Statistical Journal of the IAOS 39 (2023) 473–489
DOI 10.3233/SJI-220116
IOS Press

Integration of remote sensing data into national statistical office sampling designs for agriculture

Luis Ambrosio[a,*], Luis Iglesias[a], Carmen Marín[a] and Nicolas Deffense[b]
[a]*Universidad Politécnica de Madrid, Madrid, Spain*
[b]*Université Catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium*

Abstract. The integration of remote sensing data in agricultural statistics is a research topic with a long history. The research focus is on using statistical models to link ground and remote sensing data such that the resulting estimators are design-consistent. A design-consistent estimator assisted by linear models is well established in the literature. However, it requires enough geographic information about the boundaries of agricultural parcels to develop a simple sample of areas. Many countries use complex samples based on non-georeferenced list frames of households or farms and reduce to point data the georeferenced information required for linking ground and remote sensing data.
Data on crop acreage observed at a point are necessarily categorical because a point is dimensionless. Little work has been done on the integration of categorical ground data within complex list samples using remote sensing data. Our focus was on using multinomial logit models for this integration. Special attention was paid to evaluate the cost efficiency of remote sensing data.

Keywords: Agricultural statistics, point sampling, categorical data, multinomial logit models, cost efficiency

1. Introduction

Recently, the European Space Agency (ESA) launched the Sentinels for Agricultural Statistics (Sen4Stat) project. The aim was to facilitate the use of remote sensing information at the National Statistical Offices (NSOs) supporting agricultural statistics. Other international programs to improve agricultural statistics were begun around the same time. An additional Sen4Stat target is synergy with other initiatives, including the Living Standards Measurement Study – Integrated Survey on Agriculture (LSMS-ISA), which was launched in 2008 by the World Bank, and the Global Strategy to Improve Agricultural and Rural Statistics (GSARS), which was endorsed in 2010 by the United Nations Statistical Commission and implemented by the Food and Agriculture Organization (FAO) of the United Nations. An FAO output is the Agricultural Integrated Survey (AGRIS) method [1], together with the 50 × 2030 Initiative, which builds on work of the LSMS-ISA and AGRIS programs [2]. Four of five partner countries in Sen4Stat were also partners in other initiatives, i.e., Ecuador and Senegal of AGRIS, Malawi of LSMS-ISA, and Tanzania of GSARS. Accordingly, through existing collaborations and initiatives, the FAO coordinated the provision of ground data from respective agencies in Ecuador, Senegal and Tanzania, while ground data collection from Malawi was coordinated by the World Bank.

Remote sensing (RS) furnishes a very useful auxiliary data source to improve the cost efficiency of design-based inference, and the integration of ground and remote sensing data is a research topic with a long history [3–7]. Statistical models are used for this integration, in such a way that the estimator assisted by these models is design-consistent [8].

The focus has been on linear models, assuming that the sample design is simple. In countries where geographic information of agricultural parcel boundaries

*Corresponding author: Luis Ambrosio, Universidad Politécnica de Madrid, Madrid, Spain. E-mail: luis.ambrosio@upm.es.

is available, it is possible to use an individual or cluster of parcels as the sampling unit and to design a simple sample of areas (polygons) using area frames. A linear model is suitable for integrating ground and remote sensing data on crop acreage or yield because these are continuous variables. However, many countries use complex samples based on non-georeferenced list frames (of households or farms) and reduce to point data the georeferenced information required for linking ground and remote sensing data. Crop data observed at a point are necessarily categorical because a point is dimensionless, and multinomial logit is an appropriate statistical model to integrate remote sensing and categorical ground data [9,10].

Little work has been done on the integration of categorical ground data in complex list samples of households or farms with remote sensing data. In this paper, we focus on using multinomial logit models for this integration. We give details on statistical models to be used for crop acreage, yield and production estimation, as well as on the sample design used by five Sen4Stat partner countries that are representative of the large types of sample designs proposed in the literature. These are list, area, and point sampling [11].

Considering the literature on the applications of remote sensing to agricultural statistics, as well as expectations of the NSO partner countries for these applications, we elaborated four use cases: 1. Cost efficiency; 2. Granularity; 3. Multi-seasonal estimation; 4. Optimizing the sample design. A prototype for integrating ground and remote sensing data in each use case was developed for a large set of sampling designs, including those currently used by the partner countries. We include results of the application of these prototypes to data provided by the NSOs, for cost efficiency and granularity. Finally, we outline prototypes for multi-seasonal estimation and optimizing the sample design.

2. Methods

The approach to elaborate official statistics is well established in the literature. It is design-based in the sense that the inference is based on the sampling distribution generated by the probabilistic sampling scheme designed to select the sample [12,13]. This sampling scheme associates with each population unit, $i = 1, 2, \ldots, N$, a known and positive inclusion probability π_i, which play a key role in design-based inference. Usually, the sample size is sufficiently large to attain design consistency (the design-based distribution of the estimator is tightly concentrated around the true population value) at national and regional levels, and design-consistent estimators are used. Estimate accuracy is also based on the estimator's design-based distribution.

Our focus was on improving the cost-efficiency of the currently used sampling designs for elaborating official statistics on crop acreage, using RS as auxiliary data. The approach to using auxiliary data in design-based inference is well stablished in the literature [8], i.e., the auxiliary extra-sample information is integrated in the sampling design using statistical models, without loss of design-consistency. In minor administrative areas (municipalities) the sample size is small or nil, so that design consistency is meaningless. Thus, a model-based approach is used to get estimates, whose accuracy strongly depends on the auxiliary information reliability.

Model suitability for ground and remote sensing data integration depends on the nature of ground data. For continuous ground data a linear regression model is used, but for categorical and counts ground data, a generalized linear model is more suitable [9,10]. Both types of models allow remote sensing data of any character, continuous or discrete. The efficiency of RS as auxiliary data for crop acreage has already been extensively demonstrated in the scientific literature, but only for continuous ground data and with the assumption that the sample design is simple. To date, little work has been done on the integration of categorical ground data in complex list samples of households or farms with remote sensing data. In this paper, we focus on using multinomial logit models for this integration.

2.1. Sample designs

The design of the sampling scheme for use in sample selection is key in design-based inference because it associates with each population unit $i = 1, 2, \ldots, N$ a known and positive inclusion probability π_i, which is used to define the estimators' statistical distribution. This distribution is used to evaluate the estimators' characteristics, including design consistency and accuracy.

A sampling frame is required to select a probabilistic sample and we considered the two main types of sampling frames used in practice, area frame and list frame. The former is based on maps and/or satellite images and the latter on population censuses. Sampling units are either polygons or points in area frames and households or farms in list frames. Usually, the former is georeferenced, but not the latter [11].

The area frame has important advantages over list frames, including versatility, reduction of non-sampling errors such as coverage errors, and longevity. However, it has some inconvenience (sensitivity to outliers and sub representation of rare or minor activities) [14]. To overcome these inconveniences, a dual frame composed of an area and list frame is recommended [15]. In Section 4, we give details on the dual frame used in Senegal to integrate ground and RS data.

