Food and Agriculture Organization of the United Nations

Integrating surveys with geospatial data through small area estimation to disaggregate SDG indicators at subnational level

Case study on SDG Indicators 2.3.1 and 2.3.2

1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
THE GLOBAL GOALS
For Sustainable Development

Integrating surveys with geospatial data through small area estimation to disaggregate SDG indicators at subnational level

Case study on SDG Indicators 2.3.1 and 2.3.2

Clara Aida Khalil
Food and Agriculture Organization of the United Nations
and
Stefano di Candia
Food and Agriculture Organization of the United Nations

Published by

Food and Agriculture Organization of the United Nations
Rome, 2023

Required citation:
Khalil, C.A. & di Candia, S. 2023. *Integrating surveys with geospatial data through small area estimation to disaggregate SDG indicators at subnational level – Case study on SDG Indicators 2.3.1 and 2.3.2*. Rome, FAO. https://doi.org/10.4060/cc3944en

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

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

ISBN 978-92-5-137545-7

Contents

Tables

Figures

Acknowledgements

Clara Aida Khalil and Stefano di Candia wrote this technical report under the encouragement and supervision of FAO’s Chief Statistician Pietro Gennari.

The authors are thankful to Piero Demetrio Falorsi, Senior Statistician affiliated with the University of Rome “La Sapienza”, who provided relevant technical advice on models and tools to implement the case study presented in this report.

Abbreviations and acronyms

CV	coefficient of variation
EBLUP	empirical best linear unbiased prediction
FAO	Food and Agriculture Organization of the United Nations
HT	Horvitz-Thompson
IAEG-SDG indicators	Inter-Agency and Expert Group on Sustainable Development Goal Indicators
Istat	Italian National Institute of Statistics
kWh	kilowatt-hour
LMG	Lindeman Merenda and Gold
LSMS	living standard measurement study
MSE	mean squared error
NSO	National Statistical Office
NSS	National Statistical System
PPP	purchasing power parity
QQ	quantile-quantile
SAE	small area estimation
SDG	Sustainable Development Goal
TLU	tropical livestock unit
UNDESA	United Nations Department of Economic and Social Affairs

Introduction

Target 2.3 of the 2030 Agenda for Sustainable Development aims to double the agricultural productivity and incomes of small-scale food producers, and is monitored by Sustainable Development Goal (SDG) Indicators 2.3.1 and 2.3.2, which are both under the custodianship of the Food and Agriculture Organization of the United Nations (FAO).

While Indicator 2.3.1 measures the average value of agricultural production per labour unit, providing a measure of average partial factor productivity of agricultural holdings, Indicator 2.3.2 estimates the average income that small-scale food producers derive from their agricultural production activities. Among the mandatory disaggregation dimensions of these two indicators there are the size of the holding (small versus non-small) – which is implemented by applying the official definition of small-scale food producers developed by FAO to enhance estimates' international comparability (Khalil *et al.,* 2017) – and the sex of the holding's head.

Although disaggregation at the subnational level is not included in the set of "mandatory" dimensions for disaggregation of indicators monitoring Target 2.3, producing estimates at the local level may prove to be a much more effective and relevant tool than national-level aggregates for effective planning and decision making. In this respect, being the computation of Indicators 2.3.1 and 2.3.2 normally based on sample surveys microdata, the production of reliable granular subnational estimates is often not possible with standard estimation approaches. Indeed, despite collecting detailed information on socio-economic characteristics of target populations, most sample surveys are characterized by sampling sizes that are either not large enough to guarantee reliable direct estimates for all subpopulations of interest, or that do not cover all possible disaggregation domains (Falorsi *et al.*, 2022).

This kind of issues can be addressed at different stages of the statistical production process. During the survey design phase, they can be tackled by adopting sampling strategies guaranteeing an observed set of sampling units for every disaggregation domain. Although potentially optimal, several studies show how this approach quickly results in an exponential increase of the necessary sampling size and, consequently, survey costs and complexity (FAO, 2021; Kish, 1987). Alternatively, data disaggregation can be addressed at the data analysis stage, by adopting indirect estimation approaches that borrow strength from related disaggregation domains and/or periods, thus resulting in an increase of the actual sample size (Rao and Molina, 2015).

Small area estimation (SAE) methods are among the possible indirect (or model-based) estimation approaches that can be adopted to deal with data disaggregation at the analysis stage. These techniques allow combining survey data with auxiliary information coming from additional data sources, and are seen as cost-effective ways to produce precise estimates of disaggregated parameters. Traditionally, SAE techniques, have relied on the integration of survey microdata with information from population and agricultural censuses or administrative records through explicit models linking the variable of interest to a set of auxiliary variables retrieved from these sources. However, with more and more data made available to National Statistical Systems (NSS) from multiple innovative data sources, other types of auxiliary data can be considered for the production of small area estimates of SDG indicators. In this respect, the 2030 Agenda explicitly stresses the need for new and enhanced data integration strategies, including the exploitation of the potential contribution of geospatial information systems and other big data sources.

Within this framework, the present technical report illustrates a case study on the adoption of SAE techniques to produce granular subnational estimates of SDG Indicators 2.3.1 and 2.3.2, by integrating survey microdata with auxiliary information retrieved from various trustworthy geospatial information systems. The technical report starts from and expands results presented in Khalil *et al.* (2022) with the intent of providing practical guidance to National Statistical Offices (NSO) and other institutions wanting to implement small area estimation techniques on SDG Indicators 2.3.1 and 2.3.2 or similar indicators based on surveys microdata. The document is structured as follows. Section 1 briefly presents SDG Indicators 2.3.1 and 2.3.2, their data sources and required disaggregation dimensions, and the main challenges for their computation both at aggregate and disaggregated level. Then, the main indirect estimation approaches available to address these challenges – with a particular focus on SAE techniques– are presented in Section 2. All the steps for the implementation of an area-level SAE approach on SDG Indicators 2.3.1 and 2.3.2 are described in Section 3, discussing the data sources used, the selection and preparation of auxiliary variables, and the model estimation. Finally, Section 4 presents and discusses the key results of the application and Section 5 outlines the main conclusions.

1. Overview of SDG Indicators 2.3.1 and 2.3.2 and their disaggregation dimensions

Target 2.3 of the 2030 Agenda for Sustainable Development prescribes doubling "*the agricultural productivity and incomes of small-scale food producers, in particular women, Indigenous Peoples, family farmers, pastoralists and fishers, including through secure and equal access to land, other productive resources and inputs, knowledge, financial services, markets and opportunities for value addition and non-farm employment.*"

The target is monitored by means of two indicators under FAO's custodianship, namely:

- **Indicator 2.3.1**, measuring the value of production per labour unit by classes of farming/pastoral/forestry enterprise size; and
- **Indicator 2.3.2**, measuring the average income of small-scale food producers, by sex and Indigenous Peoples' status.

The computation of both indicators is based on a methodology proposed by the FAO and endorsed by the Inter-Agency and Expert Group on SDG Indicators (IAEG-SDG) in 2018, which is documented in the paper *"Methodology for computing and monitoring the Sustainable Development Goal Indicators 2.3.1 and 2.3.2"* (FAO, 2019).

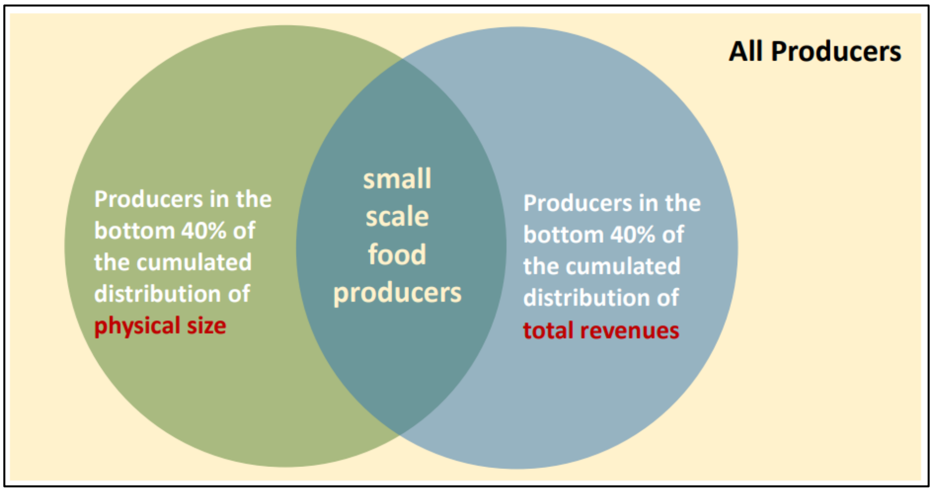
At the base of the approach, the FAO has proposed an internationally agreed definition of small-scale food producers with the objective of computing internationally comparable figures for all countries, territories and regions (Khalil *et al.*, 2017). This definition identifies small-scale food producers using a combination of two criteria, namely the **physical size**, as expressed by the amount of operated land and the number of livestock heads, and the **economic size** of the holding, as expressed by the total value of agricultural production (**Figure 1**). Both criteria are applied in relative terms, in order to enhance international comparability. In practice, small-scale food producers are farmers who:

- operate an amount of land falling in the bottom 40 percent of the cumulative distribution of households' land size at the national level, measured in hectares;
- operate a number of livestock falling in the bottom 40 percent of the cumulative distribution of the number of animals per household at the national level, measured in tropical livestock units (TLUs); and
- obtain total annual revenues from agricultural activities falling in the bottom 40 percent of the cumulative distribution of agricultural revenues per household at the national level, measured in purchasing power parity (PPP) USD.

Within the resulting set of producers identified by these criteria, an additional absolute cap is applied, to exclude producers earning a revenue higher than 34 387 PPP USD per year.[1]

[1] The addition of this threshold was one of the recommendations FAO received from a consultation of the Inter-Agency and Expert Group on Sustainable Development Goal Indicators (IAEG-SDG indicators) on the international definition of small-scale food producers.

Figure 1: Identification of small-scale food producers


All Producers
Producers in the bottom 40% of the cumulated distribution of physical size
small scale food producers
Producers in the bottom 40% of the cumulated distribution of total revenues

Source: FAO. 2019. *Methodology for computing and monitoring the Sustainable Development Goal Indicators 2.3.1 and 2.3.2*. FAO Statistics Working Paper Series, No. 18–14. Rome, FAO. https://doi.org/10.4060/cc3583en

After identifying the population of N_s small-scale food producers among the N food producers ($N_s < N$) of a country, SDG Indicators 2.3.1 and 2.3.2 can respectively be expressed as:

$$\boldsymbol{SDG\ 2.3.1} = I^t_{2.3.1} = \frac{\sum_{j=1}^{N_S}(\frac{\sum_i V^t_{ij} p^t_{ij}}{L^t_j})}{N_S} = \frac{\sum_{j=1}^{N_S} y_{231,j}}{N_S} \quad (1)$$

$$\boldsymbol{SDG\ 2.3.2} = I^t_{2.3.2} = \frac{\sum_{j=1}^{N_S}(\sum_i V^t_{ij} p^t_{ij} - C^t_{ij})}{N_S} = \frac{\sum_{j=1}^{N_S} y_{232,j}}{N_S} \quad (2)$$

Where:

- V^t_{ij} is the physical volume of agricultural product i sold or used by the small-scale food producer j (with $j = 1, \dots, N_s$) during year t;
- p^t_{ij} is the constant sale price received by the small-scale food producer j during year t for product i;
- L^t_j is the number of labour days (full time equivalent) utilized by the small-scale food producer j during year t; and
- C^t_{ij} is the production cost of agricultural product i for the small-scale food producer j during year t.

Being referred to a specific group in the population of agricultural holdings, i.e. the small-scale food producers, the ideal data source for both indicators is represented by a single agricultural or integrated household survey collecting information on all the variables needed to identify smallholders and compute the two indicators. Although Indicators 2.3.1 and 2.3.2 could also be computed using microdata collected through agricultural censuses, data collections of this kind are implemented only once or – in limited cases – twice per decade, and usually do not collect the same detail of information provided by sample surveys.

Concerning data disaggregation, indicators selected to monitor Target 2.3 are currently disaggregated by holding size, as they are produced for both small and non-small scale food producers, and by the sex of the holding head. As reported in the compilation of minimum disaggregation dimensions for SDG indicators prepared by the working group on data disaggregation of the IAEG-SDG (United Nations Department of Economic and Social Affairs, 2022), Indicator 2.3.1 should also be disaggregated by type of product (farming, pastoral, forestry, fishery) and Indicator 2.3.2 by the Indigenous Peoples' status of the holding's head. In addition, the Inter-Agency and Expert Group on Sustainable Development Goal Indicators (IAEG-SDG indicators) has identified future desirable disaggregation dimensions for the two indicators, such as, for example, the geographical level (urban/rural, agro-ecological zones, or subnational level).

