Food and Agriculture Organization of the United Nations

FAO Statistics Working Paper Series

Issue 23/31

PRODUCING AGRICULTURAL POPULATION ESTIMATES USING AN AREA FRAME

FAO Statistics Working Paper Series / 23-31

PRODUCING AGRICULTURAL POPULATION ESTIMATES USING AN AREA FRAME

Silvia Missiroli and Dramane Bako
Food and Agriculture Organization of the United Nations

and

David Salazar
Ecuador National Institute of Statistics and Censuses

Food and Agriculture Organization of the United Nations
Rome, 2023

Required citation:
Missiroli, S., Bako, D. & Salazar, D. 2023. *Producing agricultural population estimates using an area frame*. FAO Statistics Working Paper Series, No. 23-31. Rome, FAO. https://doi.org/10.4060/cc4444en

ISBN 978-92-5-137669-0

Abstract

Estimating with a high precision the number of agricultural holdings and other population-related indicators is one of the main targets of the agricultural surveys. Achieving this objective is not straightforward when using an area frame: the choice of the estimator is indeed delicate since it depends on the criterion to link the observation units (e.g. agricultural holdings or tracts) and the sampling units (e.g. segments or points). According to the rule chosen, different types of estimators are defined: the open segment estimator, the closed segment estimator, the weighted segment estimator, the Horvitz–Thompson estimator and the multiplicity estimator. Moreover, auxiliary variables correlated to the population indicators can be used to produce the estimates: this is the case of the ratio estimator. In this paper, we describe the estimators mentioned above, focusing particularly on the comparison between them when the target parameter to be estimated is the number of agricultural holdings. The advantages and disadvantages of each of them are presented in terms of sampling and non-sampling errors, costs and the characteristics of the agricultural surveys within which they are intended to be used. Finally, a comparison between the weighted segment estimator and the ratio estimator is shown using data from the Area and Agricultural Production Survey 2017 of Ecuador. Both estimators seem to work well, with a slightly better performance of the weighted segment estimator in terms of sampling error. However, there is no optimal estimator in absolute terms; the choice is survey-specific, and it should be undertaken by carefully considering the sampling strategy, the frame, the costs, the target variables gathered and the data collection methods of each survey.

Contents

Acknowledgements

The authors would like to thank Flavio Bolliger, Javier Gallego and Luis Ambrosio Florence for the valuable comments, inputs and expertise provided that enriched and refined the document.

1 Introduction

The Food and Agriculture Organization of the United Nations (FAO) recommends the use of area frames among the options for developing master sampling frames for agricultural statistics. Area frames have been used in agricultural and forestry surveys for a long time with a specific focus on estimating crop acreage and production, and on monitoring vegetation and land cover/use (FAO, 2016). However, the scope of agricultural surveys nowadays goes far beyond land area and production. FAO is promoting integrated agricultural survey programmes covering the social, economic and environmental dimensions of the agricultural sector, aiming to produce a large set of indicators to fulfil both national and international data demands. One of the main targets of agricultural surveys is to estimate population-related indicators including the number of agricultural holdings that are defined as "economic units of agricultural production under single management comprising all livestock kept and all land used wholly or partly for agricultural production purposes, without regard to title, legal form or size" (FAO, 2015b). When using an area frame in agricultural surveys, the choice of the estimator for the number of agricultural holdings deserves attention mainly due to the complexity of linking the observation unit (agricultural holding) to the sampling unit (usually segment or point).

One of the main advantages of an area frame is that it allows a complete coverage of the target population. Its use is quite frequent in many countries like Brazil, Costa Rica, Ecuador and the United States of America. According to the sampling design chosen, a sample of area units (segments or points) is selected and the information on agricultural holdings linked to the area units is collected. Since farms[1] usually overlap segments' boundaries, a rule to assign a farm to a segment should be defined when segments are sampling units. According to the rule chosen, we can define different types of estimators: the open segment estimator, the closed segment estimator, the weighted segment estimator, the Horvitz–Thompson estimator and the multiplicity estimator. We describe and analyse them in Section 3 and Section 4, after a brief overview in Section 2 of the area frame approach and its use for the estimation of population units. In Section 5, we show alternative methods to estimate the number of agricultural holdings based on information collected on them during the survey. In Section 6, we compare the estimators presented in terms of sampling error. Section 7 reports a comparison between the weighted segment estimator and the ratio estimator for the number of agricultural holdings in Ecuador, using the data collected through the Area and Agricultural Production Survey 2017 (Encuesta de Superficie y Producción Agropecuaria Continua - ESPAC 2017), an agricultural survey conducted in Ecuador since 2002 and managed by the National Institute of Statistics and Censuses (INEC).

[1] For simplicity, the concepts of farms and agricultural holdings are used interchangeably in this paper.

2 Area frame and population estimates

An area frame can be considered as a list of area units that are points or regular pieces of land called segments. The latter are identified through satellite and cartographic images, with reference to a regular grid (which is often square) or to geographic and physical boundaries, such as roads, rivers or permanent field boundaries. Gallego (1995) and Gonzáles *et al.* (1991) show that using square segments is cheaper because they produce similar sampling errors and are easily defined on the map. Usually, the size of the segment is fixed in such a way that the groundwork can be performed in one working day at most. Area frames of points have been widely used for forestry inventories (FAO, 2015a), whereas area frames of segments are preferred in agricultural contexts, with the US Department of Agriculture (USDA) as one of the main users since the 1930s. However, points frames are also adopted for agricultural and land cover surveys, particularly in Europe, showing a higher cost efficiency and smoother data collection and processes operations (FAO, 2017b).

The limitations of an area frame for reliable population estimates are well acknowledged in the literature and several works discuss suitable approaches for such estimates including Faulkenberry and Garoui (1991) and FAO (2017a). In general, area frame surveys are known for producing accurate area-related data and are considered to have a major weakness for items that are not proportionately associated with cultivated land use (FAO, 2018). When such items (including estimates on farms population) are of interest, the use of an area frame is recommended only if there are no suitable list frame for the farm survey (Gallego, 2015; FAO, 2018). Two important components of these limitations are coverage and multiplicity issues.

