\{e_{dc,i}\}\right]^T$ is the vector of residuals and $\mu_{d,i} = \left[row_{c=1}^{C}\{\mu_{dc,i}\}\right]^T$ is the vector containing the conditional expected values of y_d, with respect to the matrix of the auxiliary variables $X_{d,i}$ and to the set of effects of the model formally expressed as

$$\mu_{d,i} = \left(y_{d,i} \middle| X_{d,i},\ \beta, u_d\right). \quad (6.2)$$

In fixed effects models, $\mu_{d,i}$ depends only on the vector of the regression coefficients β, while in the case of linear mixed effects models, it also depends on the vector of random effects, u_d. In the latter case, the random components of the WM consist of the vectors of random variables *i.i.d.* $e_{d,i}$ and u_d, whose probability distributions depend on a set of unknown parameters called variance components, included in vector ω ; moreover, $X_{d,i}$ is a matrix of known constants of dimensions $[C \times (G \times C)]$, while β denotes, a vector of unknown constants of order $(G \times C)$. In the case of generalized linear models with multivariate mixed effects, which represent the more general models that will be dealt with here, the following relation is valid:

$$\eta_{d,i} = g\left(\mu_{d,i}\right) = X_{d,i}\,\beta + u_d, \qquad (6.3)$$

in which $g\,(\cdot)$ is a known invertible and differentiable function called the link function, which is chosen, generally, to assume values in the whole set of real numbers. In the case of linear regression models, $g\,(\cdot)$ coincides with the identity function. Other forms of an exponential nature of the function are introduced, instead, to consider specific generalized linear models. By replacing (6.3) in (6.1), the general expression of the WM becomes

$$y_{d,i} = g^{-1}\left(X_{d,i}\,\beta + u_d\right) + e_{d,i}. \quad (6.4)$$

The vector of the predicted values $\hat{y}_{d,i}(\hat{\omega})$ of $y_{d,i}$ based on the assumed WM is given by the following expression:

$$\hat{y}_{d,i} = g^{-1}\left(X_{d,i}\,\hat{\beta} + \hat{u}_d\right) + \hat{e}_{d,i} \quad (6.5)$$

with $\hat{\beta} = \hat{\beta}(\hat{\omega})$, and $\hat{u}_d(\hat{\omega})$ the estimators of β and u_d which, in turn, are a function of the estimator $\hat{\omega}$ of the vector of the variance component ω. Moreover, let $\hat{e}_{d,i}$ be the vector of the estimated residuals, obtained as

$$\hat{e}_{d,i} = y_{d,i} - \hat{y}_{d,i}. \quad (6.6)$$

Knowledge of the vector of the estimated residuals is important to evaluate the quality of the considered estimators. More generally, the statistical properties of the considered estimators, both in terms of variability and bias, depend both on the actual capacity of the hypothesized WM to interpret the investigated phenomena and on the inferential reference approach – design-based or model-based – with respect to which these models are derived.

Example 6.3.1. Suppose that we adopt for $y_{d,i}$ a linear multivariate WM mixed with variance components. We therefore specify (6.3) by setting $\eta_{d,i} = g\left(\mu_{d,i}\right) = \mu_{d,i}$. Then,

$$\hat{\eta}_{d,i} = X_{d,i}\,\hat{\beta} + \hat{u}_d \quad (6.7)$$

with $\hat{\eta}_{d,i} = \hat{\eta}_{d,i}(\hat{\omega})$ the Empirical Best Linear Unbiased Predictor (EBLUP) of $\eta_{d,i}$, which depends on the generalized least squares estimator, $\hat{\beta}$ =$\hat{\beta}$ $(\hat{\omega})$, of β and on the EBLUP estimator, $\hat{u}_d = \hat{u}_d$ $(\hat{\omega})$, of the random effects vector u_d. The estimators introduced above are called "two-stage" or plug-in, as they depend on the estimator of Maximum Likelihood (ML) or of Restricted Maximum Likelihood (REML), $\hat{\omega}$, of the vector of variance components ω.

Example 6.3.2. Let us return to the case discussed in the previous example, but consider the simplest linear multivariate model with fixed effects, for which

$$\hat{\mu}_{d,i} = X_{d,i}\,\hat{\beta}. \quad (6.8)$$

This is obtained as a particular case of the random effects model by setting $\Omega(\sigma_u) = 0$ for the variance and covariance matrix of the random effects. We then have $u_d = 0$ by definition and $\omega_c = [\sigma_{ec}^2\,,\ \sigma_{uc} = 0]^T$ T. The variance and covariance matrix, $\Sigma = \Sigma\,(\omega)$, of *y* therefore becomes

$$\Sigma = R\,(\sigma_e), \qquad (6.9)$$

since $Z^T\,\Omega\,(\sigma_u)\,Z = 0$, where $R\,(\sigma_e)$ is expressed as

$$R = W_N^{-1} \otimes \Sigma_e\,(\sigma_e) = diag_{d=1}^{D}\left\{diag_{i=1}^{N_d}\left\{w_{d,i}^{-1} \otimes \Sigma_e\right\}\right\} \quad (6.10)$$

The generalized least squares estimator of *β* obtained on the basis of the data of the probabilistic sample, *S*, selected is given by

$$\beta = (X_S^T R_S^{-1}\,X_S)^{-1}\,X_S^T\,R_S^{-1}\,Y_S^T \quad (6.11)$$

where $R_S = W_S^{-1} \otimes \Sigma_e\,(\sigma_e) = diag_{d=1}^{D}\left\{diag_{i=1}^{n_d}\left\{w_{d,i}^{-1} \otimes \Sigma_e\right\}\right\}$.

From a general point of view, the matrix W_S represents a matrix of individual weights linked to a pattern of heteroscedasticity residuals of the model. In the case of design-based estimators, however, this matrix can be defined based on the probabilities of inclusion in the sample of units, as follows: $W_S^{-1} = diag_{d=1}^{D}\left\{diag_{i=1}^{n_d}\{\pi_i \cdot 1_C\}\right\}$ to obtain valid inferences under the sampling design.

Example 6.3.3. In the specific case, $y_{d,i}$ is a vector of *C* binary variables for which the relation $1_C^T \cdot y_{d,i} = 1$ holds. Likewise, the corresponding vector, $\hat{y}_{d,i}$ of the synthetic values must also satisfy the condition $1_C^T \cdot \hat{y}_{d,i} = 1$. In this context, $\hat{\eta}_{d,i}$ represents an estimate of the unknown probability vector $\eta_{d,i} = [row_{c=1}^{C}\{\eta_{dc,i}\}]$ T with $\eta_{dc,i}$ $(i = 1,\dots,N_d;\ d = 1,\dots,D;\ c = 1,\dots,C)$ being the probability that the $i-$th unit of the domain U_d possesses the $c-$th $(c = 1,\dots,C)$ characteristic of interest, hereinafter called the probability of success. In this case, for example, if it is desired to adopt a logistic regression model with mixed effects, we must define in (6.3) *g* = *logit*, obtaining

$$\eta_{d,i} = logit\left(\mu_{d,i}\right) = X_{d,i}\,\beta + u_d, \qquad (6.12)$$

from which

$$\mu_{d,i} = logit^{-1}\left(X_{d,i}\,\beta + u_d\right). \quad (6.13)$$

A logit model of this type has been proposed in Scealy (2010). The work is based on a model relating to the first *C* - 1 categories conditioned by the last category, *c* = *C*, the object of estimation the absolute frequency of which can be calculated by difference. The estimation model for the first C - 1 categories is expressed as

$$\frac{log\ (\mu_{dc,i})}{log\ (\mu_{dC,i})} = X_{d,i}\,\beta + u_d. \quad (6.14)$$

It is also worth remembering that the paper aims to estimate the *C* = 3 categories "employed", "unemployed" and "inactive". It also divides the estimation domains into two types: territorial domains and age groups. Furthermore, it is assumed that the territorial domains are all observed in the sample, while for each of them, not all sex groups by age are observed in the sample.