Using area frames, it is possible to define sampling units (polygons) of the same size and simple samples with equal inclusion probabilities. In contrast, using list frames, the sample selection scheme entails two or more stages, the sampling units are usually of distinct size, and the probabilistic scheme assigns unequal (size-proportional) inclusion probabilities. Therefore, the sample is not simple but complex.

The integration of ground and RS data will be illustrated for the set of sample designs most often used in practice for collecting ground data [11]. With area frames, systematic sampling with one or more random starts and stratified random sampling are most often used as the sample schemes when geographic information on parcel boundaries is available. If the parcel boundaries are unknown, the sampling unit is usually a point, and we will illustrate how to integrate a sample of points selected from a list frame with RS data.

Agricultural data are frequently collected using national household surveys, for which the sample is selected from a list frame (a population census) in two or more stages, with unequal inclusion probabilities. The sampling unit is a household and geographic information at parcel level is rarely available. For RS data integration, georeferenced information is required at a minimum. To have at least one georeferenced point by agricultural parcel within the household sample, an additional sampling stage is used to select the parcel centroid.

2.1.1. Area sampling

Three of the five Sen4Stat partner countries use area samples, namely, Spain, Ecuador, and Tanzania. In Spain, the area frame is the national topographic map. The sampling unit is a square segment of side 700 m (49 hectares). The sample design is non-stratified systematic with three random starts. The sample of segments for crop acreage estimation is selected in a single stage with inclusion probability $\pi_i = n/N$, where n and N are the number of sampling units in the sample and population, respectively.

In Ecuador, the sampling frame is a land-use map, stratified into four strata according to the percentage of cultivated area. The sampling unit is a segment with geometric boundaries, whose size changes with strata: 9 hectares (ha) in stratum 1a, 36 ha in stratum 1b, 144 ha in stratum 2, and 576 ha in stratum 3. The inclusion probability is equal within strata, $\pi_{hi} = n_h/N_h$, where n_h and N_h are the number of sampling units in the sample and in the population of stratum h, respectively.

In Tanzania the sampling frame used to be a list frame based on an agricultural census. Recently, in the framework of the GSARS program, the Government of Tanzania in collaboration with the FAO, United States Department of Agriculture (USDA) and African Development Bank, designed an area sample [16]. The sampling frame is a map of the country stratified by land use and the sampling unit is a point. Point sampling has a long history [17,18]. In agricultural statistics, it is used for selecting a sample of farms when a sampling frame is not available; only an accurate map is required [11]. The sample of farms is used for data on a large set of items, generally by direct interview of farmers.

Some of the aforementioned quantities, including crop acreage and yield and natural resources, can be observed directly on the ground, bypassing the need to contact farmers for data collection. France is an example of countries using point sampling for crop acreage estimation, using data collected directly on the ground [19]. Furthermore, the USA is an example of countries using point sampling for a national inventory of natural resources [20,21].

The sampling design is basically the same in the two aforementioned countries, i.e., a two-stage sampling in which the primary sampling unit (PSU) is a square segment and the secondary sampling unit (SSU) is a point. In the USA National Resources Inventory survey, the segments are stratified, and the side of the typical segment is 800 meters (1/2 mile), i.e., 64 hectares, with some smaller (200 meters) and some larger (1600 meters). The primary sampling rates are generally 2% to 6% of the land area. There are about 300,000 PSUs and 844,000 points, and the second-stage sample size is between 1 and 3 points per PSU.

In the French TER-UTI, the sampling design is non-stratified, one-stage, and systematic. The national territory is divided into square blocks with side 12000 meters. Each block is divided into four square sections of side 6000 meters. In each section, a systematic sample of 36 points is selected, aligned in both row and column directions, separated by a distance of 300 meters, forming a 6×6 point grid.

This sampling scheme can be seen as two-stage, so the inclusion probability is of the form $\pi_{i_1 i_2} = \pi_{i_1}\pi_{i_2|i_1}$, where π_{i_1} and $\pi_{i_2|i_1}$ are the first and second stage inclusion probabilities, respectively.

2.1.2. List sampling

Two of the five Sen4Stat partner countries use list sampling, Malawi and Senegal. In Malawi, the ground data were collected in the framework of the LSMS-ISA program. The IHS5 sampling frame is based on the 2018 Malawi Population and Housing Census. It is stratified into rural and urban strata and each of the 27 districts were considered a separate sub-stratum, part of the main rural stratum. A two-stage sampling scheme was used to select the sample from each district. The PSU is the Enumeration Area (EA) and the first-stage sample is systematic without replacement, with probabilities proportional to the number of households in the PSUs. The SSU is a household and a systematic sample of 16 households was used to choose the second-stage sample, using equal probabilities. To integrate with RS data, a georeferenced point is selected in each parcel of the sample [22].

In Senegal, the sampling frame is based on the 2013 population census [23]. The sampling unit is a household and the sample was designed within the AGRIS program framework. This is a two-stage sampling scheme in which the PSU is the EA, and the first-stage sample is selected with replacement and probabilities proportional to the number of households in the PSUs. The SSU is a farm (households with agricultural activities) and the second-stage sample is selected without replacement and with equal probabilities among households with agricultural activities.

To integrate the data in the list sample of farms with RS data, a point is selected in each parcel within the two-stage sample. The inclusion probability of this point is $\pi_{i_1 i_2 i_3 i_4} = \pi_{i_1}\pi_{i_2|i_1}$, where π_{i_1} and $\pi_{i_2|i_1}$ are the first and second stage inclusion probabilities, respectively. This is so because if the SSU $i_1 i_2$ is included in the sample, then the set of parcels $i_1 i_2 i_3$ are also included, $\pi_{i_3|i_1 i_2} = 1$; the georeferenced point $i_1 i_2 i_3 i_4$ is not selected at random but is the parcel centroid $\pi_{i_4|i_1 i_2 i_3} = 1$.

2.2. Statistical models

Statistical models are used to integrate the auxiliary extra-sample information into the sampling design without loss of design-consistency [8]. We used linear models if field data were continuous and multinomial models if those data were categorical [9,24].

2.2.1. Linear models

We consider a population of N units, say agricultural parcels. If geographic information on parcel boundaries is available, it is possible to generate the required remote sensing (RS) data at unit level. We then consider a linear model $y_i = x_i\beta + \varepsilon_i$ relating ground data in the i^{th} unit y_i, with the values of a set of RS variables associated with this same unit, $x_i = \underset{1\leqslant l\leqslant L}{row}(x_{l_i})$. The parameter vector $\beta = \underset{1\leqslant l\leqslant L}{col}(\beta_l)$ is unknown but can be estimated. ε_i represents unobservable zero-mean random perturbations that, conditionally on x_i, account for the variability of y_i about its expected value $Ey_i = x_i\beta$.

