Food and Agriculture Organization of the United Nations

An indirect estimation approach for disaggregating SDG indicators using survey data

Case study based on SDG Indicator 2.1.2

An indirect estimation approach for disaggregating SDG indicators using survey data

Case study based on SDG Indicator 2.1.2

Food and Agriculture Organization of the United Nations
Rome, 2022

Required citation:
FAO. 2022. *An indirect estimation approach for disaggregating SDG indicators using survey data – Case study based on SDG Indicator 2.1.2*. Rome. https://doi.org/10.4060/cb8670en

ISBN 978-92-5-135785-9

Contents

Tables

Figures

Boxes

Acknowledgements

This technical report, which further expands the methods and case studies presented in the *FAO Guidelines on data disaggregation for Sustainable Development Goal (SDG) Indicators using survey data* (FAO, 2021), was prepared by the Office of Chief Statistician (OCS) of the Food and Agriculture Organization of the United Nations (FAO), under the general direction and encouragement of FAO Chief Statistician, Pietro Gennari.

Piero Demetrio Falorsi supervised the development of the entire publication, while the individual sections were drafted by the following authors:

- Section 1: Clara Aida Khalil and Piero Falorsi;
- Section 2: Piero Falorsi and Clara Aida Khalil;
- Section 3: Piero Falorsi and Clara Aida Khalil;
- Section 4: Clara Aida Khalil, Stefano Di Candia and Ayça Dönmez;
- Section 5: Piero Falorsi and Clara Aida Khalil.

The authors are particularly grateful to Carlo Cafiero, Sara Viviani, for their precious technical inputs, and suggestions for current and future improvement of the study.

Abbreviations and acronyms

FAO	Food and Agriculture Organization of the United Nations
FIES	Food Insecurity Experience Scale
FIES- SM	Food Insecurity Experience Scale Survey Module
GWP	Gallup World Poll
HT	Horvitz-Thompson
IAEG-SDG Indicators	Inter-Agency and Expert Group on Sustainable Development Goals indicators
IHS	Integrated Household Survey
IHS4	Fourth Integrated Household Survey
IRT	item response theory
NSO	National Statistical Office
OCS	Office of Chief Statistician
prob.ms	probability of moderate and severe food insecurity
prob.s	probability of severe food insecurity
SDG	Sustainable Development Goals
UNSC	United Nations Statistical Commission
WM	Working model

1. Introduction

As the custodian United Nations (UN) agency of 21 Sustainable Development Goal (SDG) indicators, and a member of the Inter-Agency and Expert Group on Sustainable Development Goals indicators (IAEG-SDGs) and the Working Group on data disaggregation, the Food and Agriculture Organization of the United Nations (FAO) has been working to support countries in reporting SDG indicators at the required disaggregation level. To this end, FAO Office of Chief Statistician (OCS) has developed ***Guidelines on data disaggregation for SDG Indicators using survey data*** (FAO, 2021; from here on referred to as "the Guidelines"), which offer methodological and practical guidance for the production of direct and indirect estimates of SDG indicators having surveys as their main or preferred data source. Starting from this work, the FAO is continuing working on data disaggregation and indirect estimation approaches, by developing practical case studies on indicators under its custodianship.

As in the Guidelines, this technical report presents a case study based on the so-called ***"projection estimator"***, allowing the integration of two independent surveys for the production of synthetic disaggregated estimates. In particular, the publication presents a practical exercise focused on the production of disaggregated estimates for SDG Indicator 2.1.2, on the ***Prevalence of Moderate or Severe Food Insecurity in the population based on the Food Insecurity Experience Scale (FIES)***. This application – based on survey microdata from Malawi – expands and enriches the brief practical exercise presented in the Guidelines by:

- Providing a step-by-step guide for the replication of the exercise and its implementation on different datasets and contexts.
- Illustrating the basic R routines developed for each of the steps, in order to promote and facilitate the adoption of the open source software in countries' national statistical offices (NSOs).
- Presenting the functional form of the projection estimator necessary to address all the SDG indicators under FAO custodianship the computation of which should be based on survey microdata.
- Providing the necessary theoretical and practical tools to assess the precision and accuracy of the produced indirect disaggregated estimates;
- Replicating the case study on survey datasets from two additional countries, namely Guatemala and South Africa, the results of which confirm the robustness of the approach and are reported in separate annexes of this publication.

Given the practical nature of this technical report, the target audience includes statistical practitioners in NSOs and international organizations wanting to adopt the projection estimator for data disaggregation of SDG indicators by means of integrating survey data with additional data sources such as other surveys, censuses, administrative records, and/or geospatial information.

The report is structured as follows. **Section 2** provides an overview of the main challenges to achieve data disaggregation for SDG indicators having survey data as their main data source. In addition, the section presents possible strategies to address these limitations at the analysis stage of the statistical production process. **Section 3** illustrates in detail the main characteristics of the indirect estimation approach based on the projection estimator, highlighting its relevance in the context of the SDG monitoring Framework.

Section 4 provides a step-by-step guide on how to implement the projection estimator, using an example based on indicator 2.1.2 and microdata from Malawi. Finally, **Section 5** presents the main conclusions from the study and outlines possible extensions and way forward. In addition to the case study presented in Section 4, **Annex A** and **B** present results for the same indirect estimation method implemented on microdata from Guatemala and South Africa.

2. Possible approaches to data disaggregation for SDG indicators based on survey data

In a sample-survey context, the estimator of a parameter of interest for a given subpopulation is said to produce a direct estimate when the estimation process 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. In addition, the sample of most surveys does not cover all possible sub-domains of the population, and there is the possibility of having disaggregation domains without any sample observation. A "small area" or "small domain" is any subpopulation for which a direct estimator with the required precision cannot be obtained with a given data source.[1]

These issues can potentially be addressed at different stages of the statistical production process. They can be tackled at the design stage, by adopting sampling strategies guaranteeing an observed set of sampling units for every subpopulation for which disaggregated data must be produced. With traditional sampling techniques, this goal implies increased survey costs and complexity and can become quickly unfeasible when dealing with multiple disaggregation domains. On the other hand, problems of this nature can be addressed at the analysis stage, by adopting approaches of indirect estimation that cope with the little information available for small areas by borrowing strength from additional data sources or domains.

The *Guidelines on data disaggregation for SDG indicators using survey data* (FAO, 2021) provide a detailed review of methods to deal with data disaggregation at the sampling design, and also discusses methods to be adopted at the analysis stage. Referring to that publication for a complete overview, the main categories of approaches to address data disaggregation during the analysis phase is provided below.

2.1 Addressing data disaggregation at the analysis stage

Data disaggregation can be addressed adopting indirect estimation approaches coping with the little information available for so-called small areas, by borrowing strength from additional domains. In particular, the integrated use of different data sources offers a powerful approach for achieving the desired level of disaggregation by preserving estimates accuracy.

Typical data sources that could be integrated with data from a particular household and/or agriculture surveys are:

- other surveys;
- censuses;
- administrative registers;
- geospatial information and big data.

[1] 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).

Indirect estimation approaches range from model-based methods to model-assisted approaches.

The **model-based approach** (such as that adopted in small area estimation techniques) assumes that the values of a variable of interest observed on the units of a population are the realization of a random variable. The model (often denoted with the term **superpopulation model**) defines a class of distributions to which this random variable is supposed to belong. In this context, the sample is interpreted as the result of a double random experiment:

1) the observed realization of the model generates the population from which the sample is drawn; and,
2) the sample units are observed in accordance with specific random selection rules incorporated in the sampling designs (Royall, 1976; Valliant *et al.*, 2000; Chambers and Clark, 2015; and Tillé, 2019).

Taking into account the fact that the sample is an ancillary statistic[2], Royall (1976) proposed to develop the inference conditionally on it. Indeed, once the sample is selected, the observed units are no longer random.

In these contexts, it is important that the model express a known and previously tested relationship. If the model adequately describes the population, inference can be conducted with respect to the model and conditional to the sample selection. In other words, when the model is correct, a model-based approach results in the optimal estimator. However, a model is always an approximate representation of reality. For instance, the model may fail in its objective to reproduce reality when the survey does not capture some relevant auxiliary variables for the phenomenon at hand. Hansen *et al.* (1983) argue that, when the model is not correctly specified, the bias may be so important to result in confidence intervals that do not include the true value of the parameter to be estimated.

The debate on the validity of model-based approaches versus those based on the properties of sampling designs is wide and interesting, even if its arguments are more philosophical than mathematical. From a statistical point of view, both theories are valid. The controversy relates to the idea that, with a model, we provide a formalization of reality that may not be the correct one. The principle of impartiality – which is one of the fundamental principles of official statistics – is a strong argument against the adoption of inferential approaches entirely based on model assumptions.

As a response to this debate, a hybrid approach – the so-called model-assisted approach – was developed, which allows producing valid inference under model assumptions and is robust to wrong specifications of the model. In this context, the model only allows exploiting auxiliary information available at the time of estimation to increase the accuracy of final estimates and deal with traditional issues such as non-response. At the same time, the estimation process is based on the inferential properties of the survey's sampling design. In model-assisted approaches, the model is often denoted as Working model (WM), thus meaning that it does not have the ambition to explain the phenomenon, but only to offer a working tool that can help improving the quality of the estimator. For more details on model-assisted approaches, the authors of this technical report refer to Särndal, Swensson and Wretman (1992), where it is demonstrated that an estimator developed under a model-assisted approach is approximatively unbiased under the

[2] An ancillary statistics is a measure extracted from a sample, the result of which does not depend on the parameter(s) of a model.

assumption of repeated sampling, irrespective of the shape of the finite population scatter. From this follows that the estimator is unbiased irrespective of whether the assumptions of the model are true or false. On the other hand, the validity of the model is a crucial factor to achieve a small variance.

3. The projection estimator

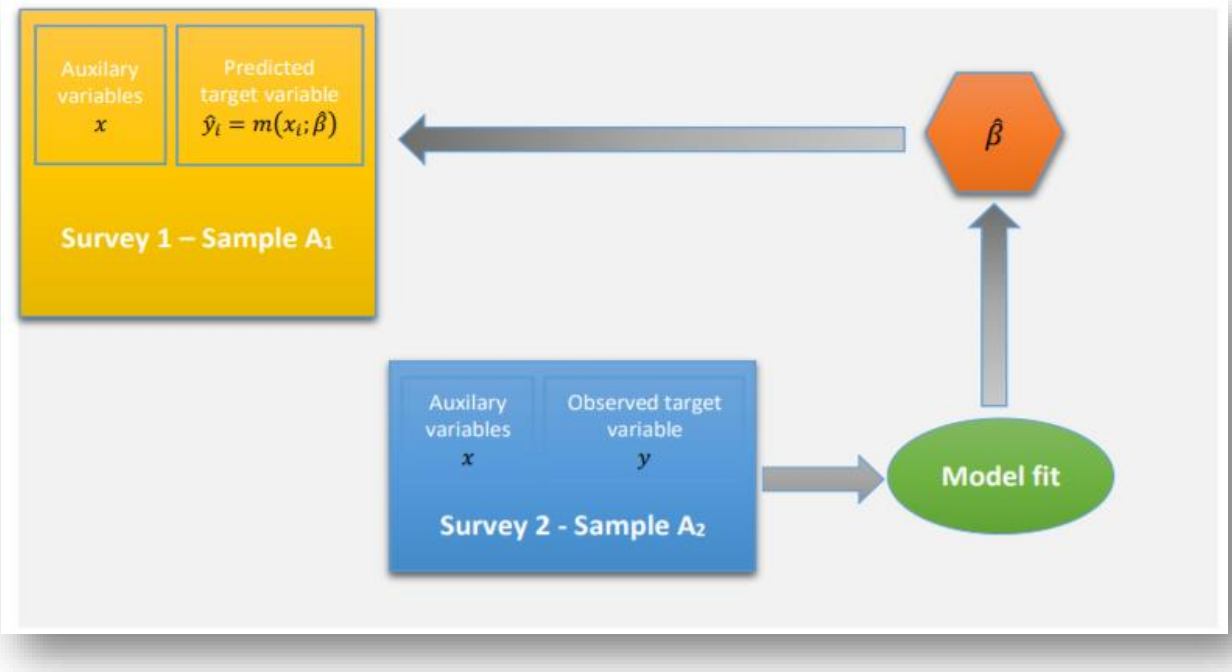
Among the various approaches available to produce indirect disaggregated estimates for SDG indicators, this technical report presents and applies a model-assisted approach based on the so-called ***projection estimator***. This method was first presented in its simpler formulation in the seminal paper "Combining data from two independent surveys: a model-assisted approach" from Kim and Rao (2012).

The paper starts from the following scenario: Let us consider two independent surveys, where the first survey is characterized by a large sample A_1, and collects only variables of general use (auxiliary variables); while the second survey has a smaller sample A_2, but collects information on a target variable y along with the same set of auxiliary variables available in A_1. The two samples A_1 and A_2 may be selected from possibly different frames exploiting distinct sampling designs. (Kim and Rao, 2012). A relevant assumption is that the auxiliary variables collected with the two surveys share similar structure and definitions.

Kim and Rao (2012) suggest a model-assisted projection method of estimation based on a WM that results in asymptotically unbiased projection estimators. With this approach, a two-steps process (**Figure 1**) generates synthetic or proxy values of a variable of interest:

- First, the WM linking the variable of interest to the common auxiliary variables is fitted on data from A_2, leading to the estimation of a set of model parameters.
- Then, the values of the variable of interest are predicted in A_1 by applying the estimated parameters to the auxiliary variables observed in the large sample.

Figure 1. The projection estimator to combine two independent surveys



Source: authors' own elaboration, 2022.

In the following paragraphs we introduce the basic theory and notation used in Kim and Rao (2012), presenting also some possible extensions considered relevant in the context of the SDG monitoring framework.

3.1 The basic approach: estimation of a population total

Let $Y = \sum_{i=1}^{N} y_i$ be a target population total, where y_i is the value of the variable of interest y on unit i and N the overall population size. A Working model is introduced according to which $y_i = m(x_i; \beta) + u_i$, 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, with β being the column vector of the model parameters. The two column vectors x_i and β must have the same number of elements.

The model expectation of u_i is equal to 0. i.e. $E_M(u_i) = 0$.

When the WM is a regression model, $m(x_i; \beta) = x_i'\beta$, with x_i' being the transpose of x_i.,

Let $\hat{\beta}$ be the estimator of β obtained from the second survey, using data $\{(y_i; x_i): i \in A_2\}$ and let $\hat{y}_i = m(x_i; \hat{\beta})$ be the predicted value of y_i, with $E_M(\hat{\beta}) = \beta$. Let E_P denote the expectation under repeated sampling and let w_{i1} be the sampling weights of the sample A_1 allowing to compute sampling-unbiased estimates.

If y_i would be available in A_1, $\hat{Y}_1 = \sum_{i \in A_1} y_i$ would be a sample unbiased estimator of Y, where the sampling unbiasedness implies that, under the repeated sampling, the expected value of $\hat{Y}$ is equal to the unknown total Y: $E_P(\hat{Y}_1 - Y) = 0$.

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

$$\hat{Y}_p = \sum_{i \in A_1} w_{i1}\hat{y}_i = \sum_{i \in A_1} w_{i1} m(x_i; \hat{\beta}) \qquad (3.1)$$

The estimator $\hat{Y}_p$ is called projection estimator (or synthetic estimator), as $\hat{y}_i = m(x_i; \hat{\beta})$ can be viewed as a projection of y_i using the auxiliary variables x_i.

Bias and variance

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

$$E_P E_M\left[\hat{Y}_p - E_M(Y)\right] = 0.$$

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

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

with β_0 denoting the estimate of β when observing the entire population, i.e. the estimation that we would get using census data.

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

$$\hat{B}ias(\hat{Y}_p) = \sum_{i \epsilon A_2} w_{i2} \left[y_i - m(x_i; \hat{\beta}) \right],$$

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

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

$$\sum_{i \epsilon A_2} w_{i2} \left[y_i - m(x_i; \hat{\beta}) \right] = 0. \qquad (3.2)$$

Therefore, to guarantee sample-unbiasedness, estimates of $\hat{\beta}$ should be obtained by respecting the condition established in Formula 3.2. For generalized linear models (such as heteroscedastic linear regression models or logistic models) to satisfy condition 3.2, 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

$$Var\left(\hat{Y}_p\right) = Var\left(\sum_{i \epsilon A_1} w_{i1}\, m(x_i; \beta_0) \right) + Var\left(\sum_{i \epsilon A_2} w_{i2} \left[y_i - m(x_i; \beta_0) \right] \right) \qquad (3.3)$$

where the first term on the right-hand side is the variance due to sampling, in survey S_1, of the population predictions (with the β_0 value), and the second term is the variance due to sampling, in survey S_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, i.e. if model *m* is sufficiently predictive.

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

$$\hat{V}ar\left(\hat{Y}_p\right) = \hat{V}ar\left(\sum_{i \epsilon A_1} w_{i1}\, m(x_i;\, \hat{\beta}) \right) + \hat{V}ar\left(\sum_{i \epsilon A_2} w_{i2} \left[y_i - m(x_i;\, \hat{\beta}) \right] \right)$$

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

3.2 First extension: domain estimation

Let *d* denote a particular domain for which disaggregated data must be produced (e.g. sex of individuals, the indigenous status, or a particular geographic location).

Let

$$Y_d = \sum_{i=1}^{N} y_i\, \gamma_{di} \qquad (3.4)$$

be the total of the target variable for the *d*-th domain, where γ_{di} is the domain membership variable such that:

$$\gamma_{di} = \begin{cases} 1 \; if \; i \in d \\ 0 \; otheriwse \end{cases}$$

The projection estimator of the total Y_d is given by:

$$\hat{Y}_{p,d} = \sum_{i \epsilon A_1} w_{i1} \, m(x_i; \hat{\beta}) \gamma_{di} \qquad (3.5)$$

The condition for sample-unbiasedness becomes

$$\sum_{i \epsilon A_2} \omega_{i2} \left[y_i - m(x_i; \hat{\beta}) \right] \gamma_{di} = 0 \qquad (3.6)$$

To satisfy Condition 3.6, vector x_i must include the γ_{di} values, which means that the model has a domain intercept. This can be fulfilled only if, in sample A_2, domain *d* has a sufficient sample size.