Let us now consider vector $y_+ = col_{c=1}^{C}\{y_c\}$, the vector of the *C* unknown population totals being

$$y_c = \sum_{i=1}^{N} y_{dc,i}$$

The vector y_+ is given by the matrix product

$$y_+ = A_+^T \cdot y, \quad (6.15)$$

where A_+^T is a suitable aggregation matrix of the elementary data vector y .

Similarly, X_+ denotes the matrix of order $[C \times (G \times C)]$ of the known population totals of the auxiliary variables, expressed as

$$X_+ = A_+^T \cdot X. \quad (6.16)$$

Let us now consider vector $y'_+ = col^D_{d=1}\{y_d\}$ of order $(D \times C)$, containing the total *C* of unknown interest for each of the *D* subpopulations U_d $(d = 1,\ldots,D)$, with

$$y_d = \left(\sum\nolimits_{i=1}^{N_d} y_{d,i} = col^C_{c=1}\{y_{dc}\}\right)$$

being the vector of the *C* totals of interest relative to the $d-$ th subpopulation U_d $(d = 1,\ldots,D)$. The vector y'_+ is given by the matrix product

$$y'_+ = A'^T_+ \cdot y, \quad (17)$$

where $A'_+ = row^D_{d=1}\{row^C_{c=1}\{a_{dc}\}\}$ is a suitable aggregation matrix of the elementary data vector *y*, where a_{dc} denotes the generic aggregation vector, of order $(N \times C)$, corresponding to the total $c-$ th $(c = 1,\ldots,C)$ of the subpopulation U_d $(d = 1,\ldots,D)$. Each of the vectors y_d $(d = 1,\ldots,D)$, components of y'_+, is thus obtained as $y_d = a_{dc} \cdot$ y and must respect the condition $y_d^T \cdot 1_C = N_d$.

Finally, it is useful to consider the matrix X'_+ of order $[(D \times C) \times (G \times C)]$, relating to the total *G* of the auxiliary variables referred to the *D* domains, obtained as

$$X'_+ = A'^T_+ \cdot X. \quad (18)$$

6.4. Construction of the vector of target variables and domains

An initial fundamental classification of the variables involved in the tabulation plan (TP) is based on their subdivision into replaceable and non-replaceable variables.

6.4.1. Replaceable and non-replaceable variables

The first class includes variables for which data for each unit of the population of interest are known. For the second class, however, only the data observed on a representative sample of units of the population of interest and/or from administrative sources on particular subsets of units may be available. Therefore, the estimates of the aggregates relating to the first type of variables can be obtained directly by aggregating the microdata relating to the elementary units belonging to the subpopulations of interest. As for the aggregates relating to non-replaceable variables, these can be obtained through univariate or multivariate estimation processes based on the imputation of missing data, or on design-based or model-based sample estimation methods.

6.4.2. Partially replaceable variables

In addition to dividing the target variables into replaceable and non-replaceable variables, there is the further category of partially replaceable variables, for which, for example, only some of the modalities of the variable of interest are known. The latter subset, however, can be traced back to the partition between replaceable and non-replaceable variables. Take, for example, the case of the variable employment condition (hereafter referred to as the Partially Replaceable Economic Condition [PREC]),

classified according to the following modes: Occupied, Person seeking employment and Inactive. The latter mode is further classified into four subcategories: Housewife, Student, Retired and Disabled. Moreover, suppose that for the data from administrative sources, relating to employees, students and retirees, it is possible to produce complete and good-quality data, for each unit of the population of interest, concerning the abovementioned methods, by applying specific control, correction and imputation processes. Then, it is possible to consider the PREC variable, which is partially replaceable with respect to two new variables. The first of the two variables is replaceable and includes the modalities, known for all individuals of the population, of Occupied and Not employed. This new variable, the Replaceable Economic Condition (REC), clearly refers to a partition of the population that is hierarchically more aggregated than that of the corresponding original variable, which is partially replaceable and considers five modalities instead of four. With respect to the existing variable, this new variable corresponds to the introduction of a new PREC classification that is more hierarchically aggregated than that of the original variable. On the other hand, the second new variable, which is not replaceable, concerns the further partition of the unemployed-other mode into the three subcategories of Person seeking employment, Housewife and Disabled.

6.4.3. Unit-level auxiliary variables

In addition to the distinction between replaceable and non-replaceable target variables, another fundamental subdivision is that between variables of interest and auxiliary unit-level variables. The latter, like the substitutable target variables, are variables for which the data for each unit of the target population are known, even if, in this case, the modalities of these variables do not coincide with any of the categories of the target variables of the TP. Therefore, the use of these variables in the estimation process has the purpose of improving the quality of the estimation process. It is clear that the quality of the estimates produced improves as the correlation between the set of non-replaceable target variables and the unit-level auxiliary variables used in the estimation procedure is greater.

The auxiliary variables also include all variables derived from the set of replaceable variables for which a different classification definition (than that established for those already defined for the TP) is required. For example, the age class variable could be requested at a more detailed classification level – for example at an annual level – than the five-year one required for the TP. Another example is the territorial level replaceable variable, for which the TP could require the three traditional hierarchical levels of municipal, provincial and regional administrative type. Exclusively for the purpose of defining the best work model, a territorial classification at the level of the local labour system could be required, a classification that entails an aggregation of the municipalities that cannot be inserted hierarchically into the three administrative classifications to be considered for the TP.

6.4.4. Unit-level auxiliary variables with error and proxy measurement

A different scenario occurs when the micro values of the auxiliary variable are available for the units existing on a date, $t - \Delta$, prior to the reference time *t* of the TP. Therefore, at time *t*, the individual values of the unit-level auxiliary variables may only be available for the units belonging to the longitudinal population intersection $U_{t-\Delta,t}$. Therefore, the co-present units belonging both to the transversal population, U_t, existing at time *t* and to that of $U_{t-\Delta}$existing at time $t - \Delta$ are available. With respect to time *t*, this type of auxiliary information is certainly affected by a coverage error, as it is not available for all units existing at time *t*, being unavailable for the units that entered the population in the period

between $t - \Delta$ and t. Furthermore, there is an additional component of a measurement error due to failure to update, as the values of the auxiliary variable are not updated on the relevant date. Hereafter, this type of variable, affected by coverage or measurement errors, will be referred to as auxiliary unit-level variables with error. When an auxiliary unit-level variable with error coincides with one of the variables of interest, it is called a unit-level proxy variable. This is the case, for example, with the Employment condition variable available from administrative sources, which is updated to approximately 16 months prior to the current time *t*. From these sources, it is possible to obtain proxy information on the employment condition (Occupied or Not employed modes), for each unit co-present in the population in the period between month $t - 16$ and month *t.*

6.4.5. Area-level auxiliary variables

Finally, it is useful to introduce a last class of auxiliary variables for which the average population values at the aggregate level (but not the corresponding micro values) refer to the various estimation domains (whose definition in the context under examination is given in the next paragraph). Here too, these variables may be affected by coverage errors, as observed on a previous population of the same type. This is the case, for example, when for each estimation domain, the average value of the variable of interest is known based on the previous census or an administrative archive. In certain situations, the average value in question, even if referred to time *t*, can also refer to a completely different population from the one of interest. Consider, for example, the case where for the Person seeking employment mode, the Employment condition variable has aggregated data from the Internet; in the case of variables linked to commuting, it might be possible to infer data on movements between different municipalities from the GPS tracks left on the Internet. On this basis, this type of information can be denoted as auxiliary area-level variables. Furthermore, using the terminology of error variables and proxies is also reasonable.