2.2.1.1. National-level estimators

The survey variable total in the population is $y_N = \sum_{i=1}^{N} y_i$ and the mean is $\bar{y}_N = \frac{1}{N}\sum_{i=1}^{N} y_i$. If x_i contains 1 (say $x_{1i} = 1$), then $y_N = x_N B_N$. Here, $x_N = \sum_{i=1}^{N} x_i$, and $B_N = \left(X_N^T X_N\right)^{-1} X_N^T y_N$ designates the vector of population regression coefficients $B_N = \underset{1\leqslant l\leqslant L}{col}(B_l)$, where $X_N = \underset{1\leqslant i\leqslant N}{row}\ \underset{1\leqslant l\leqslant L}{col}(x_{li})$ and $y_N = \underset{1\leqslant i\leqslant N}{col}(y_i)$. Since x_i is known for every population unit $i = 1, 2, \ldots, N$, then x_N and X_N are known. However, y_N is unknown and so is B_N.

To estimate B_N, we use the data collected in the sample of size n units, selected from the population with inclusion probabilities $\{\pi_i; i = 1, 2, \ldots, N\}$. The design-based estimator $\hat{B}_\pi = \left(\sum_{i=1}^{n} \frac{x_i^T x_i}{\pi_i}\right)^{-1} \sum_{i=1}^{n} \frac{x_i^T y_i}{\pi_i}$ is design-consistent for B_N and as a result, the synthetic or projective estimator $\hat{y}_N = x_N\hat{B}_\pi$ is design-consistent for y_N. This same estimator can be written in the form $\hat{y}_N = y_\pi + (x_N - x_\pi)\hat{B}_\pi$, called a generalized regression estimator, where $y_\pi = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$ and $x_\pi = \sum_{i=1}^{n} \frac{x_i}{\pi_i}$. The projective estimator can be expressed as weighted sum $\hat{y}_N = \sum_{i=1}^{n} w_i y_i$ with weight $w_i = \frac{g_i}{\pi_i}$, where $g_i = 1 + \left(\sum_{i=1}^{N} x_i - \sum_{i=1}^{n} \frac{x_i}{\pi_i}\right)\left(\sum_{i=1}^{n} \frac{x_i^T x_i}{\pi_i}\right)^{-1} x_i^T$ [12].

The error of $\hat{y}_N$ as an estimator of y_N is $\hat{y}_N - y_N = x_N\left(\hat{B}_\pi - B_N\right)$ and the sampling variance is $V(\hat{y}_N - y_N) = x_N V\left(\hat{B}_\pi - B_N\right)x_N^T$, where $V\left(\hat{B}_\pi - B_N\right) = \left(\sum_{i=1}^{N} x_i^T x_i\right)^{-1} V \sum_{i=1}^{n} \frac{x_i^T \varepsilon_{iN}}{\pi_i}\left(\sum_{i=1}^{N} x_i^T x_i\right)^{-1}$ and $\varepsilon_{iN} = y_i - x_i B_N$. A designconsistent estimator of the sampling variance is $\hat{V}(\hat{y}_N - y_N) = V\left(\sum_{i=1}^{n} \frac{\hat{\varepsilon}_{iN}}{\pi_i}\right)$, where $V(.)$ is the design-based variance and $\hat{\varepsilon}_{iN} = y_i - x_i\hat{B}_\pi$.

The asymptotic distribution of $\hat{y}_N$ is $[V(\hat{y}_N - y_N)]^{-1/2}(\hat{y}_N - y_N) \to N(0, 1)$ and can be used for constructing confidence intervals for y_N.

A design-consistent estimator for the mean is $\hat{\bar{y}}_N = \frac{1}{N}\hat{y}_N$. Its sampling variance is $V\left(\hat{\bar{y}}_N - \bar{y}_N\right) = \frac{1}{N^2}V(\hat{y}_N - y_N)$ and can be estimated using $\hat{V}(\hat{\bar{y}}_N -$

$\bar{y}_N) = \frac{1}{N^2}\hat{V}\left(\hat{y}_N - y_N\right)$. Its asymptotic distribution is the same as that of $\hat{y}_N$.

2.2.1.2. Domain-level estimators

A domain is a part of the population, for instance, a region R or province within a country. Let $N_R = \sum_{i=1}^{N} I_i$ be the domain size, where $I_i = 1$ if unit i belongs to the domain, and $I_i = 0$ otherwise. The survey variable total in the domain is $y_{N_R} = \sum_{i=1}^{N_R} y_i$ and the mean is $\bar{y}_{N_R} = \frac{1}{N_R}\sum_{i=1}^{N_R} y_i$. To estimate y_{N_R}, we use the sample s of size n selected from the population with inclusion probabilities $\{\pi_i; i = 1, 2, \ldots, N\}$ and the estimator $\hat{y}_{N_R} = x_{N_R}\hat{B}_\pi + \frac{N_R}{\hat{N}_R}\sum_{i=1}^{n_R}\frac{y_i - x_i\hat{B}_\pi}{\pi_i}$. Here, $n_R = \sum_{i=1}^{n} I_i$ is the number of units in the sample belonging to the study domain and $\hat{N}_R = \sum_{i=1}^{n}\frac{I_i}{\pi_i}$ is an estimator of N_R. If N_R is unknown, then we use the estimator $\hat{y}_{N_R} = x_{N_R}\hat{B}_\pi + \sum_{i=1}^{n_R}\frac{y_i - x_i\hat{B}_\pi}{\pi_i}$. The sampling variance is $V\left(\hat{y}_{N_R} - y_{N_R}\right) = N_R^2 V\frac{1}{\hat{N}_R}\left(\sum_{i=1}^{n_R}\frac{\varepsilon_{iN}}{\pi_i}\right)$. A design-based estimator of $V\left(\hat{y}_{N_R} - y_{N_R}\right)$ is $\hat{V}\left(\hat{y}_{N_R} - y_{N_R}\right) = N_R^2 V\frac{1}{\hat{N}_R}\left(\sum_{i=1}^{n_R}\frac{\hat{\varepsilon}_{iN}}{\pi_i}\right)$.

2.2.1.3. Model-based small area estimation

The sample was designed to achieve the required estimate accuracy at national level. However, reliable estimates in small administrative areas such as a municipality are also required without increasing sample size n. In a minor administrative area, n will always be smaller than at national level and, as a result, so will be the estimators' accuracy. A small area is a part of the population of which, owing to the small sample size, the design-based estimator is not sufficiently accurate for most uses and the design consistency requirement is meaningless.

For estimates at municipality level, we used a model-based estimator to "borrow strength" from related small areas to obtain accurate estimates for a given small area. Let $\{(y_{di}, x_{di}); i = 1, 2, \ldots, n_d; d = 1, 2, \ldots, D\}$ be the available dataset, where y_{di} represents ground data in unit i of the small area d, x_{di} is the vector of RS data, n_d is sample size in the small area d, and D is the number of small areas, so that $n = \sum_{d=1}^{D} n_d$.