However, in general, the condition established under condition 3.6 cannot be ensured in the sampling design phase for very small disaggregation domains. Therefore, it is preferable to focus on the model conditions that provide negligible bias. From Kim and Rao (2012), it can be derived that the relative bias $E_P(\hat{Y}_{p,d} - Y_d)/Y_d$ can be expressed as

$$\frac{E_P(\hat{Y}_{p,d} - Y_d)}{Y_d} = -\frac{Cov\left[\gamma_{di}, \left(y_i - m(x_i; \beta)\right)\right]}{\bar{N}_d \, \bar{Y}_d}, \qquad (3.7)$$

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 0. the model should be specified to ensure that the model residuals depend slightly on the domain membership variables:

$$Cov\left[\gamma_{di}, \left(y_i - m(x_i; \beta)\right)\right] \cong 0. \qquad (3.8)$$

This will be the case if the WM is correctly specified.

From the relationship in Formula 3.7, 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 Expression 3.3 as

$$Var\left(\hat{Y}_{p,d}\right) = Var\left(\sum_{i \epsilon A_1} w_{i1} \, m(x_i; \beta_0) \, \gamma_{di}\right) + Var\left(\sum_{i \epsilon A_2} w_{i2} \left[y_i - m(x_i; \beta_0)\right]\right), \qquad (3.9)$$

Where the sampling estimate is

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

Highlight: It is possible to produce cross-tabulations of the variable of interest y also for disaggregation domains not included in the data collection instrument used to get A_2 (sample providing information on y). For example, let's suppose to be interested in estimating a parameter related to y, disaggregated by indigenous status. Let us also assume that the information on the indigenous status of respondents is not available in A_2, but only in A_1. By projecting the values of y on A_1, it is possible to use the auxiliary information on the indigenous status to estimate the parameter of interest considering this disaggregation dimension.

3.3 Second extension: estimation of a proportion or a ratio

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

$$R_d = \frac{Y_d}{Z_d}, \quad (3.10)$$

where Y_d is defined as in Section 3.2 (formula 3.4) and

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

z_i being the value of the variable z on unit i, where the variable z is observed in the survey A_1.

In all these cases, the projection estimator can also be expressed in the form of the ratio:

$$\hat{R}_{p,d} = \frac{\hat{Y}_{p,d}}{\hat{Z}_d} \quad (3.11)$$

where $\hat{Y}_{p,d}$ is defined in Section 3.2 (formula 3.5) and

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

is the direct estimate of the total X_d from the survey A_1.

When $z_i = 1$, expression (3.11) provides the projection estimator of a proportion

$$\hat{R}_{p,d} = \frac{\hat{Y}_{p,d}}{\hat{N}_d} \quad (3.12)$$

where

$$\widehat{N}_d = \sum_{i \epsilon A_1} \omega_{i1}\, \gamma_{di}$$

is the direct estimator of the population size in domain d

$$N_d = \sum_{i \epsilon U} \gamma_{di}.$$

In order to study the asymptotic properties of estimator (3.11), we consider its linear approximation, given by the first order terms of Taylor's series approximation:

$$\hat{R}_{p,d} = R_d + \frac{1}{Z_d}\left[\left(\hat{Y}_{p,d} - Y_{p,d}\right) - R_d\left(\hat{Z}_d - Z_d\right)\right] + o_i \quad (3.13)$$

where o_i is a rest of minor order, and

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

Omitting the term o_i, we have that - for large values of *N* and *n* - the estimator $\hat{R}_{p,d}$ is approximately design unbiased, and the variance is

$$Var\left(\hat{R}_{p,d}\right) \cong \frac{1}{Z_d^2}\left[Var\left(\hat{Y}_{p,d}\right) + R_d^2 Var\left(\hat{Z}_d\right) - 2R_d Cov\left(\hat{Y}_{p,d}, \hat{Z}_d\right)\right]. \quad (3.14)$$

To have a rough evaluation of the variance of $\hat{R}_{p,d}$ as a function of the variances of the numerator and denominator, we can adopt the approximation $\hat{Y}_{p,d} \cong R_d \hat{Z}_d$.

From this it follows that:

$$Var\left(\hat{R}_{p,d}\right) \cong \frac{1}{X_d^2}\left[Var\left(\hat{Y}_{p,d}\right) + R_d^2 Var\left(\hat{Z}_d\right) - 2R_d^2 Var\left(\hat{Z}_d\right)\right]$$

$$\cong \frac{1}{X_d^2}\left[Var\left(\hat{Y}_{p,d}\right) - R_d^2 Var\left(\hat{Z}_d\right)\right].$$

In Woodruff (1971) it is demonstrated that this expression may be approximated with the variance of the total of a transformed variable z:

$$Var\left(\hat{R}_{p,d}\right) \cong Var\left(\sum_{i \epsilon A_1} \omega_{i1}\, t_{di}\right) \quad (3.15)$$

where t_{di} is the *Woodruff* transformation:

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

A plug-in estimate of t_{di} from the survey data is

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

The plug-in sampling estimate of Variance 3.15 is

$$\hat{V}ar\left(\hat{R}_{p,d}\right) \cong Var\left(\sum_{i\epsilon A_1}\omega_{i1}\,\hat{t}_{di}\right).$$

Highlight: Various SDG Indicators under FAO custodianship can be treated using the ratio extension of the projected estimator. These are:

SDG Indicator 2.1.1: Prevalence of Undernourishment;

SDG Indicator 2.1.1: Prevalence of moderate or severe food insecurity in the population based on the FIES;

SDG Indicator 2.3.1: Volume of production per labour unit by classes of farming / pastoral / forestry enterprise size;

SDG Indicator 2.3.2: Average income of small-scale food producers, by sex and indigenous status;

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

4. Step-by-step implementation of the projection estimator with a case study based on SDG indicator 2.1.2

This section illustrates the steps and the software to integrate two independent surveys to produce disaggregated estimates, by means of a practical example relying on the projection estimator discussed in Section 3. This model-assisted indirect estimation approach is applied to **SDG Indicator 2.1.2** on the **prevalence of moderate or severe food insecurity in the population based on the FIES**. Even though all the steps for implementing the projection estimator are presented with reference to a particular indicator, this approach has a much wider applicability and could be adapted to other SDG indicators based on survey and/or census data.

4.1 SDG Indicator 2.1.2

SDG Indicator 2.1.2 provides internationally comparable estimates of the percentage of individuals in the population who have experienced food insecurity at moderate or severe levels during a pre-defined reference period. The measurement of the level of severity of food insecurity is based on the FIES, which is an experience-based metric of food insecurity severity (FAO, 2022).

The scale relies on people's direct responses to questions about their experiences facing constrained access to food. The Food Insecurity Experience Scale 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 experienced any of the subjective dimensions of food insecurity captured by the scale due to limited availability of money or other resources. Responses to these eight questions are analyzed using the item response theory (IRT) to obtain cross-country comparable measures of the severity of food insecurity of individuals, treated as a "latent" – or unobservable - trait.

Unfortunately, too few countries have currently included the FIES module in their nationally representative surveys. To fill this data gap, in 2015, the FAO began collecting the FIES data by leveraging on the Gallup World Poll (GWP), a branch of Gallup, Inc. that surveys nationally representative samples of the adult population annually in nearly 150 countries, covering 90 percent of the world's population. This has enabled FAO to collect information from individual respondents at a relatively low cost and to compute country-level estimates of the prevalence of food insecurity at different levels of severity that are valid, reliable and comparable across countries.

For what concerns data disaggregation, this is particularly relevant to properly monitor the prevalence of moderate or severe food insecurity in the population. To ensure regular access to nutritious and sufficient food, and thus reduction of food insecurity, detailed and disaggregated information by age, gender, income level, and geographic location is necessary to identify priority efforts and interventions. Besides dimensions recommended by the IAEG-SDGs[3], additional disaggregation dimensions may be relevant in specific country contexts.

[3] More details on the mandatory and future dimensions for data disaggregation of SDG Indicator 2.1.2, the authors refer to the compilation of categories and dimensions of data disaggregation prepared by the IAEG-DGs (UNSD, 2022).

In order to produce direct disaggregated estimates of Indicator 2.1.2 by the above reported and other disaggregation dimensions, two conditions need to be satisfied:

- The stratification or disaggregation variables - such as sex, age, income, education, and geographic location – have to be available in the survey dataset;
- The sampling size needs to be large enough to produce accurate estimates for each disaggregation dimension.

The first condition is satisfied also when data are collected through the GWP for standard disaggregation dimensions such as sex, age, income quintile, geographic location, and education. However, given the limited number of auxiliary information available in the dataset, other dimensions that may prove relevant at the national level would not be possible. For example, information on the indigenous and/or migratory status of individuals is not available in the GWP dataset. The second aspect is often more problematic. Indeed, GWP samples are large enough to guarantee representative and accurate estimates at the national level, but often too small to ensure accuracy at more detailed disaggregation level.

These two types of issues can be addressed with the indirect estimation approach discussed in Section 3 and tested in this technical report.

4.2 Presentation of datasets used for the case study

The projection estimator presented in Section 3 has been tested on microdata from two surveys implemented in Malawi in 2016.

4.2.1 Small sample: FIES individual module collected through the Gallup World Poll (GWP)

As mentioned in Section 4.1, in 2014, the FAO started collaborating with the Gallup Inc. to implement the FIES module in over 150 countries. The FIES-SM 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 the case of Malawi, the FIES module was translated in the two local languages (Chichewa and Chitumbuka) to make sure that the intended meaning of each question was rightly expressed. The Gallup dataset for 2016 includes a sample of 1 000 individuals divided in 125 primary sampling units.

Variables in the dataset are described in Table 1 below. Variables reported in yellow cells contains dichotomous answers (Yes/No) provided by individual respondents to questions in the FIES-SM. Variables in green cells are instead those used to build auxiliary variables for fitting the projection model. Finally, variables in pink cells have been used to construct the dependent variables of the two fitted projection models (see Section 4.4). It should be noted that, compared to the exercise presented in the Guidelines (FAO, 2021), this case study relied on an extended version of the GWP dataset including additional auxiliary variables, namely the marital status of respondents and a dummy indicating whether they were in possession of a mobile phone.

Table 1. Variables included in the GWP dataset (Malawi, 2016)

Variable name	Description
Random ID	Unique respondent identifier
Worried	Whether the respondent is worried about not having enough food to eat because of a lack of money or other resources (Yes/No)
Healthy	Whether the respondent is unable to eat healthy and nutritious food because of a lack of money or other resources (Yes/No)
Fewfood	Whether the respondent ate only a few kinds of foods because of a lack of money or other resources (Yes/No)
Skipped	Whether the respondent skipped a meal because there was not enough money or other resources to get food (Yes/No)
Ateless	Whether the respondent ate less than he/she thought he/she should because of a lack of money or other resources (Yes/No)
Runout	Whether the household ran out of food because of a lack of money or other resources (Yes/No)
Hungry	Whether the respondent felt hungry but did not eat because there was not enough money or other resources for food (Yes/No)
Whlday	Whether the respondent went without eating for a whole day because of a lack of money or other resources (Yes/No)
Wt	Post-stratification sampling weights
Year	Year when the Gallup World Poll (GWP) was administered in the country
N_adults	Number of adults 15 years of age and above in household
N_child	Number of children under 15 years of age in household
Raw_score	Sum of Affirmative responses to FIES questions
Raw_score_par	Estimated person parameters using the Rasch model
Raw_score_par_error	Estimated person parameter errors using the Rasch model
Prob_Mod_Sev	Probability of being moderately or severely food insecure
Prov_Sev	Probability of being severely food insecure
Age	Age of respondent
Education	Education of respondent
Area	Area of residence of respondent
Gender	Gender of respondent
Income	Income quintile of respondent
Employment	Employment of respondent
Marital	Marital status of respondent
Cellphone	Respondent ownership of mobile phone

Source: authors' own elaboration, 2022.

The categories of above mentioned auxiliary variables are detailed in Section 4.3, where the adopted recoding procedure is also illustrated.

4.2.2 Big sample: The Fourth Integrated Household Survey (IHS4) of Malawi (2016–2017)

The HIS is implemented by the NSO of Malawi every three years to monitor and evaluate the changing conditions of Malawian households. This survey is an important source of information on the country's socio-economic indicators, which are key to the evidence-based policy formulation process and monitoring progress towards achieving the SDGs.

The IHS4 is the fourth full survey conducted under the umbrella of the World Bank's Living Standard Measurement Study – Integrated Surveys on Agriculture (LSMS-ISA), and was fielded from April 2016 to April 2017. The IHS4 collected information from a sample of 12 480 households statistically designed to be representative at national, district, and urban/rural levels.[4] The cross-sectional survey used four questionnaires: 1) the household questionnaire, 2) the agricultural questionnaire, 3) the fishery questionnaire, and 4) the community questionnaire. Full detail on the scope and structure of these four questionnaires is provided in the survey report (NSO, Malawi, 2017) and in the WB microdata catalogue (World Bank Microdata Library). Module T of the household questionnaire was designed to perform a subjective assessment of household's wellbeing, including also the FIES module at household level.

4.3 List of steps for the implementation of the case study

Exploiting microdata collected with surveys described in Section 4.2, the projection estimator has been used to produce disaggregated estimates of SDG Indicator 2.1.2, following the steps listed below:

1. **Recoding the variable of interest.** As discussed in Section 3, the synthetic values $\hat{y}_i$ are predicted through a known function $m(x_i;\ \hat{\beta})$ of the estimator $\hat{\beta}$, which is obtained from the small sample (the GWP microdata in the case of Malawi). The selection of the functional form for m relies heavily on the type of variable y considered (e.g. scale, nominal, dichotomous). Section 4.4 shows how the variable of interest of this study was recoded to then apply a multinomial logistic regression.
2. **Identifying and recoding auxiliary variables.** The implementation of the projection estimator proposed by Kim and Rao (2012) requires the availability of the same set of auxiliary variables in the two surveys to be integrated. In order to improve the efficiency of the projection, these variables need also to share common structure and definitions. Based on this prerequisite, Section 4.5 illustrates variables that have been included as potential auxiliary variables for the implementation of the projection estimators, providing the R basic commands for recoding and harmonizing them in the two datasets.
3. **Selection of variables to be included in the mode**l. Among the various statistical approaches available to select auxiliary variables to be included in a regression model, Section 4.6 of this technical report discusses and applies the Boruta feature selection method from Kursa and Rudnicki (2010). In addition, the section illustrates the *Boruta R package* (Boruta) for the implementation of such method.
4. **Definition of the function $m()$ and estimation of projection parameters.** In Section 4.8, a weighted multinomial logistic regression is used to estimate the projection parameters needed to compute synthetic values of the variable of interest. The implemented regression model includes both the auxiliary variables selected through Boruta and those representing the disaggregation

[4] The territory of Malawi is divided in regions (central, northern, and southern), which are in turn divided in a total of 28 districts.

dimensions of interest (age, sex, income and geographic location). This is done to meet the condition ensuring sample unbiasedness of the projection estimator for a specific projection domain.

5. **Computation of synthetic values.** Using the estimated projection parameter, Section 4.9 illustrates the computation of synthetic values of the variable of interest in the large dataset. This in turn, will allow producing disaggregated estimates of SDG Indicator 2.1.2 by all the considered disaggregation dimensions.
6. **Assessment of estimates accuracy.** Section 4.10 illustrates the R packages available to estimate the variance of the projected disaggregated estimates. The variance can be used as a measure of estimates accuracy.

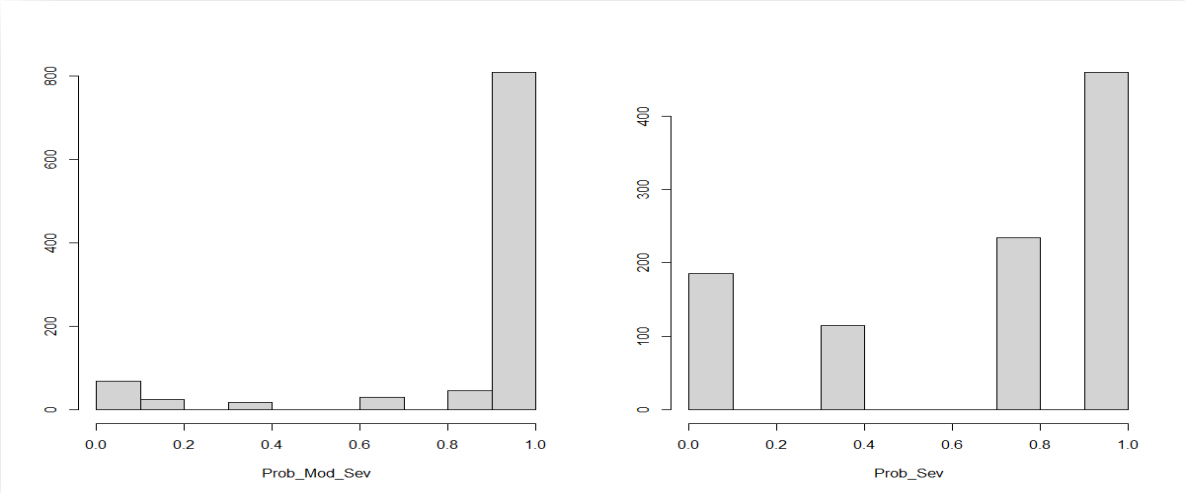
4.4 Recoding the variable of interest

This study considers the following two variables of interest **(dependent variables)**:

- ***Prob_Mod_Sev* =** The probability of being moderately or severely food insecure;
- ***Prob_Sev* =** The probability of being severely food insecure.

While the first variable is the base for computing the prevalence of moderate or severe food insecurity in the population (SDG Indicator 2.1.2), the second is another population parameter that can be estimated with the FIES module at a more severe level. Being two probabilities, these two variables could ideally take any value in the interval [0.1]. However, by construction, their distribution were concentrated around few values (Figure 2), i.e. the nine possible raw scores.