6.5. A classification of estimators

Starting from the general estimation methodology described in this chapter, we consider the options for the specific estimation procedure for the vector of the parameters of interest y'_+. Ultimately, this depends on the consideration of different factors. In particular, the main factors involved in the choice are the following:

1. how the estimator "gains strength" from the other domains, and thus the definition of a Direct (DI) or an Indirect (IN) estimator;
2. how the estimator uses the data observed in the domain, thus the definition of a Projection or a Composite estimator;
3. the reference inferential approach, from which would derive the definition of a Design-based/model-assisted (MA) or a Model-based (MB) estimator;
4. fixed effects or mixed/random effects WM (F) – the Fixed effects model (FE) or the Mixed effects model (ME);
5. linear or generalized linear WM – the Linear Model (LM) or the Generalized Linear Model (GLM) in the case of the FE, and the Linear Mixed Model (LMM) or the Generalized Linear Mixed Model (GMM) in the case of the ME;
6. the definition of random effects – Additive random effects (A) or Joint random effects (J);
7. WMs at the domain level or at the unit level – the Unit Level Model or the Domain Level Model.

Let us now briefly examine the factors of choice listed above. In this regard, it should be noted that most of these factors concern indirect estimators. In particular, Factors 2 and 6 concern only indirect estimators. Furthermore, indirect estimators that refer to fixed-effects models are called synthetic estimators, while those that refer to mixed effects models are called empirical predictors (EP). These two classes of estimators then differ further on the basis of Factors 2 to 6 included.

As for direct estimators, only some of the factors considered can be applied in choosing them. In particular, Factors 3, 4 and 5 affect direct estimators only with regard to the approach based on sampling design (that is, Design-based/model-assisted for Factor 3), the generalized linear and linear models with fixed effects (the fixed effects models, Factor 4), and the linear models and the generalized linear models, for Factor 5. As regards Factor 3, it is also specified that, in these Guidelines, only the direct estimators defined within the design-based approach are addressed, which are correct and consistent under the sampling design. In this context, the WM can play a role in reducing the sampling variability of the estimator; however, the inferential properties of the estimator – in terms of correctness and consistency – do not depend on the assumptions underlying the WM. Therefore, we speak of a design-based or model-assisted approach. These Guidelines, however, do not consider direct model-based estimators; see, for example, Chambers and Chandra (2008) on the model-based direct estimator.

The following section further examines Factors 1 and 2.

6.5.1. How the estimator gains strength from other domains

In general, direct estimators use only the sample information relating to the domain *d* of interest. Instead, indirect estimators use the sample information relating to a macro-domain d^+ that includes other domains in addition to domain *d*. It is therefore said that indirect estimators take strength from a macro-domain that includes other domains besides the one of interest. More particularly, in the context under examination, the choice of a direct or indirect estimator derives from the relationship between the set of simple indices, g_δ $(\delta = 1, \ldots, \Delta 1)$, which are part of the vector index *g* (on the basis of which the profiles of the matrix *X* are defined) and the indexes that define domains or macro-domains. To this end, it is useful to briefly summarize the concepts mentioned above. In particular, it should be remembered that the p_x index $(p_x = 1, \ldots, P_x)$ progressively numbers the different profiles, *g* = $(g_1, \ldots, G_{\Delta 1})^T$, obtainable as the simple indices $g_1, \ldots, G_{\Delta 1}$, which compose it.

Let us now return to the problem of choosing between direct or indirect estimators. Given the above, it is necessary to define the relationships existing between the elementary indices that make up vector *g* and index *d*. In particular, as regards the structure of the matrix of design *X*, it is assumed – without any loss of generality – that the first index g_1 is used to define any partition of the population linked to the domains. Therefore, the following fundamental situations are possible:

1. $g_1 \equiv d$ $\forall (g_1, d)$,
2. $g_1 \equiv \tilde{d}$ for $\tilde{d} \equiv d\tilde{g}_1 (\equiv d \times \tilde{g}_1)$ $\forall (\tilde{g}_1, d)$,
3. $g_1 \equiv d^+$ $\forall (g_1, d^+)$,
4. $g_1 \equiv \tilde{d}^+$ for $\tilde{d}^+ \equiv d^+\tilde{g}_1 (\equiv d^+ \times \tilde{g}_1)$ $\forall (\tilde{g}_1, \tilde{d}^+)$

in which the $\tilde{g}_1$ index represents a further partition within each domain, corresponding to a complete post-stratification within the domain. As regards the situations referred to in Factors 2 and 3 above, it should be noted that these are simplifications with respect to the level of complexity that can be possible in real cases. In fact, it is possible to use, at the same time in the estimation process, several aggregations of domains into macro-domains (that correspond to as many partitions of the population). For example, there may be a first macro-domain, associated with index d^+ ($d^+ = 1,\dots,D^+$), obtained from the aggregation of the first simple index that helps define the domains.

The first two situations lead to the definition of direct-type estimators. The last two cases, on the other hand, lead to the construction of indirect-type estimators.

In any case, the situations considered correspond to a specific definition of the matrix of the design *X*, which is defined on the basis of the general structure $X = row_{\delta=1}^{\Delta}\{X_\delta\}$ 11, where $X_\delta = diag_{g_\delta=1}^{G_\delta}\{1_{N_{g_d}}\}$. The same holds for the corresponding sample matrix, X_S, which has the same block structure; however, each refers to the n_{g_d} ($\delta = 1,\dots,\Delta_1$) sample units. Then, each of the situations referred to in the list above defines a specific diagonal structure of the matrix $X_{S1} = diag_{g_1=1}^{G_1}\{1_{n_{g_{d1}}}\}$. In particular, from Conditions 1 and 2, it is derived that, respectively, $X_{S1} = diag_{d=1}^{D}\{1_{n_d}\}$ and $X_{S1} = diag_{\tilde{d}=1}^{\tilde{D}}\left\{1_{n_{\tilde{d}}}\right\} = diag_{d=1}^{D}\left\{diag_{\tilde{g}1=1}^{\tilde{G}1}\left\{1_{n_{d\tilde{g}_1}}\right\}\right\}$. For Factor 3, instead, $X_{S1} = diag_{d^+=1}^{D^+}\{1_{n_{d^+}}\}$. Taking into account Kim and Rao (2012), Conditions 1 and 2 lead to the definition of estimators $\hat{y}_+^{\prime(PR)} = col_{d=1}^{D}\{\hat{y}_d^{\prime(PR)}\}$, direct projections that are correct and consistent under the design at the level of each domain d ($d = 1,\dots,D$). Conditions 3 and 4, instead, lead to the definition of estimators $\hat{y}_+^{\prime(PR)} = diag_{d^+=1}^{D^+}\{\hat{y}_{d^+}^{\prime(PR)}\}$, Projection direct estimators that are correct and consistent under the design at the level of each macro-domain d^+ ($d^+ = 1,\dots,D^+$); however, in this case, the estimates referring to domains *d*, included in the macro-domains, are indirect estimates that may be biased under the sample design.

