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

Piero Demetrio Falorsi[1]; Ayca Donmez[2]; Clara Aida Khalil[3]; Pietro Gennari[4]

1 Senior Statistician, FAO, Rome, Italy, piero.falorsi@gmail.com
2 Statistics and Monitoring Specialist, UNICEF, New York, USA, adonmez@unicef.org
3 Statistician, FAO, Rome, Italy, claraaida.Khalil@fao.org
4 Chief Statistician, FAO, Rome, Italy, Chief-Statistician@fao.org

Abstract:
The overarching principle of the 2030 Agenda for Sustainable Development – "leave no one behind" – calls for more granular and disaggregated data than currently available in most countries, in order to inform the Sustainable Development Goal (SDG) monitoring process at country-level and as a prerequisite for the formulation of national policies targeting the most disadvantaged groups. The "Guidelines on data disaggregation for SDG indicators using survey data", recently published by FAO, is one of the steps taken towards supporting Member Countries in the production of SDG indicators disaggregated by different population groups and territorial areas. They offer a comprehensive overview of survey methods and tools that member countries can adopt for the production of disaggregated estimates of SDG indicators using household surveys as the main supporting data source.

The publication addresses the limitations posed by most surveys, either having samples that are often not large enough to guarantee reliable direct estimates for all sub-populations, or that do not cover all possible disaggregation domains. The Guidelines initially set a framework for promoting a holistic approach to data disaggregation, describing standard and innovative approaches to tackle these constraints at different stages of the statistical production process. One of the proposed solutions outlines a series of actions to be taken at the sampling design stage, by resorting to alternative sampling strategies that ensure a "sufficient" number of sampling units for each disaggregation domain. The other kind of solutions, to be applied at the analysis stage, employ indirect estimation approaches that cope with the little information available for so-called small areas, by borrowing strength from other data sources or domains. In this respect, the guidelines introduce a model-assisted indirect estimation approach that allows integrating data from different surveys and censuses. The described estimator is operationalized for the production of disaggregated synthetic estimates of the SDG Indicator 2.1.2: Prevalence of Moderate and Severe Food Insecurity based on the Food Insecurity Experience Scale (FIES). Both for direct and indirect estimation approaches, the tools to assess the accuracy of the disaggregated estimates are provided.

Finally, the publication concludes with a general overview of small area estimation (SAE) methods, by presenting the key steps for their implementation, introducing the main unit-level and area-level approaches, and providing guidance to assess estimates accuracy.

Keywords:
Design-based approaches, Model-assisted approaches, Projection estimator, Estimates accuracy, Food Insecurity Experience Scale.

1. Introduction

The global spread of the novel coronavirus pandemic (COVID-19) has stalled, and sometimes reversed, global progress towards the achievement of the Sustainable Development Goals (SDGs) and has contributed to increase socio-economic inequalities within countries. Therefore, granular and disaggregated data are much more relevant now than ever before for the implementation of the 2030 Agenda. Indeed, the resolution of the UN General Assembly that endorsed the SDG Global Indicator Framework[1] strongly promotes the overarching principle of data disaggregation, stating that *"SDG Indicators should be disaggregated, where relevant, by income, sex, age, race, ethnicity, migratory status, disability and geographic location, or other characteristics in accordance with the Fundamental Principle of Official Statistics"*.

Recognizing the fundamental role played by disaggregated data and information, the United Nations Statistical Commission (UNSC), at its Forty-seventh Session, requested the Inter-Agency and Expert Group on SDG Indicators (IAEG-SDG) to form a *working group on data disaggregation*, with the objective of strengthening national capacities and developing the necessary statistical standards and tools to produce disaggregated data. This led, among other outputs, to the compilation by custodian agencies of the main categories for the disaggregation of the SDG indicators, as well as to the identification of the policy priorities targeting the most vulnerable population groups.