We consider the unit-level linear mixed model $y_{di} = x_{di}\beta + u_d + \varepsilon_{di}$, where (u_d, ε_{di}) are zero-mean independent random variables of variances $\left(\sigma_u^2, \sigma_e^2\right)$. We assume the same regression parameters β and same variance components $\left(\sigma_u^2, \sigma_e^2\right)$, through small areas.

The model-based estimator of the total survey variable in a small area $y_{N_d} = \sum_{i=1}^{N_d} y_{di}$ is $\hat{y}_{N_d} = N_d\left[\left(1 - \hat{\gamma}_d\right)\bar{x}_{N_d}\hat{\beta} + \hat{\gamma}_d\left(\bar{y}_{n_d} + \left(\bar{x}_{N_d} - \bar{x}_{n_d}\right)\hat{\beta}\right)\right]$, where $\hat{\beta} = \left(X^T\hat{V}^{-1}X\right)^{-1}X^T\hat{V}^{-1}y$ is the generalized least-square estimator of β, with $y = \underset{1\leqslant d\leqslant D}{col}\ \underset{1\leqslant i\leqslant n_d}{col}(y_{di})$, $X = \underset{1\leqslant d\leqslant D}{col}\ \underset{1\leqslant i\leqslant n_d}{col}(x_{di})$, $\hat{V}^{-1} = \underset{1\leqslant d\leqslant D}{diag}\left(\hat{V}_d^{-1}\right)$ and $\hat{V}_d^{-1} = \frac{1}{\hat{\sigma}_e^2}I_{n_d} - \frac{\hat{\gamma}_d}{n_d\hat{\sigma}_e^2}1_{n_d}1_{n_d}^T$, where $\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_d}$. Here, $\hat{\sigma}_e^2 = \frac{e^Te}{n-D-1}$ and $\hat{\sigma}_u^2 = \frac{u^Tu - (n-2)\hat{\sigma}_e^2}{n_*}$, with $n_* = \sum_{d=1}^{D} n_d\left[1 - n_d\bar{x}_d\left(X^TX\right)^{-1}\bar{x}_d^T\right]$, are unbiased estimators of the variance components, where e^Te is the sum of squared residuals in the model fitted by ordinary least squares, taking as fixed the small-area effect u_d. u^Tu is the sum of squared residuals in the model fitted by ordinary least squares, with $u_d = 0$. $\bar{x}_{N_d}$ and $\bar{x}_{n_d}$ are the population and sample means of the vector x_{di}, respectively.

An unbiased estimator of the total mean-square error estimator $MSE\left(\hat{y}_{N_d}\right)$ is [25]:

$$\hat{MSE}(\hat{y}_{N_d}) = N_d^2\left(1 - \frac{n_d}{N_d}\right)^2 [h_{1d}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) + h_{2d}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) + 2h_{3d}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)],$$

where:

$$h_{1d}\left(\hat{\sigma}_u^2, \hat{\sigma}_e^2\right) = \hat{\gamma}_d\left(\frac{\hat{\sigma}_e^2}{n_d}\right) + \left(1 - \frac{n_d}{N_d}\right)^2 \frac{(N_d - n_d)}{N_d^2}$$

$$h_{2d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) = \hat{\sigma}_e^2\left(\bar{X}_{N_d - n_d} - \hat{\gamma}_d\bar{x}_d\right)A^{-1}\left(\bar{X}_{N_d - n_d} - \hat{\gamma}_d\bar{x}_d\right)^T$$

$$h_{3d}\left(\hat{\sigma}_u^2, \hat{\sigma}_e^2\right) = \frac{1}{n_d^2}\frac{1}{\left(\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_d}\right)^3}[\left(\hat{\sigma}_e^2\right)^2 V\hat{\sigma}_u^2 + \left(\hat{\sigma}_u^2\right)^2 V\hat{\sigma}_e^2 - 2\hat{\sigma}_e^2\hat{\sigma}_u^2 Cov\left(\hat{\sigma}_e^2, \hat{\sigma}_u^2\right)],$$

where

$$A = \sum_{d=1}^{D}\left(\sum_{i=1}^{n_d} x_{di}^Tx_{di} - \hat{\gamma}_d n_d\bar{x}_{n_d}^T\bar{x}_{n_d}\right)$$

$$V\left(\hat{\sigma}_u^2\right) = \frac{2}{n_*^2}\left[\frac{1}{n - D - 1}(D - 1)(n - 2)\left(\hat{\sigma}_e^2\right)^2 + 2n_*\hat{\sigma}_e^2\hat{\sigma}_u^2 + n_{**}\left(\hat{\sigma}_u^2\right)^2\right]$$

$$V\left(\hat{\sigma}_e^2\right) = \frac{2\left(\hat{\sigma}_e^2\right)^2}{n - D - 1}$$

$$Cov\left(\hat{\sigma}_e^2, \hat{\sigma}_u^2\right) = -\frac{1}{n_*}(D - 1)V\hat{\sigma}_e^2$$

Here, $n_{**} = \sum_{d=1}^{D} n_d^2 \left(1 - n_d \bar{x}_{n_d} A_1^{-1} \bar{x}_{n_d}^T\right) + tr\left(A_1^{-1} \sum_{d=1}^{D} n_d^2 \bar{x}_{n_d}^T \bar{x}_{n_d}\right)^2$. Because $A_1 = \sum_{d=1}^{D} \sum_{i=1}^{n_d} x_{di}^T x_{di}$, n_{**} may be simplified to $n_{**} = \sum_{d=1}^{D} n_d^2 \left[1 - \bar{x}_{n_d} \left(X^T X\right)^{-1} \bar{x}_{n_d}^T\right] = n_* - n + \sum_{d=1}^{D} n_d^2$.

2.2.2. Multinomial logit models

If geographic information on parcel boundaries is unavailable, then it is not possible to generate RS data at parcel level. In this case, we use a point as the unit of observation. Because a point is dimensionless, observed crop data are necessarily categorical. For integrating categorical ground and RS data, multinomial logit is a suitable model [9,10,24].

We consider the survey vector $y_i = \underset{1 \leqslant k \leqslant K}{col}(y_{ik})$, where $y_{ik} = 1$ if crop k covers pixel i and $y_{ik} = 0$ otherwise. K is the total number of crops so that the constraint $\sum_{k=1}^{K} y_{ik} = 1$ holds. We focus on high-resolution satellite images and the population size N is the number of pixels in the study area $A = aN$, where a is the area of the piece of land represented by a pixel.

2.2.2.1. National-level estimators

The population total of the survey vector is $y_N = \sum_{i=1}^{N} y_i = \underset{1 \leqslant k \leqslant K}{col}\left(\sum_{i=1}^{N} y_{ik}\right) = \underset{1 \leqslant k \leqslant K}{col}(y_{Nk})$, where y_{Nk} is the number of pixels covered by crop k. It is assumed that a pixel is covered by only one crop (or that a class of mixed pixels is included), so that the area covered by k is $A_k = a y_{Nk}$ and the population area is $A = \sum_{k=1}^{K} A_k$. We want to estimate the total y_N or mean $\bar{y}_N = \frac{1}{N} \sum_{i=1}^{N} y_i = \underset{1 \leqslant k \leqslant K}{col}\left(\frac{1}{N} \sum_{i=1}^{N} y_{ik}\right) = \underset{1 \leqslant k \leqslant K}{col}\left(\frac{y_{Nk}}{N}\right)$, where $\frac{y_{Nk}}{N}$ is the proportion of pixels covered by k.

