Dramane Bako

Integrated sampling design for agricultural and socio-economic households surveys: a cost effective approach for agricultural and rural statistics

Dramane Bako[1]

[1] Statistician, Food and Agriculture Organization of the United Nations (FAO), Rome, Italy, Dramane.Bako@fao.org

Abstract:
The 50x2030 Initiative proposes an integrated modular agricultural and rural survey program that promotes an integration of traditional socio-economic household survey and agricultural surveys in beneficiary countries. This approach allows for the analysis of the drivers of productivity and the interactions between households' socio-economic characteristics, agricultural production methods, off-farm activities, and the environment with agricultural activities, amongst others, speaking to the needs of different data users. This integrated approach produces richer data, increases data interoperability, and is more cost-effective. An integrated sampling design is proposed to ensure that the integrated survey fulfils the measurement objectives of the traditional surveys in a cost-effective way.
This communication will present an overview of the key technical features of the proposed integrated sampling procedures including the development of sampling frames, stratification criteria, sampling size calculations and estimation procedures. The strengths, weaknesses and challenges of the methodology will be discussed.

Keywords:
Sampling; data integration; agricultural statistics; household survey;

1 Introduction:

National statistical offices are facing more and more increased surveys costs and decline of participation and response rates in traditional surveys (Groves, 2011). At the same time, there is an increased data demand from various users in public and private sectors. A solution to this challenge is the integration of data from a variety of sources, providing the potential to produce timelier, more disaggregated statistics at higher frequencies than traditional approaches alone (UNECE, 2018). Data integration is a broad and emerging topic in survey research covering the integration of survey and administrative or big data as well as integrating of data from different surveys.
The 50x2030 Initiative to Close the Agricultural Data Gap (hereinafter "50x2030 Initiative") promotes an integrated approach to the agricultural survey system. The Initiative proposes an Integrated Agricultural and Rural Survey Program aiming at integrating socio-economic and environmental data with agricultural data. This approach allows for the analysis of the drivers of productivity and the interactions between households' socio-economic characteristics, agricultural production methods, off-farm activities, and the environment with agricultural activities, amongst others, speaking to the needs of different data users. The integrated approach greatly increases the value of agricultural data beyond production of basic macro-indicators (50x30 Initiative, 2020a). Integrated survey instruments were designed (50x30 Initiative, 2020b) as well as an integrated sampling design (50x30 Initiative, 2020c) to ensure the cost effectiveness of the program. The main objective of this paper is to present and discuss key technical features of the proposed integrated sampling design.

2 Overview of surveys integration

Three possible actions could be considered for integrating data from different surveys (D'Orazio, 2021): modify/enlarge existing surveys (integrate survey instruments), set-up new sample surveys (integration samples) and Integrate data of existing surveys (microdata integration). All these options are envisaged in the 50x2030 Initiative for the Integrated Agricultural and Rural Survey Program.

2.1 Integration of surveys instruments

This is one of the most common solutions discussed in the literature on data integration. It consists in harmonising and integrating the questionnaires of the surveys: standardisation of concepts, definitions, identifiers, codification… and avoiding duplications. Main objectives include generating a coherent set of data, comparable over time and across sources when relevant, and ensuring cost-effectiveness of data production (IHSN, 2021). The length of the integrated questionnaires should be considered carefully to avoid excessive burden on respondents. The 50x2030 Initiative elaborated integrated survey instruments for the recommended integrated survey programs (see 50x30 Initiative (2020b) and https://www.50x2030.org/resources/survey-instruments)

2.2 Integration of surveys microdata through statistical matching:

The goal is to investigate the relationship between variables not jointly observed in a single data source (D'Orazio, 2019). The matching can be deterministic if the surveys adopted similar identifiers for the units. In most cases, that is not the case and probabilistic matching is considered. To provide reliable results statistical matching have a number of underlying assumptions to be fulfilled by the datasets; some of them do not hold in most of real cases. When the microdata integration is envisaged, suitable actions can be taken when designing the surveys to make the statistical matching more effective (D'Orazio, 2021).