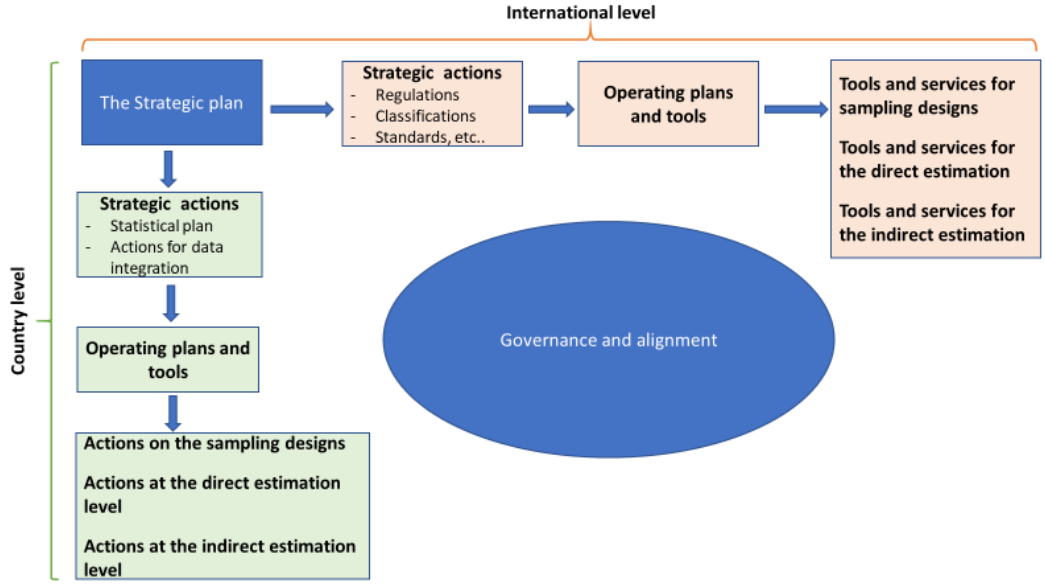
Within this framework, the FAO *Guidelines on data disaggregation for SDG indicators using survey data*, (FAO 2021), is one of the steps taken by the organization – as one of the members of the working group on data disaggregation - towards supporting Member Countries in the production of SDG indicators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

The Guidelines promote a holistic approach to data disaggregation (Figure 1), which involve both national and international actors to come up with agreed strategic plans that foresee the integrated use of various approaches, statistical methodologies and tools to be applied at different stages of the statistical production chain. These strategic plans influence and guide all actions at the technical level, such as those related to the sampling and estimation phases.

National Statistical Offices (NSOs) are the frontline actors for the success of actions at the national level; while international organizations should foster the adoption of standard methodological approaches and the implementation of common tools ensuring the international comparability and high quality of disaggregated SDG data.

[1] Resolution adopted by the General Assembly on Work of the Statistical Commission pertaining to the 2030 Agenda for Sustainable Development (A/RES/71/313).

Figure1. A holistic view of the data disaggregation process



The Guidelines start with providing the objectives of the publication and discussing the statistical challenges posed by data disaggregation in the context of the implementation of the 2030 Agenda for Sustainable Development. This is followed by the description of the characteristics of the holistic approach presented above. Subsequently, technical solutions to define sampling strategies for direct domain estimation and methods relying on the use of auxiliary information are discussed. The guidelines also propose sampling designs that guarantee a sufficient number of sampling units for every subpopulation or domain for which disaggregated data must be produced, thus allowing the calculation of direct disaggregated estimates. Moreover, methods for measuring sampling accuracy are provided. The estimation and dissemination of quality indicators assessing estimates accuracy represents a fundamental step in the production of disaggregated estimates and has the potential of increasing the transparency of NSOs and consequently the public confidence in official statistics. In addition, direct estimates presenting large sampling errors are an indication of the need to either resort to small-area techniques or revisit the sampling design.
A large section of the guidelines is dedicated to present an indirect approach for producing disaggregated estimates relying on the integrated use of two independent surveys. This method allows integrating a small survey, measuring a target variable with a small measurement error, and a more extensive survey, collecting variables of general use, at least one of which is highly correlated with the target variable (proxy variable).
The guidelines end with an overview of small area estimation (SAE) techniques, as one of the possible approaches to produce indirect disaggregated estimates. Being heavily based on model assumptions, the validation and interpretation of results obtained with SAE approaches may be challenging.