We assume that the survey vector $y_i = \underset{1 \leqslant k \leqslant K}{col}(y_{ik})$ follows a multinomial distribution $MN(1, \mu_i)$, where $\mu_i = \underset{1 \leqslant k \leqslant K}{col}(\mu_{ik})$ and μ_{ik} is the probability that unit i is of category k, i.e., the probability of $y_{ik} = 1$ and $y_{ik'} = 0$; $\forall k' \neq k$, with the constraint $\sum_{k=1}^{K} \mu_{ik} = 1$. We estimate y_N using an estimator of μ_{ik} based on a sample of pixels $\{(y_i, x_i s)\,; i = 1, 2, \ldots, n\}$ selected with inclusion probability π_i, where $x_i = \underset{1 \leqslant l \leqslant L}{row}(x_{il})$ is the vector of RS data.

To link μ_{ik} with RS data, we consider a multinomial logit model $\mu_{ik} = \frac{\exp(x_i \beta_k)}{\sum_{k=1}^{K} \exp(x_i \beta_k)}$, where $\beta_k = \underset{1 \leqslant l \leqslant L}{col}(\beta_{kl})$ is an unknown parameter vector. To estimate the probability μ_{ik} that the cover of pixel i is crop $k = 1, 2, \ldots, K$, it is sufficient to acquire estimates of β_k for $k = 1, 2, \ldots, K-1$ [9]. The probability of category K is $\mu_{iK} = 1 - \sum_{k=1}^{K-1} \mu_{ik}$ and can be estimated using the estimates $\hat{B}_{\pi k}$ for $k = 1, 2, \ldots, K-1$.

The design-based parameter estimator $\hat{B}_\pi = \underset{1 \leqslant k \leqslant K-1}{col}\left(\hat{B}_{\pi k}\right) = \underset{1 \leqslant k \leqslant K-1}{col}\ \underset{1 \leqslant l \leqslant L}{col}\left(\hat{B}_{\pi kl}\right)$ is design-consistent for the maximum likelihood estimator of $\beta = \underset{1 \leqslant k \leqslant K-1}{col}(\beta_k)$ based on the population data $\{(y_i, x_i)\,; i = 1, 2, \ldots, N\}$, considered a simple random sample of the multinomial model. This can be found iteratively using $\hat{B}_\pi^{(m+1)} = \hat{B}_\pi^{(m)} + \left[\sum_{i=1}^{n} \frac{H(y_i, \beta)}{\pi_i}\Big|_{\beta = \hat{B}_\pi^{(m)}}\right]^{-1} \sum_{i=1}^{n} \frac{b(y_i, \beta)}{\pi_i}\Big|_{\beta = \hat{B}_\pi^{(m)}}$ [13], where $b(y_i, \beta) = \underset{1 \leqslant k \leqslant K-1}{col}\ \underset{1 \leqslant l \leqslant L}{col}\left[(y_{ik} - \mu_{ik}) x_{il}\right]$ and $H(y_i, \beta) = \left[\underset{1 \leqslant k \leqslant K-1}{col}\ \underset{1 \leqslant k' \leqslant K-1}{row}\left(\delta_{kk'} \mu_{ik} - \mu_{ik} \mu_{ik'}\right)\right] \otimes x_i^T x_i$, with $\delta_{kk'} = 1$ if $k = k'$ and $\delta_{kk'} = 0$ otherwise A design-consistent estimator of $y_N = \underset{1 \leqslant k \leqslant K}{col}(y_{Nk})$ is $\hat{y}_N = \sum_{i=1}^{N} \hat{\mu}_i + \frac{N}{\hat{N}_p} \sum_{i=1}^{n} \frac{y_i - \hat{\mu}_i}{\pi_i} = \underset{1 \leqslant k \leqslant K-1}{col}(\hat{y}_{Nk})$, where $\hat{\mu}_i = \underset{1 \leqslant k \leqslant K-1}{col}(\hat{\mu}_{ik})$, $\hat{\mu}_{ik} = \frac{\exp\left(x_i \hat{B}_{\pi k}\right)}{1 + \sum_{k=1}^{K-1} \exp\left(x_i \hat{B}_{\pi k}\right)}$ for $k = 1, 2, \ldots, K-1$, $\hat{\mu}_{iK} = 1 - \sum_{k=1}^{K-1} \hat{\mu}_{ik} = \frac{1}{1 + \sum_{k=1}^{K-1} \exp\left(x_i \hat{B}_{\pi k}\right)}$, $\hat{N}_p = \sum_{i=1}^{n} \frac{1}{\pi_i}$, and $\hat{y}_{Nk} = \sum_{i=1}^{N} \hat{\mu}_{ki} + \frac{N}{\hat{N}_p} \sum_{i=1}^{n} \frac{y_{ki} - \hat{\mu}_{ki}}{\pi_i}$ is the estimator of the total number of pixels covered by crop k for $k = 1, 2, \ldots, K-1$; for category K it is $\hat{y}_{NK} = N - \sum_{k=1}^{K-1} \hat{y}_{Nk}$. The sampling covariance matrix of $\hat{y}_N$ is given approximately by $V\hat{y}_N = N^2 V \frac{1}{\hat{N}_p} \sum_{i=1}^{n} \frac{y_i - \mu_i}{\pi_i} = \underset{1 \leqslant k \leqslant K-1}{col}\ \underset{1 \leqslant k' \leqslant K-1}{row}\left(N^2 Cov\left(\hat{y}_{rk}, \hat{y}_{rk'}\right)\right)$, where $\hat{y}_{rk} = \frac{1}{\hat{N}_p} \sum_{i=1}^{n} \frac{y_{ki} - \mu_{ki}}{\pi_i}$ is a function of $\hat{N}_{pk} = \sum_{i=1}^{n} \frac{y_{ki}}{\pi_i}$, the estimators of the total number of sampling units of category k for $k = 1, 2, \ldots, K$ (because $\hat{N}_p = \sum_{k=1}^{K} \hat{N}_{pk}$), and of the estimator of the total residuals $\hat{r}_{kN} = \sum_{i=1}^{n} \frac{y_{ki} - \hat{\mu}_{ki}}{\pi_i}$ of category k. We focus on the sampling variance $V\hat{y}_{Nk}$ for $k = 1, 2, \ldots, K-1$, which are the elements along the diagonal of $V\hat{y}_N$. Let $\hat{G}_k = row(g_j)_{1 \leqslant j \leqslant K+1} = \left[\hat{r}_{kN}\ \hat{N}_{p1}\ \hat{N}_{p2} \ldots\ \hat{N}_{pK}\right]$ be a row vector whose $K+1$ components are the estimators on which $\hat{y}_{rk}$ depends. The $\hat{y}_{Nk}$ sampling variance is given approximately by $V\hat{y}_{Nk} = \underset{1 \leqslant j \leqslant K+1}{row}\left(\frac{\partial \hat{y}_{rk}}{\partial g_j}\right) V\hat{G}_k \underset{1 \leqslant j \leqslant K+1}{col}\left(\frac{\partial \hat{y}_{rk}}{\partial g_j}\right)$, where $V\hat{G}_k = \underset{1 \leqslant j \leqslant K+1}{col}\ \underset{1 \leqslant j' \leqslant K+1}{row}\left(Cov\left(g_j, g_j'\right)\right)$ is the design-based covariance matrix of $\hat{G}_k$. The sampling covariance matrix $\hat{y}_N$ is estimated replacing μ_i by $\hat{\mu}_i$ in $V\hat{y}_N$ [13].