2.1 Coverage

Although an area frame presents a complete coverage of all land in the country since the area units form a partition of the total surface under investigation (i.e. a country, a region, a province, etc.), it may exclude from the sampling process some segments or strata that are labelled as being purely non-agricultural but contain some minor agriculture. Moreover, an area frame excludes systematically the landless population of interest, such as landless agricultural holdings raising livestock. FAO (2017a) recommends complementing the area frame with a list of such a population of farms to take them into account in the estimations.

In general, an area frame is a list of area units providing an indirect access to agricultural holdings. Accordingly, there is no population information in the sampling frame to be considered in the design (e.g. stratification) and in the selection procedures (e.g. probability-proportional-to-size selection) of the area units. Thus, the final sample of farms obtained from the sample of area units usually presents an inadequate coverage of key subpopulations, such as farms growing special or rare crops, farms operated by women or young holders, or small farms. FAO (2018) notes that a sample from an area frame is considered to be less representative for small areas and for crops that are usually grown on small farms, such as tobacco and vegetables, and for orchards and vineyards.

2.2 Multiplicity

Multiplicity refers to the number of sampling units (e.g. segments or points) that would lead to the collection of data from the same observation unit (e.g. farm) (FAO, 2015a). In presence of multiplicity, the calculation of the weights of the observation units should be considered carefully, for example through direct calculations or adjustments of the design weights of the sampling units. In area frame surveys, multiplicity issues occur with both segment and point sampling and the importance of such an issue depends on agricultural practices in countries. The issue should be assessed in the population and is likely to be serious in contexts where the average number of parcels per farm is high (FAO, 2017a).

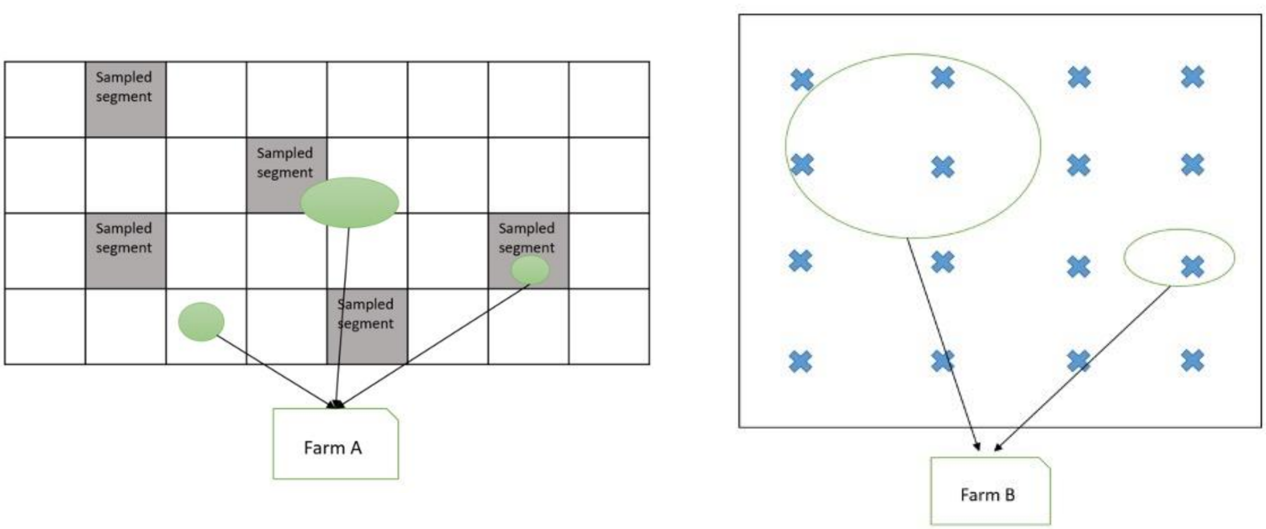
On the left side of Figure 1 with squared grid segments, Farm A is linked to 6 segments in the area frame including two sampled segments. The farm operates three parcels:

- A large parcel that intersects with four segments including one sampled segment
- Two small parcels located in two different segments including one sampled segment

On the right side of Figure 1 farm B is associated to four sampling points: the farm operates a large parcel in which three points fall and a small parcel covering one point.

Figure 1. Examples of multiplicity

In case of segments **In case of points**



Source: Authors' own elaboration.

3 Classic estimators of the number of agricultural holdings

The objective of this section is to present the classic estimators used for producing estimates of various statistics on farms in area frame surveys. Particularly, it is focused on how to estimate the total number of agricultural holdings using an area frame, and in general, on how to estimate a population total, like the planted area of a specific crop. More detailed discussions on the classic estimators can be found in FAO (2015a) and Gallego (2015).

3.1 Estimators with segments

When segments are sampling units, unbiased estimations can be obtained with variables observed at the segment level using the sampling weights of segments and the Horvitz–Thompson estimator. For instance, the total agricultural area utilized (AAU) can be estimated expanding the AAU measured in the sampled segments through the segments' sampling weights. However, for estimating farm-specific indicators, three classic estimators are proposed, mainly depending on how the agricultural holding is linked to segments: open, closed and weighted segment estimators.