2.3 Integration of surveys samples

The integration here is performed at the level of the sampling designs. An integrated sampling strategy is elaborated for the different surveys improving data consistency and facilitating cross-survey data analysis especially when well integrated survey instruments are considered. The 50x2030's integrated sampling design (50x30 Initiative, 2020c) complements the integrated questionnaires proposed by the initiative in a consistent way providing operational tools and methods to countries for a cost effective implementation of the integrated agricultural and rural survey program. Integrating sampling designs can be cost effective at an important extent when some conditions are fulfilled including:

- *Overlap among target populations*

In presence of overlap among the populations of interest of the surveys, the integration of the samples presents a number of potential advantages:

  - ✓ Reduction of cost for sampling frame development. This is an obvious advantage if the sampling populations are identical. Even in case sampling units are different, a common master sampling frame could be explored in presence of overlap between populations of interest.
  - ✓ It also offers the possibility to administer the surveys on overlapping samples, reducing the cost of survey implementation

- *Use of similar sampling method*

In case a similar sampling approach is used by the different surveys or considered adequate for them, there is a potential cost reduction for sample selection with the integration of the samples. For instance, if all surveys use multistage sampling design, there is a possibility to consider a unique sample of primary sampling units for the screening operations usually performed before selecting the final sampling units.

- *Complementarity/linkages of measurements objectives*

When the measurement objectives of the different surveys are interlinked, the integration of the samples would offer additional analytical advantages with the possibility to perform cross-survey analyses.

3 Integrated sampling design

The first step for the development of the integrated sampling design was a review of the common features of the sampling designs recommended and used by countries for the two types of surveys (agricultural and socio economic surveys). Then, features relevant for both types of surveys are discussed considering operational issues (feasibility/cost) and efficiency before deciding on the

integrated design. An important quality requirement is that the integration of the samples should not affect the reliability of the keys estimates usually expected by the users from the different surveys.

3.1 Common sampling designs used for agricultural and socio economic surveys

In a nutshell, socio economic surveys consider households as sampling and observation units. A stratified two-stage sampling design is usually used the enumeration areas (designed for population and housing censuses) as primary sampling units (PSUs). Listing operations are performed in the sampled PSUs before selecting the final sample of households. Administrative zones and urban/rural localization are the common stratification criteria.

In agricultural surveys, agricultural holdings are observation units and the sampling method depends on the type of sampling frame (list or area frame) and the sector: household sector (farms operated by households) or non-household sector (farms operated by corporations, associations…). With a list frame, the sampling units in the household sector are agricultural households usually selected through a two stage sampling method. Enumeration areas from agricultural or population census are used as PSUs and stratification criteria are usually administrative zones and urban/rural localization and agro ecological zones. In the non-household sector, a stratified single stage sampling method is considered with a list frame usually developed from registers (business/commercial farms register) and/or list from agricultural census.

The area frame covers all holdings operating agricultural land and should be complemented with a list of landless holdings raising livestock for a full coverage. With an area frame, single or multi stage sampling method are usually used for selecting segments or geographical points to reach the final sample of farms for the survey.

3.2 Proposed integrated sampling design

The integrated sampling design proposed for the integrated agricultural and rural survey program is described here. The integration is proposed in the household sector which is covered by both types of surveys. Farms operated by non-household entities (corporations, government institutions, cooperatives, etc.) shall be covered using the recommended sampling design mentioned in section 4.1 above.

3.2.1 Populations of interest

The integrated survey program aims at producing statistics on the country's agricultural sector and rural households. As already mentioned above, units observed in agricultural surveys are agricultural holdings and in the household sector they are sampled through agricultural households. Therefore the target populations, in line with the measurement objectives, are: (i) rural households, (ii) urban agricultural households and (iii) farms operated by non-household entities.

3.2.2 Sampling method and frame

A common sampling method recommended for both agricultural surveys (in the household sector) and socioeconomic surveys is the two stage sampling design using the list of enumeration areas from the most recent population and housing census (PHC) as sampling frame of PSUs. That sampling method and frame is proposed for the integrated sampling design in rural area. There is no integration in urban area as non-agricultural urban households are not part of the target population of the 50x2030 initiative. The urban area shall be covered only by agricultural surveys and in countries where urban agriculture is important. The sampling method for urban agricultural households will be country specific depending on their numbers and distribution to be checked with data from the most recent population and housing census that shall also be used for developing the sampling frame.

3.2.3 Sample size

The determination of the size of the integrated sample should be considered carefully to ensure that the final sample includes the minimum number of units required for the measurement objective of each survey as well as the integrated survey. The 50x2030 Initiative promotes the calculation of the sample size based on the analytical requirements of the survey, i.e., it ensures the reliable estimation of key variables of interest. Traditionally, agricultural surveys consider agricultural production or area while socio economic surveys consider households income or consumption when calculating the minimum sample size.