2.2.2.2. Domain-level estimators

The survey variable total in the domain R is $y_{N_R} = \sum_{i=1}^{N_R} y_i = \underset{1\leqslant k\leqslant K}{col}\left(\sum_{i=1}^{N_R} y_{ik}\right) = \underset{1\leqslant k\leqslant K}{col}\left(y_{N_R k}\right)$, where $y_{N_R k}$ is the number of pixels in R covered by crop k. To estimate y_{N_R}, we use the sample s of size n selected from the population with inclusion probabilities $\{\pi_i; i = 1, 2, \ldots, N\}$ and the estimator $\hat{y}_{N_R} = \sum_{i=1}^{N_R} \hat{\mu}_i + \frac{N_R}{\hat{N}_{Rp}} \sum_{i=1}^{n_R} \frac{y_i - \hat{\mu}_i}{\pi_i} = \underset{1\leqslant k\leqslant K}{col}\left(\hat{y}_{N_R k}\right)$. Here, $n_R = \sum_{i=1}^{n} I_i$ is the number of units in the sample belonging to the study domain, and $\hat{N}_{Rp} = \sum_{i=1}^{n_R} \frac{I_i}{\pi_i}$. The sampling variance of y_{N_R} is given approximately by $V\hat{y}_{N_R} = N_R^2 V \frac{1}{\hat{N}_{Rp}} \sum_{i=1}^{n_R} \frac{y_i - \mu_i}{\pi_i} = \underset{1\leqslant k\leqslant K-1}{col}\ \underset{1\leqslant k'\leqslant K-1}{row}\left(N_R^2 Cov\left(\hat{y}_{Rrk}, \hat{y}_{Rrk'}\right)\right)$. Let $\hat{y}_{Rrk} = \frac{1}{\hat{N}_{Rp}} \sum_{i=1}^{n_R} \frac{y_{ki} - \mu_{ki}}{\pi_i}$ and $\hat{G}_{Rk} = \underset{1\leqslant j\leqslant K+1}{row}\left(g_{Rj}\right) = [\hat{r}_{kN_R}\ \hat{N}_{N_R p1}\ \hat{N}_{N_R p2} \ldots\ \hat{N}_{N_R pK}]$. The $\hat{y}_{N_R k}$ sampling variance is given approximately by $V\hat{y}_{N_R k} = \underset{1\leqslant j\leqslant K+1}{row}\left(\frac{\partial \hat{y}_{Rrk}}{\partial g_{Rj}}\right) V\hat{G}_{Rk}\ \underset{1\leqslant j\leqslant K+1}{col}\left(\frac{\partial \hat{y}_{Rrk}}{\partial g_{Rj}}\right)$, where $V\hat{G}_{Rk} = \underset{1\leqslant j\leqslant K+1}{col}\ \underset{1\leqslant j'\leqslant K+1}{row}\left(Cov\left(g_{Rj}, g_{Rj'}\right)\right)$. The sampling covariance matrix $\hat{y}_{N_R}$ is estimated replacing μ_i by $\hat{\mu}_i$ in $V\hat{y}_{N_R}$.

Works in the literature [26–28] for small area estimation based on multinomial mixed models follow a penalized quasi-likelihood approach. As pointed out by McCulloch and Searle [24], these methods are not completely satisfactory in practice. Those authors recommended instead a linearization of the non-linear multinomial models and used linear mixed models. Additional research is required to achieve a completely satisfactory solution to this problem

2.2.3. Cost efficiency

Cost efficiency is the usual criterion for comparing a set of sampling strategies developed for estimating the same characteristic in the same population. Let C_{GD} be the cost and V_{GD} the sampling error of the current sampling strategy using only ground data, and let C_{GD+RS} and V_{GD+RS} be the cost and sampling error, respectively, of the strategy integrating ground and RS data. The cost efficiency is $C_{GD}V_{GD}$ for the former and $C_{GD+RS}V_{GD+RS}$ for the latter.

Although the Sentinel images and software required for the integration of ground and RS data are provided by ESA for free, there are costs for the NSOs such as the time of experts and commercial cloud services required for storage and computations. Because estimation of these other costs is a difficult task, we assume in the following that $C_{GD+RS} \simeq C_{GD}$, so the comparison criterion reduces to the efficiency.

The relative efficiency of RS data with respect to ground data is $RE_{RS} = V_{GD}\left(V_{GD+RS}\right)^{-1}$. If $RE_{RS} > 1$, then, using RS data, the current sampling error V_{GD} may be reduced to $V_{GD} - V_{GD+RS} = RS_{effect}V_{GD}$ without increasing the survey cost. Here, $RS_{effect} = 1 - RE_{RS}^{-1}$ is the effect of RS data on the sampling efficiency. In other words, using RS data, the current sample size n can be reduced without loss of accuracy in a quantity equal to $n - n_{RS} = RS_{effect}n$. Thus, using RS data, the ground sample size can be $n_{RS} = (1 - RS_{effect})\, n = RE_{RS}^{-1} n$.

We evaluate the effect of RS data on the cost efficiency using either linear or multinomial models. Spain and Ecuador are partner countries using area samples selected in only one stage, with equal inclusion probabilities, and we consider linear models for RS data integration. In Spain, the sampling frame is not stratified and the inclusion probability is $\pi_i = n/N$. In Ecuador, the sampling frame is stratified and the inclusion probability is $\pi_{hi} = n_h/N_h$ for every sampling unit of the same stratum $i = 1, 2, \ldots, N_h$, with $h = 1, 2, \ldots, H$.

The sampling variance using only ground data is $V_G = V\hat{y}_N = V\sum_{i=1}^{n} \frac{y_i}{\pi_i} = V\sum_{i=1}^{n} \frac{y_i}{n/N} = N^2 1/n \sum_{i=1}^{n} y_i = N^2\left(1 - n/N\right) 1/n\left(n-1\right) \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$ in Spain, and $V_G = V\hat{y}_N = V\sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{y_{hi}}{\pi_{hi}} = V\sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{y_{hi}}{n_h/N_h} = \sum_{h=1}^{N_h} N_h^2 1/n_h \sum_{i=1}^{n_h} y_{hi} = \sum_{h=1}^{H} N_h^2\left(1 - n_h/N_h\right) 1/n_h\left(n_h - 1\right) \sum_{i=1}^{n_h} \left(y_{hi} - \bar{y}_h\right)^2$ in Ecuador.