Figure 2. Histograms of the probability of being 1) moderately or severely food insecure and 2) severely food insecure



Source: authors' own elaboration, 2022.

Hence, before implementing the projection estimator, it was decided to recode the dependent variables into categorical ones. Initially, the dependent variables were grouped into three categories to apply an ordinal regression model. However, using ordinal regression models brought more complexity into the overall estimation process, without (in this specific case) a significant improvement of final estimates. In addition, the results obtained with this type of regression model are not of easy interpretation.

As the objective here is to develop a flexible method of disaggregation that is also easy to implement, it was decided to use logistic regression with binary dependent variables. Hence, the two probabilities have been recoded into binary categorical variables, taking value 1 for probabilities higher than or equal to 0.5, and 0 otherwise. It must be noted that the so built individual variables approximately lead to the prevalence of food insecurity indicators at the two levels of severity when aggregated up for the full sample; however, they do not replicate exactly the value of the indicators as they are based on a transformation of the original respondent-level food insecurity variables. Box 1 shows the R syntax used to recode the two probabilities in the GWP dataset. Despite presenting basic R commands, this and the following boxes provided in this report, are intended to facilitate the reproducibility of results, and allow implementing the projection estimators on different datasets with a limited amount of modifications.

Box 1. Recoding the variables of interest

```
library(openxlsx) # library to open excel files

### Importing GWP microdata: small survey
dfGWP<- data.frame(read.csv("FIES_data_Malawi1.csv", header= TRUE))

### Select data for 2016 (same year of large survey)
dfGWP<- dfGWP[dfGWP$year==2016,]


### Recode the probability  of being moderately or severely food insecure into a dummy
dfGWP$prob.ms<- 99999 #NAs
dfGWP$prob.ms[dfGWP$Prob_Mod_Sev<0.5]<- 0
dfGWP$prob.ms[dfGWP$Prob_Mod_Sev>=0.5]<- 1


### Recode the probability  of being severely food insecure into a dummy
dfGWP$prob.s<- 99999 #NAs
dfGWP$prob.s[dfGWP$Prob_sev<0.5]<- 0
dfGWP$prob.s[dfGWP$Prob_sev>=0.5]<- 1

### Delete records with missing values in the variable of interest
dfGWP<- subset(dfGWP, prob.ms<99999)
```

Source: authors' own elaboration, 2022.

Table 2. Cross-tabulation of the probability of being moderately or severely food insecure with the probability of being severely food insecure

	Probability of being moderately or severely food insecure		
Probability of being severely food insecure	**0**	**1**	**Total**
0	110	191	301
1	0	694	694
Total	110	885	995

Source: authors' own elaboration, 2022.

Out of the 1 000 respondents, five of them had a missing value for both probabilities, and were removed from the dataset.

4.5 Identifying and recoding potential auxiliary variables

As seen in Section 3, the implementation of the indirect estimation approach proposed by Kim and Rao (2012) requires having the same set of auxiliary variables in the two samples. In the small sample, this group of variables is used to fit a regression model for the estimation of a set of projection parameters. In turn, the auxiliary variables – along with the estimated projection parameters – allow predicting the synthetic values of the variable of interest in the big sample.

Concerning the two datasets used for this case study, the IHS4 provided access to a vast range of information collected with its four questionnaires. In contrast, the GWP only collects the FIES module at individual level along with a limited number of demographic, social and economic variables. Auxiliary variables that could be retrieved from both samples were: 1) sex; 2) age; 3) education; 4) employment status; 5) income quintile; 6) number of adults in the household; 7) number of children in the household; 8) marital; and 9) cellphone. Hence, the principal element in the selection of auxiliary variables for this case study was data availability.

One of the conditions to be satisfied by auxiliary variables before implementing the projection approach, is for these to share similar definitions and structure in the two samples. Hence, before implementing the indirect estimation approach, all the selected auxiliary variables have been recoded and harmonized across the two surveys. Boxes presented in this section provide examples of R syntax that could be used to perform variables recoding operations.

Recoding the sex of respondent

In both datasets, the variable sex has been recoded into the categorical dummy variable ***female*** taking value 1 for female respondents and 0 for male respondents.

Box 2. Recoding the variable sex in the two datasets (small and large sample)

```
### GWP dataset: recode the variable gender into the female dummy with:
### female= 1 , if the individual is a female
### female= 0 , if the individual is a male.

table(dfGWP$Gender) ## original variable
dfGWP$Gender<-as.factor(dfGWP$Gender)
dfGWP$female<-as.numeric(unclass(dfGWP$Gender)) # original categories --- 1: "Female"; 2: "Male"
dfGWP$female[dfGWP$female==2]<- 0               # new categories      --- 1: "Female"; 0: "Male"
dfGWP$female<-as.factor(dfGWP$female)

### IHS4 dataset: recode the variable sex into the same female dummy

dfLSMS<- read.csv("IHS4.csv", header= TRUE) ## Import data
table(dfLSMS$sex) ## original categories --- 1: "Male"; 2: "Female"

dfLSMS$female[dfLSMS$sex==2]<- 1 # 1: Female
dfLSMS$female[dfLSMS$sex==2]<- 0 # 0: Male
```

Source: authors' own elaboration, 2022.

Recoding the age of respondent

The GWP datasets contains individual answers to questions in the FIES module only for individuals who are 15 years old or older. Hence, the first step consisted in removing the observations corresponding to people below 15 years of age from the IHS4 dataset. Subsequently, the variable age has been recoded in both datasets according to the following age classes:

- agecat_1: 15-24 (youth)
- agecat_2: 25-49
- agecat_3: 50-64
- agecat_4: 65 and above.

Box 3. Recoding the variable age in the two datasets (small and large sample)

```
### GWP data: small sample

summary(dfGWP$Age) ## original variable
dfGWP$agecat[dfGWP$Age<25] <- 1                  ## Age 15-24
dfGWP$agecat[25<=dfGWP$Age & dfGWP$Age<=49] <- 2 ## Age 25-49
dfGWP$agecat[50<=dfGWP$Age & dfGWP$Age<=64] <- 3 ## Age 50-64
dfGWP$agecat[dfGWP$Age>=65] <- 4                 ## Age 65 and above
dfGWP$agecat<-as.factor(dfGWP$agecat)


### IHS4 data: large sample

summary(dfLSMS$age) ## original variable

dfLSMS$agecat <-00000 #0000 to drop this level in glm (NA)
dfLSMS$agecat[dfLSMS$age<25] <- 1                  ## Age 15-24
dfLSMS$agecat[25<=dfLSMS$age & dfLSMS$age<=49] <- 2 ## Age 25-49
dfLSMS$agecat[50<=dfLSMS$age & dfLSMS$age<=64] <- 3 ## Age 50-64
dfLSMS$agecat[dfLSMS$age >= 65] <- 4               ## Age 65 and above
```

Source: authors' own elaboration, 2022.

Recoding education information

Information on education were collected with a very different level of detail in the two surveys. More precisely, while the IHS4 provided very granular information on household members' education level, the GWP only distinguished between people with:

- educat_1: completed elementary education or less (up to 8 years of basic education);
- educat_2: completed secondary/three-year tertiary education and some education beyond secondary (9-15 years of education);
- educat_3: completed four years of education beyond high school and/or with a four-year college degree.

In addition, nine of the 1 000 respondents included in the GWP either refused to provide information on education or reported to not know that information. Given the limited number of cases and to simplify the recoding process, these two answer options have been included in educat_1 (which was the group with the highest frequency). For what concerns the education variable in the IHS4, Table 3 shows how initial categories have been recoded into educat_1, educat_2, and educat_3.

Table 3. Recoding of education categories of the Fourth Integrated Household Survey

Initial category	Recoded category
None	educat_1
PSLC: Primary School Leaving Certificate – Primary School Leaving Exam assesses academic achievement at the Primary School level (ages 13–14)	educat_1
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)	educat_2
MSCE: The Malawi School Certificate of Education exam, taken during the last year of secondary school (ages 17–18)	educat_2
Non-university diploma	educat_3
University diploma, degree	educat_3
Post-graduate degree	educat_3

Source: authors' own elaboration, 2022.

Box 4. Recoding the variable education in the two datasets (small and large sample)

```
### GWP data: small sample
table(dfGWP$Education)         ### Original variable

dfGWP$Education<-as.factor(dfGWP$Education)
dfGWP$educat<-as.numeric(unclass(dfGWP$Education))

dfGWP$educat[dfGWP$educat==1]<- 1  ### Don't know
dfGWP$educat[dfGWP$educat==2]<- 1  ### Refused
dfGWP$educat[dfGWP$educat==3]<- 1  ### 0-8
dfGWP$educat[dfGWP$educat==4]<- 3  ### above 15
dfGWP$educat[dfGWP$educat==5]<- 2  ### 9-15
dfGWP$educat<-as.factor(dfGWP$educat)

### IHS4 data: large sample
table(dfLSMS$edulevel)       ### Original variable

dfLSMS$educat[dfLSMS$edulevel==1 | dfLSMS$edulevel==2]<- 1  ### NONE or PSLC
dfLSMS$educat[dfLSMS$edulevel==3 | dfLSMS$edulevel==4]<- 2  ### JCE or MSCE
dfLSMS$educat[dfLSMS$edulevel>=5]<- 3                       ### ABOVE MSCE
```

Source: authors' own elaboration, 2022.

Recoding employment information

The variable Employment in the GWP dataset was originally coded as follows:

- employed full-time for an employer;
- full-time self-employed;
- employed part-time, wants full-time;
- employed part-time, does not want full-time;
- unemployed;
- out of workforce.

Variables on employment based on IHS4 microdata were extracted from the FAO Rural Livelihoods Information System (RuLIS) database. In this database, the variable "tot_employment" is coded as follows:

- 0 for not employed (inactive or unemployed);
- 1 for employed.

Hence, in order to harmonize information available in the two samples, the more detailed variable available in the GWP dataset has been recoded in the following dummy:

- empcat = 1: employed full-time for an employer; employed full-time for self; employed part-time, wants full-time; employed part-time, does not want full-time.
- empcat = 0: unemployed, out of workforce.

Box 5. Recoding the variable employment in the two datasets (small and large sample)

```
### GWP data: small sample

table(dfGWP$Employment)   ### Original variable
dfGWP$Employment<-as.factor(dfGWP$Employment)

dfGWP$empcat<-as.numeric(unclass(dfGWP$Employment))
dfGWP$empcat[dfGWP$empcat==2]<- 1
dfGWP$empcat[dfGWP$empcat==3]<- 1
dfGWP$empcat[dfGWP$empcat==4]<- 1
dfGWP$empcat[dfGWP$empcat==5]<- 0
dfGWP$empcat[dfGWP$empcat==6]<- 0
dfGWP$empcat<-as.factor(dfGWP$empcat)

### IHS4 data: large sample
dfLSMS$empcat[dfLSMS$tot_employment==0]<- 0 # Unemployed or inactive
dfLSMS$empcat[dfLSMS$tot_employment==1]<- 1 # Employed
```

Source: authors' own elaboration, 2022.

Recoding the geographic location

In both datasets, the dummy rural has been created, taking value 1 for individuals from towns and rural areas and 0 for individuals from urban areas and suburbs.

Box 6. Recoding the variable rural in the two datasets (small and large sample)

```
### GWP data: small sample
### Recode the variable "Area" into a dummy, such that:
### rural = 1, if Area = 1 (town/rural)
### rural = 0, if Area = 2 (urban/suburbs)

table(dfGWP$Area)                          ## Original variable
dfGWP$Area<-as.factor(dfGWP$Area)
dfGWP$rural<-as.numeric(unclass(dfGWP$Area))  ## rural = 1
dfGWP$rural[which(dfGWP$rural==2)]<- 0        ## rural = 0
dfGWP$rural<-as.factor(dfGWP$rural)


### IHS4 data: large sample
### Recode the variable "reside" into a dummy, such that:
### rural = 1, if reside = 2
### rural = 0, if reside = 1

table(dfLSMS$reside)                        ## Original variable
dfLSMS$rural[dfLSMS$reside==2]<- 1           ## rural = 1
dfLSMS$rural[dfLSMS$reside==1]<- 0           ## urban = 0
```

Source: authors' own elaboration, 2022.

Recoding the variable household size

The variable household size has been created in different ways in the two datasets. For what concerns the GWP dataset, the total number of individuals in the household has been computed by summing the number of adults with the number of children in the household. On the other hand, for the IHS4, the household size has been extracted from the RuLIS by counting the number of individuals reported in the household roster by household.

Recoding the variable income quintile

The GWP dataset already provides a variable indicating the income quintile to which each individual belongs. On the contrary, income quintiles needed to be computed in the IHS4 dataset. As income aggregates are already produced in the context of the RuLIS, for this exercise we referred to the variable "tot_income". Income at the individual level was estimated by dividing it by the number of adults in the household. As a result, RuLIS estimates for IHS4 are converted into categorical variables following the GWP definitions:

- 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%.

Box 7. Recoding the variable income quintile in the two datasets (small and large sample)

```
### GWP data: small sample

table(dfGWP$Income)                ### Original variable

###          1           2           3          4           5
###     Fourth 20%  Middle 20% Poorest 20% Richest 20%  Second 20%
###         208         183         172         250         187
dfGWP$Income<-as.factor(dfGWP$Income)

dfGWP$inccat<-as.numeric(unclass(dfGWP$Income))
dfGWP$inccat[dfGWP$inccat==3]<- 11 ## "Poorest 20%"-1
dfGWP$inccat[dfGWP$inccat==5]<- 12 ## "Second  20%"-2
dfGWP$inccat[dfGWP$inccat==2]<- 13 ## "Middle  20%"-3
dfGWP$inccat[dfGWP$inccat==1]<- 14 ## "Fourth  20%"-4
dfGWP$inccat[dfGWP$inccat==4]<- 15 ## "Richest 20%"-5
dfGWP$inccat <- dfGWP$inccat - 10
dfGWP$inccat<-as.factor(dfGWP$inccat)

### IHS4 data: large sample

dfLSMS$totincome <- ifelse(dfLSMS$totincome<0,0,dfLSMS$totincome)

dfLSMS <- dfLSMS %>%
          group_by(HHID) %>%
          mutate(adult_sizeHH=n()) %>%              ## Getting the number of adults in each household
          mutate(totincome1=totincome/adult_sizeHH) ## Dividing the income by the number of adults
dfLSMS <- as.data.frame(dfLSMS)

quintiles <- quantile(dfLSMS$totincome1, probs = seq(0, 1, 1/5), na.rm = TRUE)
quintiles
quintiles<-unname(quintiles) ## Get only the quantile values
dfLSMS$inccat <- 99999
dfLSMS$inccat[dfLSMS$totincome1<=quintiles[2]] <- 1                                          ## "Poorest 20%"-1
dfLSMS$inccat[quintiles[2]<dfLSMS$totincome1 & dfLSMS$totincome1<=quintiles[3]] <- 2 ## "Second  20%"-2
dfLSMS$inccat[quintiles[3]<dfLSMS$totincome1 & dfLSMS$totincome1<=quintiles[4]] <- 3 ## "Middle  20%"-3
dfLSMS$inccat[quintiles[4]<dfLSMS$totincome1 & dfLSMS$totincome1<=quintiles[5]] <- 4 ## "Fourth  20%"-4
dfLSMS$inccat[dfLSMS$totincome1>quintiles[5]] <- 5                                           ## "Richest 20%"-5
```

Source: authors' own elaboration, 2022.

Recoding information on the marital status

Both the GWP and the IHS4 collected information on the marital status of household members. For what concerns the GWP, the variable Marital was initially coded as follows:

- 1 - Single/Never been married;
- 2 - Married;
- 3 - Separated;
- 4 - Divorced;
- 5 - Widowed;
- 6 - Domestic partner.

Similarly, the variable marital included in the IHS4 dataset had the following categories:

- 1 - Monogamous married;
- 2- Polygamous married;
- 3- Separated;
- 4 - Divorced;
- 5 - Widow or widower;
- 6 - Never married.

In order to harmonize the information in the two datasets, variables categories have been recoded in three classes, namely 1) never married; 2) married; 3) other. The recoding R syntax is illustrated in Box 8. It should be noted that, despite a more granular categorization of the variable marital was possible, in both cases, the categories 3) separated and 4) divorced were reported by very few respondents. For this reason, the single category other, including widowed, divorced, and separated individuals was created.

Box 8. Recoding the variable marital in the two datasets (small and large sample)

```
### GWP data: small sample

table(dfGWP$marital)
# 1 Single/Never been married  -> 1 Not Married
# 2 Married                    -> 2 Married
# 3 Separated                  -> 3 Other
# 4 Divorced                   -> 3 Other
# 5 Widowed                    -> 3 Other
# 6 Domestic partner           -> 3 Other

dfGWP$marital <- as.numeric(unclass(dfGWP$marital))
dfGWP[dfGWP$marital==4,]$marital <- 3
dfGWP[dfGWP$marital==5,]$marital <- 3
dfGWP[dfGWP$marital==6,]$marital <- 3
dfGWP$marital <- as.factor(dfGWP$marital)

### IHS4 data: large sample

table(dfLSMS$marital)
# 1 MONOGAMOUS MARRIED OR NON-FORMAL UNION  -> 2 Married
# 2 POLYGAMOUS MARRIED OR NON-FORMAL UNION  -> 2 Married
# 3 Separated                               -> 3 Other
# 4 Divorced                                -> 3 Other
# 5 Widow or Widower                        -> 3 Other
# 6 Never Married                           -> 1 Not Married

dfLSMS$marital <- as.numeric(unclass(dfGWP$marital))
dfLSMS[dfGWP$marital==1,]$marital <- 2
dfLSMS[dfGWP$marital==4,]$marital <- 3
dfLSMS[dfGWP$marital==5,]$marital <- 3
dfLSMS[dfGWP$marital==6,]$marital <- 1
dfLSMS$marital <- as.factor(dfLSMS$marital)
```

Source: authors' own elaboration, 2022.