Unfortunately, some limitations exist towards the production of both recommended and additional disaggregation dimensions for Indicators 2.3.1 and 2.3.2. For example, the disaggregation by type of product is often not possible given that most agricultural surveys only cover cropping and livestock activities of agricultural holdings, leaving the fishery and forestry sector out of the target. Similarly, also the disaggregation by Indigenous Peoples' status is challenging – if not impossible – in most countries. First of all, most agricultural surveys collect minimal socio-demographic information on holding members which normally do not include the membership of an Indigenous Peoples' group. This issue could partly be addressed by resorting to integrated household surveys collecting a brother spectrum of demographic information, which are – however – implemented only by few countries in the world.

Concerning data disaggregation by geographic location, which is the object of the case study presented in this technical report, several considerations are in order.

A first level of geographical disaggregation is that distinguishing between the urban or rural location of the household/holding. The information on whether the household main dwelling or holding establishment is located in an urban or rural area is available in most surveys, making the production of indicators monitoring Target 2.3 by this dimension possible in the majority of cases. However, different countries use different criteria to define urban and rural areas, which reflect their various perspectives as to what constitute different levels of urbanization. It is clear that individual countries need to have their own national definitions to be implemented within their NSSs, and used to disaggregate indicators by urban and rural areas for their own national policy purposes. Nonetheless, in order to have meaningful international comparisons of statistical indicators by urban and rural areas there is also an undisputed need for an internationally comparable definition. For this reason, six international organizations[2] have worked together to develop a harmonize methodology to define the degree or urbanization of a specific area, distinguishing between cities, towns and semi-dense areas, and rural areas. The application of the degree of urbanization to SDG Indicators 2.3.1 and 2.3.2 will be discussed in a separate technical report.

Still in the domain of disaggregation by geographic location, SDG estimates by subnational administrative divisions (e.g. by regions, provinces, municipalities depending on the specific country context) may prove to be very relevant and valuable for actual planning and policymaking at country level. However, the limited size of most survey's samples makes the production of reliable subnational estimates in unplanned estimation domains[3] a challenging task. As briefly mentioned in the introduction, these issues can be addressed at different stages of the statistical production process.

Theoretically speaking, data disaggregation at subnational level can be addressed at the survey design stage, adopting sampling strategies guaranteeing an observed set of sampling units for every subnational division for which disaggregated estimates must be produced. Various sampling approaches are available to ensure a sufficient sampling size in every subpopulation, such as oversampling, deeper stratification, and multiphase sampling with screening of respondents. FAO (2021) offers a complete overview of these approaches and the packages of statistical software that could be used for their implementation. These methods, which are not mutually exclusive and could be implemented jointly, come with an increase in terms of costs and complexity of survey operations, and may become quickly unfeasible when considering very small estimation domains or multiple disaggregation dimensions. In addition, these approaches are not relevant in contexts where the objective is to make the best possible use of already available data, without repeating and redesigning the data collection exercise.

A different strategy, which has been adopted for the case study presented in this report, consists in addressing the limitations arising from the limited sample size of surveys at the analysis stage. This can be done by resorting to indirect estimation methods coping with the little sampling information available for so-called small areas by borrowing strength from additional data sources that are not

[2] The European Commission, the Food and Agriculture Organization of the United Nations, the International Labour Organization, the Organisation for Economic Co-operation and Development, the United Nations Human Settlements Programme and the Word Bank.

[3] Unplanned estimation domains are those disaggregation domains that are not planned at the sampling design stage and, consequently, may not have a sufficient sample size to yield precise estimates (i.e. estimates with a low sampling variance).

affected by sampling errors. SAE methods may be useful when the direct estimation of a parameter of interest in a given disaggregation domain does not reach the pre-specified precision level. Furthermore, SAE methods allow obtaining predictions for domains where no sample information is available. To achieve a higher reliability then the direct estimator, small area estimation methods combine the survey data with an additional data source such as population or agricultural censuses, administrative registers, but also GIS and other sources of big data.

The fundamental concepts for SAE along with the main classes of approaches available in the literature are discussed in the next section.

2. Principles of small area estimation approaches

Sample surveys, which are cost-effective means to collect detailed information on target phenomena at relatively high frequency over time, are normally used to produce estimates for the overall population or broad disaggregation domains. In this context, a direct disaggregated estimate of a target parameter is a statistics produced using exclusively sample data belonging to the related disaggregation domain. Direct estimates are produced using so-called direct estimators that, making use of sampling weights, are also known as design-based estimators. Indeed, the inferential performance and properties of direct estimators are entirely determined by the underlying sampling design (FAO, 2021).

A key requirement to achieve reliable direct estimates for a given disaggregation domain is the presence of a sufficient domain sample size to yield adequate precision, or – in other terms – a small sampling variance. When this circumstance is not verified, we are in the presence of so-called small areas, i.e. disaggregation domains where too few or no sampling observations are available.

In practical situations, it is quite uncommon to have an overall sampling size that is large enough to guarantee a sufficient number of observations for every possible disaggregation domain. Therefore, the use of indirect estimation techniques to borrow strength from auxiliary information on the population of interest is often necessary. The range of possible methods to produce indirect estimators is vast and goes from the implementation of design-based model assisted approaches, such as the generalized regression estimator (Särndal *et al.*, 1992) or the projection estimator (Kim and Rao, 2012; FAO, 2022), to model-based approaches such as SAE (Rao and Molina, 2015). SAE approaches bring the additional information into the estimation process through an explicit model linking the variable of interest to a set of auxiliary variables. More precisely, SAE methods rely on mixed models including domain-specific random effects to account for the variability between different areas that is not explained by the auxiliary variables. Given the role played by models in the small area estimation process, the resulting estimators are often referred to as model-based estimators.

The notation framework needed to discuss most common SAE approaches is introduced in the following paragraphs. Let us consider a finite population U of N units, partitioned into D estimation domains $U_1, U_2, \dots, U_D$ of sizes $N_1,\ N_2, \dots, N_D$. With d we denote the d^{th} disaggregation domain, while j specifies the j^{th} unit of the population.

Let us now consider a random sample $s \in S$ of size n and probability $p(s)$, where S represents the set of all possible sample s of size n that can be drawn from population U. Units in s can be used to produce direct estimates $\hat{\theta}_d^{dir}$ of disaggregation parameters θ_d related to a given variable of interest y. When considering continuous variables (e.g. the total quantity produced of a given crop, or the household income), typical examples of disaggregation parameters θ_d are the domain total $Y_d = \sum_{j\in U_d} y_j$ and mean $\bar{Y}_d = Y_d/N_d$.

In this framework, the most common direct estimators of Y_d and $\bar{Y}_d$ are the corresponding Horvitz-Thompson (HT) estimators (Horvitz and Thompson, 1952), which can be expressed respectively as $\hat{Y}_d = \sum_{j\in s_d} w_j y_j$ and $\hat{\bar{Y}}_d = \bar{Y}_d / \sum_{j\in s_d} w_j$. In these last two expressions, $w_j = 1/\pi_j$ denotes the sampling weights and $\pi_j = \sum_{\{s: j\in s\}} p(s)$ the inclusion probability of unit j.[4]

The HT estimator of Y_d is design unbiased, while the one for $\bar{Y}_d$ is affected by a bias that tends to 0 with increasing values of n_d. This means that their expected values are or tend to be equal to the parameter to be estimated (Cochran, 1977). As a consequence, their reliability is assessed only in terms of their precision, i.e. by the extent of their variance. Direct HT estimators are usually characterized by unknown variance $V(\hat{\theta}_d)$ that needs to be estimated with adequate estimators $\hat{v}(\hat{\theta}_d)$, for a complete overview of which we refer to Cochran (1977) or Wolter (2007).

In this framework, a measure of precision that is often used to assess direct estimates is the coefficient of variation (CV), which can be expressed as $CV(\hat{\theta}_d) = \frac{\sqrt{V(\hat{\theta}_d)}}{\hat{\theta}_d}$. When the estimated CV is unacceptably high, SAE and other indirect estimation approaches can be used to increase estimates precision. Model-based SAE approaches allow considering unexplained heterogeneity among domains, and have the potential of providing more precise estimates than those produced with direct methods. In addition, by resorting to SAE, it is possible to predict the value of indicators also in out of sample domains, meaning domains with sampling size equal to 0. It should be noted that, being based on models, SAE estimators are no longer unbiased, and their reliability needs to be assessed in terms of their mean squared error (MSE), which provides a joint measure of both accuracy (bias) and precision (variance). Hence, the CV of model-based estimators is obtained as $CV(\hat{\theta}_D^{SAE}) = \frac{\sqrt{MSE(\hat{\theta}_D^{SAE})}}{\hat{\theta}_D^{SAE}}$.

The literature on SAE classifies its models into two broad categories identified as **area-level** and **unit-level** models, which are briefly discussed in the two sections below. While area-level approaches relate a small area direct estimator $\hat{\theta}_d$ to area-specific auxiliary information and can be adopted also

[4] It should be noted that $\hat{\bar{Y}}_d$ has the functional form of a ratio estimator, as both its numerator and denominator are sampling estimates.

when unit-level data is not available, unit-level models require access to microdata at the unit level, as they relate the unit values y_j to unit-specific covariates (Rao and Molina, 2015).

2.1 Area-level models

The Fay-Herriot model (Fay and Herriot, 1979), which is by far the most popular area-level SAE approach, is often used for the production of small area estimates in official statistics and research thanks to its intuitive application and interpretation. This approach combines a sampling model, assuming that the unknown parameter θ_d and the direct estimate $\hat{\theta}_d^{dir}$ differ by a sampling error e_d and a linking model specifying a relationship between the population value θ_d and a set of domain-level auxiliary information.

The sampling model can be formulated as

$$\hat{\theta}_d^{dir} = \theta_d + e_d,\ d = 1, \dots, D \qquad (3)$$

where $\hat{\theta}_d^{dir}$ is a design-unbiased direct estimator and the sampling error e_d has mean 0 and variance $\sigma_{e_d}^2$.

On the other hand, the linking model can be expressed as

$$\theta_d = x_d^T \beta + u_d,\ d = 1, \dots, D \qquad (4)$$

where $\beta = (\beta_1, \dots, \beta_P)$ is the vector of unknown regression parameters and u_d are domain-specific random effects, which are supposed to be normally distributed with mean 0 and variance σ_u^2.

The combination of the sampling and the linking models leads to a special case of linear mixed area-level model, which can be formalized as follows:

$$\hat{\theta}_d^{dir} = x_d^T \beta + u_d + e_d,\ d = 1, \dots, D \quad (5)$$

The unknown parameters of (5) to be estimated are:

- the fixed-effects parameters β; and
- the variance of the random effects σ_u^2.

Common approaches used to estimate these unknown quantities are the empirical best linear unbiased prediction (EBLUP), the empirical Bayesian and the hierarchical Bayesian methods. In particular, the EBLUP estimator (Harville, 1991), which is implemented under the classical frequentist framework, can be expressed as a weighted average of a direct estimator and a regression synthetic component. In symbols:

$$\hat{\theta}_d^{EBLUP} = \hat{\gamma}_d \hat{\theta}_d + (1 - \hat{\gamma}_d) x_d^T \hat{\beta} \qquad (6)$$

where $\hat{\beta}$ is the weighted least squares estimator of the regression parameter, and $\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \sigma_{e_d}^2}$ is the so-called shrinkage factor for domain d, which weights the direct estimate and the regression synthetic component, and decreases with increasing sampling variance $\hat{\sigma}_{e,d}^2$. It should be noted that, when $n_d = 0$ – i.e. in correspondence of out-of-sample domains – $\hat{\gamma}_d$ is also equal to 0, and the corresponding SAE estimates are produced using only the regression synthetic component $x_d^T \hat{\beta}$ of $\hat{\theta}_d^{EBLUP}$.

The basic FH model is based on two fundamental hypothesis:

1) The error terms e_d and the random effects u_d follow a normal distribution with mean 0 and variance $\sigma_{e,d}^2$ and σ_u^2 respectively. In particular, the random effects are supposed to be independent and identically distributed.
2) The auxiliary variables are measured without errors.