6.5.2. How the estimator uses the data observed in the domain

The general formulation of the projection estimator of the vector of totals y_d is given by the sum of vectors $\hat{y}_{d,i}$, ($i = 1,\dots,N_d$; $d = 1,\dots,D$) of predicted values, on the basis of the WM, for all units of the target population belonging to domain *d*:

$$\hat{\boldsymbol{y}}_d^{(PR)} = \sum_{i=1}^{N_d} \hat{\boldsymbol{y}}_{d,i}. \quad (6.19)$$

The composite estimator is a particular form of projection estimator in which the predicted values are used only for the subset *R* of units not included in the sample; for the sample units *S*, the values $y_{d,i}$ directly observed with the survey are used. Therefore, the composite estimator of the total y_d is

$$\hat{y}_d^{(CO)} = \sum_{i=1}^{n_d} y_{d,i} + \hat{y}_{R,d}^{(PR)}, \quad (6.20)$$

where $\hat{y}_{R,d}^{(PR)}$ denotes the projection estimator referred to the set of population units not observed in the sample. The estimate is obtained, therefore, by applying the sum (20) to the $N_{R,d}$ of which the set *R* is composed. It is also useful to give the following alternative expression of the composite estimator

$$\hat{y}_d^{(CO)} = f_d \bar{y}_{S,d} + (1 - f_d)\hat{\bar{y}}_{R,d}^{(PR)} \quad (6.21)$$

obtained, for each domain *d* (*d = 1,..., D*), as a convex linear combination between the vector $\bar{y}_{S,d}$ of the sample mean values and the vector $\hat{\bar{y}}_{R,d}^{(PR)}$ of the predicted mean values, relative to the units of set *R*, where $f_d = n_d/N_d$. It is worth noting that where $f_d \cong 0$, the projection and composite estimators produce very similar results. On the other hand, where $f_d \gg 0$, the results obtained with the synthetic and the composite estimator can also differ greatly.

6.5.3. The classification adopted

This chapter adopts a basic classification of estimators that refers to the functional form of WM considered in Factors 4 and 5 above. The following four basic estimators are therefore defined: LM, GLM, LMM, GMM. Estimators referring to the generalized linear models, GLM and GMM, have the advantage of ensuring the production of contingency table estimates in which the estimated values of cell frequencies are always greater than 0. These estimators require, for their construction, knowledge of the elementary value vectors $x_{d,i}$ and $z_{d,i}$. This knowledge, however, is not required for the corresponding estimators based on the LM and LMM models. Furthermore, it should be remembered that the chapter only deals with estimators based on unit-level versions of the models. In this context, of course, some of the auxiliary variables adopted by the model may only be available at the aggregate domain level. The first set of estimation methods, referred to hereafter as Group 1, contains direct and indirect design-based estimators. In particular, for direct estimation methods, the following two projection estimators are defined: Direct-LM and Direct-GLM. Similarly, as regards indirect estimation methods, we have the projection estimators Synthetic-LM and Synthetic-GLM. For these last two estimation methods, the corresponding composite estimators are also defined: Composite-LM and Composite-GLM. A detailed discussion, relating to the case of univariate estimation, of the estimators belonging to Group 1, can be found in Rao (2003; Chapters 2, 3 and 4).

A particular case of the estimators belonging to this group occurs when $\Delta = \Delta_1 = 1$.

This situation leads to the definition of direct and post-stratified ratio estimators

synthetics. These are well-known design-based estimation methods that are used in

the practice of sample surveys. In particular, estimators of this type are among those

in which a specific type of linear WM, LM, is adopted, namely the WM Ratio (R) or the

Post-stratified ratio (P). We then extend the notation introduced so far to include the

estimators in question as special cases of the LM estimator, indicating as LMR and LMP

the estimator based on the R model and on the P model, respectively. Each of the two estimators considered differs, then, depending on Factors 1 and 2. Thus, for example, for the LMR estimator, there

are three possible versions: Direct-LMR, Synthetic LMR and Compound-LMR. Concerning the case of univariate direct estimation, these estimators are treated as specific cases in Rao (2003; Section 2.4.3), relating to the situation in which, for the generalized regression estimator, there are known totals of auxiliary domain variables.

The second set of estimation methods, referred to below as Group 2, contains the indirect model-based estimators. The projection estimators are the following: Synthetic LM, Synthetic GLM, Synthetic LMM, Synthetic GMM, Predictor LMM and Predictor GMM.

The LMM and GMM estimators of the synthetic and empirical predictor type are based on the same estimator of the vector of the regression coefficients. In particular, this estimate is obtained applying the generalized least squares method with weights given by the components of estimated variance. For each of the previous estimation methods, the corresponding composite estimators are also defined.

As for the Group 2 estimators, defined on the basis of mixed-effects models, that is the LMM and GMM predictors, this chapter considers two versions that differ based on the assumptions underlying the random components of the model. In particular, we denote the following predictors: additive random effects (ARE), that are estimators based on a mixed-effects model, in which several vectors of additive and independent random effects are considered; and joint random effects (JRE), when a single vector of random effects is defined, the possible values of which vary as the index d $(d = 1, \ldots, D)$.

Table 6.1 summarizes the projection estimators introduced above, which arise from the different definition of Factors 1, 3, 4, 5 and 6 described at the beginning of this paragraph. For all estimators of this type, Factor 2, relating to how it uses the data observed in the domain, assumes the projection mode. Furthermore, as projection estimators are all unit-level estimators, Factor 7, which differentiates between the estimators based on the level of data aggregation with which the WM is defined, is not considered in the table.

It is emphasized that if a composite estimator is adopted, this is to be reported explicitly in the denomination of the estimator itself. For example, when referring to the Design-based synthetic projection LM estimator, listed in the second row of the table, the corresponding composite estimator will be denoted as a design-based synthetic LM composite.

Table 6.1. Summary table of the projection estimators considered

Estimator	(1)	(3)	(4)	(5)	(6)
LM Direct Design-based	DI	DE	FE	LM	-
LM Synthetic Design-based	IN	DE	FE	LM	-
GLM Synthetic Design- based	IN	DE	FE	GLM	-
LMM Synthetic	IN	MO	FE	LMM	J
GLMM Synthetic	IN	MO	FE	GLMM	J
LMM	IN	MO	ME	LMM	J
GLMM	IN	MO	ME	GLMM	J
LMM with additive random effects	IN	MO	ME	LMM	A
GLMM with additive random effects	IN	MO	ME	GLMM	A

6.6. Multivariate projection estimators

6.6.1. Preamble

Let us consider a set of interest totals referred to target population U_h of size N_h. Let us also suppose that two or more probabilistic samples S_1, S_2, have been selected from U_h, of dimensions n_1 and $n_2 < n_1$, respectively. Therefore S_1 can be called the large sample and S_2 the small sample. A first class of projection estimators of the totals of interest can be obtained from the weighted sum, using the sample weights ω_{i1}, $(i = 1, \dots, n_1)$, of the vectors of the predicted values $\hat{y}_i$ $(i = 1, \dots, n_1)$, referred to the variables of interest y_i on all n_1 units of S_1. The estimate $\hat{\omega}$ of the unknown parameters ω, on which the calculation of the predicted values, $\hat{y}_i(\hat{\omega})$ depends, is obtained starting from the information available on the variables of interest and the auxiliary variables, considered by the WM, observed on the n_2 units of the small sample S_2. This type of projection estimator has been explored in Kim and Rao (2012).