4.6 Selecting the auxiliary variables to be included in the model

One of the fundamental steps to ensure the quality of results obtained with the projection estimator is the identification of suitable auxiliary variables x_i. In similar contexts, especially when big surveys collecting a multitude of information are considered, the use of variable selection methods can be helpful, and issues such as multicollinearity should be carefully assessed.

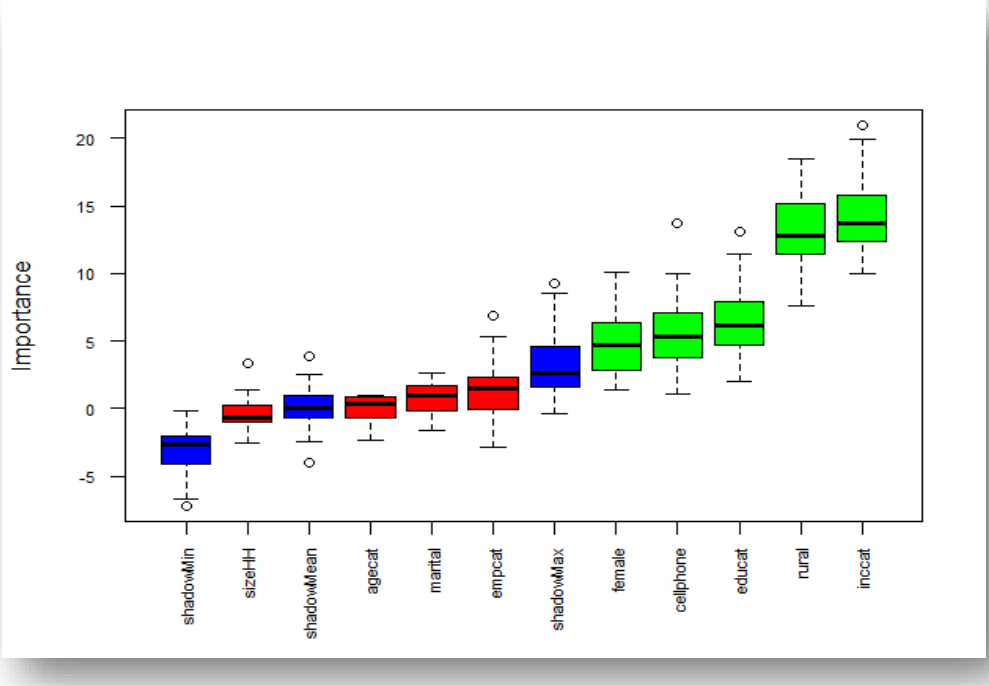
As mentioned in the Guidelines (FAO, 2021), the literature on variable selection approaches is very ample, with authors such Ryan (2008) and Harrel (2015) providing comprehensive summaries of the common methods used in this field.

In this study, despite the availability of a relatively small number of auxiliary variables common to the two datasets, we illustrate the use of the Boruta feature selection method, proposed in Kursa and Rudnicki (2010). For more details on the theory behind this selected approach, the authors refer to the Guidelines (FAO, 2021) besides, of course, the original paper presenting the approach (Kursa and Rudnicki, 2010).

For this application, all the auxiliary variables available in the GWP dataset were plugged into the Boruta algorithm to assess their relevance. The algorithm was implemented separately for the probability of being moderately or severely food insecure (Figure 3) and the probability of being severely food insecure (Figure 4).

In Figures 3 and 4, the boxplots of different colors represent various Boruta outputs: the red, yellow and green boxplots represent the scores of the rejected (unimportant), tentative and confirmed (important) variables respectively, while the color blue was assigned to shadow features. Tentative variables are those for which Boruta could not indicate a clear decision concerning their relevance, as their importance level was not significantly different from their best shadow features.

Figure 3. Level of importance of the auxiliary variables for moderate or severe food insecurity



Source: authors' own elaboration, 2022.

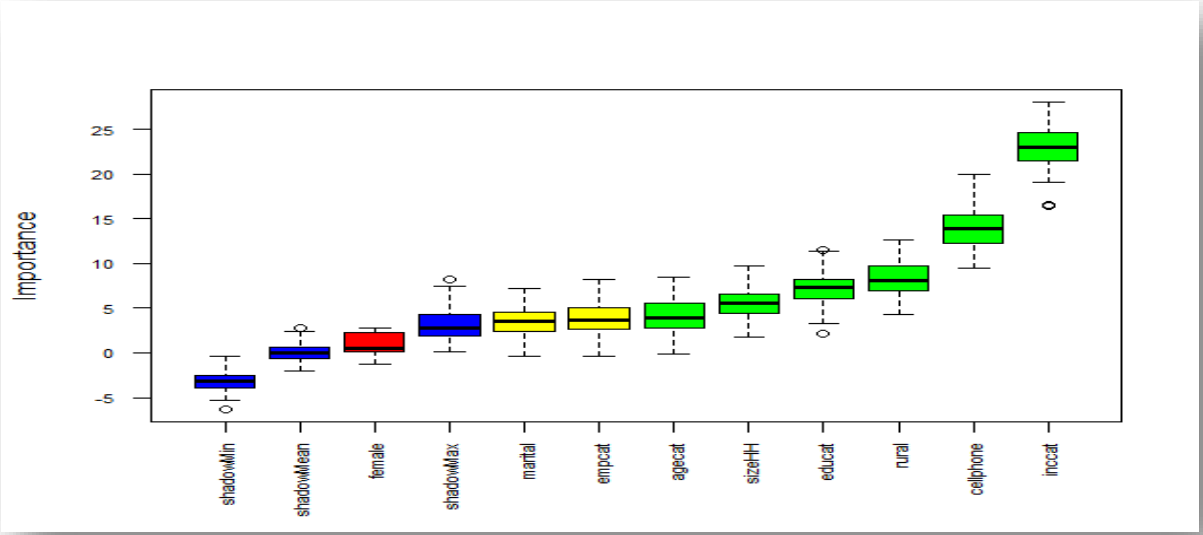
Box 9. Implementing Boruta with R for the probability of being moderately or severely food insecure

```
### Create a data frame with all possible auxiliary variables + the variable of interest
### to implement the Boruta algorithm

dfms<-dfGWP
keep_ms <- c("agecat","educat","rural","female","cellphone",
             "marital","empcat","inccat","sizeHH","prob.ms")
dfms <- dfms[ , (names(dfms) %in% keep_ms)]

### Load Boruta algorithm
library(Boruta)

### Apply the Boruta algorithm
boruta_output_prob.ms <- Boruta(prob.ms ~ ., data=dfms)

# formula= ~… formula describing model to be analysed.
# data= … data frame of predictors.

### print the output from Boruta containing the variables which have been
### identified as important, unimportant or tentative
print(boruta_output_prob.ms)

### The plot() option shows box plot of all the attributes listed in order of
### their importance
plot(boruta_output_prob.ms, cex.axis=.7, las=2, xlab="",
     main=sprintf("Variable Importance for 'prob.ms'"))
```

Source: authors' own elaboration, 2022.

From Figure 3 emerges that the most important auxiliary variables to explain the probability of being moderately or severely food insecure are the income quintile, the geographic (urban/rural) location, the education level, the ownership of a mobile phone, and the sex of respondents listed in order of importance. On the other hand, information on the employment status, the age, and the household size are identified as unimportant to explain the variability of the variable of interest. In this case, Boruta did not identify any tentative feature.

Figure 4. Level of importance of the auxiliary variables for severe food insecurity



Source: authors' own elaboration, 2022.

Similarly, the most important auxiliary variables for the probability of severe food insecurity identified by Boruta are the income quintile, the ownership of a mobile phone, the geographic location, the education level and the size of the household. The variables agecat and empcat are indicated as tentative attributes, while the variable female is classified as not relevant (Figure 4).

Box 10. Implementing Boruta with R for the probability of being severely food insecure

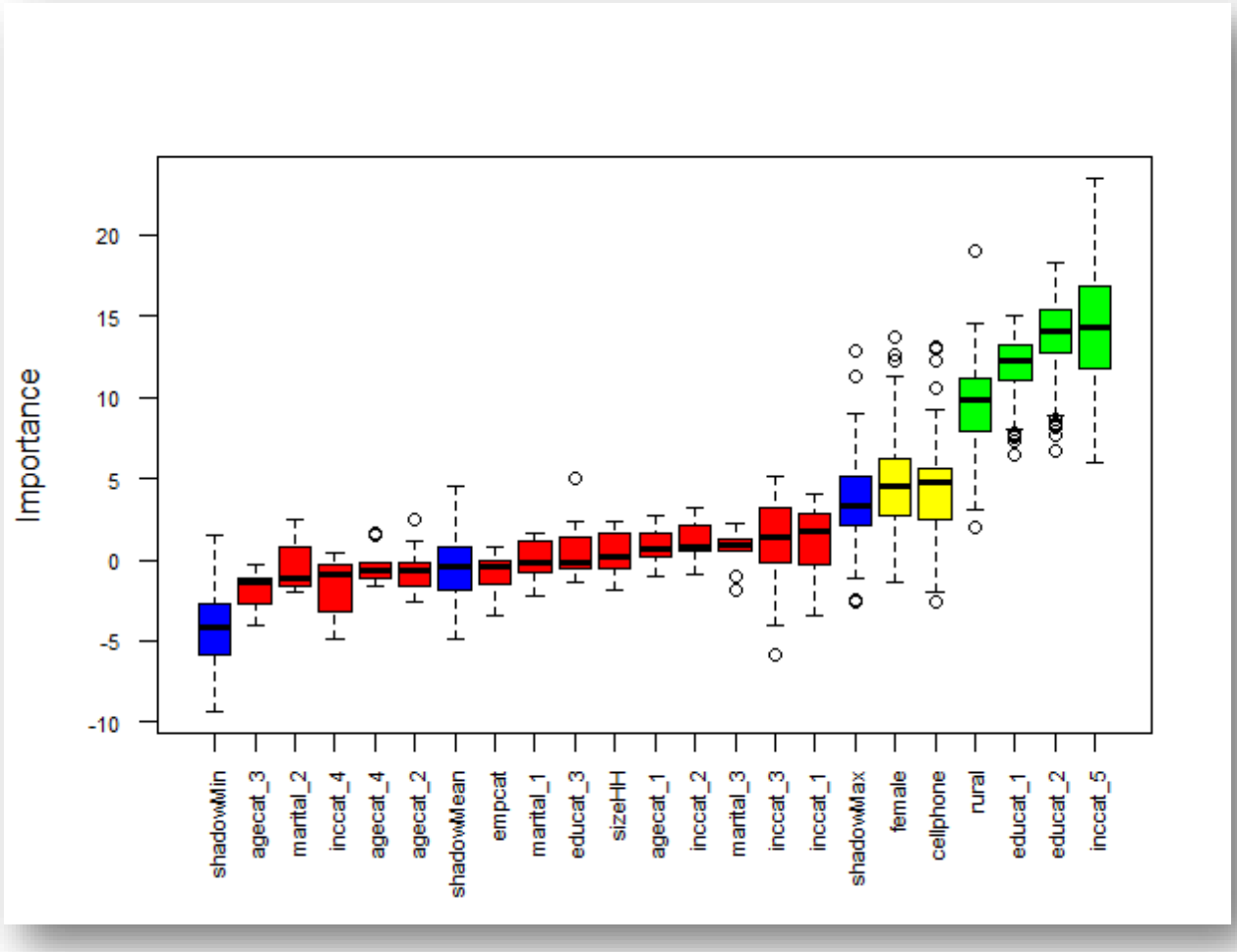
```
### Create a data frame with all possible auxiliary variables + the variable of interest
### to implement the Boruta algorithm

dfs<-dfGWP
keep_s <- c("agecat","educat","rural","female","cellphone",
            "marital","empcat","inccat","sizeHH","prob.s")
dfs <- dfs[ , (names(dfs) %in% keep_s)]

### Load Boruta library
library(Boruta)

### Apply the Boruta algorithm
boruta_output_prob.s <- Boruta(prob.s ~ ., data=dfs)

# - formula= ~… formula describing model to be analysed.
# - data= …     data frame of predictors.

### print the output from Boruta containing the variables which have been
### identified as important, unimportant or tentative
print(boruta_output_prob.s)

### The plot() option shows box plots of all the attributes listed in order of
### their importance
plot(boruta_output_prob.s, cex.axis=.7, las=2, xlab="",
     main=sprintf("Variable Importance for 'prob.s'"))
```

Source: authors' own elaboration, 2022.

In order to perform a more accurate analysis and identify levels of categorical variables with greater influence, the implementation of the Boruta algorithm was repeated on the various levels of available categorical auxiliary variables (see Section 4.5 for the definition of each level). To do so, a series of dummy variables, one for each category of all auxiliary variables, were created in both datasets.

Figures 5 and 6 below present the results of applying Boruta on these dummies for the two variables of interest. Concerning R, the syntax to be used is identical, except for the creation of dummy variables.

Figure 5. Importance of different levels of auxiliary variables for moderate or severe food insecurity

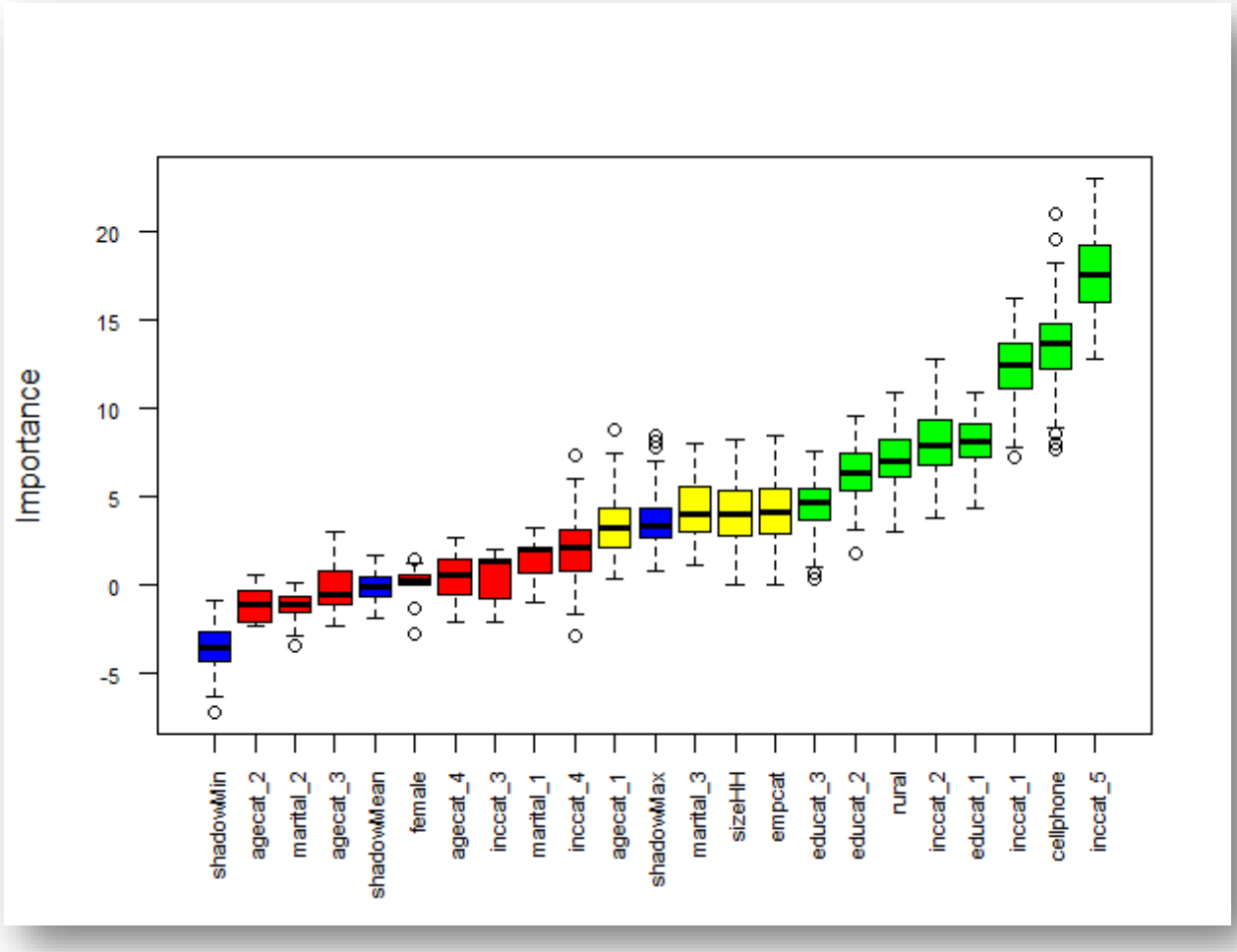


Source: authors' own elaboration, 2022.

Figure 5 allows assessing which are the important levels of auxiliary variables confirmed by Boruta. Levels identified as important to explain the variability of the probability of being moderately or severely food insecure are:

- the rural dummy;
- educat_2 and educat_1 dummies; and
- the inccat_5 dummy.

Figure 6. Importance of different levels of auxiliary variables for severe food insecurity



Source: authors' own elaboration, 2022.

Similarly, the levels of auxiliary variables identified as important by Boruta are:

- the inccat_5, inccat_1 and inccat_2 dummies;
- the educat_1, educat_2, and educat_3 dummies;
- the rural dummy; and
- the cellphone dummy.

In this case, Boruta identified as tentative the dummy variables agecat_1, marital_3, sizeHH, and empcat.

As illustrated in Section 4.8, all the levels of auxiliary variables identified as tentative or important by Boruta have been used to fit a logistic regression on the two variables of interest. In addition, all the relevant dimensions for data disaggregation (sex, age class, income, rural/urban location) have also been included in the regression model, in order to increase the sample unbiasedness of the projection domain estimator (see Section 3 for theoretical justification).

4.8 Estimating the projection parameters in the small sample

Two weighted multivariate logistic regressions have been implemented in the small sample to estimate the projection parameters $\hat{\beta}$ to be used to predict the values of the two variables of interest in the large survey.

Let us indicate with $\hat{p}_{ms,i}$ the probability of being moderately or severely food insecure for the $i-$th individual in the small sample, and with $\hat{p}_{s,i}$ the probability of being severely food insecure for the same individual. These probabilities were estimated using GWP data collected with the FIES individual module.