Several extensions of this basic approach are available in the literature to address special situations, such as the presence of spatial (Petrucci and Salvati, 2006) or spatio-temporal (Marhuenda *et al.*, 2013) correlation, heteroscedasticity of random effects (Breidenbach *et al.*, 2018), influential outliers (Schoch, 2012), and auxiliary variables affected by measurement errors (Ybarra and Lohr, 2012).

In addition, it should be noted that the FH assumes the sampling variance $\sigma_{e,d}^2$ to be known. However, in practical applications also this component needs to be estimated by means of standard direct estimators $\hat{\sigma}_{e_d}^2$.

2.2 Unit-level models

Contrarily to area-level approaches, unit-level SAE models require the availability of unit-level microdata for both the variable of interest y_j and the set of auxiliary variables x_j considered to have a good predictive power with respect to the phenomenon of interest. Unit-level models are popular in poverty mapping, which is one of the most common applications of small area estimation (Bedi *et al.*, 2007). The basic unit-level model, also known as nested error linear regression model (Battese *et al.*, 1988), has the following structure:

$$y_{dj} = x_{dj}^T \beta + u_d + e_{dj}\,; \quad d = 1, \dots, D; \;\; j = 1, \dots, n_d \quad (7)$$

where $x_{dj} = (x_{1,dj}, \dots, x_{p,dj}, \dots, x_{P,dj})$ is the vector of P auxiliary variables for unit j.

The model in (7) contains independent and identically distributed domain-specific random effects u_d, with $u_d \sim N(0, \sigma_u^2)$, and unit-level error terms $e_{dj} \sim N(0, \sigma_e^2)$. As for area-level models, besides the error variance σ_e^2, the unknown parameters are the fixed effect parameters β and the variance of random effects σ_u^2, which can be estimated with EBLUP, empirical Bayes, and hierarchical Bayes methods.

Under the EBLUP approach, the SAE estimator can be formalized as a linear combination of the survey regression estimator and a regression-synthetic component:

$$\hat{\theta}_d^{EBLUP} = \hat{\gamma}_d\left[\bar{y}_d + \left(\bar{X}_d^T\hat{\beta} \; - \bar{x}_d^T\hat{\beta} \;\right)\right] + (1 - \hat{y}_d)\bar{X}_d^T\hat{\beta}$$

where $\bar{y}_d$ is the sample mean of the variable of interest for domain d, $\bar{X}_d^T$ and $\bar{x}_d^T$ are the means of the auxiliary information from the additional data source and the survey, respectively, and $\hat{\beta}$, $\hat{\sigma}_u^2$, and $\hat{\sigma}_e^2$ are the estimated parameters. The weight $\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2+\frac{\hat{\sigma}_e^2}{n_d}}$ measures the amount of unexplained between-area variability to the total variability, and gives more importance to the survey regression component of the estimator with increasing domain sample size n_d.

Similarly to what seen for area-level models, various extensions of the unit-level approach are available in the literature. In particular, while the model in (7) only supports the estimation of means and totals, approaches relying on nested error linear regression models allow the estimation of non-linear indicators (Elbers *et al.*, 2003; Molina and Rao, 2010). These extensions are particularly relevant in the context of the SDG monitoring framework, where many of the indicators are expressed as ratios and proportions. Additional extensions allow to include sampling weights in the estimation process (You and Rao, 2002), address the presence of heteroscedasticity in the error term (Breidenbach *et al.*, 2018), and produce estimates which are robust to influential outliers (Schoch, 2012).

2.3 Practical considerations for small area estimation implementation

Despite their increasing popularity, SAE should not be seen as the solution to any data disaggregation problem, and there are various considerations that an NSO should make before engaging in the production of indirect estimates.

First, model-based approaches have stricter data requirements than direct estimation methods, with unit-level models being more data intensive than area-level ones. In this respect, the confidentiality concerns that may limit the access to microdata on individual units should be taken into account before embarking in the process of producing SAE.

Being based on models, assumptions underlying implemented SAE techniques need to be carefully validated through adequate diagnostic methods. In addition, the bias of small area estimates needs to be measured to assess the reliability of final estimates.

An important prerequisite for the construction of SAE models with satisfactory predictive power is the availability of auxiliary variables of good quality. Traditional sources of this additional information are population and agricultural census as well as administrative records. Census data have the advantage of providing a complete coverage of target populations and can offer valid socio-economic predictors of the variable of interest. However, the low frequency at which censuses are usually implemented limits their use for the production of disaggregated statistics on an annual basis. Administrative records, which are often generated as side product of government operations, do not suffer from this drawback. However, data of this kind are not produced with the primary purpose of computing official statistics, and, as a consequence, their accuracy, coverage, content, and characteristics need to be carefully assessed before them being used for statistical purposes (Erciulescu *et al.*, 2021). Some examples of applications of SAE based on administrative records are given in Rao and Molina (2015), Erciulescu *et al.* (2021), and Zhang and Giusti (2016).

The huge amount of digital and geospatial information produced by a wide range of tools and technologies nowadays offers good alternative sources of auxiliary variables for SAE production. These rich large-scale datasets, also referred to as big data, generally cover a vast portion of the population within a territory, often reaching nationwide coverage. Potential sources of big data are social networks, GISs, and records generated by human transactions and interactions. These “new” or “alternative” data sources can complement traditional surveys and censuses to reduce the time and resources needed for data production, hence contributing to fill the SDG data gap. Examples of studies relying on the use of big data and geospatial information for the implementation of SAE techniques are widely available in the literature. In particular, Marchetti *et al.* (2015) discuss the challenges opened by the extension of SAE covariates to include variables generated by big data sources and provides some solutions to address them. Specifically, besides requiring the availability of advanced statistical and information technology knowhow, the quality of data from these “new” data sources is often uncertain and rarely documented in comprehensive metadata files. In this respect, attention should be paid to the fact that basic SAE approaches are implemented under the assumption that auxiliary variables are measured without error, or, in other words, that they are available for all areas and they come from archives covering the entire population of interest. However, data coming from big data sources are often affected by measurement errors (e.g. under-coverage and over-coverage) and bias. To mitigate these issues, various authors – such as Ybarra and Lohr (2012), and Arima *et al.* (2018) – have developed SAE approaches accounting for the presence of measurement errors in the covariates.

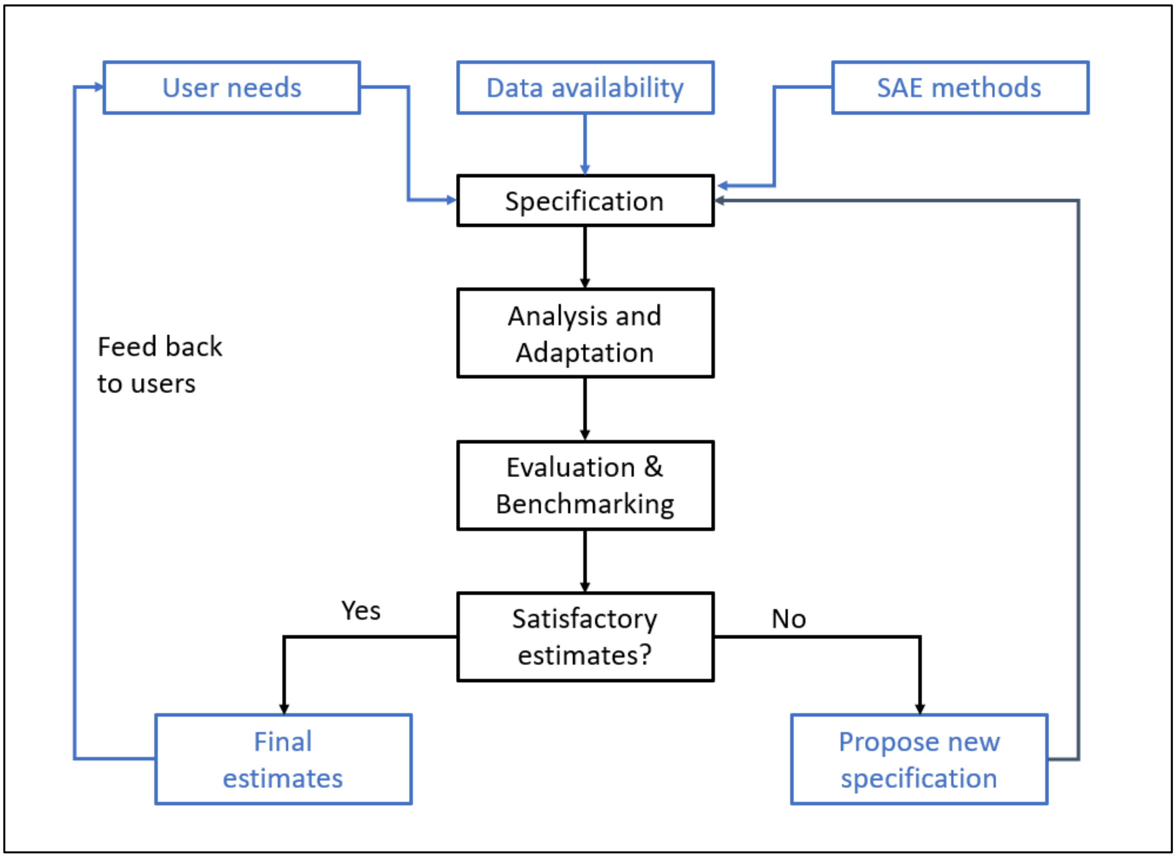
3. Study description

After introducing the two main classes of SAE approaches in Section 2, this section of the technical report describes the various phases and components of a case study implemented to explore the application of a FH area-level SAE approach to the production of subnational estimates of SDG Indicators 2.3.1 and 2.3.2 presented in Section 1. In particular, the report discusses the estimation of the labour productivity and average income of small-scale food producers at the level of second administrative divisions in Mali, considering the integration of microdata from an integrated household survey with area-level auxiliary information retrieved from multiple geospatial information systems.

3.1 Specification of the small area estimation problem

The production of small area estimates is normally implemented following a series of interconnected steps that can be repeated iteratively until the obtained result is of satisfactory quality (**Figure 2**).

Figure 2: Main steps for the implementation of small area estimation



Source: United Nations Department of Economic and Social Affairs (UNDESA). 2022. Producing small area estimation. In: *UN Statistics wiki*. New York, UNDESA. Cited 01 December 2022. https://unstats.un.org/wiki/display/SAE4SDG/Producing+SAE

At the input (or specification) stage of any SAE problem, three main elements should be clarified:

- The **user needs** that, from a statistical stand point, determine the indicator to be estimated and the disaggregation dimensions to be produced;
- The **data availability** including the survey to estimate the indicator and the additional sources providing relevant auxiliary variables; and
- The **SAE method** to be implemented taking into account the identified user needs and the available data. Indeed, the SAE approach to be selected will highly depend on the functional form of the selected indicator and the level of aggregation/disaggregation at which auxiliary variables are available.

User needs

The objective of the presented case study was to produce estimates of SDG Indicators 2.3.1 and 2.3.2 at the level of second administrative units in Mali. Both Indicator 2.3.1, measuring the average labour productivity of small-scale food producers, and Indicator 2.3.2, measuring the average income of the same group of farmers, are statistics expressed as population means (see expressions (1) and (2) reported in **Section 1**). With similar functional forms, standard SAE models – such as the area-level FH model (**Section 2.1**) or the unit-level nested-error linear regression model (**Section 2.2**) can be considered.

Data sources

The SAE application was implemented with microdata from the Agricultural Survey Integrated to Households Living Conditions 2017. The Agricultural Survey Integrated to Households Living Conditions is a multi-thematic cross-sectional household survey implemented under the World Bank's living standard measurement study (LSMS) programme, and is based on a nationally representative sample of about 8 390 households and with a specific focus on agriculture. In 2017, sampling units were divided into two groups, one of 3 813 households that received the full questionnaire, and one with the remaining households that received a light version of the same questionnaire. Given that some of the variables needed to estimate SDG Indicators 2.3.1 and 2.3.2 were collected with the larger questionnaire, only households that completed the full interview could be considered for the present study. In addition, considering that the two indicators already have a disaggregation dimension embedded in their definition, i.e. the size of the farm, the sample that could be actually used to produce small area estimates for small-scale food producers included only 1 637 households.[5]

Since 2016, Mali is divided into nine regions and one capital district (Bamako), where each region bears the name of its capital. All regions are further divided into 53 circles, which were the

[5] Full description of the survey and the microdata can be found on the World Bank Microdata Catalogue (World Bank, 2023).

disaggregation domains considered for the study. **Table 1** below provides a summary of the sample size by Malian regions and circles for Indicator 2.3.1, and the total number of out of sample circles (i.e. the circles with a sample size equal to 0). In particular, the region of Kidal was left out of the sample due to security reasons. In addition, the new region of Menaka had not been officially announced at the time of the survey and, for this reason, was not included in the sample.