We call Y the variable of interest and $\hat{Y}$ its estimated total in the population. We also suppose for simplicity that all the information within the sampled segments is collected (i.e. all the farms considered as within the segment are interviewed). The estimator for the total can be expressed in terms of segments or in terms of farms. We consider the second option since the farms are the reporting units. In the specific case of this paper, our focus is the total number of farms *F* that we can estimate through $\hat{F}$, a special value of $\hat{Y}$ when Y is equal to one for each farm. If every farm is completely included within a segment, we would have a one-to-one correspondence and the Y-value at the segment level would be simply the sum of the Y-values over all farms within the segment. Such a situation would correspond to a cluster sampling of farms where the sampling weight of each farm would be the same as the sampled segment to which it belongs. Since this is not a realistic scenario because farms usually overlap segments' boundaries, rules should be set to link farms to segments. Each rule gives rise to a different estimator. Faulkenberry and Garoui (1991) derived a general form for the estimator of the total in case the segments are selected with a simple random sampling design. It can be generalized to a stratified sampling design and to more complex designs like stratified two-stage sampling.

The general expression for the estimator of the total is:

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{M_h} \sum_{j=1}^{F_i} y_j \delta_{ij} w_{ih} \tag{1}$$

where y_j is the value of the variable of interest Y for farm j, δ_{ij} is a variable that expresses the "link" between farm j and segment i and it varies in [0,1], w_i is the sampling weight of segment i in stratum h, for j=*1,…,* F_i, *i=1,…,* M_h and *h=1,…,H*, with F_i the total number of farms overlapping with segment *i,* M_h the total number of sampled segments in stratum *h*, *H* the total number of strata.

The same estimator is used to compute the number of farms and it becomes:

$$\hat{F} = \sum_{h=1}^{H} \sum_{i=1}^{M_h} \sum_{j=1}^{F_i} \delta_{ij} w_{ih} \quad (2)$$

where y_j is replaced by one in (1).

The behaviour of the variable δ_{ij} is given by the criterion used to assign farms to segments that, consequently, determines the estimator. According to the rule chosen, we can distinguish among different types of estimators: the open segment estimator, the closed segment estimator, the Horvitz–Thompson estimator, the weighted segment estimator and the multiplicity estimator. The details of these estimators are given below.

3.1.1 Open segment estimator

The open segment method assigns a farm to the segment in which its headquarters are located.

In the *open segment estimator* δ_{ij} is equal to 1 if the headquarters of farm j are in segment i, 0 otherwise. In case we want to estimate the total of a variable Y, the y_j in $\hat{Y}$ refers to the Y-value of farm j inside and outside the sampled segment i that contains the farm headquarters. It is an unbiased estimator if the sampling weights are the inverse of the inclusion probabilities of the segments.

The open segment estimator was widely used in the past and it is recommended when holders live on the farms that generally have a small size. When holders instead live in the villages and go to irregular-shaped fields, it can be difficult to define the headquarters of the farm and this estimator becomes less efficient and precise (Ford *et al.*, 1986). In general, it is not always easy to give a clear definition of headquarters or to identify them in the field. Moreover, if many holders live in urban areas that were excluded from the frame, the open segment estimator would not consider their holdings in the count, producing an underestimation of the number of farms. Additionally, in countries where the size of agricultural holdings varies a lot, the open segment estimator might not provide precise estimates of farms' characteristics proportional to size.

3.1.2 Closed segment estimator

The *closed segment estimator* can be used to estimate totals when it is possible to know the proportion of the variable of interest Y within the segment or directly measure the value of Y included in the segment. For coherence with other estimators, δ_{ij} is here the proportion of y_j included in segment i, where y_j denotes the value of the variable Y for farm j. The product $y_j \delta_{ij}$ can be substituted with the value of y_j included in the segment, when it is directly measured/reported, as frequently happens with crop measurements. The closed segment estimator is unbiased and is usually used when the reporting unit is the tract, i.e. the total land operated by one operator in the segment.

The closed segment estimator is not adaptable to all variables that are reported at the farm level, like equipment or labour costs. This is why it will not be used to estimate the number of farms and will not be considered in the analysis of Ecuador data in Section 7.

3.1.3 Weighted segment estimator

In the *weighted segment estimator* the value y_j of the variable Y for farm j is weighted for the proportion of the area of farm j in the sampled segments, hence δ_{ij} represents the proportion p_{ij} of the total area of farm j included in segment i. This estimator is unbiased and is commonly used to estimate the number of farms. Bethel (1985) studied the use of the main crop area instead of the total farm area, with the objective of reducing the non-sampling errors. Indeed, the area of the farm is a variable that may not be precisely reported, particularly if it is collected through the farmer's declaration and not through direct measurement. However, the weighted segment estimator is widely used in agriculture surveys that adopt an area frame.

3.2 Estimators in point sampling

The farm sampling approach through points can be seen as a variant of the weighted segment method, although with some significant differences. In particular, computing the area of a particular farm that belongs to the sampled segments is not needed (Gallego, 2015). When a single-stage points sampling is applied and a point falls on a farm j, the farm is considered sampled with a probability proportional to its area, say x_j. If we suppose that farms do not cross strata boundaries and that farm j is located in stratum h, if n_h is the number of points sampled in stratum h and D_h is the total area of stratum h, the probability of selection of the farm j is:

$$\pi_{jh} = n_h \frac{x_j}{D_h} = \frac{1}{w_{jh}} \tag{3}$$

The number of farms is estimated using the Horvitz–Thompson estimator:

$$\hat{F} = \sum_{h=1}^{H} \sum_{j=1}^{f_h} w_{jh}, \tag{4}$$

with H the total number of strata and f_h the number of farms where the points in stratum h fall.

A similar approach can be used in two-stage point sampling that is usually applied when segments are large pieces of territory or contain a high number of fields/farms. To reduce the fieldwork per segment, farms are subsampled by points, i.e. data are collected only for those farms corresponding to points falling on their land. For regular segments, it is recommended to use a fixed pattern of points – such as that illustrated in Figure 2 – with one central point and four points close to the corners (FAO 2015a).