As seen in Section 4.4, $\hat{p}_{ms,i}$ and $\hat{p}_{s,i}$ were concentrated around few discrete values in the [0.1] interval. For this reason, they were recoded into two dummy variables $y_{ms,i}$, and $y_{s,i}$, where:

- $y_{ms,i} = 1$ if $\hat{p}_{ms,i} \geq 0.5$, and $y_{ms,i} = 0$ otherwise;
- $y_{s,i} = 1$ if $\hat{p}_{s,i} \geq 0.5$, and $y_{s,i} = 0$ otherwise.

Then, the $y_{l,i}$ values (with $l = ms$ or $l = s$) were modeled with a multivariate logistic function of the set of discrete categorical auxiliary variables $x_i' = (1, x_{i1}, x_{i2}, \dots, x_{ik})$:

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

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

It is worth noting that $P(y_{l,i} = 1|x_i)$ represents the odds of being food insecure. The natural log of the odds – also known as the logit – is as follows:

$$\ln\left[\frac{P(y_{l,i} = 1|x_i)}{1 - \ P(y_{l,i} = 1|x_i)}\right] = logit[P(y_{l,i} = 1|x_i)]$$

As a result, the logits are modeled with a multivariate linear regression:

$$logit\left[P\left(y_{l,i}=1|x_i\right)\right]=\beta_{l,0}+\beta_{l,1}x_{1,i}+\beta_{l,2}x_{2,i}+\cdots.+\beta_{l,k}x_{k,i}+\varepsilon_{l,i}$$

where the error term $\varepsilon_{l,i}$ takes only two values: $\varepsilon_{l,i}=-x_i'\beta_l$ when $y_{l,i}=0.$ and $\varepsilon_{l,i}=1-x_i'\beta_l$ when $y_{l,i}=1$. Therefore, we cannot assume a normal distribution for the error term.

For both cases ($y_{ms,i}, and\ y_{s,i}$), the multinomial regression model included both the levels of auxiliary variables identified as important by Boruta, and those representing important dimensions for data disaggregation (e.g. sex, age classes, income quintiles and geographic location of individuals). This was done in order to meet the condition of sample unbiasedness of the projection estimator in each disaggregation domain, as detailed in Section 3.

Table 4. Results of logistic regression for the probability of moderate and severe food insecurity

Variable	Coefficient	Std. Error	t value	Prob. (>|t|)	OR
(Intercept)	3.5569	0.8120	4.381	2.79e-05***	
Educational level	educat1 as reference class to assess the parameters				
educat2:	-0.3680	0.2671	-1.378	0.171232	0.69
educat3:	-1.3122	0.6770	-1.938	0.055248	0.27
Geographic location	urban location as reference class to assess the parameter				
rural1: Rural	0.2049	0.4740	0.432	0.666460	1.23
Gender	male as reference class to assess the parameter				
female1: Female	-0.1964	0.2826	-0.695	0.488757	0.82
Age class	agecat1 as reference class to assess the parameters				
agecat2: 25-49	0.5045	0.3087	1.634	0.105196	1.66
agecat3: 50-64	-0.2097	0.4497	-0.466	0.641883	0.81
agecat4: 65+	1.3230	0.7956	1.663	0.099274	3.75
Income group	incat1 as reference class to assess the parameters				
inccat2:	-0.2751	0.6464	-0.426	0.671294	0.76
inccat3:	0.2501	0.6375	0.392	0.695651	1.28
inccat4:	-1.2772	0.5756	-2.219	0.028633*	0.28
inccat5:	-1.9372	0.5238	-3.698	0.000346***	0.14
Cellphone	cellphone0 as reference class to assess the parameter				
cellphone1:	-0.7903	0.3473	-2.275	0.024891 *	0.45
Codes to assess results: 0 '*' - 0.001 '**' - 0.01 '*' - 0.05 '.' - 0.1 ' '**					

Source: authors' own elaboration, 2022.

Table 4 presents results of the logistic regression model for the probability of being moderately or severely food insecure. Coefficients can be better interpreted in terms of e^{β}, which is the odds ratio (OR), holding all other variables constant. In this context, negative coefficients correspond to odds ratios lower than one, while positive coefficients to odds ratios greater than one.

As it can be seen, individuals belonging to the second and third level of education (between 9 and 15 years of education or with more than 15 years of education) are less likely to be moderately or severely food-insecure as shown by the negative signs of their coefficients. In fact, the odds of being severely or moderately food insecure of people in educat3 is 0.27 (=exp(-1.3122)) times that of individuals belonging to educat1.

From the same table, it is possible to observe that the probability of being moderately or severely food insecure is expected to decrease when income increases. Indeed, people in the fourth- and fifth-income quintiles show odds smaller than one.

Finally, another aspect that can be noted is that some of the variables identified as important by Boruta are not flagged as significant by the regression model. This is mainly due to the fact that Boruta also identifies non-linear relationships between variables, which are instead not considered by the adopted regression approach. This specific issue could be addressed in future extensions of this study.

Box 11. Implementing weighted logistic regression with R

```
### Specification of the survey design
library(ReGenesees)
pdesign_GWP_2<- e.svydesign(data = dfGWP, ids = ~psu, strata= ~sampl.strata, weights = ~wt_rep)

### PROBABILITY OF BEING MODERATELY OR SEVERELY FOOD INSECURE
### Fitting a generalised linear model to data from a complex survey design
lmfit_w_1<- svyglm(formula=prob.ms ~ educat+rural+female+agecat+inccat+cellphone, design = pdesign_GWP_2, family = "binomial")

# - formula= ~… model formula with the form response ~ terms where response is the response vector and terms is
#              a series of terms which specifies a linear predictor for response.
# - design= … survey design from svydesign (or svrepdesign). Must contain all variables in the formula.
# - family= … a description of the error distribution and link function to be used in the model.


### PROBABILITY OF BEING SEVERELY FOOD INSECURE
### Fitting a generalised linear model to data from a complex survey design
lmfit_w_2 <- svyglm(prob.s ~ educat+rural+female+agecat+inccat+sizeHH+cellphone+marital+empcat, design = pdesign_GWP_2, family = "binomial")
```

Source: authors' own elaboration, 2022.

Table 5 below presents results of the logistic regression model for the probability of being severely food insecure. As seen for prob.ms, individuals with higher education level (between 9 and 15 years of education or with more than 15 years of education) are less likely to be severely food insecure (negative signs of their estimated regression parameters). In this case, the odds have similar values to those found with the first model.

The effect of income is even clearer. Indeed, individuals falling in the higher income quintiles (Incat4 and Incat5) have a lower chance of being severely food insecure compared to individuals belonging to the first income quintile.

Table 5. Results of logistic regression for the probability of severe food insecurity

Variable	Coefficient	Std. Error	t value	Prob.(>|t|)	OR
(Intercept)	2.538185	0.581346	4.366	3.04e-05***	
Educational level	Educat1 as reference class to assess the parameters				
educat2:	-0.291948	0.193782	-1.507	0.135009	0.75
educat3:	-1.319640	0.978926	-1.348	0.180629	0.27
Geographic location	Urban location as reference class to assess the parameter				
rural1: Rural	0.009173	0.324677	0.028	0.977515	1.01
Gender	Male as reference class to assess the parameter				
female1: Female	0.173854	0.182922	0.950	0.344143	1.19
Age class	Agecat1 as reference class to assess the parameter				
agecat2: 25-49	0.494922	0.237736	2.082	0.039863 *	1.64
agecat3: 50-64	0.453487	0.429486	1.056	0.293516	1.57
agecat4: 65+	0.926711	0.626119	1.480	0.141932	2.53
Income group	Incat1 as reference class to assess the parameter				
inccat2:	-0.610043	0.357247	-1.708	0.090749	0.54
inccat3:	-0.874457	0.372752	-2.346	0.020913 *	0.42
inccat4:	-1.320689	0.363015	-3.638	0.000433***	0.27
inccat5:	-1.951083	0.366965	-5.317	6.25e-07***	0.14
Household size					
Size of the household	-0.067980	0.042826	-1.587	0.115527	0.93
Employment status	Empcat0 as reference class to assess the parameter				

Variable	Coefficient	Std. Error	t value	Prob.(>|t|)	OR
Empcat1	-0.084853	0.185886	-0.456	0.649016	0.92
Marital status	Marital1 as reference class to assess the parameters				
Marital2	-0.268987	0.258382	-1.041	0.300318	0.76
Marital3	0.095752	0.416277	0.230	0.818537	1.10
Cellphone	Cellphone0 as reference class to assess the parameter				
Cellphone1	-0.606360	0.215278	-2.817	0.005827 **	0.55
Codes to assess results: 0 '*' - 0.001 '**' - 0.01 '*' - 0.05 '.' - 0.1 ' '**					

Source: authors' own elaboration, 2022.

After fitting a regression model, a relevant issue is that of verifying model assumptions and performance. Obviously, methods selected to make this assessment will depend on the type of model at hand – e.g. ordinary regression, generalized linear regression, etc. This section presents some of the methods considered for the empirical exercise presented in the technical report, with the purpose of illustrating all the technical steps to implement the projection estimator.

A common approach adopted in this or similar contexts is the Hosmer and Lemeshow's goodness of fit (GOF)[5], which 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). 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, such as the F-adjusted mean residual test, which can be performed in Stata. The F-corrected Wald test was applied to both logistic regression models (implemented in the GWP dataset) leading to:

- A p-value close to 0 for the probability of being moderately or severely food-insecure, indicating a poor performance of the model;
- A p-value of 0.414 for the probability of being severely food-insecure, indicating that there is no statistical evidence to say that the model results in a poor fit.

As illustrated in previous sections, Boruta identified very few relevant auxiliary variables for the probability of being moderately or severely food-insecure, compared to the probability of being severely food-insecure. This may be an explanation of why the model poorly explains the distribution of the dependent variable. Furthermore, another aspect that could negatively affect the accuracy of the model is that variables identified as not important by Boruta, but needed for data disaggregation, have also been used to fit the regression to ensure unbiasedness. In general terms, having so few auxiliary variables available in the small sample (GWP) is certainly a limitation for the identification of a good model.

[5] In addition to the goodness of fit, measures of the extent of variation explained by the model could be considered to assess its performance. In case of logistic regressions, measure of pseudo R-squares – especially McFadden (1974), Cox and Snell (1989) and Nagelkerke (1991) – are possible alternatives. The psrsq() function in R produces the Nagelkerke and Cox–Snell pseudo R-square estimates for survey sample data.

However, it should be stressed once again that this approach –being model assisted and not model based – is robust with respect to wrong model specifications.

4.9 Computing the synthetic values in the large sample

Having obtained the estimates $\hat{\beta} = (\hat{\beta}_{l,0}, \hat{\beta}_{l,1}, \hat{\beta}_{l,2}, \dots, \hat{\beta}_{l,k})$ of the parameters β, as illustrated in Section 4.8, the predicted probabilities are obtained as

$$P(\hat{y}_{l,i} = 1|x_i) = \frac{\exp(\hat{\beta}_{l,0} + \hat{\beta}_{l,1}x_{i1} + \hat{\beta}_{l,2}x_{i2} + \ \dots + \ \hat{\beta}_{l,k}x_{ik})}{1 + \exp(\hat{\beta}_{l,0} + \hat{\beta}_{l,1}x_{i1} + \hat{\beta}_{l,2}x_{i2} + \ \dots + \ \hat{\beta}_{l,k}x_{ik})} \quad (4.1)$$

The logit of the estimated probabilities $P(\hat{y}_{l,i} = 1|x_i)$, with $l = ms$ or $l = s$, can also be estimated with:

$$\ln\left[\frac{P(\hat{y}_{l,i} = 1|x_i)}{1 - P(\hat{y}_{l,i} = 1|x_i)}\right] = logit\left[P(\hat{y}_{l,i} = 1|x_i)\right] = \hat{\beta}_{l,0} + \hat{\beta}_{l,1}x_{i1} + \hat{\beta}_{l,2}x_{i2} + \ \dots + \ \hat{\beta}_{l,k}x_{ik},$$

Which becomes:

$$\begin{aligned}\ln\left[\frac{P(\hat{y}_{ms,i} = 1|x_i)}{1 - P(\hat{y}_{ms,i} = 1|x_i)}\right] &= logit\left[P(\hat{y}_{ms,i} = 1|x_i)\right] \\ &= 3.5569 - 0.3680 * \text{educat}_{2,\text{i}} - 1.3122 * \text{educat}_{3,\text{i}} + 0.2049 * \ \text{rural}_i - 0.1964 \\ &* \text{female}_i + 0.5045 * \text{agecat}_{2,\text{i}} - 0.2097 * \text{agecat}_{3,\text{i}} + 1.3230 * \text{agecat}_{4,\text{i}} - 0.2751 \\ &* \text{inccat}_{2,\text{i}} + 0.2501 * \text{inccat}_{3,\text{i}} - 1.2772 * \text{inccat}_{4,\text{i}} - 1.9372 * \text{inccat}_{5,\text{i}} \\ &- 0.7903 * \text{cellphone}_{0.\text{i}} \quad (4.2)\end{aligned}$$

for the probability of being moderately or severely food insecure, and

$$\begin{aligned}\ln\left[\frac{P(\hat{y}_{s,i} = 1|x_i)}{1 - P(\hat{y}_{s,i} = 1|x_i)}\right] &= logit\left[P(\hat{y}_{s,i} = 1|x_i)\right] = \\ &= 2.538185 - 0.291948 * \text{educat}_{2,\text{i}} - 1.319640 * \text{educat}_{3,\text{i}} + 0.009173 * \text{rural}_i \\ &+ 0.173854 * \text{female}_i + 0.494922 * \text{agecat}_{2,\text{i}} + 0.453487 * \text{agecat}_{3,\text{i}} + 0.926711 \\ &* \text{agecat}_{4,\text{i}} - 0.610043 * \text{inccat}_{2,\text{i}} - 0.874457 * \text{inccat}_{3,\text{i}} - 1.320689 * \text{inccat}_{4,\text{i}} \\ &- 1.951083 * \text{inccat}_{5,\text{i}} - 0.067980 * \text{sizeHH}_i - 0.084853 * \text{empcat}_{1,i} - 0.268987 \\ &* \text{marital}_{2,i} + 0.095752 * \text{marital}_{3,i} - 0.606360 * \text{cellphone}_{1,i} \quad (4.3)\end{aligned}$$

for the probability of being severely food insecure.[6]

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

$$\hat{Y}_{p,l} = \sum_{i \in A_1} w_{i1} P(\hat{y}_{l,i} = 1|x_i)$$

for the total in the target population, and

$$\hat{R}_{p,l} = \frac{\sum_{i \in A_1} w_{i1} P(\hat{y}_{l,i} = 1|x_i)}{\sum_{i \in A_1} w_{i1}}$$

for the proportion in the target population.

[6] For the definitions of dummies used as auxiliary variables, the reader should refer to Section 4.5 of this technical report.

Box 12. Projecting the synthetic values in the large sample (Fourth Integrated Household Survey) with R

```
### Computing synthetic values of the variables using parameters estimated with the
### logistic regression model

### PROBABILITY OF BEING MODERATELY OR SEVERELY FOOD INSECURE
dfLSMS$Yhat_ms1 <- predict(lmfit_w_1, newdata = dfLSMS, type="response")

# - object= …   the class inheriting from the model object for which prediction is desired.
# - newdata= …  input data to predict the values.
# - interval= … type of interval calculation. By putting "response", the probabilities are computed;
#               the default would compute the logits.

### PROBABILITY OF BEING SEVERELY FOOD INSECURE
dfLSMS$Yhat_s1 <- predict(lmfit_w_2, newdata = dfLSMS, type="response")
```

Source: authors' own elaboration, 2022.

4.10 Disaggregated estimates and the assessment of their accuracy

Being indicator 2.1.2 obtained as realization of ratio-type estimators, the R function used to estimate the target parameter and its accuracy measures was svystatR() included in the package ReGenesees (Istat, Regenessees). This function allows producing estimates, standard errors and confidence intervals for ratio-type estimators, taking into account all relevant disaggregation dimensions (e.g by sex, age, income quintile and urban/rural location).

Variance estimation was performed by adopting the approach presented in Section 3 (see formula 3.3). The two components of the variance were estimated separately and then added together. Boxes 13 and 14 show how the function svystatR() was used to obtain disaggregated estimates and a measure of their accuracy for the two probability of interest.

Despite the non-optimal performance of the model, the use of the Horvitz-Thompson estimator guarantees robustness against its mis-specification. In other words, the HT estimator will be efficient if the specified model achieves a good fit, but maintains desirable properties such as design unbiasedness and design consistency even if the model is false.