Similar information is provided in **Table 2** for Indicator 2.3.2.

Table 1: Sampling size of the Agricultural Survey Integrated to Households Living Conditions (2017) – SDG Indicator 2.3.1

Region	Number of circles	Sample size by region	Sample size of small-scale food producers by region	Average number of sampled small-scale food producers by circle
Kayes	7	431	384	55
Koulikoro	7	381	282	40
Sikasso	7	368	206	29
Segou	7	436	295	42
Mopti	8	323	221	28
Tombouctou	5	137	126	25
Gao	3	110	101	34
Kidal	4 (out of sample)	-	-	-
Menaka	4 (out of sample)	-	-	-
*** Bamako**	1 (Bamako)	22	22	22

Source: FAO. 2023. *Integrating surveys with geospatial data through small area estimation to disaggregate SDG indicators at subnational level. Case study on SDG Indicators 2.3.1 and 2.3.2.* Rome.

Table 2: Sampling size of the Agricultural Survey Integrated to Households Living Conditions (2017) – SDG Indicator 2.3.2

Region	Number of circles	Sample size by region	Sample size of small-scale food producers by region	Average number of sampled small-scale food producers by circle
Kayes	7	572	413	59
Koulikoro	7	517	309	44
Sikasso	7	526	229	33
Segou	7	595	331	47
Mopti	8	429	265	33
Tombouctou	5	219	161	32
Gao	3	180	141	46
Kidal	4 (out of sample)	-	-	-
Menaka	4 (out of sample)	-	-	-
Bamako	1 (Bamako)	760	71	71

Source: FAO. 2023. *Integrating surveys with geospatial data through small area estimation to disaggregate SDG indicators at subnational level. Case study on SDG Indicators 2.3.1 and 2.3.2.* Rome.

For the implementation of the discussed case study, area-level auxiliary variables retrieved from various geospatial information systems were considered. The initial set of potential independent variables was selected among the vast amount of publicly available candidates according to their potential capability of being good predictors of the average labour productivity and income in agriculture. In particular, considered covariates included in the first stage of selection were providing information on the following domains:

- **soil characteristics:** for example, the volume fraction of coarse fragments, the nitrogen content, and the concentration of other soil components such as salt, silt, clary, organic carbon, etc.;
- **weather and climate:** minimum and maximum temperature, precipitation quantity, direct normal irradiation, diffuse horizontal irradiation, air temperature, and vegetation indexes;
- **land cover:** elevation, cover fraction of cropland, bare ground and extent of built up areas;
- **harvested area and production** of major crops (cotton, rice, sorghum, and wheat);

Table 3 presents the spatial and temporal resolution of each auxiliary variable along with the related source.

Table 3: Spatial-temporal resolution and sources of geospatial area-level covariates

Variable name	Spatial resolution	Temporal resolution	Source
Volume fraction of coarse fragments	~1×1 km	Static	**ISRIC: World Soil Information**
Nitrogen	~1×1 km	Static	
Sand	~1×1 km	Static	
Silt	~1×1 km	Static	
Clay	~1×1 km	Static	
Soil organic carbon	~1×1 km	Static	
Minimum temperature	~4.5×4.5 km	Monthly	**WorldCilm: Historical monthly weather data**
Maximum temperature	~4.5×4.5 km	Monthly	
Precipitation	~4.5×4.5 km	Monthly	
Direct normal irradiation (Long-term yearly average)	~0.3×0.3 km	1994–2018	**Solargis**
Diffuse horizontal irradiation (Long-term yearly average)	~0.3×0.3 km	1994–2018	
Air temperature (Long-term yearly average)	~1×1 km	1999–2020	
Vegetation indexes	~5.5×5.5 km	Monthly	**NASA EarthData**
Elevation	~1×1 km	Static	**CGIAR CSI**
Cropland	~1×1 km	Annual	**Zenodo**
Bare ground	~1×1 km	Annual	
Built-up	~1×1 km	Annual	
Harvested area (major crops)	~1×1 km	Annual	**MAPSPAM**
Production (major crops)	~1×1 km	Annual	

Source: FAO. 2023. *Integrating surveys with geospatial data through small area estimation to disaggregate SDG indicators at subnational level. Case study on SDG Indicators 2.3.1 and 2.3.2.* Rome.

Small area estimation method

Considering the functional form of the two indicators monitoring Target 2.3 and the level of aggregation of auxiliary variables, the area-level FH model presented in Section 2.1 has been adopted for this case study. It should be noted that the use of unit-level SAE to disaggregate SDG indicators monitoring Target 2.3 may be challenging. Indeed, as introduced in Section 2.2, the implementation of unit-level models requires being able to identify the target population (the small-scale food producers) in both the survey and the considered source of unit-level auxiliary data. This means that, if – for example – an agricultural census is to be considered as the source of additional data, this should include all the variables needed to identify small-scale food producers and these variables should share the same definitions with those included in the survey.

3.2 Analysis and adaptation

The specification stage of an SAE problem is followed by the analysis and adaptation phase, which deals with model specification by selecting auxiliary variables, implementation of direct and model-based estimation, and checking of model assumptions.

Selection of auxiliary variables

The values of considered geospatial predictors were initially available at the level of the cells of regular grids of different resolutions (spanning from 1×1 km to 5.5×5.5 km). Hence, being the FH approach based on auxiliary information referred to the small area of interest, data needed to be pre-processed in order to produce aggregates (totals or means depending on the considered variable) referred to the irregular polygons defining Mali's circles.[6]

The initial set of potential predictors introduced in Section 3.1 was reduced adopting two stepwise regressions implemented using the area-level direct estimates of Indicators 2.3.1 and 2.3.2 respectively as dependent variables and the area-level geospatial variables as covariates.[7]

As results of the two step-wise regressions, only 8 auxiliary variables were retained for Indicator 2.3.1 and 3 for Indicator 2.3.2 (see tables 4.4 and 4.5 respectively) according to the Lindeman Merenda and Gold (LMG) factor, which represents a measure of the relative contribution of each predictor to the overall R square of the model.

[6] All data pre-processing operations were performed using the R package "raster". More details on this package and related functions are provided in Annex 2.

[7] An initial set of variables was eliminated due to their high correlation – above 0.9 – with other variables considered important for the prediction of the parameters of interest, in order to avoid multi-collinearity issues.

Table 4: Results of step-wise regression for SDG Indicator 2.3.1

Variable name	Unit of measure	LMG (%)
Average cotton production	(Metric tonne)	24.0
Mean direct normal irradiation	(kWh/m^2)	16.1
Average wheat production	(Metric tonne)	14.6
Average rice production	(Metric tonne)	11.7
Average sorghum production	(Metric tonne)	11.4
Mean vol. fraction of coarse fragments	(%)	9.8
Mean soil organic carbon	(g kg-1)	8.7
Average rice harvested area	(hectare)	3.7

Source: FAO. 2023. *Integrating surveys with geospatial data through small area estimation to disaggregate SDG indicators at subnational level. Case study on SDG Indicators 2.3.1 and 2.3.2.* Rome.

Table 5: Results of step-wise regression for SDG Indicator 2.3.2

Variable name	Unit of measure	LMG (%)
Minimum temperature	(Celsius °C)	69.2
Average cotton production	(Metric tonne)	16.5
Total population	(Number)	14.3

Source: FAO. 2023. *Integrating surveys with geospatial data through small area estimation to disaggregate SDG indicators at subnational level. Case study on SDG Indicators 2.3.1 and 2.3.2.* Rome.

It is interesting to notice that most of the covariates considered as important by the selection approach provide information on either the quantity produced or the area harvested of Mali's major crops. Other variables retained by the stepwise procedure measure the average direct normal irradiation, the average volume fraction of coarse fragments, the average quantity of organic carbon in the soil, the average minimum temperature, and the total population.

Estimation

The estimation stage of most SAE problems starts from the production of small area direct estimates along with a measure of their precision (e.g. the CV). Direct estimates are not only assessed against model-based estimates produced at later stages of the process, but they also represent an important component of area-level estimation models such as the FH presented in expression (5).

Going back to the notation introduced in Section 1, the two variables of interest needed for the estimation of Indicators 2.3.1 and 2.3.2 are the $y_{231,j}$ and $y_{232,j}$ $(j = 1, \ldots., n)$ introduced in expressions (1) and (2). Under this framework, the direct HT estimator of the two indicators for the $d-th$ small area can be expressed as:

$$\hat{\bar{y}}_{l,d}^{dir} = \frac{\sum_{j\in s_d} w_j y_{l,j}}{\sum_{j\in s_d} w_j} \qquad \text{with } l = 231 \; or \; 232 \text{ and } d = 1, \dots D$$

Similarly, the area-level FH model introduced with expression (5) can be formulated as

$$\hat{\bar{y}}_{l,d}^{dir} = x_{l,d}^T \beta_d + u_{l,d} + e_{l,d} \qquad \text{with } l = 231 \; or \; 232 \text{ and } d = 1, \dots D$$

which leads to the following area-level EBLUP estimation :

$$\hat{\bar{y}}_{l,d}^{EBLUP} = \hat{\gamma}_{l,d} \hat{\bar{y}}_{l,d}^{dir} + \left(1 - \hat{\gamma}_{l,d}\right) x_{l,d}^T \hat{\beta}_l \quad \text{with } l = 231 \; or \; 232 \text{ and } d = 1, \dots D$$

Where $\hat{\beta}_l$ is the weighted least squares estimator of the regression parameters and $\hat{\gamma}_{l,d}$ is the shrinkage factor for estimator ($l = 231 \; or \; 232$).

Assessment of estimates accuracy and model assumptions

As introduced in Section 2, the basic FH model assumes that the error term $e_{l,d}$ and the random effect $u_{l,d}$ follow a normal distribution with mean 0 and variances $\sigma_{e_{l,d}}^2$ and $\sigma_{u_l}^2$ respectively. These assumption will have to be validated when assessing the model output, by means of quantile-quantile (QQ) plots or normality tests (e.g. the Shapiro-Wilk test).

In addition, the FH model assumes that the sampling variances $\sigma_{e_{l,d}}^2 \; (d = 1, \dots, D)$ are known. However, in practical applications these are unknown quantities that need to be estimated by means of direct estimators. Due to the limited sample size of small areas, the variance estimates might need to be stabilized though suitable smoothing techniques.

In this particular case study we adopted the smoothing approach based on the generalized variance functions. First of all, given that the parameter of interest is a ratio, a linearization approach of the variance is adopted (Wolter, 2007). Indicating with $\hat{\mathrm{R}} = \frac{\hat{\mathrm{X}}}{\hat{\mathrm{Y}}}$ the ratio estimator, where $\hat{\mathrm{X}}$ and $\hat{\mathrm{Y}}$ denote estimators of X and Y, the Taylor series estimator of the variance of $\hat{\mathrm{R}}$ is:

$$v(\hat{R}) = \frac{1}{\hat{Y}^2}\left[v(\hat{X}) + \hat{R}^2 v(\hat{Y}) - 2\hat{R}\hat{\rho}_{XY}\sqrt{v(\hat{X})}\sqrt{v(\hat{Y})}\right]$$

where $v(\hat{X})$, $v(\hat{Y})$ and $\hat{\rho}_{XY}$ denote estimators of the variance of $\hat{X}$, the variance of $\hat{Y}$ and the linear correlation between $\hat{X}$ and $\hat{Y}$ respectively.

The smoothed variance of $\hat{R}$ can be obtained by substituting to $v(\hat{X})$, $v(\hat{Y})$ their smoothed versions $\tilde{v}(\hat{X})$, $\tilde{v}(\hat{Y})$:

$$\tilde{\sigma}^2_{ea} = \tilde{v}(\hat{R}) = \frac{1}{\hat{Y}^2}\left[\tilde{v}(\hat{X}) + \hat{R}^2\tilde{v}(\hat{Y}) - 2\hat{R}\hat{\rho}_{XY}\sqrt{\tilde{v}(\hat{X})}\sqrt{\tilde{v}(\hat{Y})}\right]$$

The smoothing models chosen for $\mathrm{v}(\hat{\mathrm{X}})$ and $\mathrm{v}(\hat{\mathrm{Y}})$ are the following:

$$\log\left(v(\hat{X})\right) = \alpha + \beta\log(\hat{X})\ ,\ \log\left(v(\hat{Y})\right) = \alpha + \beta\log(\hat{Y})$$

The smoothed variances obtained $\tilde{\sigma}^2_{ea}$ were bias-corrected by applying a ratio adjustment of a factor equal to $\Delta = \frac{\sum \hat{\sigma}^2_{ea}}{\sum \tilde{\sigma}^2_{ea}}$ avoiding in this way overestimation or underestimation of the direct variances $\hat{\sigma}^2_{ea}$ (Beaumont and Bocci, 2016).