Using RS data, the sampling variance is $V_{GD+RS} = V\hat{y}_N = V\sum_{i=1}^{n} \frac{y_i - x_i\hat{B}}{n/N} = N^2 V 1/n \sum_{i=1}^{n} \left(y_i - x_i\hat{B}\right) = N^2\left(1 - n/N\right) 1/n\left(n-2\right) \sum_{i=1}^{n} \left(y_i - x_i\hat{B}\right)^2$ in Spain, where $\hat{B} = \left(\sum_{i=1}^{n} x_i^T x_i\right)^{-1} \sum_{i=1}^{n} x_i^T y_i$, and $V_{GD+RS} = V\hat{y}_N = \sum_{h=1}^{H} N_h^2\left(1 - n_h/N_h\right)\left[1/n_h\left(n_h - 2\right)\right] \sum_{i=1}^{n_h} \left(y_{hi} - x_{hi}\hat{B}_\pi\right)^2$ in Ecuador, where $\hat{B}_\pi = \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{x_{hi}^T x_{hi}}{\pi_{hi}}\right)^{-1} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{x_{hi}^T y_{hi}}{\pi_{hi}}$.

Using RS data, the current sampling error V_{GD} can be reduced to a quantity $V_{GD} - V_{GD+RS} = RS_{effect}V_{GD}$. In Spain, $RE_{RS} \simeq \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 \left(\sum_{i=1}^{n} \left(y_i - x_i\hat{B}\right)^2\right)^{-1}$ and $RS_{effect} = 1 - RE_{RS}^{-1}$. In Ecuador, $RS_{effect} = \sum_{h=1}^{H} c_h n_h V_{Gh} RS_{effect,h} \left(\sum_{h=1}^{H} c_h n_h V_{Gh}\right)^{-1}$, where $RS_{effect,h} = 1 - RE_{RS,h}^{-1}$, $RE_{RS,h} \simeq \sum_{i=1}^{n_h} \left(y_{hi} - \bar{y}_h\right)^2 \left(\sum_{i=1}^{n_h} \left(y_{hi} - x_{hi}\hat{B}_\pi\right)^2\right)^{-1}$, and c_h is the cost per sampling unit in stratum h.

Senegal, Malawi, and Tanzania use point sampling with unequal inclusion probabilities $\{\pi_i; i = 1, 2, \ldots, N\}$, and we consider multinomial models for RS data integration. The sampling variance of the y_{Nk} estimator using only ground data is approximately $V_G =$

$N^2 V \frac{1}{\hat{N}_p} \sum_{i=1}^{n} \frac{y_{ki}}{\pi_i}$. Using ground and RS data, the sampling error is $V_{G+RS} = N^2 V \frac{1}{\hat{N}_p} \sum_{i=1}^{n} \frac{y_{ki}-\mu_{ki}}{\pi_i}$. The relative efficiency of RS data for estimating the total number of pixels covered by crop k is $RE_{RSk} = V \frac{1}{\hat{N}_p} \sum_{i=1}^{n} \frac{y_{ki}}{\pi_i} \times \left(V \frac{1}{\hat{N}_p} \sum_{i=1}^{n} \frac{y_{ki}-\mu_{ki}}{\pi_i} \right)^{-1}$.

RS cost-efficiency relies on the RS contribution to reduce the sampling variance of the residuals $y_{ki} - \mu_{ki}$ with respect to the sampling variance of the ground data y_{ki}. The sampling variance reduction depends on the reliability of the RS data to estimate the y_{ki} expected values, $Ey_{ki} = \mu_{ki}$. Consequently, the accuracy and timeliness of the x_i datasets required to estimate μ_i are key to achieve a significant gain of cost-efficiency.

2.2.4. Other applications

As mentioned in the Introduction, "multi-seasonal estimation" and "optimizing the sample design" are applications of RS data useful for agricultural statistics. We developed prototypes for these two applications but we cannot test them here because the required data are unavailable. Thus, we are limited to outlining the developed prototypes.

2.2.4.1. Multi-seasonal estimation

We consider the annual agricultural cycle divided into, say, four seasons. We use the subscript t_m to refer to season $m = 1, 2, 3, 4$ of year $t = 1, 2, \ldots$. We want to estimate the total $y_{N_{t_m}}$ of the survey variable in m of t and its annual aggregate y_{N_t}.

2.2.4.1.1. The sample

We examine a two-phase sampling strategy. The first-phase sample is the NSO sample already in place and used for agricultural statistics. We use the subscript 1 to refer to this sample, for instance, n_1 denotes its size. We select a secondphase sample specific to each season from among the sampling units within n_1. We choose a splitpanel or supplementary panel sampling design consisting of a panel component and specific component [29].

For the first season, the panel component is a sample of n_{2p} sampling units selected from n_1, and the specific sample for that season is a sample of $n_{2e_{t_1}}$ sampling units selected from among those included in n_1 but not included in n_{2p}. In the second season, the panel component is the same as in season 1, n_{2p}, and the specific sample is a number $n_{2e_{t_2}}$ of sampling units chosen from among those included in n_1 but not included in either n_{2p} or $n_{2e_{t_1}}$. For the third season, the panel component is the same as in seasons 1 and 2 (n_{2p}) and the specific sample is a number $n_{2e_{t_3}}$ of sampling units selected from among those included in n_1 but not included in n_{2p}, $n_{2e_{t_1}}$, or $n_{2e_{t_2}}$. In the fourth season, the panel component is the same as in seasons 1, 2 and 3, and the specific sample is a number $n_{2e_{t_4}}$ of sampling units chosen from among those included in n_1 but not included in n_{2p}, $n_{2e_{t_1}}$, $n_{2e_{t_2}}$, or $n_{2e_{t_3}}$.

2.2.4.1.2. Composite estimators

A composite estimator is a function of single estimators, each defined separately for panel and specific samples. We consider the linear model $y_{t_m i} = x_{t_m i}\beta_{t_m} + \varepsilon_{t_m i}$, referring to season t_m and both panel and specific samples. Model parameters are estimated using the weigthed estimator and data from the secondphase sample n_{2st_m} (with $s = p$ for the panel sample and $s = e$ for the specific sample), with $\hat{B}_{\pi 2 s_{t_m}} = \left(\sum_{i=1}^{n_{2j_{t_m}}} \frac{x_{s_{t_m} i}^T x_{s_{t_m} i}}{\pi_{2 s_{t_m} i}} \right)^{-1} \sum_{i=1}^{n_{2j_{t_m}}} \frac{x_{s_{t_m} i}^T y_{s_{t_m} i}}{\pi_{2 s_{t_m} i}}$. The seasonal survey variable total is estimated separately, using data from the panel and specific samples and projective estimator $\hat{Y}_{N s_{t_m}} = x_{N_{t_m}} \hat{B}_{\pi 2 s_{t_m}}$. A design-consistent estimator of the sampling error variance is $\hat{V}\left(\hat{y}_{N s_{t_m}} - y_{N s_{t_m}}\right) = V \sum_{i=1}^{n_{2 s_{t_m}}} \frac{\hat{\varepsilon}_{i N s_{t_m}}}{\pi_{2 s_{t_m} i}}$, where $\hat{\varepsilon}_{i 2 N s_{t_m}} = y_{2 s_{t_m} i} - x_{2 s_{t_m} i} \hat{B}_{\pi 2 s_{t_m}}$.