When point subsampling is applied within a segment, it is not necessary to record the farm land within the segment; indeed the estimator in (1) for the total of the variable Y with the weighted approach becomes:

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{M_h} \sum_{j=1}^{F_i} y_j \frac{p_{ij} x_j}{x_j} \frac{S_i}{p_{ij} x_j n_i} w_{ih} = \sum_{h=1}^{H} \sum_{i=1}^{M_h} \sum_{j=1}^{F_i} y_j \frac{S_i}{x_j n_i} w_{ih}, \tag{5}$$

where F_i is the number of farms in segment i where the sampling points fall, $\frac{p_{ij} x_j}{x_j} = p_{ij}$ is equal to δ_{ij} for the weighted segment estimator and $\frac{S_i}{p_{ij} x_j n_i}$ is the inverse of the probability of selection of farm j in segment i which is equal to the number n_i of points sampled in segment i multiplied by the ratio between the farm area within the segment and the total area S_i of segment i. The other quantities are exactly defined as per the estimator in (1).

Similarly, the estimator for the number of farms is:

$$\hat{F} = \sum_{h=1}^{H} \sum_{i=1}^{M_h} \sum_{j=1}^{F_i} \frac{S_i}{x_j n_i} w_{ih} \quad (6)$$

It is worthy to mention that both $\hat{Y}$ and $\hat{F}$ in this case do not depend on the farm area inside the segment, but only on the total farm area and on the area of each sampled segment. In this context, the area can indicate the agricultural land of the farm and of the segment. Before data collection, it is important to establish a clear definition of "agricultural land" and be consistent during the whole survey process.

Figure 2. Subsampling farms inside a square segment

Source: FAO. 2015. *Handbook on master sampling frames for agricultural statistics*. Rome. https://www.fao.org/3/ca6398en/ca6398en.pdf

4 Addressing multiplicity bias in population estimates

In the presence of multiplicity, the classic estimators presented above are biased, as for any survey without one-to-one correspondence between the sampling unit (e.g. segments, points) and the observation unit (e.g. farms, tracts). In fact, the actual probabilities of selection of farms will not coincide with those associated to points or segments (no matter the way farms are associated to segments). This issue can be addressed with a multiplicity estimator or the Horvitz–Thompson estimator using the actual probabilities of selection of farms when it is possible to calculate them.

4.1 Multiplicity estimator

Many multiplicity estimators are proposed in the literature including Sirken (1970), Huang (1984) and Ernst (1989). Lavallée (2007) proposed a generalized method addressing any kind of multiplicity: the generalized weight share method (GWSM). Following the GWSM, farm j will be assigned a weight $w_j = \sum_{i=1}^{m_{js}} w_i / m_j$ where w_i is the sampling weight of the segment or point i associated to farm *j;* m_{js} and m_j are the number of segments that farm *j* intersects or the number of points associated to farm *j* in the sample and in the population respectively (FAO, 2017a).

In the case of segment sampling, if the sampling weights are equal for all the segments in each stratum the *multiplicity estimator* would consist simply in using δ_{ij} equal to $\frac{m_{js}}{m_j}$ in equation (2) (Faulkenberry and Garoui, 1991). Hence, to compute this estimator, we need an extra information: the number of segments not in the sample that farms intersect. This information can be costly, and for this reason the multiplicity estimator is not frequently used in practice (Garoui, 1985) even though it can be very efficient in terms of variance respect to the other estimators, as we will see in Section 6. The multiplicity estimator is also unbiased.

4.2 Horvitz–Thompson estimator

From (1) the classic *Horvitz–Thompson estimator* can be computed associating to each agricultural holding a probability π_{jh} of inclusion in the sample and in stratum *h*, for *j=1, …,* $f, h = 1, \dots, H,$ with f the number of farms totally or partially included in the sampled segments and H the total number of strata. The estimator becomes:

$$\hat{Y} = \sum_{j=1}^{f} \sum_{h=1}^{H} y_j w_{jh}, \tag{7}$$

where $w_{jh} = \frac{1}{\pi_{jh}}$, y_j is the value of the variable Y for farm *j, H* is the total number of strata. The Horvitz–Thompson estimator of the number of agricultural holdings becomes:

$$\hat{F} = \sum_{j=1}^{f} \sum_{h=1}^{H} w_{jh}. \tag{8}$$

Supposing that segments are selected with a simple random sampling in each stratum, the inclusion probability of farm *j* in stratum *h* is given by:

$$\pi_{jh} = \left(\binom{M_h}{m_h} - \binom{M_h - M_{hj}}{m_h}\right) \Big/ \binom{M_h}{m_h}, \tag{9}$$

where M_h is the total number of segments in stratum *h*, m_h is the total number of sampled segments in stratum *h and* M_{hj} is the number of segments that farm *j* intersects in stratum *h*. Like for the multiplicity estimator, the Horvitz–Thompson estimator requires to know M_{hj}. In contexts where a typical agricultural holding operates many parcels in different places, collecting M_{hj} may be costly as it would require the geo-referenced coordinates of the different parcels to assess the number of overlapping segments. This estimator is unbiased.

5 Estimating the number of farms using additional information

It is possible to use additional variables collected through the survey and correlated to the character of interest Y to derive other types of estimators, in particular *the ratio estimator* $\hat{Y}_R$ for the total (Faulkenberry and Garoui 1991), which is given by:

$$\hat{Y}_R = \sum_{h=1}^{H} \sum_{j=1}^{f_h} \frac{\sum_{i=1}^{M_h} \sum_{j=1}^{F_i} w_i x_j p_{ij}}{\sum_{j=1}^{f_h} x_j} y_j \tag{10}$$

where y_j is the value of the variable Y for farm *j*, w_i is the sampling weight of segment *i*, p_{ij} is the proportion of the area of farm *j* inside the segment *i* and x_j is the total area of farm *j*, for j=*1,…,* F_i or *j=1,…,* f_h, *i=1,…,* M_h and *h=1,…,H*, with F_i the total number of farms intersecting segment *i*, f_h the number of farms totally or partially included in the sampled segments in stratum *h*, M_h the total number of sampled segments in stratum *h*, *H* the total number of strata.