4. Estimation results

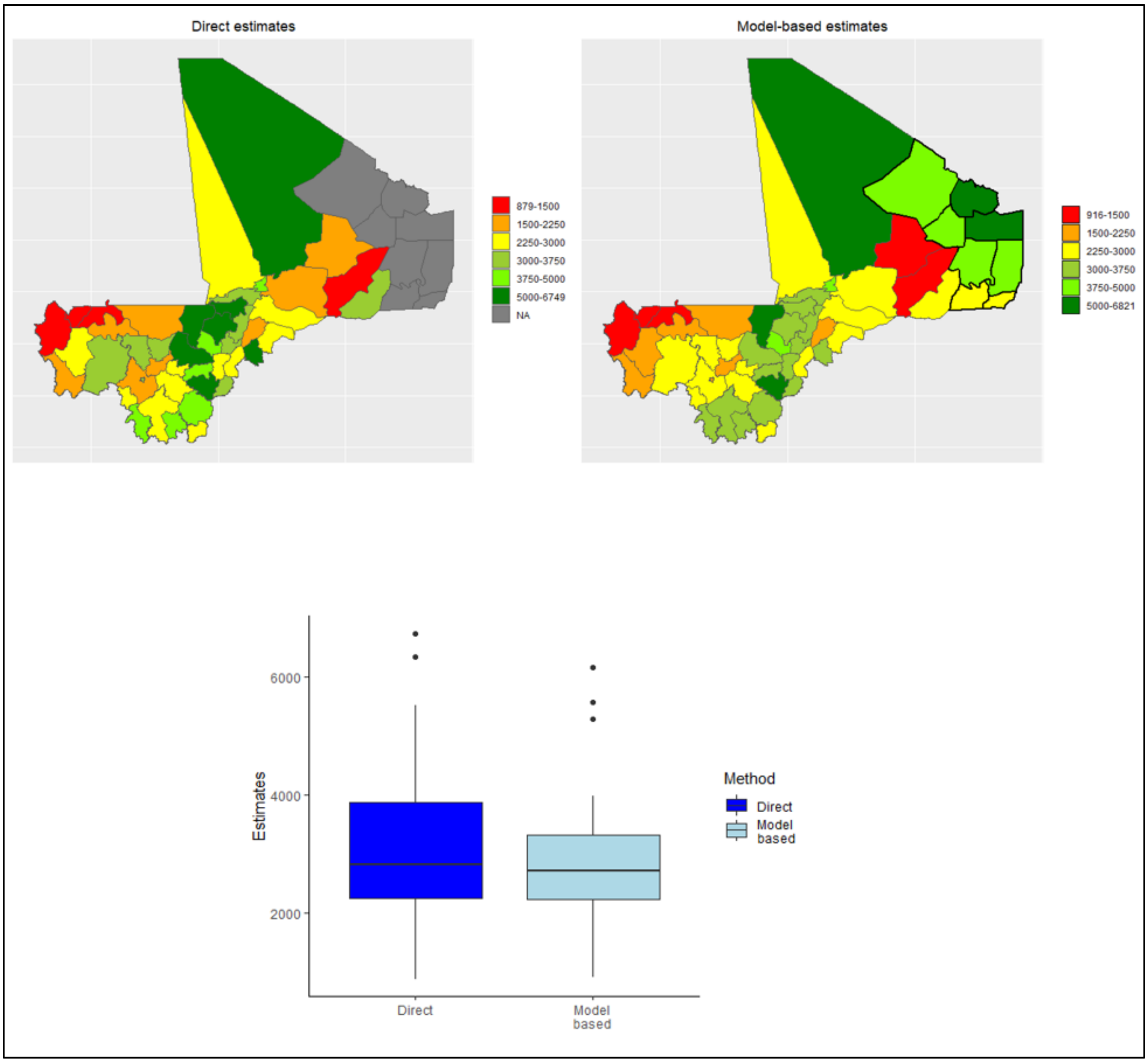
The main results of the study presented in Section 3 are here discussed by comparing the performance of direct and model-based estimators, and producing evidence on the validity of the model. The present section is complemented by Annex 1, displaying direct and model-based small area estimates of SDG Indicators 2.3.1 and 2.3.2 in Mali's circles along with the associated measures of reliability. In addition, Annex 2 provides the main R packages used for the analysis. For a step-by-step tutorial on the implementation of the case study, the authors refer to the recordings of a virtual training on data disaggregation and small area estimation for SDG indicators organized by the Office of the Chief Statistician of FAO in November 2022.[8]

Figure 3 and 4 respectively present the maps of direct and model-based estimates of Indicators 2.3.1 and 2.3.2 by circle. Both maps show how – using the FH – it is possible to produce predictions in out of sample domains.[9] Each figure also presents the distribution of direct and indirect estimates by means of boxplots. In this respect, the boxplots provide evidence of the fact that the implementation of model-based SAE techniques results in a reduction of the variability between estimates referred to different small areas or, in other terms, a smoothing of the variability of the phenomena between small areas.

[8] See the recordings of Day 3 of the Virtual Training on Data Disaggregation and Small Area Estimation for SDG Indicators that was organized by FAO, from 22 to 25 November 2022 (FAO, 2023).

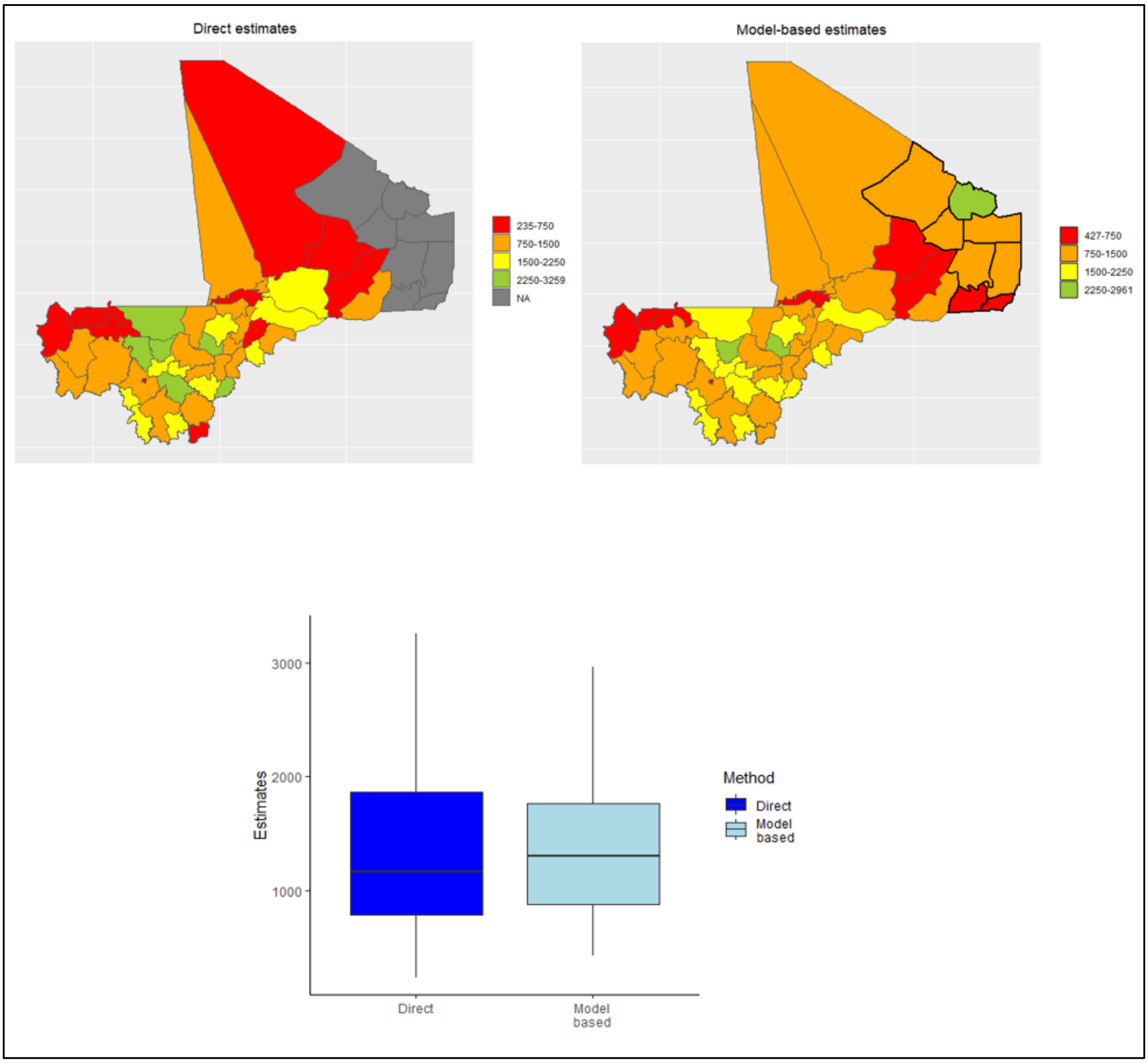
[9] Out of sample areas are colored with grey in the map of direct estimates and have ticker borders in the map of model-based estimates.

Figure 3: Direct and model-based estimates of SDG Indicator 2.3.1 by circle



Source: FAO. 2023. *Integrating surveys with geospatial data through small area estimation to disaggregate SDG indicators at subnational level. Case study on SDG Indicators 2.3.1 and 2.3.2.* Rome.

Figure 4: Direct and model-based estimates of SDG Indicator 2.3.2

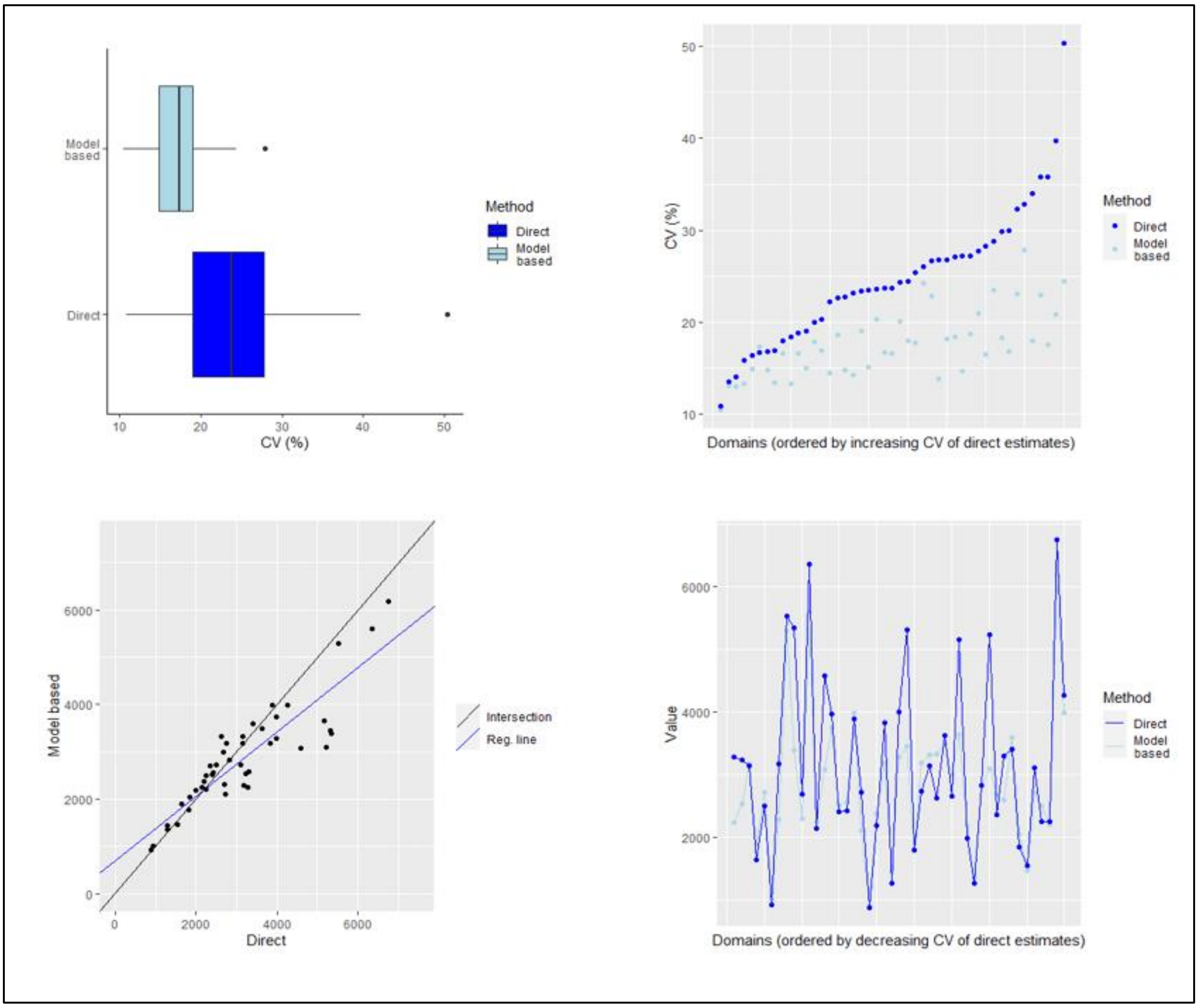


Source: FAO. 2023. *Integrating surveys with geospatial data through small area estimation to disaggregate SDG indicators at subnational level. Case study on SDG Indicators 2.3.1 and 2.3.2.* Rome.