The second class of projection estimators is obtained through the sum of the predicted values of the variables of interest for all N_h units of the population itself. In this case, the unknown parameters are estimated starting from the information available on the variables of interest and on the auxiliary variables, considered by the WM, observed on the units of S_1, S_2 or both samples. In the first two cases, the choice of the sample to be used for the estimation of the parameters will fall on the one containing the information required to compute the estimate, that is, sample S_1 in the first case and sample S_2 in the second. On the other hand, in the third case, where both samples considered have detected the variables of interest and the auxiliary variables, two choices are possible. The first is to use only the data from the large sample. The second option, however, involves using the pooled sample of the two, $S = S_1 \cup S_2$. A general formulation of the projection estimator, that includes the two classes described above, is given by

$$\hat{y}_+^{(PR)} = \sum_{i \in A} \hat{y}_i(\hat{\omega}_B)\, w_{A,i}, \quad (6.22)$$

where $w_{A,i}$ is a known constant assigned to the $i-$ th unit of the collective A $(i \in A)$; moreover, $\hat{\omega}_B$ is an estimator of the unknown parameters ω of the WM obtained on the basis of the sample data *B*. The projection estimator is therefore expressed as the weighted sum of the elementary vectors of synthetic values, also called imputed values, for all units of *A*. The projection estimators – which use the vectors of imputed values $\hat{y}_i$ for all units of the collective of interest instead of the corresponding vectors of true values y_i – are widely used, implicitly or explicitly, in the practice of sample surveys.

The first class of projection estimators is obtained by setting $A = S_1$, $B = S_2$ and $w_{A,i} = \omega_{i1}$

$(i \in S_1)$. Regarding this class of estimators, already discussed in Kim and Rao (2012) and in Chapter 3 of these Guidelines, it is noted that one of the main advantages of this approach is the possibility to use a single sample weight for all variables of interest also when different WMs are adopted for each. Thus, the imputed values of the variables of interest may depend on different choices of the vectors of the auxiliary variables associated with the variables of interest themselves.

The second class of Projection estimators is obtained, instead, by setting $A = U_h$, $B = S_2$ and

$w_{A,i}$ $(i \in A_1)$.

6.6.2. Alternative strategies for defining the auxiliary variables for direct and indirect estimators

In the next two sections, reference will be made to the projection estimators belonging to the second class where $U = A$, in the case of the estimate of y_+, or $U_d = A$ for $d = 1 \ldots D$, in the case of the estimate of y'_+. Furthermore, $S = S_1$ and the sampling weights are equal to ω_{i1}.

To illustrate some important characteristics of the estimator under consideration, let us take the case of the estimate of vectors y_d, $(d = 1, \ldots, D)$ included in y'_+, the generic form of which contains the total *C* of interest relative to the $d-$ th domain $(d = 1, \ldots, D)$). The expression of the estimator is, then,

$$\hat{y}_d^{(PR)} = \sum\nolimits_{i \in U_d} \hat{y}_{di}\left(\hat{\omega}_{S_1}\right). \quad (6.23)$$

The generic vector of predicted values $\hat{y}_{di}$ depends on the strategy to define the model and the auxiliary variables adopted. A particular strategy at the single-mode level allows for the definition of different models – each potentially based on a different set of auxiliary variables – for each of the *C* modes of vector $\hat{y}_{di}(\hat{\omega}_{S_1})$. At the opposite extreme, a general strategy at the TP level provides for the definition of the same model-based approach, therefore, on the same set of auxiliary variables – for all aforementioned modalities. Obviously, each strategy carries with it the ability to obtain estimates that are consistent with one other and different levels of quality of the estimates themselves.

Another important aspect to consider – one that certainly affects the consistency and quality of the estimates produced – is the choice between model-assisted and model-based estimators. The model-assisted $\hat{y}_d^{(PR)}$ $(d = 1 \ldots, D)$ estimator is based on a fixed-effects linear regression model. Moreover, to guarantee correctness under the sample design, the generic matrix $X_{S,d}$ – relating to the n_d sample units of the *d*th domain $(d = 1.., D)$ – must contain a vector of intercepts 1_{n_d}. This corresponds to the case where the macro-domains coincide with the domains, with different types of post-stratification (complete, incomplete or mixed). The sample matrix X_S is formed by the sequencing of *D* matrices $X_{S,d}$ $(d = 1 \ldots, D)$, upon the inclusion of the submatrix X_{1S}=$diag_{d=1}^{D}\{1_{n_d}\}$, with $X_S = row_g^G\{X_{g,S}\}$ see Kim and Rao (2012).

6.6.3. Projection estimator of y_+

The compact expression of the estimator of vector $\hat{y}_+^{(P\,R)}\left(= \sum_{d=1}^{D} \sum_{i \in U_d} \hat{y}_{d,i} = col_{c=1}^{C}\{\hat{y}_c\}\right)$ is given by

$$\hat{y}_+^{(P\,R)} = A_+^T \cdot \hat{y}, \quad (6.24)$$

with $\hat{y} = col_{d=1}^{D}\left\{col_{i=1}^{N_d}\{\hat{y}_{di}\}\right\}$ being the vector of order $(N \times C)$ containing the *N* elementary vectors of predicted values. The explicit expression of the vector of the predicted values depends on the WM adopted. If a linear mixed-effects model is chosen,

$$\hat{y} = X\hat{\beta} + Z\hat{u} \quad (6.25)$$

and the corresponding vector of the estimated residuals are given by

$$\hat{e} = y - \left(X\hat{\beta} + Z\hat{u}\right). \quad (6.26)$$

If, instead, a linear model with fixed effects is adopted, the above equations are modified by eliminating the term $Z\hat{u}$ which is equal to $0_{N \times C}$.

6.6.4. Projection estimator of $\mathbf{y}'_+$

The equations introduced in the previous paragraph are easily adapted to the case in question, relating to the estimate of the vector of domain totals y'_+. Indeed, the general expression of the projection estimator $\hat{y}_+^{\prime(P\,R)} = col_{d=1}^{D}\left\{col_{c=1}^{C}\{\hat{y}_{dc}\}\right\}$ of vector y'_+ is given by

$$\hat{y}_+^{\prime(P\,R)} = A_+^{\prime T} \cdot \hat{y}. \quad (6.27)$$

Let us consider now the multivariate linear WM with mixed effects, for which

$$\hat{y}_{d,i} = X_{d,i}\hat{\beta} + Z\hat{u}. \qquad (6.28)$$

In this case the expressions for the projection estimators of the totals *y* + and *y* 0 + are given by, respectively,

$$\hat{y}_+^{(P\,R)} = X_+\hat{\beta} + Z_+\hat{u} \quad (6.29)$$

and

$$\hat{y}_+^{\prime(P\,R)} = X'_+\hat{\beta} + Z'_+\hat{u} \quad (6.30)$$

where $Z_+ = A_+^T Z$ and $Z'_+ = A_+^{\prime T} Z$.