2. Planning for data disaggregation at the survey design phase

In order to produce direct disaggregated estimates, the sample should be designed in a way to ensure the presence of sampling units in each disaggregation domain. This will also ensure the production of more accurate indirect estimates through a substantial reduction of the model bias. When the number of people belonging to a rare sub-population can be determined from the sampling frame, selecting the required sample size for the domain is relatively straightforward. In this case, the main issue is the extent of oversampling to apply. On the other hand, extracting a sample from rare domains whose members cannot be identified in advance is more challenging. A variety of methods have been used in these situations. In addition to large-scale screening, methods such as disproportionate stratified sampling, two-

stage sampling, multiple frames, multiplicity sampling, and location sampling can be used. Traditional sampling techniques address data disaggregation by oversampling, deeper stratification, or by introducing multistage designs with screening of respondents. However, for small domains, or segregated and hard-to-reach populations, standard techniques are generally not feasible as they tend to produce an exponential increase of survey costs.
More sophisticated techniques allow for improving sampling designs by geographically spreading the sample units and diminishing the level of clustering. More recent approaches – such as marginal stratification techniques, indirect sampling, multisource and balanced sampling - allow overcoming some of the abovementioned limitations. However, the main drawback of these techniques is the fact that they are not generally known and applied in NISs, and their adoption would require the provision of technical assistance and capacity development programs. Strengths and weaknesses of all the described methods are extensively discussed in the Guidelines. In addition, the publication provides a useful appendix with software packages to be used in empirical applications. Finally, methods and tools to estimate the accuracy of direct disaggregated estimates are presented.

3. Addressing data disaggregation at the analysis stage

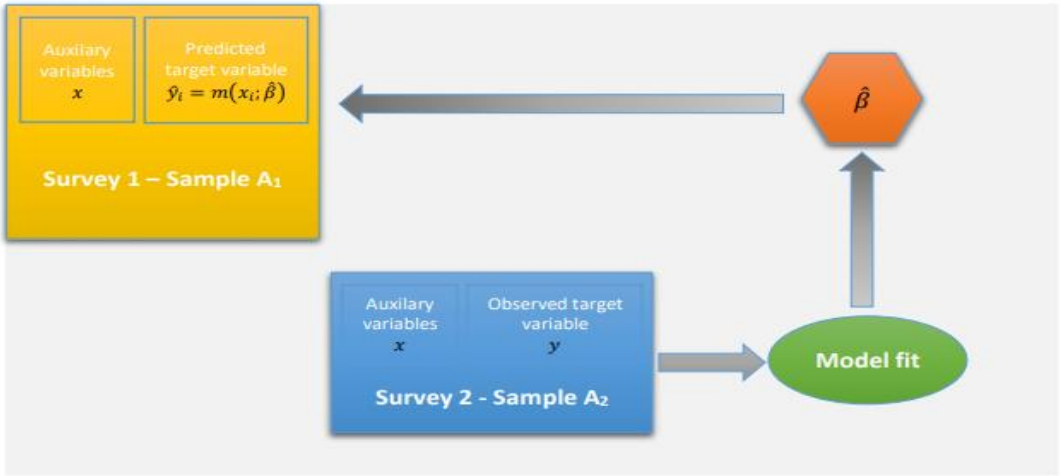
At the analysis stage, data disaggregation can be addressed adopting indirect estimation approaches - including SAE techniques - coping with the little information available for so-called small areas by borrowing strength from additional sources. In particular, the integrated use of different data sources offers a powerful approach for achieving the desired level of disaggregation.

Among the various methods available to produce indirect estimates, the Guidelines present the so-called "*Projection estimator"* (Kim and Rao, 2012). This approach (Figure 2) allows integrating data from two sample surveys – or a sample survey and a census – where the first survey, is characterized by a large sample A_1, but only collects auxiliary information or variables of general use (e.g. socio-economic variables); while the second survey has a smaller sample A_2 but collects information on the target variable y, along with the same set of auxiliary variables available from A_1. In this statistical setting the total of variable y in the disaggregation domain d can be obtained as

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

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

Figure 2: Implementation of the projection estimator



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

In the Guidelines, the proposed methodology has been applied to produce synthetic disaggregated estimates for one of the SDG Indicators under FAO custodianship, namely Indicator 2.1.2 on the prevalence of moderate and severe food insecurity in the population, based on the Food Insecurity Experience Scale (FIES). For the empirical application, two data sources have been integrated:

- The Fourth Integrated Household Survey (IHS4) 2016-17 of Malawi, implemented by the country's NSO under the umbrella of the Living Standard Measurement Study (LSMS). The IHS4 has been used as large survey (S_1) for the projection.
- The FIES module for Malawi collected through the Gallup World Poll (GWP) on 1000 individuals in 2016.