Figure 3 and 4 provide complementary representations of estimates reliability by comparing the CV of direct and model-based estimates. The two boxplots in the top-left quadrant of both figures display the distribution of CVs of direct and model-based estimates and highlight the higher accuracy of small area estimates compared to their direct counterpart. Indeed, for what concerns Indicator 2.3.1, the CV of model-based estimates is below 0.2 in the 75 percent of the cases, while the same threshold is surpassed by more than the 50 percent of direct estimates. Similarly, the CV of small area estimates of Indicator 2.3.2 is below 0.3 in the 75 percent of the cases, while most direct estimates have a CV above this threshold.

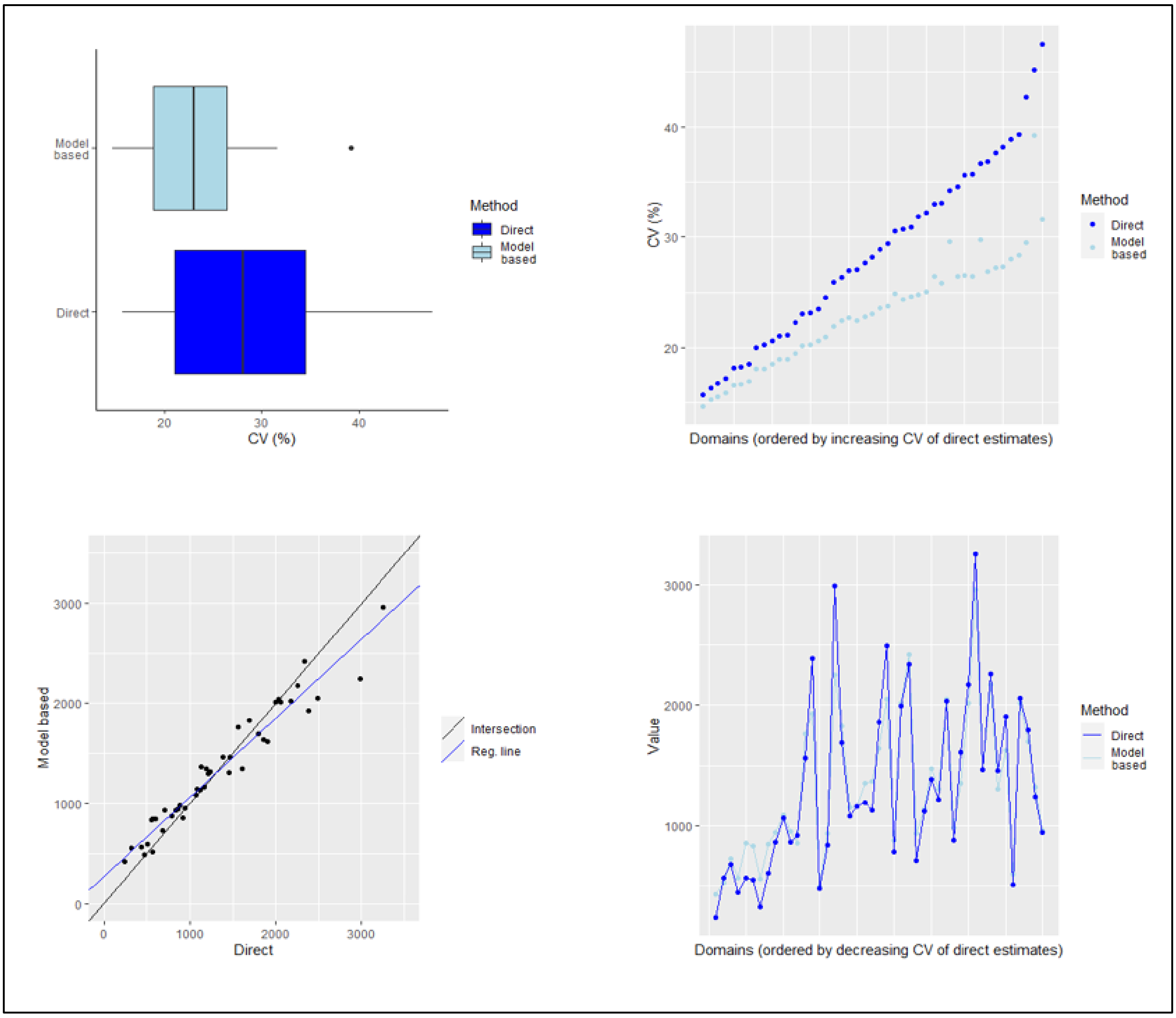
The plot in the top-right corner provides similar evidence, with direct and indirect estimates plotted by increasing values of their CV. The picture gives a visual indication of the fact that the CV of small area estimates falls always below the same variability measure referred to direct estimates, except in the very few cases where the domain direct estimates were already showing a high accuracy. The two graphs presented in the second row of Figure 5 and Figure 6 allow assessing the linear relationship between direct and indirect estimates. Generally speaking, and especially in correspondence of domains with higher sampling size, direct and indirect estimates are expected to be correlated. In other terms, it is highly desirable that the two approaches produce similar estimation results. In the considered case, the graphs illustrate a strong linear relationship between estimates produced with the two different approaches, with a correlation equal to 0.89 and 0.97 for Indicator 2.3.1 and 2.3.2 respectively.

Figure 5: Assessment of direct and model-based estimates – SDG Indicator 2.3.1



Source: FAO. 2023. *Integrating surveys with geospatial data through small area estimation to disaggregate SDG indicators at subnational level. Case study on SDG Indicators 2.3.1 and 2.3.2.* Rome.

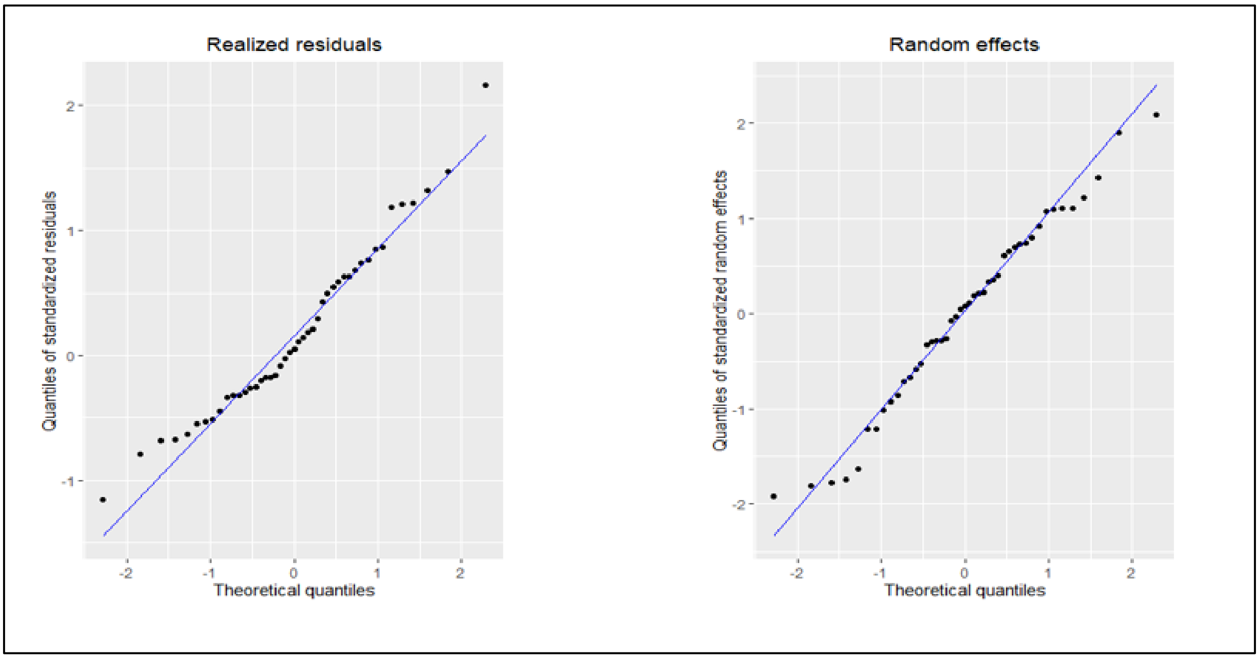
Figure 6: Assessment of direct and model-based estimates – SDG Indicator 2.3.2



Source: FAO. 2023. *Integrating surveys with geospatial data through small area estimation to disaggregate SDG indicators at subnational level. Case study on SDG Indicators 2.3.1 and 2.3.2.* Rome.

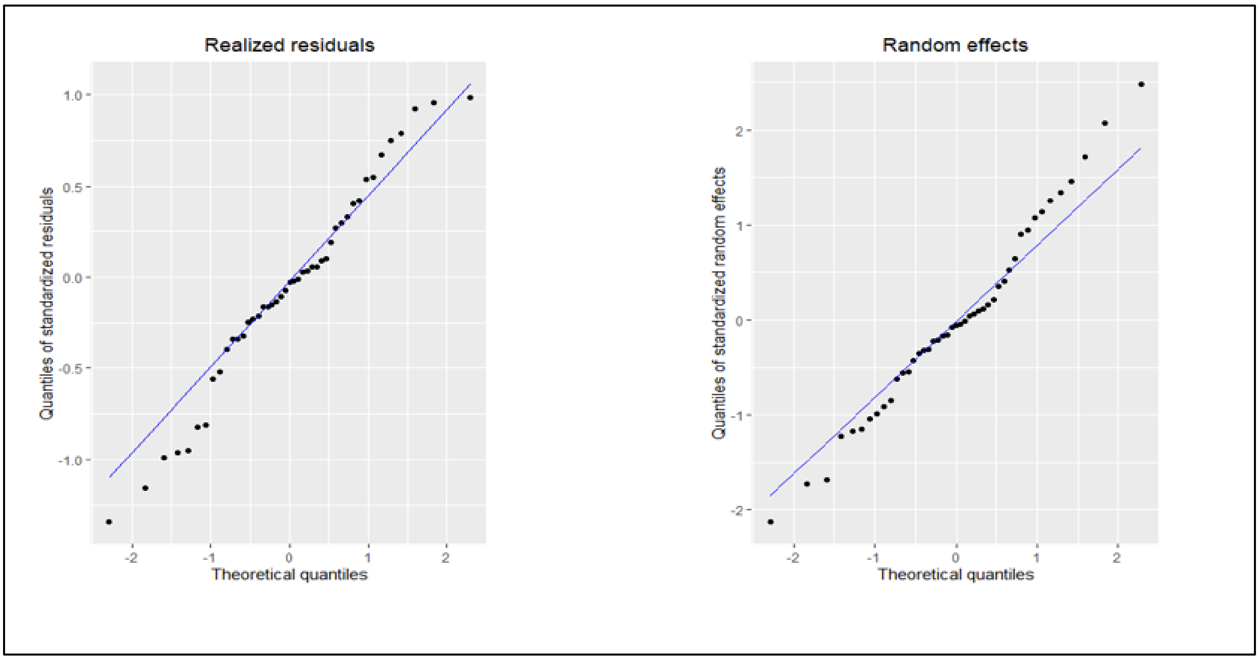
After assessing estimates reliability, an important component of SAE is the validation of the fundamental assumptions underlying the model, i.e. the normality of residuals and random effects. To that purpose, Figure 7 and Figure 8 present the QQ plots of both the error term and the random effects, which does not provide any significant proof of deviation from the normality assumption. This was also confirmed also by the Shapiro-Wilk test, which resulted in p-values above 0.05 for both the residuals and the random effects, leading to accept the null hypothesis of normality.