The ratio estimator $\hat{Y}_R$ can also be seen as:

$$\hat{Y}_R = \sum_{h=1}^{H} \frac{\sum_{j=1}^{f_h} y_j}{\sum_{j=1}^{f_h} x_j} \hat{X}_h, \tag{11}$$

where $\hat{X}_h$ is the closed segment estimator for the total farm area in stratum *h*.

If we want to estimate the number of farms, we fix y_j equal to one for j=*1, …,* F_i and get:

$$\hat{F}_R = \sum_{h=1}^{H} \sum_{j=1}^{f_h} \frac{\sum_{i=1}^{M_h} \sum_{j=1}^{F_i} w_i x_j p_{ij}}{\sum_{j=1}^{f_h} x_j} \tag{12}$$

$\hat{Y}_R$ and $\hat{F}_R$ are special cases of $\hat{Y}$ in (1) and $\hat{F}$ in (2) respectively when δ_{ij} is equal to $\frac{\sum_{i=1}^{M_h} \sum_{j=1}^{F_i} x_j p_{ij}}{\sum_{j=1}^{f_h} x_j}$ and the w_i are all equal within each stratum for *i=1,…,* M_h.

From the estimator $\hat{F}_R$ in (12), with some algebraic manipulation we get:

$$\hat{F}_R = \sum_{h=1}^{H} \sum_{i=1}^{M_h} \sum_{j=1}^{F_i} w_i \frac{x_j p_{ij}}{\bar{x}_h} \tag{13}$$

where $\bar{x}_h$ is the simple mean estimate of the farms' land area in stratum *h*. This estimator was used by the National Institute of Statistics and Census (INEC) of Ecuador to estimate the number of farms with the data collected through the Area and Agricultural Production Survey 2017.

It can be also simply noted that fixing y_j equal to one in equation (11) gives us:

$$\hat{F}_R = \sum_{h=1}^{H} \frac{f_h}{\sum_{j=1}^{f_h} x_j} \hat{X}_h = \sum_{h=1}^{H} \frac{\hat{X}_h}{\bar{x}_h} \tag{14}$$

The ratio estimator simply estimates the number of farms in a stratum by dividing the total estimated farm area in the stratum by the estimated mean of the total farm area in that stratum.

As we will see in Section 7, the estimators presented throughout this paper can be used also to compute the number of farms by size category. For instance, if we group farms according to their measure of land area and if we apply $\hat{F}_R$ to each group, we can get the number of farms by group. Similarly, $\hat{F}_R$ can be used to estimate the number of farms raising a particular livestock species. In this case f_h is the number of farms in the sampled segments that raise that specific species and the auxiliary variable x refers to the number of animals.

The INEC estimator for the number of farms growing a specific crop $\hat{F}_c$ is given by:

$$\hat{F}_c = \sum_{h=1}^{H} \sum_{i=1}^{M_h} \sum_{j=1}^{F_i} w_i \frac{z_{ij}}{\bar{x}_h} \qquad (15)$$

where w_i is the theoretical weight of segment *i*, z_{ij} is the planted area within the segment *i* of the specific crop field of farm *j* and $\bar{x}_h$ is the simple mean estimate of the land area of the farms growing the specific crop, for j=*1,…,* F_i *, i=1,…,* M_h and *h=1,…,H*, with F_i the total number of farms in segment *i* growing the specific crop*,* M_h the total number of sampled segments in stratum *h*, *H* the total number of strata. It is recommended to calculate this estimator by domain and by farm size.

6 Comparing estimators in terms of precision

The estimators described in Section 3 and Section 4 are unbiased and their sampling error can be measured in terms of variance, whereas the estimators listed in Section 5 are generally biased, hence the variance is not enough to express their ability of estimating the parameter of interest. However non-sampling errors should be also considered carefully to keep the total survey error at a minimum level. If each farm of the population is contained within a segment, δ_{ij} is equal to one for all the estimators. Hence, estimators and variances are all equal. When instead farms overlap segments' boundaries, some estimators can be more efficient than others. It is difficult to determine which estimator is the best because it depends on the overlap pattern and on the values of the variable of interest. Faulkenberry and Garoui (1991) showed that the multiplicity estimator and the Horvitz–Thompson estimator have lower variances respect to the open segment estimator if the headquarters of farm j are equally likely to be located in any of the m_j segments that farm j intersects and if at least one farm overlaps segments' boundaries. For large samples, the Horvitz–Thompson estimator is proved to be more efficient than all the other estimators presented above, including the multiplicity estimator (Faulkenberry and Garoui, 1991). The drawback of this estimator is that it requires to know the segments that sampled farms intersect and this information is usually not available or very costly.

If the segments are accurately stratified in terms of the main agricultural variables and segment sizes are approximately equal, the closed segment estimator has a good precision for estimating the total area and if y_j is proportional to x_j, the ratio estimator $\hat{Y}_R$ is unbiased with the same coefficient of variation (CV) as the closed segment estimator for total farm area (Faulkenberry and Garoui, 1991). In Section 7 it will be shown, using country data and a bootstrapping technique, that the ratio estimator produces CVs for estimates of the number of farms globally similar to those provided by the weighted segment estimator, except when outliers are present in the sample data.

The open segment estimator is usually the least precise and often underestimates the number of agricultural holdings due to the difficulty of screening for holders in densely populated areas when the farm headquarters coincide with the holder's residence, as it happens in most of the cases. Moreover, the number of agricultural holdings per segment and, consequently, the total of the agricultural variables collected in each segment can be very heterogeneous, increasing the sampling error.

7 Country example: estimating the number of farms in Ecuador

To compare the estimators presented in the previous sections, we use data collected through the Area and Agricultural Production Survey 2017 (Encuesta de Superficie y Producción Agropecuaria Continua - ESPAC 2017), an agricultural survey conducted in Ecuador since 2002 and managed by INEC. The aim is to estimate the number of farms in Ecuador in 2017 using the weighted segment estimator and the ratio estimator reported in (13), showing the differences. It is not possible to analyse the behaviour of the open segment estimator since the location of the headquarters of the holdings was not collected.