Box 13. Indirect estimation of the probability of being moderately or severely food insecure and its variance with R

```r
### Estimation of Ratio for prob.ms
library(ReGenesees)

# Denominator of the proportion
dfLSMS$ref_pop <- rep(1,dim(dfLSMS)[1])

# Specification of the survey design
pdesign_LSMS <- e.svydesign(id =~psu+HHID, data = dfLSMS, weights = ~weight_ind)

# Estimation
LSMS_ms_prop <- svystatR(design=pdesign_LSMS, ~Yhat_ms1, ~ref_pop, vartype = c("se", "cv", "var"),
                         conf.int = TRUE, conf.lev = 0.95)

# - design= …       analytical object specified with the function e.svydesign().
# - num= ~…         formula defining the numerator variable for the ratio to be estimated.
# - den= ~…         formula defining the denominator variable for the ratio to be estimated.
#                   This assumes value one for each unit in the dataset in case of proportions.
# - by= ~…          formula specifying the estimation domains. This parameter is set to NULL
#                   by default, producing estimates for the entire population.
# - vartype= c(…)   vector of characters specifying the selected variability estimators. It is possible
#                   to choose one or more of: standard error ("se"), coefficient of variation ("cv"),
#                   percent coefficient of variation ("cvpct"), or variance ("var").
# - conf.int= …     is a logical vector set as FALSE by default. If set equal to TRUE, confidence intervals
#                   for estimates are computed.
# - conf.lev= …     is a probability specifying the desired confidence level. Its default value is 0.95.

LSMS_ms_prop$Ratio ### Ratio estimate
LSMS_ms_prop$VAR   ### First component of the variance

### The projected estimates are computed with direct estimates in the small sample (GWP)
# Compute the residuals
dfGWP$res_ms <- residuals(lmfit_w_1,type = "response")

# Denominator of the proportion
dfGWP$ref_pop <- rep(1,dim(dfGWP)[1])
# Survey design
pdesign_GWP_2 <- e.svydesign(data = dfGWP, ids = ~psu, strata= ~sampl.strata, weights = ~wt_rep)
# Compute estimate
GWP_ms_prop <- svystatR(design=pdesign_GWP_2, ~res_ms, ~ref_pop, vartype = c("se", "cv", "var"),
                        conf.int = TRUE, conf.lev = 0.95)

GWP_ms_prop$VAR                                  ### Second component of the variance
var_ms_prop <- LSMS_ms_prop$VAR + GWP_ms_prop$VAR ### Sum of the two variance terms that gives the total variance
sqrt(var_ms_prop)/ LSMS_ms_prop$Ratio            ### Coefficient of variation

### Lower and Upper bound of 95% confidence interval
c(LSMS_ms_prop$Ratio - sqrt(var_ms_prop)*qnorm(0.975), LSMS_ms_prop$Ratio + sqrt(var_ms_prop)*qnorm(0.975))
```

Source: authors' own elaboration, 2022.

Box 14. Indirect estimation of the probability of being moderately or severely food insecure and its variance with R

```r
### Estimation of prob.s in the large sample (IHS4)
# Compute estimate
LSMS_s_prop <- svystatR(design=pdesign_LSMS, ~Yhat_s1, ~ref_pop, vartype = c("se", "cv", "var"),
                        conf.int = TRUE, conf.lev = 0.95)

LSMS_s_prop$Ratio ### Ratio estimate - proportion in this case
LSMS_s_prop$VAR   ### First component of the variance

#### GWP ####

# Compute the residuals
dfGWP$res_s <- residuals(lmfit_w_2,type = "response")

# Compute estimate
GWP_s_prop <- svystatR(design=pdesign_GWP_2, ~res_s, ~ref_pop, vartype = c("se", "cv", "var"),
                       conf.int = TRUE, conf.lev = 0.95)

GWP_s_prop$VAR                                  ### Second component of the variance
var_s_prop <- LSMS_s_prop$VAR + GWP_s_prop$VAR ### Sum of the two variance terms that gives the total variance
sqrt(var_s_prop)/ LSMS_s_prop$Ratio            ### Coefficient of variation

### Lower and Upper bound of 95% confidence interval
c(LSMS_s_prop$Ratio - sqrt(var_s_prop)*qnorm(0.975), LSMS_s_prop$Ratio + sqrt(var_s_prop)*qnorm(0.975))
```

Source: authors' own elaboration, 2022.

Tables 6 and 7 present the disaggregated estimates along with their accuracy measures.

Table 6. Projected versus direct estimates of the probability of being moderately or severely food insecure

Moderate or severe food insecurity					
		Prob.ms	CV (%)	Lower_CI	Upper_CI
Fourth Integrated Household Survey (IHS4)	Total	0.91	1.3	0.88	0.93
Gallup World Poll (GWP)		0.91	1.3	0.89	0.93
IHS4	Female	0.90	1.4	0.88	0.93
GWP		0.90	1.5	0.89	0.94
IHS4	Male	0.91	2.0	0.87	0.94
GWP		0.91	2.0	0.87	0.94
IHS4	Rural	0.93	1.2	0.91	0.95
GWP		0.92	1.3	0.90	0.94
IHS4	Urban	0.81	6.1	0.72	0.91
GWP		0.82	5.9	0.74	0.93
IHS4	15-24	0.90	2.0	0.86	0.93
GWP		0.89	2.1	0.85	0.93
IHS4	25-49	0.91	1.6	0.88	0.93
GWP		0.92	1.6	0.89	0.95
IHS4	50-64	0.87	3.7	0.81	0.93
GWP		0.90	3.5	0.84	0.96
IHS4	65+	0.98	1.6	0.94	1
GWP		0.98	1.7	0.95	1
IHS4	Inc_1	0.96	1.5	0.93	0.98
GWP		0.97	1.5	0.94	1
IHS4	Inc_2	0.96	1.5	0.93	0.99
GWP		0.96	1.6	0.93	0.99
IHS4	Inc_3	0.97	1.1	0.95	0.99
GWP		0.97	1.1	0.95	0.99
IHS4	Inc_4	0.89	3.6	0.82	0.95
GWP		0.88	3.7	0.82	0.94
IHS4	Inc_5	0.74	3.8	0.70	0.79
GWP		0.76	3.8	0.71	0.82

Source: authors' own elaboration, 2022.

Table 7. Projected versus direct estimates of the probability of being severely food insecure

Severe food insecurity					
		Prob.s	CV (%)	Lower_CI	Upper_CI
Fourth Integrated Household Survey (IHS4)	**Total**	0.72	2.4	0.69	0.76
GWP (Gallup World Poll)		0.71	2.8	0.67	0.75
IHS4	**Female**	0.75	2.7	0.71	0.79
GWP		0.75	3,1	0.71	0.80
IHS4	**Male**	0.69	3,6	0.65	0.74
GWP		0.67	4,2	0.61	0.73
IHS4	**Rural**	0.75	2.3	0.72	0.79
GWP		0.72	2.9	0.68	0.76
IHS4	**Urban**	0.61	9.9	0.49	0.73
GWP		0.63	9.2	0.52	0.75
IHS4	**15-24**	0.69	3.9	0.64	0.75
GWP		0.67	4.5	0.61	0.73
IHS4	**25-49**	0.71	3.1	0.67	0.76
GWP		0.72	3.6	0.67	0.77
IHS4	**50-64**	0.74	7.2	0.64	0.85
GWP		0.75	7.1	0.65	0.86
IHS4	**65+**	0.87	5.5	0.76	0.98
GWP		0.87	5.8	0.78	0.98
IHS4	**Inc_1**	0.87	3.3	0.81	0.93
GWP		0.88	3.4	0.83	0.94
IHS4	**Inc_2**	0.81	4.1	0.74	0.87
GWP		0.81	4.2	0.75	0.88
IHS4	**Inc_3**	0.77	5.3	0.69	0.85
GWP		0.75	5.3	0.67	0.83
IHS4	**Inc_4**	0.68	5.8	0.60	0.75
GWP		0.64	6.5	0.56	0.72
IHS4	**Inc_5**	0.47	8.4	0.40	0.55
GWP		0.48	8.4	0.40	0.56

Source: authors' own elaboration, 2022.

The comparison of projected and direct estimates in terms of their coefficient of variation (CV) shows that the former have greater or same accuracy than the latter in almost all cases. The only exception for both models is represented by the disaggregated estimate by geographic location (urban) and agecat_3 (50–

64 years of age), where the CV of the indirect estimate is slightly higher than the one of the direct estimate. This was probably due to the fact that variables urban and agecat_3 were not important auxiliary variables in the two implemented regression models.

Similar steps to those illustrated in this section were implemented and comparable results were achieved using microdata from two additional countries – Guatemala and South Africa – proving the robustness of the proposed approach. The results of these two additional experiments are presented in separated annexes.

5. Conclusions and way forward

The indirect estimation approach presented in this report covers a great deal of interesting and relevant empirical applications for the production of disaggregated data for SDG (and other) indicators. In particular, most countries can normally rely on auxiliary variables provided by large-scale surveys, censuses, administrative records, or geospatial information. In this context, some of the target phenomena for SDG monitoring and data disaggregation are often too costly or complex to be incorporated in large-scale data collection campaigns. The presented approach allows measuring the variable of interest with a small-scale survey, on the sample of which the parameters of a regression-type statistical model can be estimated, by linking this variable to a set of auxiliary variables. Based on these parameters, the values of the target variable can be predicted on a larger-scale data source collecting the auxiliary information used to fit the model. Relying on a larger sample allows increasing the accuracy of disaggregated estimates and consider disaggregation domains that are not available in the small survey. In addition, predicting a variable of interest on the sample of a more extensive survey from which most national official statistics are produced, allows improving estimates' consistency.

An additional aspect that should be highlighted is that the proposed strategy could be easily extended to other empirical contexts where, instead of integrating two independent surveys, the small survey could be integrated with auxiliary information coming from other types of data, such as censuses, administrative registers, and/or earth observation data. Furthermore, the extension of the projection estimator presented in section 3 of this report allows applying this approach to many other FAO-relevant SDG Indicators besides 2.1.2, such as:

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

The three case studies presented in Section 4 and the Annexes considering SDG Indicator 2.1.2, show how – by using this approach - it is possible to increase the accuracy of disaggregated estimates in various disaggregation domains. However, for future extensions of the study or practical implementations in countries, the adoption of different methodological solutions will be considered. For example, a beta regression model could be tested instead of the logistic one. In addition, instead of modelling the probability of being moderately or severely food insecure, the prediction could be based on models fitted directly on the eight dichotomous variables collected with t

Annexes

Annex A: Projection estimator on ....crodata from Guatemala

This annex presents results based on the use of the projection estimator to produce disaggregated estimates of the probability of being moderately or severely food insecure on microdata from Guatemala.

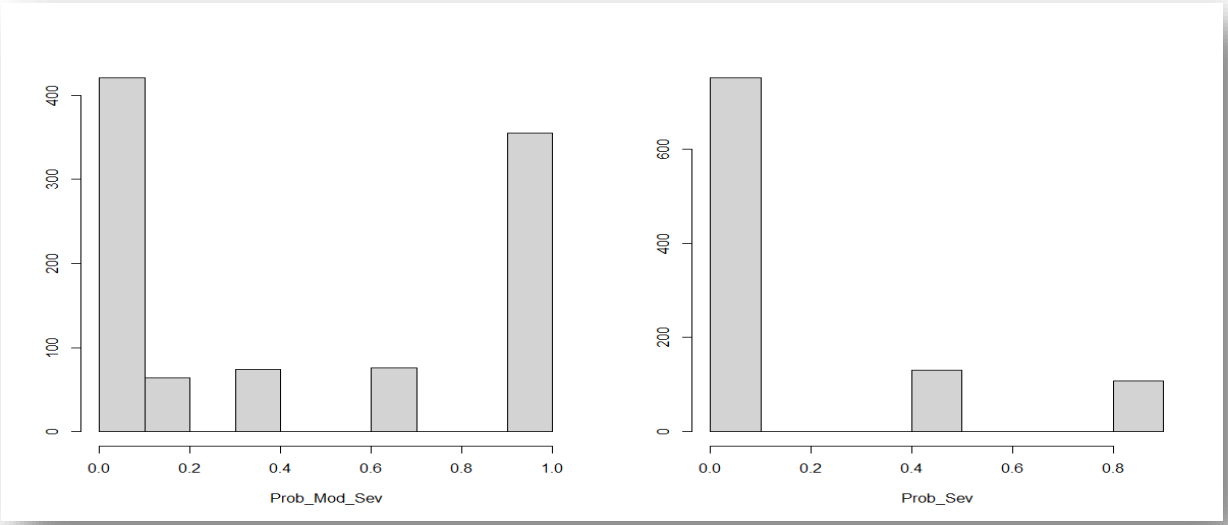
The two surveys used to implement the proposed estimation approach are:

- **Small sample: FIES individual module collected through the GWP.** The Guatemala GWP dataset provides FIES data for a sample of 1 000 individuals divided into 125 enumeration areas. This dataset contains the same variables described in Table 2, with the exception of the variable on the employment and the marital status of respondents, and their ownership of a mobile phone. Given the availability of a larger-scale survey in the same year, the GWP dataset from 2014 was considered for the case study.
- **Big sample: Encuesta Nacional de Condiciones de Vida (ENCOVI) 2014.** The ENCOVI, implemented by the National Statistical Institute of Guatemala (INE) in 2014, collected information on 11 536 households and 35 069 individuals in a total of 1 037 primary sampling units.

Step 1: Recoding the variable of interest

Also in this case, the two probabilities of interest (prob.ms and prob.s) have been recoded into binary categorical values taking value 1 for probabilities higher than 0.5 and 0 otherwise.

Figure 7. Histogram of the probability of being 1) moderately or severely food insecure and 2) severely food insecure (Guatemala)



Source: authors' own elaboration, 2022.

Table 8. Cross-tabulation of the probability of being moderately or severely food insecure with the probability of being severely food insecure (Guatemala)

Prob.ms			
Prob.s	**0**	**1**	**Total**
0	559	324	883
1	0	107	107
Total	559	431	990

Source: authors' own elaboration, 2022.

Out of 1 000 sampling observations, 10 reported a missing value for both probabilities and were removed from the dataset.

Step 2: Recoding auxiliary variables

This second case study considered the same set of auxiliary variables included in the Malawi's study (with the exception of marital and cellphone). These variables were recoded adopting the approach detailed in Section 4.5. For what concerns respondents' education level, information collected with the Encuesta Nacional de Condiciones de Vida (ENCOVI) were recoded as illustrated below in Table 9.

Table 9. Recoding of ENCOVI's education categories (Guatemala)

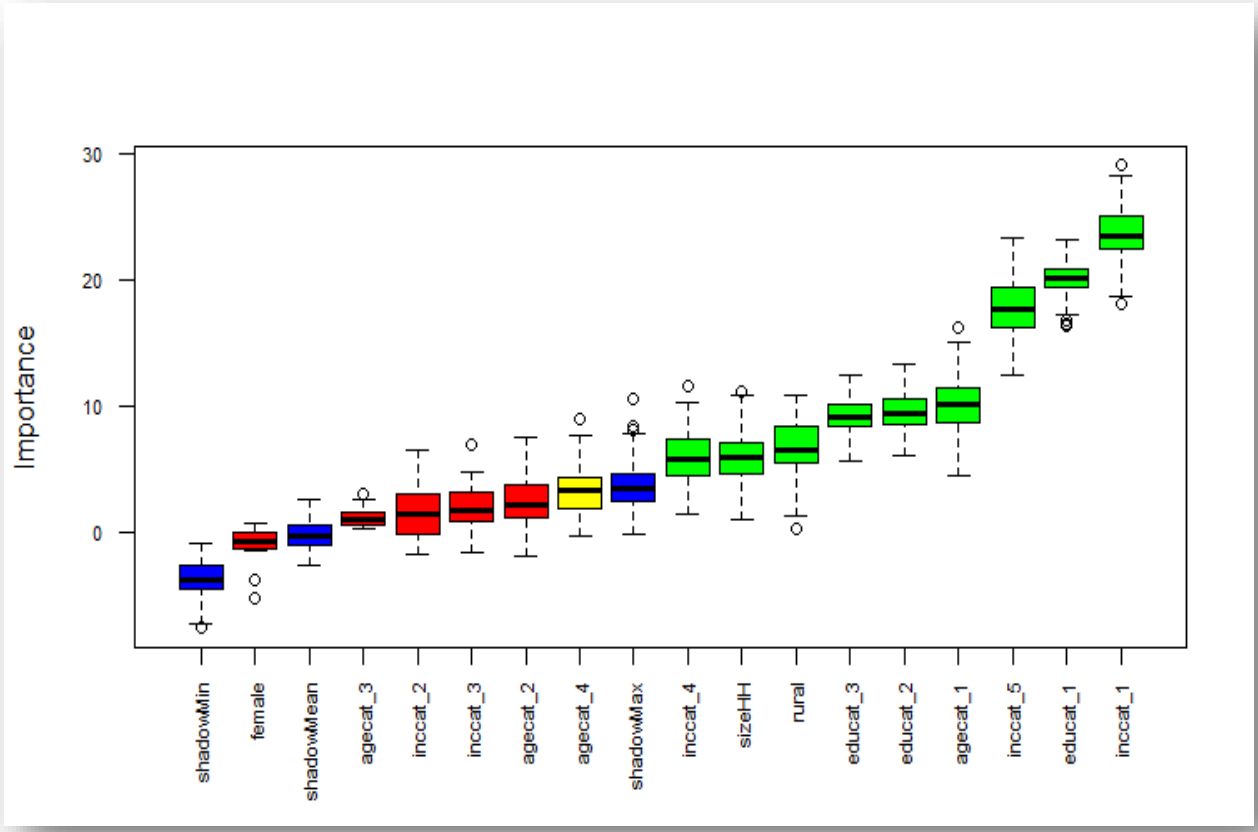
Initial category	Recoded category
Preprimary (ages 5-6)	educat_1
Primary (ages 7-12)	educat_1
Basic (ages 13-15)	educat_2
Diversified (ages 16-18)	educat_2
University	educat_3
Post-graduate degree	educat_3

Source: authors' own elaboration, 2022.

Step 3: Selecting auxiliary variables to be included in the model

All auxiliary variables available in the GWP dataset were plugged into the Boruta algorithm to assess their relevance. The algorithm was implemented separately for the probability of being moderately or severely food insecure (Figure 8) and the probability of being severely food insecure (Figure 9). For brevity, in this annex we only report results of Boruta applied on dummies representing all levels of auxiliary variables.