Figure 7: Quantile-quantile plots of residuals and random effects – SDG Indicator 2.3.1



Source: FAO. 2023. *Integrating surveys with geospatial data through small area estimation to disaggregate SDG indicators at subnational level. Case study on SDG Indicators 2.3.1 and 2.3.2.* Rome.

Figure 8: Quantile-quantile plots of residuals and random effects – SDG Indicator 2.3.2



Source: FAO. 2023. *Integrating surveys with geospatial data through small area estimation to disaggregate SDG indicators at subnational level. Case study on SDG Indicators 2.3.1 and 2.3.2.* Rome.

5. Conclusions

Monitoring the implementation of the 2030 Agenda for Sustainable Development and its overarching pledge to leave no one behind calls for more disaggregated data and SDG indicators than what is currently available in most countries. In this context, sample surveys are the preferred data source for about the 30 percent of indicators in the SDG monitoring framework and can offer valuable information to measure the social, economic and environmental dimensions of sustainable development. However, traditional household and agricultural surveys are usually characterized by sampling sizes that are either too small to produce precise estimates, or that do not cover all disaggregation domains of interest. Hence, indirect estimation approaches such as SAE techniques can represent a valuable tool for NSOs and international organizations to produce timely and granular disaggregated estimates of SDG indicators, without having to increase the survey sampling size. In particular, with the proliferation of new data sources such as geospatial and big data information systems, SAE models can be implemented by combining survey data with a vast amount of auxiliary information available at no or limited cost and at high frequency. In this respect, the body of literature and the number of case studies on SAE techniques applied to SDG indicators can still be expanded.

After a brief review of the methodology behind the computation of SDG Indicators 2.3.1 and 2.3.2 and the main SAE approaches available in the literature, this technical report presents a case study based on the FH area-level SAE model to produce subnational estimates of SDG indicators monitoring Target 2.3 at the level of Mali's circles . The case study considers the integration of survey data with area-level auxiliary information retrieved from multiple geospatial information systems. It shows how the model-based estimates of Indicators 2.3.1 and 2.3.2 in Mali's circles reach greater precision compared to direct estimates at the same level of disaggregation. In addition, adopting the considered indirect estimation approach, estimates for out of sample areas can also be produced.

The FH area-level model was selected in place of unit-level methods in order to provide a simple example of SAE based on SDG indicators related to the agricultural sector development, only requiring access to area-level direct estimates and auxiliary information. In addition, using an indirect estimator – such as the area-level EBLUP – expressed as a linear combination of area-level direct and synthetic estimates, the SAE approach can intuitively be interpreted as a way of improving direct estimates through a synthetic component based on external information correlated with the phenomenon of interest.

Annexes

Annex 1: Small area estimation results

Table A.1.1: Small area estimates for Indicator 2.3.1

Domain	Direct estimates	MSE direct estimates	CV direct estimates (%)	FH estimates	MSE FH estimates	CV FH estimates (%)
Kayes	879.25	52 525.79	26.07	916.28	4 9362.21	24.25
Bafoulabe	2 722.70	528 121.06	26.69	2 105.13	23 0635.73	22.81
Diéma	1 803.64	181 418.22	23.62	1 782.85	131 652.75	20.35
Kéniéba	1 987.45	162 480.24	20.28	2 194.30	138 108.17	16.94
Kita	3 174.13	1 049 586.32	32.28	2 290.58	280 169.35	23.11
Nioro	1 275.43	64 736.56	19.95	1 357.82	58 633.06	17.83
Yélimané	927.54	92 773.91	32.84	1 018.06	80 504.19	27.87
Koulikoro	2 247.26	100 429.23	14.10	2 211.84	82 113.23	12.96
Banamba	3 235.60	1 654 409.56	39.75	2 535.67	279 047.59	20.83
Dioila	2 404.33	424 290.69	27.09	2 520.92	214 831.38	18.39
Kangaba	2 255.09	128 281.46	15.88	2 494.96	109 626.55	13.27
Kati	2 182.93	308 577.72	25.45	2 371.24	176 672.90	17.73
Kolokani	3 292.31	352 179.23	18.03	2 588.48	185 101.39	16.62
Nara	1 638.25	343 174.52	35.76	1 897.77	189 090.28	22.91
Sikasso	3 991.86	899 223.46	23.76	3 286.56	299 263.88	16.65
Bougouni	2 743.19	414 279.55	23.46	3 182.48	231 932.27	15.13
Kadiolo	2 685.79	597 819.82	28.79	2 304.40	293 642.58	23.52
Kolondieba	3 822.95	871 867.45	24.42	3 183.60	328 096.76	17.99
Koutiala	5 529.24	2 746 170.85	29.97	5 286.18	794 217.75	16.86
Yanfolila	4 580.34	1 556 252.63	27.24	3 074.41	331 240.05	18.72
Yorosso	3 632.50	682 022.92	22.73	3 482.54	265 936.48	14.81
Ségou	5 154.36	1 311 467.91	22.22	3 646.34	278 586.79	14.48
Baraouéli	2 354.04	186 814.54	18.36	2 696.25	129 402.63	13.34
Bla	3 970.85	1 168 115.65	27.22	3 742.05	303 476.57	14.72
Macina	3 889.31	1 086 334.99	26.80	3 980.23	302 544.26	13.82
Niono	6 749.44	835 636.73	13.54	6 175.95	653 657.28	13.09
San	2 625.37	368 601.69	23.13	3 326.38	223 720.20	14.22
Tominian	2 500.89	720 785.39	33.95	2 719.89	239 764.38	18.00
Mopti	3 137.34	538 398.40	23.39	3 320.67	397 949.76	19.00
Bandiagara	1 851.83	96 949.19	16.81	2 047.44	91 960.26	14.81
Bankass	5 225.01	968 579.80	18.84	3 094.39	265 334.09	16.65
Djenné	3 143.19	1 263 571.44	35.76	3 176.30	309 613.19	17.52
Douentza	2 834.63	290 544.50	19.02	2 821.84	179 433.65	15.01
Koro	2 429.92	424 038.14	26.80	2 561.20	217 875.49	18.22
Ténenkou	5 310.81	1 591 321.05	23.75	3 457.43	335 049.91	16.74
Youwarou	5 339.18	2 534 326.10	29.82	3 391.29	384 643.21	18.29
Tombouctou	6 353.80	3 236 874.08	28.32	5 587.84	854 164.54	16.54

Domain	Direct estimates	MSE direct estimates	CV direct estimates (%)	FH estimates	MSE FH estimates	CV FH estimates (%)
Diré	4 259.60	212 647.17	10.83	3 987.93	173 318.61	10.44
Goundam	2 664.81	365 218.19	22.68	2 993.96	310 487.02	18.61
Gourma-Rharous	2 147.94	355 846.19	27.77	2 252.71	223 577.59	20.99
Niafunké	3 412.11	335 054.40	16.96	3 596.81	233 457.27	13.43
Gao	1 277.29	96 537.53	24.33	1 442.53	84 414.52	20.14
Ansongo	3 106.52	259 545.30	16.40	2 716.57	163 420.34	14.88
Bourem	1 548.99	66 653.68	16.67	1 473.40	65 667.05	17.39
Kidal				4 607.21	969 029.27	21.37
Abeibara				6 821.14	2 288 037.18	22.18
Tessalit				4 946.98	1 200 921.75	22.15
Tin-Essako				5 215.33	1 266 131.64	21.58
Bamako	3 283.24	2 735 010.33	50.37	2 244.01	299 783.86	24.40
Ménaka				2 880.45	410 563.84	22.24
Anderamboukane				2 911.18	406 467.37	21.90
Inekar				3 913.16	638 171.44	20.41
Tidermene				4 032.51	678 249.17	20.42

Source: FAO. 2023. *Integrating surveys with geospatial data through small area estimation to disaggregate SDG indicators at subnational level. Case study on SDG Indicators 2.3.1 and 2.3.2.* Rome.

Table A.1.2: Small area estimates for Indicator 2.3.2

Domain	Direct estimates	MSE direct estimates	CV direct estimates	FH estimates	MSE FH estimates	CV FH estimates
Kayes	472.93	2 4343.90	32.99	491.24	16 870.37	26.44
Bafoulabe	1 459.02	7 3130.86	18.53	1 306.32	48 985.30	16.94
Diéma	550.15	4 4084.29	38.16	832.96	51 701.41	27.30
Kéniéba	941.93	2 1761.07	15.66	954.94	19 568.20	14.65
Kita	1 079.49	110 390.07	30.78	1 149.48	78 476.24	24.37
Nioro	323.41	14 811.69	37.63	556.83	23 071.93	27.28
Yélimané	234.54	12 437.98	47.55	427.37	18284.94	31.64
Koulikoro	2 032.36	219 922.43	23.07	2 046.77	170 801.25	20.19
Banamba	2 343.43	381 386.89	26.35	2 420.38	294 550.94	22.42
Dioila	2 490.22	475 867.38	27.70	2 053.95	219 249.06	22.80
Kangaba	1 997.15	290 805.17	27.00	2 018.61	209 907.92	22.70
Kati	858.17	98 937.78	36.65	941.23	78 351.42	29.74
Kolokani	2 258.88	203 174.47	19.95	2 175.14	153 865.52	18.03
Nara	2 988.76	907 199.98	31.87	2 249.58	310 606.33	24.77

Domain	Direct estimates	MSE direct estimates	CV direct estimates	FH estimates	MSE FH estimates	CV FH estimates
Sikasso	1 161.56	126 366.55	30.60	1 161.69	83 685.25	24.90
Bougouni	1385.27	105 824.70	23.48	1 468.65	91 160.15	20.56
Kadiolo	563.64	47 902.64	38.83	851.73	57 024.62	28.04
Kolondieba	2 169.39	207 672.49	21.01	2 021.37	146 490.68	18.93
Koutiala	1 560.56	285 312.79	34.23	1 762.21	271 821.77	29.59
Yanfolila	1 692.69	273 493.61	30.90	1 827.88	202 075.99	24.59
Yorosso	2 389.03	626 358.81	33.13	1 926.56	247 751.46	25.84
Ségou	1 215.53	79 002.36	23.12	1 302.43	69 634.36	20.26
Baraouéli	2 057.48	124 702.70	17.16	2 017.03	102 379.10	15.86
Bla	1 129.04	106 581.61	28.92	1 364.33	103 268.28	23.55
Macina	3 258.70	450 138.01	20.59	2 960.97	300 609.69	18.52
Niono	879.75	38 235.09	22.23	979.61	36 245.05	19.43
San	1 236.31	40 672.24	16.31	1 322.56	40 578.22	15.23
Tominian	1 467.44	88 116.00	20.23	1 464.91	70 184.08	18.08
Mopti	1 120.48	75 428.29	24.51	1 135.31	56 434.17	20.92
Bandiagara	710.83	33 965.77	25.93	931.38	41 554.70	21.89
Bankass	1 796.73	90 669.08	16.76	1 699.49	69 542.77	15.52
Djenné	1 187.30	121 712.27	29.38	1 349.65	102 411.02	23.71
Douentza	1 906.06	120 759.89	18.23	1 622.94	73 467.85	16.70
Koro	785.10	45 187.84	27.08	878.23	38 721.03	22.41
Ténenkou	1 860.58	274 561.45	28.16	1 641.76	142 819.29	23.02
Youwarou	863.00	94 416.81	35.61	946.07	62 938.62	26.52
Tombouctou	605.84	49 764.32	36.82	846.58	51 685.38	26.85
Diré	836.94	72 553.08	32.18	933.78	54 704.84	25.05
Goundam	1 064.83	144 952.02	35.75	1 090.64	83 376.15	26.48
Gourma-Rharous	1 610.39	115 448.68	21.10	1 348.43	64 919.01	18.90
Niafunké	507.50	8 500.73	18.17	597.18	9 835.36	16.61
Gao	441.17	30 042.63	39.29	562.94	25 537.35	28.39
Ansongo	917.92	100 583.68	34.55	857.53	51 291.79	26.41
Bourem	676.95	83 618.27	42.72	728.82	46 210.89	29.50
Kidal				967.36	135 221.52	38.01
Abeibara				2 325.77	854 680.31	39.75
Tessalit				1 190.85	200 708.47	37.62
Tin-Essako				1 396.80	275 956.15	37.61
Bamako	562.12	64 399.12	45.14	522.29	41 906.64	39.20
Ménaka				625.26	66 143.55	41.13
Anderamboukane				602.84	62 715.76	41.54
Inekar				791.48	96 786.01	39.31
Tidermene				808.70	100 345.02	39.17

Source: FAO. 2023. *Integrating surveys with geospatial data through small area estimation to disaggregate SDG indicators at subnational level. Case study on SDG Indicators 2.3.1 and 2.3.2.* Rome.