ESPAC 2017 uses a multiple frame that combines an area frame and a list frame. The former is obtained from cartographic material and satellite images, guaranteeing a full geographical coverage of the country. The latter includes the largest producers of concentrated crops like industrial or horticultural crops. The sampling units of the area frame are segments with geometrical boundaries of around 10 parcels each that have been stratified before selection according to the agricultural intensity. In this context, a parcel is defined as a continuous field with the same use of land that can also belong to different owners. The sample size is equal to 5 678 segments. An additional 3 969 agricultural holdings included in the list frame received an interview. To avoid duplications, the farms in the sampled segments are removed from the list frame. The segments are selected with a stratified sampling procedure and allocations proportional to the strata agricultural intensity. Replicated sampling is also implemented in support of sample rotation. Substrata or zones are used to form five replicates. Within each sampled segment, all the parcels have been tracked and all the agricultural holders of these parcels have been interviewed. A dataset is created for each chapter of the questionnaire.[2]

To calculate the total number of agricultural holdings, both the weighted segment estimator and the ratio estimator in (13) are used and a comparison is presented. To compute the former, we need δ_{ij}, i.e. the proportion of farm j contained in segment i of stratum h, for $j=1,\ldots,F_i$, $i=1,\ldots,M_h$, that is calculated dividing the total area of the farm within the segment (i.e. the sum of the values of the variable "supertotal" in the *Usnac2017* dataset over the records having the same "identificador" that indicates the farm) by the total area of the farm located inside and outside the segment (i.e. the value of the variable "eu_superficie_ha" in the *Eunac2017* dataset for the same "identificador"). This proportion is then multiplied by the theoretical sampling weight of the segment (i.e. the variable "fact_exp_fin" of the *Usnac2017* dataset) and summed over all the farms within sampled segments, excluding the agricultural holdings in the list frame and adjusting the sampling weights of those farms that are in both the list frame and the area frame. The results of the estimated number of farms are shown in Table 1, which also presents the values by farm size categories and the relative CVs that were calculated by applying a bootstrap technique with 500 bootstrap samples. Calculating the number of farms raising cattle is straightforward because the weights contained in the *glnac2017* dataset (farms with cattle) have already been weighted by the proportion of the farm contained inside the segment. It is enough to sum up the values of the weights in the *glnac2017* dataset to get the total number of farms raising cattle, excluding those belonging to the list frame and adjusting the sampling weights of those farms that are in both the list frame and the area frame. Results are also shown in Table 2.

---

[2] The data and other information can be found at https://www.ecuadorencifras.gob.ec/encuesta-de-superficie-y-produccion-agropecuaria-continua-2015-2016-2017-2/. A dictionary is provided to explain the variables contained in each dataset.

The ratio estimator in (13) is also used to estimate the number of farms. The numerator of the weighted segment estimator is here multiplied by the farmland area (i.e. the variable “eu_superficie_ha” in the *Eunac2017* dataset) and divided by its simple mean within the stratum. This quantity is then summed over all the farms totally or partially included in the sampled segments. Table 1 shows the results, presenting the estimated number of agricultural holdings also by farm size categories. Table 2 reports the estimate of the number of farms raising cattle, showing the results also by herd-size categories. In this case, x_j and $\bar{x}_h$ of the estimator in (13) are respectively the number of cattle of farm j and the simple stratum mean of the number of cattle, calculated within each herd-size category.

Table 1. Estimates and relative CVs of the number of farms by farm size, using the weighted segment and the ratio estimators with ESPAC 2017 data

Farm categories by land size (hectares)	Number of farms (weighted estimator)		Number of farms (ratio estimator)	
	Estimate	CV (percent)	Estimate	CV (percent)
Total	**850 482**	**0.36**	**802 911**	**0.35**
< 1	289 223	0.90	245 405	0.87
1–2	125 139	1.00	125 688	1.00
2–3	79 972	1.18	80 656	1.18
3–5	99 088	1.08	99 366	1.08
5–10	101 591	0.98	100 813	0.99
10– 20	68 658	1.16	68 264	1.17
20–50	54 341	1.23	53 060	1.26
50–100	22 269	1.73	22 130	1.77
100–200	6 978	2.66	6 852	2.70
> 200	3 223	3.17	677	13.84

Source: Authors’ own elaboration.

Comparing results in Table 1, we can notice good precisions of the estimates of the total numbers of farms from both the weighted and the ratio estimators: the CVs are low and almost equal (0.36 percent and 0.35 percent). The precision of the estimates of the number of farms by subpopulations (categorized by land size) for both estimators is also globally good and not much different (all CVs are less than 5 percent except for the estimate of the number of farms larger than 200 hectares [ha] with the ratio estimator).

The comparisons of the number of farms (total and by subpopulations) show differences between estimates from the two estimators (weighted and ratio estimators). The highest differences appear in groups 1 (area less than 1 ha) and 10 (area greater than 200 ha). This can be due to outliers in these groups because the mean at the denominator of the ratio estimator (13) is affected by extreme values of the farm area. For instance, high values would cause a general decrease of the number of farms; this is the case in group 10 where three holdings with large farmland outside the segment are present.

Table 2 below also shows good and similar precisions of the estimates of the total number of farms raising cattle from the two estimators. The sampling error of the two estimators seems to be quite similar, with a slightly better performance of the weighted segment estimator if the farm categories are not considered. The estimated number of farms raising cattle by cattle herd size from the ratio estimators are all lower than those presented from the weighted estimator. This may happen because the farm area can be not precisely proportional to its number of livestock, mainly for the farms with intense agricultural activities. From ESPAC 2017 data, the correlation between farmland and the number of cattle seems indeed to be quite low.