6.7. Summary of the main recommendations

The main advice provided in this chapter is the following.

1. In many surveys, the sample size is not sufficiently large to guarantee reliable estimates for all target subpopulations.
2. When direct estimates cannot be disseminated because they are of unsatisfactory quality, the SAE methods allow for the problem to be overcome, borrowing strength from the sample information belonging to other domains and resulting in an increase in the effective sample size for each small area.
3. The SAE techniques are strongly model-dependent, and the model parameters can be estimated only by the observed sample data. Thus, if the true model is domain-dependent, the SAE techniques would induce substantial bias in the estimates. Therefore, one should always proceed with caution when applying these techniques for the production of regular official statistics.
4. A standardized approach should always be adopted. The process flow should follow these three main steps: (1) clarification, for the identification and prioritization of the needs and uses of small-area estimates; (2) calculation of direct estimates together with basic design smoothing techniques, i.e. synthetic and composite estimators calculated under a design-based approach; and (3) enhancement of basic design smoothing techniques, via SAE techniques.

References

Alleva, G., Falorsi, P.D., Petrarca, F. & Righi, P. (forthcoming, 2021). Measuring the Accuracy of Aggregates Computed from a Statistical Register, *Journal of Official Statistics.*

Alleva, G., Arbia, G., Falorsi, P.D., Nardelli, V. & Zuliani, A. 2020. A sample approach to the estimation of the critical parameters of the SARS-CoV-2 epidemics: an operational design. *Cornell University arXiv.org* [online]. https://arxiv.org/abs/2004.06068.

Archer, K., Lemeshow, S. & Hosmer, D.W. 2007. Goodness-of-fit test for logistic regression models when data are collected using a complex sample design. *Computational Statistics & Data Analysis*, 51(9): 4450–4464.

Ardilly, P. & Le Blanc, D. 2001. Sampling and weighting a survey of homeless persons: a French example. *Survey Methodology*, 27(1): 109–118.

Barcaroli, G., Falorsi, P., Fasano, A. & Mignolli, N. 2015. A Business Architecture Model to support the Modernisation Project within the Italian National Institute of Statistics. Paper presented at the "Sixtieth World Statistics Congress – ISI2015" July 2015, Rio de Janeiro, Brazil.

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.

Breidt, J.F. & Opsomer. J.D. 2017. Model-Assisted Survey Estimation with Modern Prediction Techniques. *Statistical Science. 2016.* 32(2): 190–235.

Box, G.E.P. & Draper, N.R. 1987. *Empirical Model Building and Response Surface*. New York City, USA, John Wiley & Sons.

Breiman, L. 2001. Random forests. *Machine Learning*, 45(1): 5–32.

Chambers, R.L. & Clark, R.G. 2015. *An Introduction to Model-Based Sampling with Applications*. Oxford Statistical Science Series No. 37. New York City, USA, Oxford University Press.

Chambers, R. & Chandra, H. 2008. *Improved Direct Estimators for Small Areas*. Centre for Statistical and Survey Methodology Working Paper Series No. 03-08. Wollongong, Australia, University of Wollongong. (available at http://ro.uow.edu.au/cssmwp/3).

Chauvet, G. & Tillé, Y. 2006. A fast algorithm for balanced sampling. *Computational Statistics* 21: 53–62.

Cochran, W.G. 1977. *Sampling Techniques*. New York City, USA, John Wiley & Sons.

Common Statistical Production Architecture (CSPA). 2018. CSPA Global Artefacts Catalogue. [online]. https://statswiki.unece.org/display/CSPA/CSPA+Global+Artefacts+Catalogue

Cox, D.R. & Snell, E.J. 1989. *Analysis of Binary Data*. Second edition. London, Chapman & Hall/CRC Press.

Datta, G.S., Lahiri, P., Maiti, T. & Lu, K.L. 1999. Hierarchical Bayes estimation of unemployment rates for the US states. *Journal of the American Statistical Association*, 94(448): 1074–1082.

Deville, J.-C. & Tillé, Y. 2004. Efficient Balanced Sampling: the Cube Method, *Biometrika*, 91(4): 893–912.

De Vitiis, C., Falorsi, S. & Inglese, F. 2014. Implementing the First ISTAT Survey of Homeless Population by Indirect Sampling and Weight Sharing Method. In F. Mecatti, P.L. Conti & M.G. Ranalli, eds. *Contributions to Sampling Statistics*, pp. 119-138. Switzerland, Springer International Publishing.

De Vitiis, C., Righi, P. & Tuoto T. 2008. Joint use of Balanced Sampling and Calibration for Multivariate and Multi-Domain Sample Designs, *Invited Sessions - Proceedings of XLIV Scientific Meeting of the Italian Statistical Society* (SIS), Università della Calabria, 25-27 June 2008, pp. 363-370.

Deville, J.-C. & Tillé, Y. 2005. Variance approximation under balanced sampling, *Journal of Statistical Planning and Inference*, 128: 569–591.

ESSnet on Small Area Estimation. 2012. *Report On Workpackage 6: Guidelines, Final Version*. (available at https://ec.europa.eu/eurostat/cros/system/files/WP6-Report.pdf)

European Commission. 2018. Bucharest Memorandum on Official Statistics in a Datafied Society, as adopted by the European Statistical System Committee (ESSC) meeting on 12 October 2018, https://ec.europa.eu/eurostat/documents/7330775/7339482/The+Bucharest+Memorandum+on+Trusted+Smart+Statistics+FINAL.pdf/59a1a348-a97c-4803-be45-6140af08e4d7.

European Commission. 2020. *White Paper On Artificial Intelligence – A European approach to excellence and trust*. Brussels, European Commission. (available at https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf).

Falorsi, P.D., Righi, P. & Lavallée, P. 2019. Optimal Sampling for the Integrated Observation of Different Populations. *Survey methodology*, 45(3): 485–511.

Falorsi, P.D. & Righi, P. 2008. Balanced Sampling Approach for Multi-way Stratification Designs for Small Area Estimation. *Survey Methodology*, 34(2): 223–234.

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. & Orsini, D. 2006. Balanced and coordinated sampling designs for small domain estimation. *Statistics In Transition*, vol. 7 – 4: 805-830.

Food and Agriculture Organization of the United Nations (FAO). 2014. *The Global Strategy to Improve Agricultural and Rural Statistics. Technical Report on the Integrated Survey Framework*, Technical Report Series GO-02-2014. Rome, FAO. (also available at http://gsars.org/wp-content/uploads/2014/07/Technical_report_on-ISF-Final.pdf).

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. 2019. *Our priorities – The Strategic Objectives of FAO*. Rome, FAO. (also available at http://www.fao.org/3/I8580EN/i8580en.pdf).

FAO. 2020a. *The State of Food Security and Nutrition in the World 2020: Transforming food systems for affordable healthy diets*. FAO, IFAD, UNICEF, WFP and WHO. Rome, FAO. (also available at http://www.fao.org/3/ca9692en/CA9692EN.pdf).

FAO. 2020b. *World Programme for the Census of Agriculture 2020. Volume 1: Programme, concepts and definitions*. Rome, FAO. (also available at http://www.fao.org/3/a-i4913e.pdf).