In rural areas, the Integrated Agricultural and Rural Survey Program has two main estimation goals: producing estimates for the whole population of rural households, and estimates for the subset of agricultural households at the national and sub-national level. To meet these objectives, the optimal sampling strategy would require a complete list of rural households from a recent PHC, classed according to whether they are agricultural (denoted as A from now on) or non-agricultural (denoted as B).

In the integrated survey, the household-sector sample size should ensure reliable estimation of a key household-related variable (e.g., income) in the population of rural households (A and B), and reliable estimation of a key agricultural variable (e.g., agricultural area) from the sub-population of agricultural households (A) as households in subpopulation B do not operate agricultural land. To calculate the minimum sample size of households needed to fulfil this goal, the usual approximate formula based on the coefficient of variation can be used.

Let us consider for each estimation domain U_d:

- M_{Ad} and M_{Bd} is the total number of households respectively of type A and B.
- cv^2_{Aincd} and cv^2_{Bincd} is the coefficient of variation of income of households of type A and B, respectively
- cv^2_{Aland} is the coefficient of variation of agricultural area of the agricultural household
- cv^{*2}_d is the maximum acceptable relative error for estimating the total (average) income and agricultural area
- $\widetilde{deff}_{Aincd}$, $\widetilde{deff}_{Bincd}$ and $\widetilde{deff}_{land}$ are estimates of the design effect for income of households of type A and B and agricultural area, respectively
- g is the expected response rate

The minimum sample size of households (m_d) in the domain U_d is:

$$m_d = \frac{1}{g}\left[Max\left(\widetilde{deff}_{land}\frac{cv^2_{Aland}}{cv^{*2}_d + \frac{cv^2_{Aland}}{M_{Ad}}}, \widetilde{deff}_{Aincd}\frac{cv^2_{Aincd}}{cv^{*2}_d + \frac{cv^2_{Aincd}}{M_{Ad}}}\right) + \widetilde{deff}_{Bincd}\frac{cv^2_{Bincd}}{cv^{*2}_d + \frac{cv^2_{Bincd}}{M_{Bd}}}\right]$$

Or:

$$m_d = \max\left(m_{dA,inc}, m_{dA,land}\right) + m_{dB,inc} = m_{dA} + m_{dB,inc}$$

This procedure requires having all the variables in the formula for household types A and B (agricultural and non-agricultural rural households) in each domain *d*. However, it may happen that the coefficient of variation of the income cannot be estimated for each subpopulation if the exercise is undertaken with data from a household survey that did not cover agricultural activities. In such case, if $m_{d,inc}$ is the overall minimum size of rural households for a reliable estimate of the income, we have:

$$m_{d,inc} = \frac{1}{g}\widetilde{deff}_{incd}\frac{cv^2_{incd}}{cv^{*2}_d + \frac{cv^2_{incd}}{M_{Ad} + M_{Bd}}}$$

And:

$$m_d = \max\left(\widetilde{W}_{Ad}m_{d,inc}, m_{d,land}\right) + (1 - \widetilde{W}_{Ad})m_{d,inc}$$

Where:

- cv^2_{incd} is the coefficient of variation of the income of rural households in the domain *d*
- $\widetilde{deff}_{incd}$ is an estimate of the design effect for the income of rural households
- $\widetilde{W}_{Ad}$ is an estimate of the proportion of agricultural households in the domain *d*.

3.2.4 Stratification

Stratification can contribute at an important extent to improve the accuracy of estimates. There is usually a distinction between design strata (used mainly for improving estimates) and analytical strata also called domains of inference (usually administrative zones considered for reporting purposes). In the framework of integrating surveys, if they do not have similar design or analytical strata, considering all different stratification criteria in the integrated survey would lead to too many strata, which is unnecessary (Cochran, 1977). A solution is to identify stratification criteria that are suitable for the different surveys.

Area units like enumeration areas or villages usually present relatively low within variance of key households' variables because of geographical proximity. When used as PSU in multi stage sampling, an important proportion of the sampling variance would consist in the variance between the PSUs. A proper stratification of the PSUs is therefore important to reduce sampling variance. FAO (2017) recommends a stratification of the EAs by administrative zones (e.g., regions, provinces, etc.) and agro-ecological zones. This should happen prior to the first-stage selection, in order to improve the estimates of agricultural statistics. Stratification of PSUs should be carefully controlled, since having too many strata is not desirable (an independent sample has to be selected in each stratum). To avoid too many strata, explicit stratification can be coupled with implicit stratification. This consists of sorting the sampling frame by relevant criteria (usually geographical) in each stratum and selecting an independent sample in each stratum with systematic sampling.