The proportion of farmland inside the segment with respect to the total farmland is around 50 percent, with an average of 2.5 parcels within each segment per farm and an average area of 0.5 ha per parcel. These values may imply the presence of multiplicity. However, if we compare the estimated total number of farms raising cattle reported in Table 2 with the ratio estimator (275 306) with the one (259 071) derived from an administrative source (AGROCALIDAD 2016) that resulted from a widely robust vaccination programme with an excellent coverage in field, we get similar values, indicating that the ratio estimator performs quite well.

Table 2. Estimates and relative CVs of the number of farms raising cattle, by herd size category, using the weighted segment estimator and the ratio estimator with ESPAC 2017 data

Farm categories by cattle herd size	Number of farms raising cattle (weighted estimator)		Number of farms raising cattle (ratio estimator)	
	Estimate	CV (percent)	Estimate	CV (percent)
Total	327 782	0.53	275 306	0.59
Less than 30 cattle	301 088	0.57	249 484	0.65
From 30 to 50 cattle	13 603	2.25	13 431	2.24
From 50 to 100 cattle	8 609	2.65	8 497	2.68
From 100 to 500 cattle	4 207	3.23	3 707	3.17
More than 500 cattle	275	11.10	187	9.26

Source: Authors' own elaboration.

When producing survey estimates, it is fundamental to make a comparison with other reliable sources, like administrative registers or agricultural censuses. In the case of population estimates, this is even more useful because the information obtained from auxiliary sources can be used to calibrate the sampling weights and, consequently, increase the estimators' precision.

The cost of computing the weighted segment estimator and the ratio estimator is the same: both require the total farm area and the farm area inside the segment. The more precise the area measurements, the more precise the estimators. In ESPAC 2017, the farm area within segments is measured by digitalizing the orthophoto of the segments, whereas the farm total area is declared by the holder. The measurements are checked at different steps, both by field supervisors and by analysts at headquarters. This generally ensures the precision of area measurements, but at a cost that may be high. It can be interesting to verify whether the cost for identifying the farm headquarters is lower, so that the use of the open segment estimator might be preferred: an assessment of the precision should be performed before through a pilot study to appreciate the cost-effectiveness. The costs for the detection of the number of segments intersected by each sampled farm can also be evaluated, together with the use of the multiplicity estimator. Before every survey, it is always worthy to conduct a cost–benefit analysis to choose the most suitable estimator in terms of expected precision and costs that are survey-specific and depend on several aspects including the sampling design, the data collection methods and the field work organization.

8 Conclusions and recommendations

Each of the estimators of the number of agricultural holdings described in the previous sections has relative strengths and weaknesses, and there is no optimal estimator in absolute terms. The choice depends on the type of survey, the frame, the target variables gathered and the data collection methods. For instance, the weighted segment estimator can be preferred if particular attention is given to the collection methods of field measurements and if there are not too many holders that graze livestock on public or communal land. The use of the weighted segment estimator allows to discard urban areas or forest/bare land from the frame of an agricultural survey, since the rule to link holdings to segments is based on land. However, if the measurement error of a holding's land area inside and outside the segment is expected to be high, the weighted segment estimator is not recommended to estimate the number of agricultural holdings. The same is true for the ratio estimator and for the derived estimators in (13) and (15) that make use of the holding's land area inside and outside the segment, variables that should be collected carefully if the objective is, among others, the estimation of the number of agricultural holdings. Ford *et al.* (1986) presented the results of a study conducted by the US Statistical Reporting Service where it was shown that farmers usually underreport the total land area, particularly omitting woodland and idle land. This generates a positive bias of the above-mentioned estimators since the denominator is often lower than the real value. The same evidence resulted from more recent experiments conducted by the World Bank (under the Living Standards Measurement Study programme) in Ethiopia, Nigeria and the United Republic of Tanzania: self-reported estimates of land area are subject to very large biases and systematic measurement errors (the mean self-reported and compass and rope measurements differ by as much as 143 percent on average in the United Republic of Tanzania); the area of the smaller plots provided by farmers is severely overestimated whereas the area of the larger plots is generally underestimated, resulting in a lower mean estimate of the total farm area (World Bank, 2016) that appears at the denominator of estimators (13) and (15).

However, from the analysis conducted in Ecuador with ESPAC 2017 data, the estimator in (13) was shown to perform quite well, so it can be considered when estimating the number of farms. Indeed, its CV, estimated through a bootstrapping method, was very similar or even lower than that generated by the weighted segment estimator, except in presence of outliers in terms of farmland.

The open segment method instead requires first to locate the headquarters of the agricultural holding that usually coincide with the residence of the holder and, consequently, to also include in the sampling frame the urban areas. In this case, the holder must be uniquely identified, which may create issues when a holding is managed by more households or farmers. Open segment estimators were the first type of estimators used in area sampling and they are still used in some countries. In the agricultural surveys conducted in the United States of America, the open segment estimator has been substituted by the weighted segment estimator when the reporting unit is the holding, as it happens when estimating the number of farms, and by the closed segment estimator when the reporting unit is the tract (FAO, 1996). The closed segment estimator cannot be used to estimate the number of farms or other variables linked to the totality of the holding, like agricultural production, livestock inventories, farm labour, planting intentions or grain stocks, but it is effective to estimate the planted crop area, mainly for major crops. In this case, the reporting units are the tracts within the sampled segments whose information can be collected even using photographic enlargements or cartographic material, without the need to link them with the relative farmer or agricultural holding. Therefore, the coverage and response errors are quite low. The latter affects more the estimates generated by the open segment estimator, the weighted segment estimator and the ratio estimator. In general, the closed segment estimator is not recommended in multipurpose agricultural surveys in favour of the weighted segment estimator.

Ultimately, the choice of the estimator for the number of agricultural holdings depends on different aspects of the survey considered: its purposes, the variables collected, the structure of the agricultural sector, the geographical distribution and the main typologies of farms in the country, and the data collection methods.