Annex 2: R packages used for the case study

A2.1 R packages to treat geospatial variables

Geospatial information and remote sensing data are normally referred to spatially continuous phenomena typically represented through the so-called raster data model. This model allows partitioning the world into a grid of equally sized rectangles, referred to as cells or pixels. The open software R includes several packages allowing to read and process raster data. In particular, the present study relied on the ***raster*** package (Harmening *et al.*, 2022), which provides functions to implement the following steps:

- Crop and mask to select the portion of raster data that falls within the region under study (i.e. Mali), using the functions ***mask()*** and ***crop()***, respectively;
- Identify the cells falling in each circle using the function ***cellnumbers()*** from package ***tabularaster*** (Sumner, 2022) and then extract and aggregate the values associated to the cells by circle using the function ***extract()*** and ***groupby()*** followed by ***summarise()***, these last belonging to the ***dplyr*** package (Cran.r project, 2022).

A2.2 R packages to select auxiliary variables

The selection of auxiliary variables to be included in the SAE model has been performed using a stepwise selection approach. This method can be implemented using the *step* function included in the R package ***stats*** (ETHzurich, 2022), which, by default, allows selecting the best model based on the Akaike Information criterion (AIC). Among the various arguments to be specified in the function ***step()***, the “object” argument represents the model to be used by the stepwise selection approach as initial model. This object can be defined either with the function *lm*, for linear models, and *glm*, for generalized linear models.

A2.3 R packages to produce direct estimates

In order to produce direct estimates, one of the prerequisite actions is the selection of a suitable software package, allowing to consider complex sampling designs (such as those typically adopted in household and agricultural surveys) and producing estimates along with related accuracy measures. In this framework, a good option is provided by the R package **Regenesees** – **R evolved generalized software for sampling estimates and errors in surveys** – developed by the Italian National Statistical Institute. This package allows implementing design-based and model assisted analysis of complex surveys, and achieves a dramatic reduction in user workload for the production of estimates and error measures.

The package can be downloaded from the website of Italian National Institute of Statistics (Istat, 2023) along with a graphic user interface – ReGenesses.GUI – that enhance the usability of the package for non-expert R programmers.

Being Indicators 2.3.1 and 2.3.2 obtained as realization of ratio-type estimators, specific functions included in the ReGenesees package should be used. In particular, the function ***e.svydesign()*** has been used to specify the survey sampling design to be considered during the estimation process. After this step, the function ***svystatR()*** has been used to compute estimates, standard errors and confidence intervals for ratio-type estimators, such as Indicators 2.3.1 and 2.3.2.

A2.4 R packages to produce area-level SAE

The software package chosen to estimate the parameters of the FH model was the R package ***emdi*** (Harmening *et al.*, 2022)***.*** The function ***fh()*** of this package allows producing the area-level EBLUP estimates and their mean squared error. Additionally, the package provides function to implement various extension of the basic FH.

References

Ambrosio, F.L. & Iglesias Martínez L. 2000. Land cover estimation in small areas using ground survey and remote sensing. *Remote Sensing of Environment*, 74(2): 240–248.

Arima, S., Bell, WR, Datta, G.S., Franco C. & Liseo, B. 2018. Multivariate Fay-Herriot Bayesian estimation of small area means under functional measurement error model. *Journal of the Royal Statistical Society*, 180(4): 1191–1209.

Battese, G.E., Harter, R.M. & Fuller, W.A. 1998. An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*. 83(401): 28–36.

Bedi, T., Coudouel, A. & Simler, K. 2007. *More Than a Pretty Picture: Using Poverty Maps to Design Better Policies and Interventions.* Washington, DC, World Bank.

Beaumont, J.F. & Bocci, C. 2016. *Small area estimation in the Labour Force Survey.* Ottawa, Advisory Committee on Statistical Methods, Statistics Canada.

Brackstone, G.J. 1987. Small Area Data: Policy Issues and Technical Challenges. In: R. Platek, J.N.K. Rao, C.-E. Sarndall, & M.P. Singh, eds. *Small Area Statistics*, pp. 3–20. New York, John Wiley & Sons, Inc.

Breidenbach, J., Magnussen, S., Rahlfa, J. & Astrupa, R. 2018. Unit-level and area-level small area estimation under heteroscedasticity using digital aerial photogrammetry data. *Remote Sensing of Environment*, 2(12): 199–211.

CGIAR Consortium for Spatial Information (CGIAR-CSI). 2022. CGIAR-CSI homepage. In: *CGIAR-CSI website*. Cited 01 December 2022. https://cgiarcsi.community/

Cochran, W.G. 1977. *Sampling Techniques*. New York, John Wiley & Sons.

Cran.r project. 2022. A Grammar of Data Manipulation. In: *Package 'dplyr'*. Cited 01 December 2022. https://cran.r-project.org/web/packages/dplyr/dplyr.pdf

Elbers, C., Lanjouw, J. & Lanjouw, P. 2003. Micro-level estimation of poverty and inequality. *Econometrica*, 7(1): 355–364.

Erciulescu, A.L, Franco, C. & Lahiri, P. 2021. Use of administrative records in small area estimation. Administrative records for survey methodology. *John Wiley & Sons*, 231–267.

ETHzurich. 2022. The R Stats Package. In: *Stat.Ethz.* Cited 01 December 2022. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html

Eurostat (European Commission). 2019. *Guidelines on small area estimation for city statistics and other functional geographies.* Luxembourg.

Fay, R.E. & Herriot, R.A. 1979. Estimates of income for small places: An application of James-Stein procedures to census data. Journal of the American Statistical Association. 74(366): 269–277.

Falorsi, P.D., Donmez, A., Khalil, C.A., Di Candia, S. & Gennari, P. 2022. Alternative Methods for Disaggregating Sustainable Development Goal Indicators Using Survey Data. *Statistical Journal of the IAOS.* 1(22): 611–623.

FAO. 2019. *Methodology for computing and monitoring the Sustainable Development Goal Indicators 2.3.1 and 2.3.2*. FAO Statistics Working Paper Series, No. 18–14. Rome, FAO.

FAO. 2021. *Guidelines for data disaggregation of SDG indicators using survey data.* Rome.

FAO. *An indirect estimation approach for disaggregating SDG indicators using survey data. A case study based on SDG Indicator 2.1.2*. Rome.

FAO. 2023. Virtual Training on Data Disaggregation and Small Area Estimation for SDG indicators (22-25 November 2022). In: *FAO Sustainable Development Goals website*. Cited 01 December 2022. https://www.fao.org/sustainable-development-goals/events/detail/en/c/1618214/

Harmening, S., Kreutzmann, A.K., Pannier, S., Skarke, F., Rojas-Perilla, N., Salvati, N., Schmid, T., Templ, M., Tzavidis, N. & Würz, N. 2022. Estimating and Mapping Disaggregated Indicators. In: *Package 'emdi'.* Cited 01 December 2022. https://cran.r-project.org/web/packages/emdi/emdi.pdf

Harville, D.A. 1991. That BLUP is a Good Thing: The Estimation of Random Effects. Comment. *Statistical Science*, 6: 35–39.

Horvitz, D.G. & Thompson, D.L. 1952. A generalization of sampling without replacement from finite universe. *Journal of the American Statistical Association*, 47(260): 663–685.

Isric World Soil Information. 2022. SoilGrids, global gridded soil information. In: *ISRIC World Soil Information*. Wageningen, The Netherlands. Cited 01 December 2022.
https://www.isric.org/explore/soilgrids

Italian National Institute of Statistics (Istat). 2023. Regenesees (r evolved generalised software for sampling estimates and errors in surveys). In: *Regenesees.* Rome, Istat. Cited 01 December 2022. https://www.istat.it/en/methods-and-tools/methods-and-it-tools/process/processing-tools/regenesees

Khalil, C.A, Conforti P., Ergin, I. & Gennari P. 2017. *Defining small-scale food producers to monitor target 2.3 of the 2030 Agenda for Sustainable Development.* FAO Statistics Working Paper Series. No. 17–12. Rome, FAO.

Khalil, C.A., di Candia, S., Falorsi, P.D. & Gennari, P. 2022. Integrating surveys with geospatial data through small area estimation to disaggregate SDG indicators: A practical application on SDG Indicator 2.3.1. *Statistical Journal of the IAOS,* 38(22): 879-891.

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

Kish, L. 1987. *Statistical Design for Research*. New York, John Wiley & Sons.

Mapsam. 2022. Mapsam homepage. In: *Mapsam*. Cited 01 December 2022.
https://www.mapspam.info/data/

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giovannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L. & Gabrielli, L. 2015. Small area model-based estimation using big data sources. *Journal of Official Statistics*. 31(2): 263–281.

Marhuenda, Y., Molina, I. & Morales, D. 2013. Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis*, 58: 308–325.

Molina, I. & Rao, J.N.K. 2010. Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, 38: 369–385.

NASA EarthData. 2022. Your Gateway to NASA Earth Observation Data. In: NASA. Cited 01 December 2022. https://www.earthdata.nasa.gov/

Petrucci, A., Salvati, N. 2006. Small Area Estimation for spatial correlation in watershed erosion assessment. *Journal of Agricultural, Biological and Environmental Statistics*. 11(2): 169–182.

Porter, A.T., Holan, S.H., Wikle, C.K. & Cressie, N. 2014. Spatial Fay-Herriot models for small area estimation with functional covariates. *Spatial Statistics*, 14(10): 27–42.

Rao, J.N.K. 2003. *Small Area Estimation*. Wiley Series in Survey Methodology. New York, Wiley.

Rao, J.N.K. & Molina, I. 2015. *Small Area Estimation*. New York, Wiley.

Särndal, C.E., Swensson, B. & Wretman, J. 1992. *Model Assisted Survey Sampling*. Berlin, Springer-Verlag.

Schoch, T. 2022. Robust unit-level small area estimation: A fast algorithm for large data sets. *Austrian Journal of Statistics.* 41(4): 243–265.

Singh, R., Semwal, D.P., Rai, A. & Chhikara, R.S. 2022. Small area estimation of crop yield using remote sensing satellite data. *International Journal of Remote Sensing*. 23(1):49-56.

Solargis. 2022. Solargis homepage. In: *Solargis website*. Cited 01 December 2022. https://solargis.com/

Sumner, M.D, 2022. Tidy Tools for 'Raster' Data. In: *Package 'tabularaster'*. Cited 01 December 2022. https://cran.r-project.org/web/packages/tabularaster/tabularaster.pdf

United Nations Department of Economic and Social Affairs (UNDESA). 2022. Data Disaggregation for the SDG Indicators. In: *Inter-Agency and Expert Group on Sustainable Development Goal Indicators*. New York, UNDESA. Cited 01 December 2022. https://unstats.un.org/sdgs/iaeg-sdgs/disaggregation/

UNDESA. 2022. Producing small area estimation. In: *UN Statistics wiki*. New York, UNDESA. Cited 01 December 2022. https://unstats.un.org/wiki/display/SAE4SDG/Producing+SAE

Wolter, K.M. 2007. *Introduction to Variance Estimation. Second Edition.* Statistics for Social and Behavioral Sciences. Springer Series in Statistics. New York, Springer.

World Bank. 2023. World Bank Microdata Catalogue. In: *The World Bank*. Washington, DC. Cited 01 December 2022. https://microdata.worldbank.org/index.php/catalog/3409

WorldClim. 2022. Historical monthly weather data. In: *WorldClim website*. Cited 01 December 2022. https://www.worldclim.org/data/monthlywth.html

Ybarra, L.M.R. & Lohr SL. Small area estimation when auxiliary information is measured with error. Biometrika. 2012; 95(4): 919–93.

You, Y. & Rao, J.N.K. 2002. A pseudo empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*. 30: 431–439.

Zenodo. 2022. Zenodo Recent uploads. In: *Zenodo website*. Cited 01 December 2022. https://zenodo.org/

Zhang, L.C. & Giusti C. 2016. Small Area Methods and Administrative Data Integration. New York, John Wiley & Sons.

ISBN 978-92-5-137545-7

9 789251 375457

CC3944EN/1/01.23