As previously stated, a two stage sampling method is suggested for the integrated sampling design with the list of enumeration areas from the most recent PHC as sampling frame of PSU. In most cases, the list of households from the most recent PHC would be outdated (or difficult to obtain in some countries). Therefore, the actual structure of the households within the sampled PSUs can be known only after a fresh listing of households in these PSUs. A major drawback is the lack of control over the final sample, especially the number of agricultural households required in the domain (as calculated in section 4.2.3). Since the selection is made at the level of PSUs, it may show a varying situation in terms of the proportion of agricultural households.

To maintain control of the final sample size by household type (A and B), it is preferable to make a first-level stratification of the EAs in terms of the proportion of agricultural households in each of them, estimated from the latest PHC or other suitable source.

Even if the PHC data is considered outdated, this structural information (proportion of agricultural households) is not likely to vary much in all PSUs and could be helpful for stratification purposes. The first-level stratification below may be considered using a proportion threshold ρ $(\frac{1}{2} < \rho < 1)$.

First-level PSU strata	**Definition**
Agricultural	$Proportion\ of\ agricultural\ households\ in\ the\ PSU\ \geq\ \rho$
Mixed	$1-\rho <\ Proportion\ of\ agricultural\ households\ in\ the\ PSU\ <\rho$
Non-agricultural	$Proportion\ of\ agricultural\ households\ in\ the\ PSU\ \leq\ 1-\rho$

The sample of PSUs in the domain d (n_d) can be allocated using parameters θ_a, θ_m and θ_{na} with $\theta_a + \theta_m + \theta_{na} = 1$

First-level allocation	
First-level PSU strata	**Allocation of the sample of PSU**
Agricultural	$\theta_a n_d$
Mixed	$\theta_m n_d$
Non-agricultural	$\theta_{na} n_d$

If m_0 households will be selected in each sampled PSU using a systematic or simple random sampling without replacement, the expected number of agricultural households in the final sample (m_{dAexp}) is:

$$m_{dAexp} = \rho m_0 \theta_a n_d + (1-\rho) m_0 \theta_m n_d + \delta m_0 \theta_{na} n_d = (\rho\theta_a + (1-\rho)\theta_m) m_0 n_d + \delta\theta_{na} m_0 n_d$$

$\delta < 1$ is unknown before the selection of the sample of households, contrary to the other parameters that are fixed by the sample designer.

Let us consider τ the proportion of agricultural households in the planned sample:

$$\tau = \frac{m_{dA}}{m_d} = \frac{m_{dA}}{m_0 n_d} \Longrightarrow m_{dA} = \tau m_0 n_d$$

To ensure the achievement of the planned sample of agricultural households in the final sample of households, parameters θ_a, θ_m and θ_{na} could be fixed to have $m_{dAexp} \geq m_{dA}$. That corresponds to:

$$(\rho\theta_a + (1-\rho)\theta_m)m_0 n_d + \delta\theta_{na} m_0 n_d \geq \tau m_0 n_d$$

δ being unknown, parameters θ_a, θ_m and θ_{na} can therefore be fixed under the following conditions:

$$\rho\theta_a + (1-\rho)\theta_m \geq \tau$$

$$\theta_{na} = 1 - (\theta_a + \theta_m)$$

This first level stratification criterion is obviously relevant for agricultural aggregates and would be suitable for socio economic surveys in most cases. In fact, an important stratification criterion for those late surveys is the urban/rural localisation and proportions of agricultural households tend to be high in rural areas and low in urban ones. In any case, an assessment of the correlation between the proportions of agricultural households in EAs and their localisation in urban/rural area would help to confirm the suitability of that proposed first-level stratification for the socio economic survey as well. If not suitable stratification should be considered at a second level for that survey.