If population estimates are among the key measurement objectives of the agricultural survey, the use of a list frame is recommended particularly in a context where farms operate many parcels of land in different locations. In case an area frame is preferred, population estimates can be improved by following specific recommendations on both survey design and estimations:

- complement the area frame with specific list frames in a multiple frame survey perspective or take into account landless, special farms and commercial farms. The main condition for a smooth use of mixed frames is that the surveyor can easily distinguish or check if a given farm belongs to the list frame.
- When segments are sampled in a context where the average area of the parcels/tracks is rather low, experts advise subsampling farms in the segments as a cost-effective option. Discussions on subsampling procedures of farms inside a segment can be found in Gallego (2015).
- There is no optimal estimator suitable for any kind of population statistics. Countries may consider the discussions on the suitability and accuracy of various estimators performed in this paper.
- Multiplicity should be carefully assessed. In cases where it represents a serious issue, actions should be taken to collect suitable data required to address it with estimations procedures recommended here.
- To increase the precision of the estimators, it is recommended to calibrate the sampling weights using for example information on the number of farms, agricultural population, or number of livestock from reliable auxiliary sources like:
    - previous agricultural census or national agricultural surveys of good quality and reliability;
    - farmers population projections; and
    - sector-specific administrative sources (e.g. administrative records coming from vaccination programmes with a good coverage in the country).

References

Bethel, J. 1985. *A New Approach to Weighted Area Frame Estimates*. Washington, DC, National Agricultural Statistics Service, US Department of Agriculture.

Cochran, W.G. 1977. *Sampling Techniques*. 3rd Edition. New York, USA, John Wiley & Sons.

Ernst, L. 1989. Weighting issues for longitudinal household and family estimates. In: D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh, eds. *Panel Surveys*, pp. 135-159. New York, USA, John Wiley & Sons.

Faulkenberry, G.D. & Garoui, A. 1991. Estimating a Population Total Using an Area Frame. *Journal of the American Statistical Association*, Vol. 86, No. 414, pp. 445-449.

FAO. 1996. *Multiple Frame Agricultural Surveys: Volume I: Current Surveys based on area and list sampling methods*. FAO Statistical Development Series n.7. Rome. http://www.fao.org/3/ca6106en/CA6106EN.pdf

FAO. 2015a. *Handbook on master sampling frames for agricultural statistics.* Rome. https://www.fao.org/3/ca6398en/ca6398en.pdf

FAO. 2015b. *World Programme for the Census of Agriculture 2020, Vol. 1: Programme, concepts and definitions*. FAO Statistical Development Series 15. Rome. http://www.fao.org/world-census-agriculture/wcarounds/wca2020/en/

FAO. 2016. *Information on Land in the Context of Agricultural Statistics*. Rome. http://www.fao.org/3/ca6419en/ca6419en.pdf

FAO. 2017a. *Handbook on the Agricultural Integrated Survey (AGRIS)*. Rome. http://gsars.org/wp-content/uploads/2017/12/AGRIS-HANDBOOK.pdf

FAO. 2017b. *Handbook on Remote Sensing for Agricultural Statistics.* Rome. http://gsars.org/en/handbook-on-remote-sensing-for-agricultural-statistics/

FAO. 2018. *Master Sampling Frames for Agriculture: Supplement on selected country experiences*. Rome. https://www.fao.org/3/ca6431en/ca6431en.pdf

Ford, B.L., Nealon, J., & Tortora, R.D. 1986. Area Frame Estimators in Agricultural Surveys: Sampling Versus Nonsampling Errors. *Agricultural Economics Research*, 38, 1-10.

Gallego, F.J. 1995. *Sampling Frames of Square Segments*. Ispra, European Commission Joint Research Centre.

Gallego, F.J. 2015. *Area Sampling Frames for Agricultural and Environmental Statistics*. doi:10.2788/88253

Garoui, A. 1985. *Estimating Populations Totals from Area Frame Samples.* PhD Thesis. Oregon State University.

González, F., López, S. & Cuevas, J.M. 1991. Comparing Two Methodologies for Crop Area Estimation in Spain Using Landsat TM Images and Ground Gathered Data. *Remote sensing of the environment*, 35:29-36.

Huang, H. 1984. Obtaining Cross-Sectional Estimates from a Longitudinal Survey: Experiences of the Income Survey Development Program. Survey Research Section of the Proceedings of the American Statistical Association, 1984, pp. 670-675.

INEC. 2017a. *Encuesta de Superficie y Producción Agropecuaria Continua. Metodologia 2017.* Quito. https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_agropecuarias/espac/espac_2017/Metodologia_ESPAC2017.pdf

INEC. 2017b. Encuesta de Superficie y Producción Agropecuaria Continua. Bases de datos 2017. In: *INEC*. Quito. Cited 15 January 2021. https://www.ecuadorencifras.gob.ec/encuesta-de-superficie-y-produccion-agropecuaria-continua-2015-2016-2017-2/

Lavallée, P. 2007. *Indirect Sampling*. Ottawa, Springer.

Sirken, M.G. 1970. Household Surveys with Multiplicity. *Journal of the American Statistical Association*, Vol. 65, No. 329, pp. 257-266.

World Bank. 2016. Land Area Measurement in Household Surveys. Washington, DC. https://documents1.worldbank.org/curated/en/606691587036985925/pdf/Land-Area-Measurement-in-Household-Surveys-Empirical-Evidence-and-Practical-Guidance-for-Effective-Data-Collection.pdf.

Contact:

Statistics Division – Economic and Social Development
FAO-statistics@fao.org
www.fao.org/food-agriculture-statistics/resources/publications/working-papers/en/

Food and Agriculture Organization of the United Nations
Rome, Italy

ISBN 978-92-5-137669-0

9 789251 376690

CC4444EN/1/02.23