The main results of the experiment are illustrated in Table 1 below. It should be highlighted that the projected values are very close to the actual values for all disaggregation domain, supporting the conclusion that the proposed method performs well for this data set. The reasons for the subpar performance of the projection estimator for the lowest income level are still being explored. This result could most probably depend on the fact that Income is defined differently in the two surveys. For an extensive presentation of the implemented case study, the reader should refer to Chapter 5 of the Guidelines on data disaggregation (FAO, 2021).

Table 1: Estimates for prevalence of moderate and severe food insecurity (prob.ms)

	ALL	
prob.ms	Actual	Predicted
0	6 037	5 548.857
1	23 696	24 184.143

	FEMALE		MALE	
prob.ms	Actual	Predicted	Actual	Predicted
0	3 123	2 927.641	2 914	2 621.215
1	12 651	12 846.359	11 045	11 337.785

	RURAL		URBAN	
prob.ms	Actual	Predicted	Actual	Predicted
0	3 382	4 128.118	2 655	1 420.739
1	20 496	19 749.882	3 200	4 434.261

	AGE<25		AGE>=65	
prob.ms	Actual	Predicted	Actual	Predicted
0	2 134	1 875.174	332	394.91
1	8 490	8 748.826	1 956	1 893.09

	INCOME – Poorest (1st Q)		INCOME – Richest (5th Q)	
prob.ms	Actual	Predicted	Actual	Predicted
0	1 224	882.4605	2 382	2 143.883
1	5 094	5 435.5395	2 973	3 211.117

Source: FAO, 2021

4. Conclusions and way forward

This paper provides a broad overview on the FAO Guidelines on data disaggregation for SDG Indicators using survey data (FAO, 2021). The Guidelines offer readers a comprehensive account of different approaches to data disaggregation and provide practical guidance to produce direct and indirect disaggregated estimates of SDG indicators having surveys as their main or preferred supporting data source. In addition, the document provides practical examples and applications, as well as a list of software packages to be employed at different stages of the statistical production process. The plan for the next few months is to analyse other case studies using the methodological tools described in the guidelines. In particular, the FAO is applying the projection estimator and small-area estimation techniques to other countries and other FAO-relevant SDG indicators (i.e. 2.3.1 = productivity of small-scale food producers, 2.3.2 = income of small-scale food producers, and 5.a.1 = women's land tenure rights) to verify the robustness of the results and to further enhance and extend the methodology for the disaggregation of the SDG Indicators.

References:

1. FAO (2021). Guidelines on data disaggregation for SDG Indicators using survey data. http://www.fao.org/documents/card/en/c/cb3253en
2. Kim, J.K. & Rao, J.N.K. (2012). Combining data from two independent surveys: a model-assisted approach. Biometrika, 99(1): 85–100.
3. Rao, J.N.K. (2003). Small Area Estimation. New York City, USA, John Wiley & Sons
4. Särndal, C.-E., Swensson, B. & Wretman, J. (1992). Model Assisted Survey Sampling. New York City, USA, Springer-Verlag.
5. Valliant, R., Dorfmann, A.H & Royall, R.M. (2000). Finite Population Sampling and Inference: A Prediction Approach. New York City, USA, John Wiley & Sons.
6. UNSD (2020). Report on the results of the UNSD survey on 2020 round population and housing censuses. Background document presented at the Fifty-first session of the United Nations Statistical Commission, 3–6 March 2020, New York City, USA (https://unstats.un.org/unsd/statcom/51stsession/documents/BG-Item3j-Survey-E.pdf).