A second-level stratification of PSUs may be performed inside the first-level strata. Common stratification criteria for improving estimates in agricultural and household surveys are: agro-ecological zones; urban/rural localisation, land use classes; size categories based on population; agricultural area; intensity of agricultural activity, etc.). The allocation in these second-level strata can follow different criteria. Typically, in household surveys an allocation proportional to the population in the strata is considered. FAO (2017) recommends the Neyman's optimum allocation for agricultural surveys. Kish (1987, page 228) suggests a compromise solution between equal and proportional allocation:

$$n_{dh} = n_d \times \frac{\theta_{dh}}{\sum_{h=1}^{H_d} \theta_{dh}}$$

Where:

$$\theta_{dh} = \sqrt{\left(W_{dh}^2 + \frac{1}{H_d^2}\right)}$$

H_d is the number of strata in the domain *d*, while W_{dh} is the relative size of stratum *h* in domain *d*, it can be the proportion of PSUs in stratum *h* compared to the domain total, $W_{dh} = N_{dh}/N_d$, (relative size in terms of population). A multivariate stratification and allocation (Barcaroli, 2020) or compromise power allocation (Bankier, 1988) could also be explored if the frame contains relevant variables correlated with households' income or agricultural area (household size, livestock, agricultural production, etc.) at PSU level.

The table below presents the main components of the integrated design discussed in this section.

Table 2. Summary of the major elements of the sampling design for the Integrated Program

Items	**Populations of interest**		
	Household (rural)	**Household (urban)**	**Non-household sector**
Observation units	- Households - Agricultural holdings	Agricultural holdings	Agricultural holdings
Final sampling units	Households	Households	Agricultural holdings

Frames	List of households from population census or list of EAs from population census and microcensuses in sampled EAs	List of households from population census	List of non-household farms developed from registers and/or field operations
Sampling method	Stratified two-stage	Country specific: Stratified one-stage or two-stage	Stratified one-stage
Stratification	Country specific: - PSU-level strata: administrative zones; agro-ecological zones; intensity of agricultural activity using land use data; proportion of agricultural households - SSU-level strata (intra-PSU): practice of agriculture	Country specific: administrative zones; agro-ecological zones	Country specific: administrative zones; production systems (crop/livestock/mixed); ad-hoc categorization, e.g., strata based on a measure of size (e.g., value of production)
Sampling scheme	1st stage: PPS of PSUs (EAs) 2nd stage: Systematic or Simple random sampling without replacement of Households	Country specific: depending on the sampling method adopted	Systematic or Simple random sampling without replacement within each stratum

References:

- 50x30 Initiative. 2020a. An introduction to the 50x2030 Initiative. Technical Paper Series #1. World Bank, Rome.
- 50x30 Initiative. 2020b. A Guide to the 50x2030 Data Collection Approach: Questionnaire Design. Technical Paper Series #2. Rome.
- 50x30 Initiative. 2020c. A Guide to sampling. Technical Paper Series. Rome.
- Bankier, M.D. 1988. "Power allocations: determining sample sizes for subnational areas." The American Statistician 42, 174-177.
- Barcaroli, G. Ballin, M. Odendaal, H. Pagliuca, D. Willighagen, E. Zardetto D. 2020. SamplingStrata: optimal stratification of sampling frames for multipurpose sampling surveys, R package. Version 1.5-1.
- Cochran, W.G. 1977. Sampling Techniques. 3rd Edition. John Wiley & Sons: New York, USA.
- D'Orazio, M. 2019. Statistical learning in official statistics: The case of statistical matching. Statistical Journal of the IAOS, 35(3), pp. 435-441. DOI: 10.3233/SJI-190518.
- D'Orazio, M. 2021. Integration of Household Survey Data through Statistical Matching: where we stand. Communication at LABFAM Seminar- 16 March 2021.
- FAO. 2015. Handbook on Master Sampling Frames for Agricultural Statistics: Frame Development, Sample Design and Estimation. Global Strategy Handbook: Rome.
- FAO. 2017. Handbook on the Agricultural Integrated Survey (AGRIS). Global Strategy to improve Agricultural and Rural Statistics. Rome.
- Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly, Vol. 75, No. 5, pp. 861–871. doi10.1093/poq/nfr057.
- International Household Survey Network (IHSN). 2021. Guidelines on Integration of survey instruments. Available at: http://www.ihsn.org/node/124. Accessed on July 12, 2021.
- Kish, L. (1987). Statistical design for research. New York, NY: John Wiley & Sons.
- UN Economic Commission for Europe (UNECE). 2018. A Guide to Data Integration for Official Statistics. https://statswiki.unece.org/spaces/flyingpdf/pdfpageexport.action?pageId=129171769.