Figure 8. Importance of various levels of auxiliary variables for moderate or severe food insecurity (Guatemala)



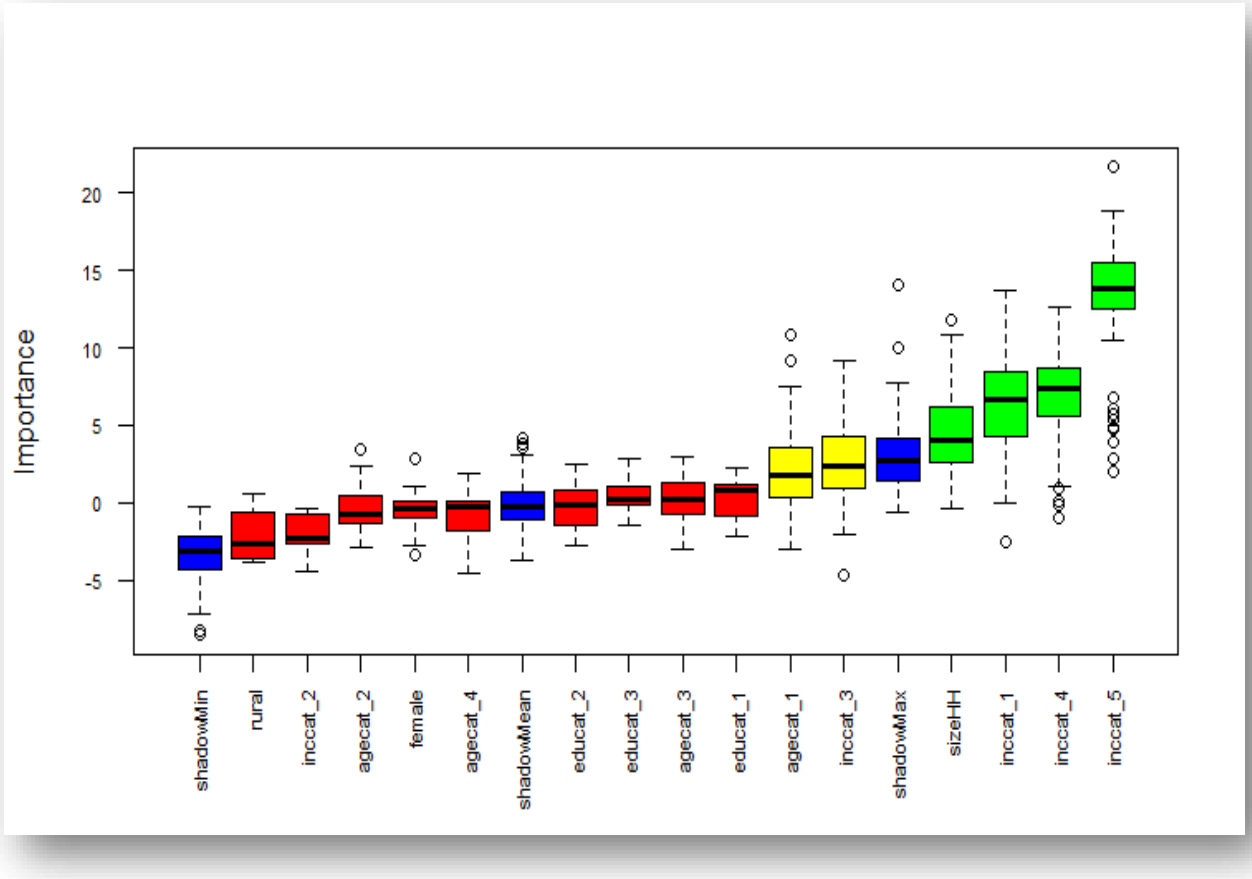
Source: authors' own elaboration, 2022.

Levels identified as important by Boruta:

- Inccat_1, Inccat_5 and Inccat_4 dummies.
- Educat_1, educat_2 and educat_3 dummies;
- Agecat_1 dummy;
- Rural dummy.

On the other hand, Boruta could not take a clear decision on agecat_4 and classified it as tentative.

Figure 9. Importance of various levels of auxiliary variables for severe food insecurity (Guatemala)



Source: authors' own elaboration, 2022.

The levels of auxiliary variables identified as important for the probability of being severely food insecure are:

- Inccat_5, Inccat_4, Inccat_2 dummies;
- sizeHH.

On the other hand, Boruta classified Inccat_3 and agecat_1 as tentative. Also in this case, the logistic regression models for the two probabilities of interest have been fitted using all levels of auxiliary variables identified as important or tentative by Boruta, along with all the relevant dimensions for data disaggregation.

Step 4: Estimating the projection parameters

The results of the two multinomial logistic regressions are reported respectively in Tables 10 and 11.

Table 10. Results of logistic regression for the probability of being moderately or severely food insecure (Guatemala)

Variable	Coefficient	Std Error	t value	Pr(>|t|)	OR
(Intercept)	1.10899	0.41486	2.673	0.008679 **	
Educational level	Educat1 as reference class to assess the parameters				
educat2:	- 0.69125	0.19135	- 3.613	0.000462 ***	0.50
educat3:	- 1.38440	0.35687	- 3.879	0.000181 ***	0.25
Geographic location	Urban location as reference class to assess the parameter				
rural1: Rural	0.38280	0.20477	1.869	0.064273	1.47
Gender	Male as reference class to assess the parameter				
female1: Female	- 0.02569	0.16269	- 0.158	0.874826	0.97
Age class	Agecat1 as reference class to assess the parameters				
agecat2: 25-49	0.61701	0.19843	3.109	0.002397 **	1.85
agecat3: 50-64	0.76291	0.25522	2.989	0.003464 **	2.14
agecat4: 65+	1.20902	0.37413	3.232	0.001632 **	3.35
Income group	Incat1 as reference class to assess the parameters				
inccat2:	- 1.13007	0.26567	- 4.254	4.49e-05 ***	0.32
inccat3:	- 1.56466	0.26539	-5.896	4.32e-08 ***	0.21
inccat4:	- 1.86398	0.28913	- 6.447	3.29e-09 ***	0.16
inccat5:	- 2.44643	0.31880	- -7.674	7.90e-12 ***	0.09
Household size					

Variable	Coefficient	Std Error	t value	Pr(>|t|)	OR
Size of the household	- 0.09233	0.03735	- 2.472	0.014996 *	0.91
Signif. codes: 0 '*' - 0.001 '**' - 0.01 '*' - 0.05 '.' - 0.1 ' ' 1**					

Source: authors' own elaboration, 2022.

Table 11. Results of logistic regression for being severely food insecure (Guatemala)

Variable	Coefficient	Std Error	t value	Pr(>|t|)	OR
(Intercept)	- 0.75576	0.61629	- 1.226	0.222754	
Educational level	Educat1 as reference class to assess the parameters				
educat2:	- 1.08924	0.33355	- 3.266	0.001464 **	0.34
educat3:	- 1.50909	0.74676	- 2.021	0.045769 *	0.22
Geographic location	Urban location as reference class to assess the parameter				
rural1: Rural	0.17745	0.38901	0.456	0.649195	1.19
Gender	Male as reference class to assess the parameter				
female1: Female	0.01624	0.27021	0.060	0.952177	1.02
Age class	Agecat1 as reference class to assess the parameter				
agecat2: 25-49	0.20606	0.31783	0.648	0.518144	1.23
agecat3: 50-64	0.74657	0.40076	1.863	0.065196	2.11
agecat4: 65+	-0.12151	0.52884	-0.230	0.818703	0.89
Income group	Incat1 as reference class to assess the parameter				
inccat2:	- 0.80250	0.32453	- 2.473	0.014965 *	0.45
inccat3:	- 1.25605	0.33500	- 3.749	0.000287 ***	0.28
inccat4:	- 1.77843	0.41233	- 4.313	3.57e-05 ***	0.17
inccat5:	- 2.80772	0.63392	- 4.429	2.28e-05 ***	0.06
Household size					
Size of the household	- 0.07242	0.05879	- 1.232	0.220631	0.93
Signif. codes: 0 '*' - 0.001 '**' - 0.01 '*' - 0.05 '.' - 0.1 ' ' 1**					

Source: authors' own elaboration, 2022.

The goodness of fit, the F-adjusted mean residual test, was applied to both logistic regression models (implemented in the GWP dataset) leading to a p-value of 0.991 for the probability of being moderately or severely food-insecure, and 0 for the probability of being severely food-insecure. The low p-value obtained for the second model is justified by the fact that very few significant auxiliary variables were available.

Step 5: Producing disaggregated estimates and assessing their accuracy

Table 12 presents a comparison of direct estimates – obtained with the GWP dataset – and indirect estimates – obtained by using the projection estimator on the ENCOVI – for the prevalence of moderate or severe food insecurity in the population. Measures of accuracy for the two probabilities are provided, for different disaggregation levels.

As it can be seen from the table, indirect estimates are systematically more accurate than direct ones. In addition, in most cases, projected estimates are close to direct ones. The greatest difference between direct and indirect estimates can be observed for the urban/rural disaggregation.

Table 12. Projected versus direct estimates of the prevalence of moderate or severe food insecurity (Guatemala)

Moderate or severe food insecurity					
		Prob.ms	**CV (%)**	**Lower_CI**	**Upper_CI**
Encuesta Nacional de Condiciones de Vida (ENCOVI)	**Total**	0.40	3.9	0.37	0.43
GWP		0.39	4.9	0.35	0.42
ENCOVI	**Female**	0.40	5.4	0.36	0.45
GWP		0.41	5.8	0.37	0.46
ENCOVI	**Male**	0.39	5.4	0.35	0.44
GWP		0.36	7.5	0.31	0.41
ENCOVI	**Rural**	0.50	3.6	0.47	0.54
GWP		0.43	5.0	0.39	0.47
ENCOVI	**Urban**	0.31	8.9	0.25	0.36
GWP		0.24	10.5	0.20	0.30
ENCOVI	**15-24**	0.30	8.5	0.25	0.35
GWP		0.27	10.6	0.22	0.33
ENCOVI	**25-49**	0.40	5.5	0.36	0.45
GWP		0.41	6.9	0.35	0.46
ENCOVI	**50-64**	0.49	9.2	0.40	0.58
GWP		0.51	9.6	0.41	0.61
ENCOVI	**65+**	0.64	9.1	0.53	0.76
GWP		0.64	10.1	0.51	0.77

Moderate or severe food insecurity					
		Prob.ms	**CV (%)**	**Lower_CI**	**Upper_CI**
ENCOVI	**Inc_1**	0.69	4.9	0.66	0.80
GWP		0.65	7.1	0.56	0.74
ENCOVI	**Inc_2**	0.48	7.4	0.41	0.55
GWP		0.44	8.1	0.37	0.51
ENCOVI	**Inc_3**	0.36	9.8	0.29	0.43
GWP		0.38	9.9	0.31	0.45
ENCOVI	**Inc_4**	0.30	9.7	0.24	0.35
GWP		0.26	12.5	0.19	0.32
ENCOVI	**Inc_5**	0.18	12.5	0.14	0.22
GWP		0.20	14.1	0.15	0.26

Source: authors' own elaboration, 2022.

Similarly, Table 13 presents a comparison of direct and indirect estimates – along with their accuracy measures - for the prevalence of severe food insecurity in the population.

Table 13. Projected versus direct estimates of the prevalence of severe food insecurity (Guatemala)

Severe food insecurity					
		Prob.s	**CV (%)**	**Lower_CI**	**Upper_CI**
Encuesta Nacional de Condiciones de Vida (ENCOVI)	**Total**	0.11	9.2	0.09	0.12
GWP		0.10	10.9	0.08	0.12
ENCOVI	**Female**	0.11	12.7	0.08	0.14
GWP		0.11	13.3	0.08	0.14
ENCOVI	**Male**	0.10	13.3	0.08	0.13
GWP		0.08	18.5	0.05	0.11
ENCOVI	**Rural**	0.14	8.2	0.12	0.17
GWP		0.11	11.3	0.09	0.13
ENCOVI	**Urban**	0.07	21.1	0.04	0.10
GWP		0.05	31.3	0.02	0.08
ENCOVI	**15-24**	0.08	18.5	0.05	0.11
GWP		0.06	24.1	0.03	0.09
ENCOVI	**25-49**	0.10	13.0	0.07	0.12
GWP		0.10	15.0	0.07	0.12
ENCOVI	**50-64**	0.18	21.3	0.11	0.26
GWP		0.18	21.2	0.11	0.26

Severe food insecurity					
		Prob.s	**CV (%)**	**Lower_CI**	**Upper_CI**
ENCOVI	**65+**	0.12	33.5	0.04	0.20
GWP		0.11	35.1	0.04	0.20
ENCOVI	**Inc_1**	0.26	12.7	0.19	0.32
GWP		0.23	14.1	0.17	0.30
ENCOVI	**Inc_2**	0.13	16.6	0.09	0.18
GWP		0.11	20.0	0.07	0.16
ENCOVI	**Inc_3**	0.09	22.7	0.05	0.13
GWP		0.08	24.0	0.04	0.12
ENCOVI	**Inc_4**	0.05	23.9	0.03	0.08
GWP		0.04	33.5	0.01	0.06
ENCOVI	**Inc_5**	0.02	37.3	0.004	0.03
GWP		0.012	52.2	0.000	0.02

Source: authors' own elaboration, 2022.

One aspect that emerge from Tables 12 and 13 is that, while for the prevalence of moderate or severe food insecurity the CV of disaggregated estimates is always below 15 percent, the same cannot be said for the prevalence of severe food insecurity. The high variance of disaggregated estimates presented in Table 13 can be explained by the fact that, in the small sample, the variable of interest is very unbalanced (883 values equal to zero, 107 values equal to 1).

Annex B: Projection estimator on microdata from South Africa

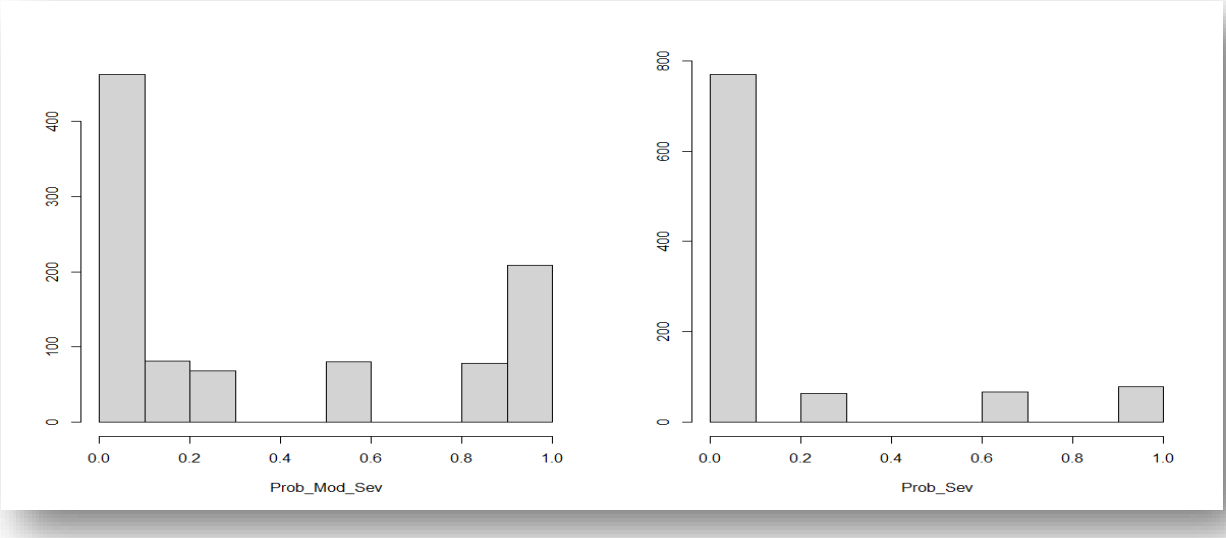
This annex presents results from the application of the projection estimator to produce disaggregated estimates of the probability of being moderately or severely food insecure on microdata from South Africa. The two surveys used to implement the proposed approach are:

- **Small Sample: FIES individual module collected through the GWP.** The South Africa 2015 GWP dataset provides FIES data for a sample of 1 000 individuals divided into 125 enumeration areas. This dataset contains the same variables described in Table 2, with the exception of variables on the employment and marital status of respondents, and the ownership of a mobile phone.
- **Big Sample: National income dynamics study (NIDS) 2015.** The National income dynamics study is a panel survey implemented by the South Africa Labour and development research unit at the University of Cape Town. The survey was designed to provide information on poverty and well-being, at individual and household level, as well as wealth in terms of income and expenditures, demographic dynamics, education and employment. NIDS data are representative at the national level. The survey was first implemented in 2008 with a sample of over 28 000 individuals in 7 300 households across the country. The survey is repeated every two years with these same household members, who are called Continuing sample members (CSMs). The survey is designed to follow people who are CSMs, wherever they may be in South Africa at the time of interview. The NIDS data is therefore, by design, not representative provincially or at a lower level of territorial disaggregation.

Step 1: Recoding the variable of interest

The two probabilities (prob.ms and prob.s) have been recoded into binary categorical values taking value 1 for probabilities higher or equal to 0.5 and 0 otherwise.

Figure 10. Histogram of the probability of being 1) moderately or severely food insecure and 2) severely food insecure (South Africa)



Source: authors' own elaboration, 2022.

Table 14. Cross-tabulation of the probability of being moderately or severely food insecure with the probability of being severely food insecure (South Africa)

	Prob.ms		
Prob.s	**0**	**1**	**Total**
0	612	221	833
1	0	146	146
Total	612	367	979

Source: authors' own elaboration, 2022.

Out of 1 000 sampling observations, 21 reported a missing value for both probabilities and were removed from the dataset.

Step 2: Recoding auxiliary variables

The same set of auxiliary variables used for Malawi was considered for the South African case study. These variables were recoded adopting the approach detailed in Section 4.5. For what concerns respondents' education level, information collected with the NIDS were recoded as illustrated in Table 15.

Table 15. Recoding of National income dynamics study education categories (South Africa)

Initial category	Recoded category
Other[7]	educat_1
Preprimary (grade 0. ages 5-6)	educat_1
Primary (grades 1 to 9, ages 7-15)	educat_1
Secondary (grades 10 to 12, ages 16)	educat_2
Post-secondary (national certificates)	educat_2
University	educat_3
Post-graduate degree	educat_3

Source: authors' own elaboration, 2022.

[7] The initial category "Other" was mainly selected by individuals with disabilities attending special trainings.