FAO. 2020c. World Programme for the Census of Agriculture 2020. *Operational Guidelines.* Rome, FAO. (also available at http://www.fao.org/3/CA1963EN/ca1963en.pdf).

Fox, J. & Monette, G. 1992. Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417): 178–183.

Global Strategy. 2018. *AGRIS Handbook on the Agricultural Integrated Survey*. http://gsars.org/wp-content/uploads/2018/05/AGRIS-HANDBOOK-completo-02-24.pdf.

Goodman, L.A. 1961. Snowball Sampling. *Annals of Mathematical Statistics*, 32(1): 148–170.

Grosh, M.E. & Munoz, J. 1996. *A manual for planning and implementing the living standards measurement study survey.* No. LSM126. Washington, D.C. The World Bank.

Grafström, A., Lundström, N.L.P. & Schelin, L. 2012, Spatially Balanced Sampling through the Pivotal Method, *Biometrics,* 68: 514, 520.

Hartley, H.O. 1974. Multiple Frame Methodology and Selected Applications, *Sankhya*, Series C, 36: 99–118.

Harrell Jr., 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.

Horvitz, D.G. & Thompson, D.L. 1952. A generalization of sampling without replacement from finite-universe*. Journal of the American Statistical Association*, 47(260): 663–685.

Hosmer, D.W. & Lemeshow, S. 2013. *Applied Logistic Regression*. New York City, USA, John Wiley & Sons.

Integrated Household Panel Survey (IHPS). Integrated Household Panel Survey 2016 microdata [online]. https://microdata.worldbank.org/index.php/catalog/2936

Insee. [online] http://www.insee.fr/fr/nom_df_met/outils_stat/cube/accueil_cube.htm (last accessed September 2019).

Inter-Agency and Expert Group on Sustainable Development Goal Indicators (IAEG-SDG). 2019. Data Disaggregation and SDG Indicators: Policy Priorities and Current and Future Disaggregation Plans.

Background document presented at the Fiftieth Session of the United Nations Statistical Commission, 5–8 March 2019, New York City, USA. (available at https://unstats.un.org/sdgs/files/meetings/iaeg-sdgs-meeting-09/BG-Item3a-Data-Disaggregation-E.pdf).

Italian National Institute of Statistics (Istat). Stat. Regenesees (r evolved generalised software for sampling estimates and errors in surveys) [online] https://www.istat.it/it/metodi-e-strumenti/metodi-e-strumenti-it/elaborazione/strumenti-di-elaborazione/regenesees (last accessed September 2019).

Istat. Mauss-r (multivariate allocation of units in sampling surveys) [online]. https://www.istat.it/en/methods-and-tools/methods-and-it-tools/design/design-tools/mauss-r (last accessed September 2019).

Jessen, R.J. 1978. *Statistical Survey Techniques.* New York City, USA, John Wiley & Sons.

Kalton, G. 2009. Methods for oversampling rare subpopulations in social surveys. *Survey methodology*, 35(2): 125–142.

Kendell, M.G. & Stuart, A. 1976. *The Advanced Theory of Statistics, Vol. 3: Design and analysis, and time-series*. New York City, USA, and London, Hafner.

Kim, J.K. & Rao, J.N.K. 2012. Combining data from two independent surveys: a model-assisted approach. *Biometrika*, 99(1): 85–100.

Kish, L. 1987. *Statistical Design for Research*. New York City, USA, John Wiley & Sons.

Khan, M.G.M., Mati, T. & Ahsan, M.J. 2010. An optimal Multivariate stratified sampling design using auxiliary information: An integer solution using goal programming approach. *Journal of Official Statistics*, 26(4): 695–708.

Kursa, M.B. & Rudnicki, W.R. 2010. Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11): 1–13.

Lavallée, P. 2007. *Indirect sampling.* New York City, USA, Springer.

Lehtonen, R. & Veijanen, A. 1998. Logistic Generalized Regression Estimators. *Survey Methodology*, 24(1): 51–55.

Lu, W. & Sitter, R.R. 2002. *Multi-way Stratification by Linear Programming Made Practical, Survey Methodology,* 2: 199–207.

Lumley, T. 2019. *Analysis of Complex Survey Samples - R package 'survey'* [online]. Seattle, USA. https://cran.r-project.org/web/packages/survey/survey.pdf

Manyamba, C. 2013. Voices of the Hungry Project – Piloting the Global Food Insecurity Experience Scale for the Gallup World Poll in Malawi: Linguistic adaptation in Chichewa and Chitumbuka. Rome, FAO. (available at http://www.fao.org/fileadmin/templates/ess/voh/MALAWI___FIES_Language_Adaptation_Report_Aug2013.pdf).

McCullagh, P. 1985. On the asymptotic distribution of Pearson's statistics in linear exponential family models. *International Statistical Review*, 53(1): 61–67.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In P. Zarembka, ed., *Frontiers in Econometrics*, pp. 105–142. New York City, USA, Academic Press.

Montanari, G.E. & Ranalli, G. 2002, Asymptotically Efficient Generalized Regression Estimator. *Journal of Official Statistics*, 18(4): 577–589.

Narain, R.D. 1951. On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3: 169–174.

Nedyalkova, D. & Tillé, Y. 2008. Optimal sampling and estimation strategies under the linear model. *Biometrika*, 95(3): 521–537.

National Statistical Office (Malawi). 2017. *Malawi Integrated Household Panel Survey (IHPS) 2016 – Basic Information Document*. National Statistical Office of Malawi. https://microdata.worldbank.org/index.php/catalog/2939/download/48116

Pfeffermann, D. 2002. Small Area Estimation: New Developments and Directions. *International Statistical Review*, 70(1): 125–143.

Pfeffermann, D. 2013. New Important Developments in Small Area Estimation. *Statistical Science*, 28(1): 40–68.

Rdocumentation [online] https://www.rdocumentation.org/packages/car/versions/3.0-9/topics/vif (last accessed January 2021).

Rao, J.N.K. 2003. *Small Area Estimation*. New York City, USA, John Wiley & Sons.

Ryan, T.P. 2008. *Modern Regression Methods*, Second edition. Wiley Series in Probability and Statistics. New York City, USA, John Wiley & Sons Book Series.

Sanders, L.L. & Kalsbeek, W.D. 1990. Network sampling as an approach to sampling pregnant women. In *Joint Statistical Meetings Proceedings, Survey Research Methods Section*, pp. 326–331. Alexandria, VA, USA, American Statistical Association.

Scealy, J. 2009. *Small area estimation using multinomial logit mixed model with category specific random effects.* Research Paper 1351.0.55.029. Canberra, Australian Bureau of Statistics.

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.

Singh, C.A. & Mohl, A.C. 1996. Understanding Calibration Estimators in Survey Sampling. *Survey Methodology*.22(2): 107–115.

Smith, M.D., Rabbitt, M.P. & Coleman-Jensen, A. 2017. Who are the World's Food Insecure? New Evidence from the Food and Agriculture Organization's Food Insecurity Experience Scale. *World Development*, 93: 402–412.

Startfishers Malawi [online] http://starfishers.starfishmalawi.com/secondary-schools-in-malawi/ (Last accessed: January 2021].

Sudman, S., Monroe, G., Sirken, M.G. & Cowan, C.D. 1988. Sampling Rare and Elusive Populations. *Science*, 240(4855): 991–996.

Särndal, C.-E. & Lundström, S. 2005. *Estimation in Surveys with Nonresponse.* 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.

Thompson, S.K. & Seber, G.A.F. 1996. *Adaptive sampling*. New York City, USA, John Wiley & Sons.

United Nations. Youth website [online] https://www.un.org/en/sections/issues-depth/youth-0/index.html (last accessed: January 2021).

United Nations Economic Commission for Europe (UNECE). 2020. Modernization of official statistics. In: *UNECE* [online]. Geneva, Switzerland. https://www.unece.org/stats/mos.html.

UNECE. UNECE Big Data Inventory Home [online] https://statswiki.unece.org/display/BDI/UNECE+Big+Data+Inventory+Home.

United Nations Statistical Commission (UNSC). 2016. *Report on the forty-seventh session* (8-11 March 2016), *United Nations Statistical Commission*, Decision no. 47/101 (para. (n)), E/2014/24-E/CN.3/2014/35, New York City, USA.

UNSD. 2019. Annex I: Compilation of Data Disaggregation Dimensions and Categories for Global SDG Indicators. In: *Supplementing the United Nations Fundamental Principles of Official Statistics: Implementation Guidelines*. Fiftieth Session of the United Nations Statistical Commission, New York City, USA, 5–8 March 2019. https://unstats.un.org/sdgs/files/Annex%201%20-%20Disaggregation%20Compilation.xlsx

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/51st-session/documents/BG-Item3j-Survey-E.pdf).

Valliant, R., Dorfmann, A.H & Royall, R.M. 2000. *Finite Population Sampling and Inference: A Predicition Approach.* New York City, USA, John Wiley & Sons.

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).

Verma, V. 2008. *Sampling for household-based surveys of child labour*. http://www.ilo.org/ipecinfo/product/download.do?type=document&id=8770

Wolter, K.M. 1985. *Introduction to Variance Estimation*, New York City, USA, Springer-Verlag.

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.

World Bank Microdata Library (Integrated Household Panel Survey 2010-2013-2016 (Long-Term Panel, 102 EAs) [online]. https://microdata.worldbank.org/index.php/catalog/2939/related-materials (last accessed January 2021).

World Bank, FAO & UN. 2011. *Global Strategy to Improve Agricultural and Rural Statistics*. Report No. 56719-GLB. Washington, D.C.: World Bank. (also available at http://www.fao.org/fileadmin/templates/ess/documents/meetings_and_workshops/ICAS5/Ag_Statistics_Strategy_Final.pdf)

Annex 1: R packages for data disaggregation

This Annex provides a list of R packages that can be used to implement the statistical methods for data disaggregation presented in Chapters 3, 4 and 5 of these Guidelines.

R package	Use and functions	Documentation
R2BEAT	Determine optimal sample allocation in the multivariate and multidomains case of estimates for two-stage stratified samples (for two-stage stratified samples, the design effects values are required as further input).	https://cran.r-project.org/web/packages/R2BEAT/index.html
Stratification	Produces univariate stratification of survey populations with a generalization of the Lavallée-Hidiroglou method of stratum construction. This package might be useful in a second-stage screening sample design. The package defines the take-all stratum for large units, a take-none stratum for small units, and a certainty stratum to ensure that some specific units are in the sample. The take-all stratum (inclusion probability =1) defines the threshold for defining the screening variable.	https://cran.r-project.org/web/packages/stratification/index.html

R package	Use and functions	Documentation
PracTools	Extensive package implementing several phases of the sampling design. It contains functions for sample size calculation for survey samples, using stratified or clustered one-, two- and three-stage sample designs. Other functions compute variance components for multistage designs and sample sizes in two-phase designs. Among the specific functions of the package are the computation of: - the optimum number of sample elements per PSU for a fixed set of PSUs; - various types of design effects; - sample sizes at each phase of a two-phase design in which strata are created using the first phase; - separate nonresponse adjustments in a set of classes; - optimal values of the first-phase sample size and the second-phase sampling fraction in a two-phase sample; and - the proportional, Neyman, cost-constrained and variance-constrained allocations in a stratified simple random sample (univariate and single-domain).	https://cran.r-project.org/web/packages/PracTools/index.html

R package	Use and functions	Documentation
sampling	Produces the sample selection using several sampling designs (one-stage and two-stage) with uniform and variable inclusion probabilities. It includes balanced sampling, which is a generalization of the most common designs (simple random sample, stratified design, PPS design).	https://cran.r-project.org/web/packages/sampling/index.html
BalancedSampling	Select balanced and spatially balanced probability samples in multidimensional spaces with any prescribed inclusion probabilities. This package is an extension of the sampling package.	https://cran.r-project.org/web/packages/BalancedSampling/index.html
survey	Summary statistics, two-sample tests, rank tests, generalized linear models, cumulative link models, Cox models, loglinear models, and general maximum pseudolikelihood estimation for multistage stratified, cluster-sampled, unequally weighted survey samples. Variances by Taylor series linearization or replicate weights. Post-stratification, calibration and raking. Two-phase subsampling designs. Graphics. PPS sampling without replacement. Principal components, factor analysis.	https://cran.r-project.org/web/packages/survey/index.html

R package	Use and functions	Documentation
Regenesees	Regenesees (R Evolved Generalized Software for Sampling Estimates and Errors in Surveys) is a fully fledged R software for design-based and model-assisted analysis of complex sample surveys. This package replicates almost all functions included in the survey package, while being more flexible and user-friendly.	https://www.istat.it/en/methods-and-tools/methods-and-it-tools/process/processing-tools/regenesees
convey	Enables variance estimation on indicators of income concentration and poverty using complex sample survey designs.	https://cran.r-project.org/web/packages/convey/index.html
Stats	This package includes various statistical functions, including the function glm() to estimate generalized linear models, including logistic regression. It also includes the function AIC() to implement the Akaike information criterion.	https://www.rdocumentation.org/packages/stats/versions/3.6.2 https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm
randomForest	Breiman and Cutler's Random Forests for Classification and Regression.	https://cran.r-project.org/web/packages/randomForest/index.html
Boruta	Boruta feature selection method of Kursa and Rudnicki (2010).	https://cran.r-project.org/web/packages/Boruta/index.html
glmnet	Fits a generalized linear model via penalized maximum likelihood and therefore fits lasso regression to select variable.	https://cran.r-project.org/web/packages/glmnet/

R package	Use and functions	Documentation
ResourceSelection	This package includes the function hoslem.test() to estimate Hosmer-Lemeshow GOF.	https://cran.r-project.org/web/packages/ResourceSelection/
generalhoslem	This package contains a series of functions to assess the GOF of binary, multinomial and ordinal logistic models.	https://cran.r-project.org/web/packages/generalhoslem/
pscl (political science computation laboratory)	This package includes the function pR2() to estimate pseudo R-squares of McFadden (1974) and Cox and Snell (1989) measuring explained variation to assess the model performance for logistic regression.	https://cran.r-project.org/web/packages/pscl/
car	This package includes the function vif() to estimate the generalized VIF of Fox and Monette (1992) for testing multicollinearity.	https://cran.r-project.org/web/packages/car/index.html

Guidelines on data disaggregation for SDG Indicators using survey data

ISBN 978-92-5-133942-8

9 789251 339428

CB3253EN/1/02.21