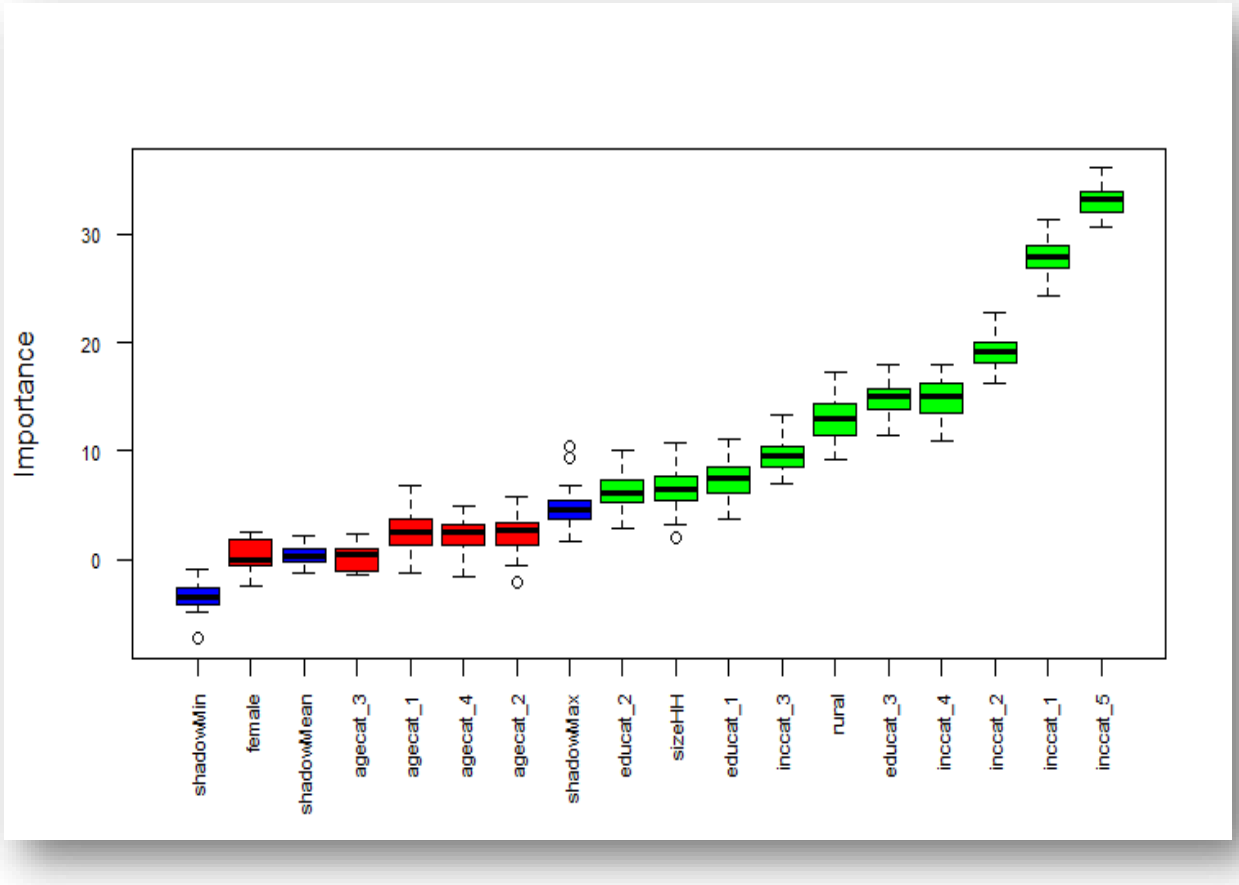
Step 3: Selecting the auxiliary variables for the model

All the auxiliary variables available in the GWP dataset were plugged into the Boruta algorithm to assess their relevance. The algorithm was implemented separately for the probability of being moderately or severely food insecure (Figure 11) and the probability of being severely food insecure (Figure 12).

From Figure 11 it can be seen that levels identified as important by Boruta were:

- Inccat_5, Inccat_1, inccat_2 dummies;
- All levels of the education variable;
- Rural dummy.

Figure 11. Importance of various levels of auxiliary variables for moderate or severe food insecurity (South Africa)

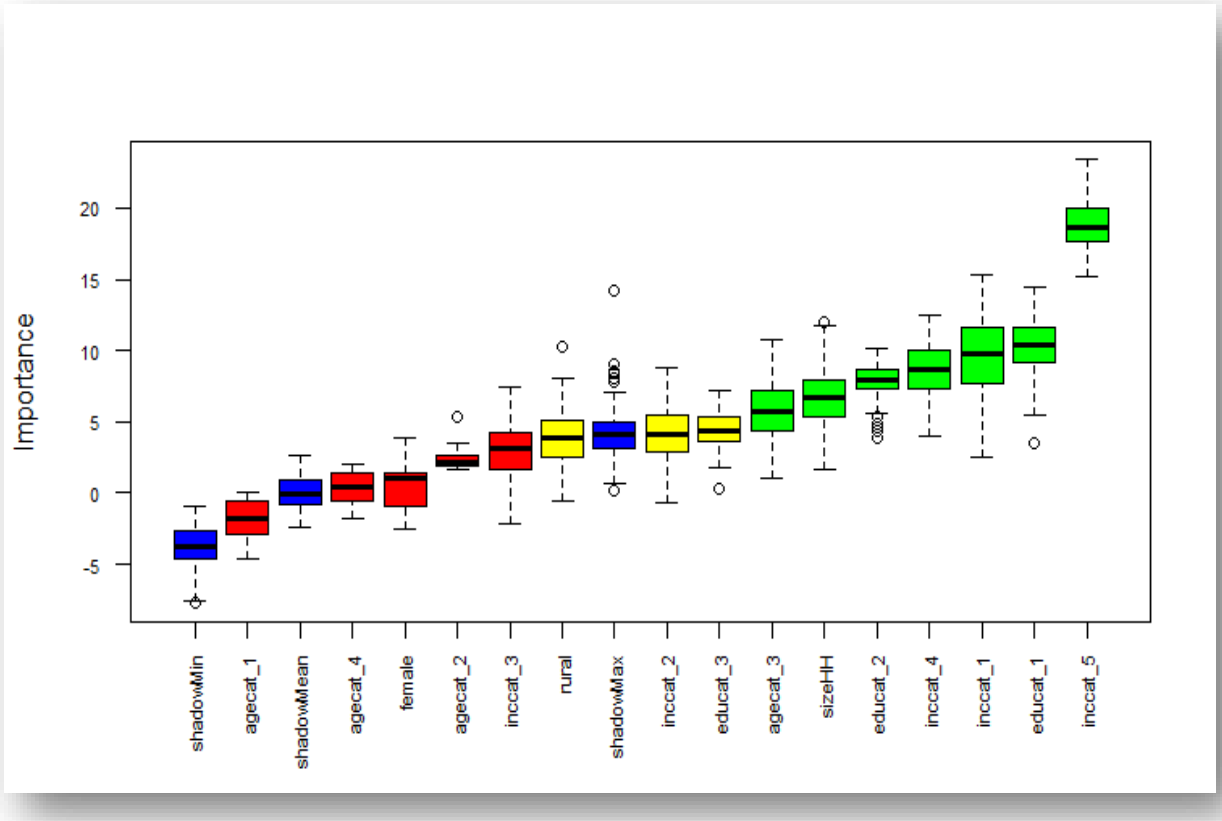


Source: authors' own elaboration, 2022.

Similarly, Figure 12 shows the important levels of auxiliary variables according to Boruta:

- Incat_5, Incat_1 and Incat_4 dummies.
- Educat_1 and educat_2 dummies.
- Agecat_3 dummy.

Figure 12. Importance of the levels of auxiliary variables for severe food insecurity



Source: authors' own elaboration, 2022.

Step 4: Estimating the projection parameters

Two weighted multinomial logistic regressions were implemented to estimate the projection parameters to predict the values of the two variables of interest in the large survey (Tables 16 and 17).

Table 16. Results of logistic regression for the probability of being moderately or severely food insecure (South Africa)

Variable	Coefficient	Std Error	t value	Pr(>|t|)	OR
(Intercept)	1.35185	0.55361	2.442	0.016735 *	
Educational level	Educat1 as reference class to assess the parameters				
educat2:	-0.86001	0.31209	-2.756	0.007198 **	0.42
educat3:	-1.89325	0.46594	-4.063	0.000109 ***	0.15
Geographic location	Urban location as reference class to assess the parameter				
rural1: Rural	0.66948	0.29162	2.296	0.024213*	1.95
Gender	Male as reference class to assess the parameter				
female1: Female	0.15742	0.19522	0.806	0.422344	1.17
Age class	Agecat1 as reference class to assess the parameters				
agecat2: 25-49	0.16396	0.22551	0.727	0.469239	1.18
agecat3: 50-64	-0.83135	0.36581	-2.273	0.025635 *	0.44
agecat4: 65+	-0.69149	0.38624	-1.790	0.077048	0.50
Income group	Incat1 as reference class to assess the parameters				
inccat2:	-0.58079	0.28439	-2.042	0.044305 *	0.56
inccat3:	-1.43706	0.30707	-4.680	1.10e-05 ***	0.24
inccat4:	-1.95097	0.35048	-5.567	3.12e-07***	0.14
inccat5:	-2.82958	0.40805	-6.934	8.20e-10***	0.06
Household size					
Size of the household	-0.01978	0.03069	-0.645	0.520981	0.98
Signif. codes: 0 '*' - 0.001 '**' - 0.01 '*' - 0.05 '.' - 0.1 ' ' 1**					

Source: authors' own elaboration, 2022.

Table 17. Results of logistic regression for the probability of severe food insecurity (South Africa)

Variable	Coefficient	Std Error	t value	Pr(>|t|)	OR
(Intercept)	0.15660	0.69092	0.227	0.821245	
Educational level	Educat1 as reference class to assess the parameters				
educat2:	-1.63771	0.33044	-4.956	3.75e-06***	0.19
educat3:	-1.76301	0.51144	-3.447	0.000892 ***	0.17
Geographic location	Urban location as reference class to assess the parameter				
rural1: Rural	1.06995	0.44904	2.383	0.019470*	2.92
Gender	Male as reference class to assess the parameter				
female1: Female	0.09558	0.24236	0.394	0.694308	1.10
Age class	Agecat1 as reference class to assess the parameter				
agecat2: 25-49	0.09585	0.31449	0.305	0.761303	1.10
agecat3: 50-64	-1.50896	0.47225	-3.195	0.001976 **	0.22
agecat4: 65+	-0.45573	0.43786	-1.041	0.300987	0.63
Income group	Incat1 as reference class to assess the parameter				
inccat2:	-0.76808	0.34745	-2.211	0.029816 *	0.46
inccat3:	-1.12622	0.33010	-3.412	0.001000***	0.32
inccat4:	-1.61427	0.42053	-3.839	0.000241***	0.20
inccat5:	-3.68159	0.80030	-4.600	1.50e-05***	0.03
Household size					
Size of the household	-0.03642	0.03526	-1.033	0.304599	0.96
Signif. codes: 0 '*' - 0.001 '**' - 0.01 '*' - 0.05 '.' - 0.1 ' ' 1**					

Source: authors' own elaboration, 2022.

The goodness of fit, the F-adjusted mean residual test, was applied to both logistic regression models leading to a p-value of 0.093 for the probability of being moderately or severely food-insecure, and 0 for the probability of being severely food-insecure.

Step 5: Producing disaggregated estimates and assessing their accuracy

After using the parameters estimated in Section B.5 to calculate the synthetic values of the two variables of interest in the large LSMS sample, estimates are produced alongside their coefficient of variation and confidence intervals considering all relevant disaggregation dimensions (e.g by sex, age_class, income quintile and urban/rural location).

Tables 18 and 19 show that most CVs of indirect estimates are lower than those of direct estimates.

Table 18. Projected versus direct estimates of the prevalence of moderate or severe food insecurity (South Africa)

Moderate or severe food insecurity					
		Prob.ms	CV (%)	Lower_CI	Upper_CI
National income dynamics study (NIDS)	Total	0.41	5.3	0.36	0.45
Gallup World Poll (GWP)		0.43	6.8	0.38	0.49
NIDS	Female	0.43	6.5	0.37	0.48
GWP		0.47	7.9	0.40	0.54
NIDS	Male	0.38	7.2	0.33	0.44
GWP		0.40	8.9	0.33	0.46
NIDS	Rural	0.54	4.9	0.49	0.60
GWP		0.52	6.6	0.45	0.59
NIDS	Urban	0.32	11.2	0.25	0.39
GWP		0.22	18.5	0.14	0.31
NIDS	15-24	0.47	7.3	0.41	0.54
GWP		0.41	9.6	0.34	0.49
NIDS	25-49	0.40	7.0	0.35	0.46
GWP		0.46	8.1	0.39	0.53
NIDS	50-64	0.32	16.1	0.22	0.42
GWP		0.38	15.1	0.27	0.49
NIDS	65+	0.38	19.4	0.23	0.52
GWP		0.44	20.0	0.27	0.62
NIDS	Inc_1	0.72	5.3	0.64	0.79
GWP		0.75	5.8	0.67	0.84
NIDS	Inc_2	0.63	6.3	0.55	0.71
GWP		0.64	6.5	0.56	0.72
NIDS	Inc_3	0.41	11.6	0.32	0.50
GWP		0.39	12.5	0.29	0.48
NIDS	Inc_4	0.28	15.2	0.20	0.37
GWP		0.28	15.8	0.19	0.37
NIDS	Inc_5	0.13	21.4	0.08	0.18
GWP		0.11	27.0	0.05	0.17

Source: authors' own elaboration, 2022.

Table 19. Projected versus direct estimates of the prevalence of severe food insecurity (South Africa)

Severe food insecurity					
		Prob.s	**CV (%)**	**Lower_CI**	**Upper_CI**
National income dynamics study (NIDS)	**Total**	0.18	10.1	0.14	0.21
Gallup World Poll (GWP)		0.19	11.8	0.14	0.23
NIDS	**Female**	0.18	12.3	0.14	0.23
GWP		0.21	13,0	0.16	0.26
NIDS	**Male**	0.17	13.5	0.13	0.22
GWP		0.16	16.5	0.11	0.22
NIDS	**Rural**	0.30	8,0	0.25	0.35
GWP		0.24	11.4	0.19	0.30
NIDS	**Urban**	0.10	18.9	0.06	0.14
GWP		0.05	40.5	0.01	0.10
NIDS	**15-24**	0.22	13.3	0.17	0.28
GWP		0.15	19.5	0.09	0.21
NIDS	**25-49**	0.17	13.6	0.12	0.22
GWP		0.21	13.9	0.15	0.27
NIDS	**50-64**	0.09	37.5	0.02	0.16
GWP		0.11	32.7	0.04	0.20
NIDS	**65+**	0.22	30.4	0.09	0.35
GWP		0.29	26.8	0.14	0.45
NIDS	**Inc_1**	0.36	14.2	0.26	0.46
GWP		0.41	12.8	0.31	0.52
NIDS	**Inc_2**	0.28	14.8	0.20	0.36
GWP		0.25	16.8	0.56	0.72
NIDS	**Inc_3**	0.19	15.0	0.14	0.25
GWP		0.16	19.9	0.10	0.22
NIDS	**Inc_4**	0.12	24.1	0.06	0.18
GWP		0.10	29.1	0.04	0.16
NIDS	**Inc_5**	0.01	62.0	-0.003	0.03
GWP		0.01	74.4	-0.005	0.03

Source: authors' own elaboration, 2022.

References

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. Breiman, L. 2001. Random forests. Machine learning, 45(1): 5–32 (also available at https://www.sciencedirect.com/science/article/pii/S0167947306002167)

Boruta. Boruta feature selection method of Kursa and Rudnicki (2010). [online] https://cran.r-project.org/web/packages/Boruta/Boruta.pdf [last accessed 2 February 2022]

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.
Cox, D.R. & Snell, E.J. 1989. *Analysis of Binary Data*. Second edition. London, Chapman & Hall/CRC Press.

FAO. 2022. [online] The Food Insecurity Experience Scale (FIES) webpage. https://www.fao.org/in-action/voices-of-the-hungry/fies/en/ (last accessed 2 February 2022).

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

Hansen, M.H., Madow, W.G. & Tepping, B.J. 1983. An Evaluation of Model Dependent and Probability-Sampling Inferences in Sample Survey (with discussion and rejoinder). *Journal of the American Statistical Association* 78. 776-807.

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.

Italian National Institute of Statistics (Istat). Stat. Regenesees [online] https://www.istat.it/it/files//2021/09/ReGenesees.pdf (last accessed 2 February 2022)

Kim, J.K. & Rao, J.N.K. 2012. Combining data from two independent surveys: a model-assisted approach. *Biometrika*, 99(1): 85–100.

Kursa, M. & Rudnicki, W., 2010. Feature Selection with the Boruta Package. *Journal of Statistical Software*. September 2010. Volume 36, Issue 11. https://www.jstatsoft.org/article/view/v036i11

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_Aug 2013.pdf).

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.

Nagelkerke, N.J.D. 1991. A Note on a General Definition of the Coefficient of Determination. Biometrika, 78, 691-692. https://doi.org/10.1093/biomet/78.3.691

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

Royall, R.M. 1976. The linear least squared prediction to two stage sampling. *Journal of The American statistical Association*. 71. 657-674.

Ryan, T.P. 2008. *Modern Regression Methods*, Second edition. Wiley Series in Probability and Statistics. New York City, USA, John Wiley & Sons Book Series.

Särndal, C.-E., Swensson, B. & Wretman, J. 1992. Model Assisted Survey Sampling. New York City, USA, Springer-Verlag.

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.

Tillé, Y. 2019. Sampling and Estimation from Finite Populations. Wiley.

United Nations Statistical Division. 2022. [online]. Data Disaggregation for the SDG Indicators website. https://unstats.un.org/sdgs/iaeg-sdgs/disaggregation/ (last accessed: 2 February 2022).

Valliant, R., Dorfmann, A.H & Royall, R.M. 2000. *Finite Population Sampling and Inference: A Predicition Approach.* New York City, USA, John Wiley & Sons.

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. (Fourth Integrated Household Survey 2016-2017) [online]. https://microdata.worldbank.org/index.php/catalog/2936 (last accessed 2 February 2022).

Zardetto D. 2015. "ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys" (extended version). *Journal of Official Statistics*, 31(2):177-203.

An indirect estimation approach for disaggregating SDG indicators using survey data

Case study based on SDG Indicator 2.1.2

ISBN 978-92-5-135785-9

9 789251 357859

CB8670EN/1/02.22