Let $y_{N_t} = \underset{1 \leqslant m \leqslant 4}{col}\left(y_{N_{t_m}}\right)$ be the (4×1) vector of survey variable totals in the four seasons of year t, and let $\hat{y}_{N_t} = \underset{1 \leqslant m \leqslant 4}{col} \underset{j=e,p}{col} \left(\hat{y}_{N j_{t_m}}\right)$ be the (8×1) vector of total estimators based on the panel and specific samples. We consider the model $\hat{y}_{N_t} = Z y_{N_t} + e_{N_t}$, where $e_{N_t} = \hat{y}_{N_t} - y_{N_t}$ is the vector of sampling errors and $Z = I_4 \otimes 1_2$ is an (8×4) matrix indicator of the panel or specific sample.

We assume that panel and specific samples in the same season are independent and, as a result, in the error covariance matrix Ve_{N_t} the covariances between $\hat{y}_{N p_{t_m}}$ and $\hat{y}_{N e'_{t_m}}$ are nil. A composite estimator of y_{N_t} is $\hat{y}_{CN_t} = (Z^T (Ve_{N_t})^{-1} Z)^{-1} Z^T (Ve_{N_t})^{-1} \hat{y}_{N_t}$ and its covariance matrix is $V\hat{y}_{CN_t} = (Z^T (Ve_{N_t})^{-1} Z)^{-1}$. The composite estimator of the annual total $y_{N_t} = 1_4^T y_{N_t}$ is $\hat{y}_{CN_t} = 1_4^T \hat{y}_{CN_t}$ and its variance is $V\hat{y}_{CN_t} = 1_4^T V\hat{y}_{CN_t} 1_4$.

To estimate Ve_{N_t}, we use the empirical correlation and an autoregressive model $y_{it_m} = \bar{Y}_{t_m} + e_{it_m}$, where $\bar{Y}_{t_m} = \frac{1}{N_{t_m}} \sum_{i=1}^{N_{t_m}} y_{it_m}$ and e_{it_m} are modeled using $e_{it_m} = u_{it_m} + \varepsilon_{it_m}$. Here, $(u_{it_m}, \varepsilon_{it_m})$ designates the zeromean independent random variable $Cov\left(u_{it_m}, \varepsilon_{it_m}\right) = 0$ and the variance of ε_{it_m} is σ_ε^2. $u_{it_m} = \rho u_{i,t_{m-1}} + \eta_{it_m}$ is an AR(1) stationary process,

where η_{it_m} is a random zeromean perturbation term whose variance is σ^2 [30].

As a result, we have the linear mixed model $y_{it_m} = \bar{Y}_{t_m} + u_{it_m} + \varepsilon_{it_m}$, whose temporal random component $u_{it_m} = \rho u_{i,t_{m-1}} + \eta_{it_m}$ is autocorrelated and $Cov(\eta_{it_m}, \varepsilon_{it_m}) = 0$. In this last model, y_{it_m} represents the observed data in sampling unit i in season t_m and the right term in the equation is interpreted as follows: $\bar{Y}_{t_m} + u_{it_m}$ is the true value of the survey variable and the sum of the population mean $\bar{Y}_{t_m}$ and u_{it_m}. The latter is a specific component associated with sampling unit i in season t_m and represents deviation from $\bar{Y}_{t_m}$ Finally, ε_{it_m} is the measurement error of the aforementioned true value.

The autocovariance function of the process u_{it_m} is $Cov\left(u_{it_m}, u_{it'_m}\right) = C_u(s)$, where $s = |m' - m|$ and $C_u(s) = \sigma^2 \rho^s \left(1-\rho^2\right)^{-1}$. The variance of u_{it_m} is $Vu_{it} = C_u(0) = \sigma^2\left(1-\rho^2\right)^{-1}$ and the autocorrelation function is $Corr\left(u_{it_m}, u_{it'_m}\right) = C_u(s)/Vu_{it_m} = \rho^s$. The autocovariance function of the process e_{it} is $Cov\left(e_{it_m}, e_{it'_m}\right) = Cov\left(u_{it_m}, u_{it'_m}\right) = C_u(s)$ and the variance is $Ve_{it_m} = Vu_{it_m} + V\varepsilon_{it_m} = \sigma^2\left(1-\rho^2\right)^{-1} + \sigma^2_{\varepsilon}$. The autocorrelation function for the process e_{it_m} is $\rho_e(s) = Corr\left(e_{it_m}, e_{it'_m}\right) = C_u(s)/Ve_{it_m} = I_{(s=0)} + \rho^s\left(1+\left(1-\rho^2\right)\kappa\right)^{-1} I_{(s\neq 0)}$, where $\kappa = \sigma^2_{\varepsilon}/\sigma^2$ and $I_{(.)}$ is an indicator variable whose value is 1 if the argument is true and zero otherwise.

The ratio $v = \sigma^2_{\varepsilon}/Ve_{it_m} = \kappa(1-\rho^2)(1+\kappa(1-\rho^2))^{-1}$ is a measure of the weight of the measurement error relative to the total error. The parameters ρ and κ can be estimated through fitting by nonlinear minimum least squares the theoretical correlation model $\rho_e(s) = I_{(s=0)} + \rho^s\left(1+\left(1-\rho^2\right)\kappa\right)^{-1} I_{(s\neq 0)}$ to the empirical correlations in Ve_{N_t}.

2.2.4.2. Optimizing the sampling design

The sample design is optimized to find the design variable values that minimize the sampling variance subject to cost (budget) constraints. The sampling variance depends on the spatial correlation structure of the survey variable and we follow a superpopulation approach to identify this structure. We limit ourselves to area samples and assume that the population values of the survey variable y_i constitute a sample generated according to a second-order stationary random process with the following characteristics [31]: the mean is $Ey_i = \mu$, variance is $Vy_i = \sigma^2$, and the covariance $Cov(y_i, y'_i) = \sigma^2 \rho_y(dist(s_i, s'_i))$ between two elementary observations (y_i, y'_i) at the points of coordinates s_i and s'_i is positive, decreasing when the distance between these points $d = dist(s_i, s'_i)$ increases.

To assess the correlation structure $\rho_y(d)$, theoretical variogram and correlogram models have been proposed in the literature. Two frequently used correlogram models are the exponential, $\rho(u, v|a, \tau) = (1-\tau)e^{-d/a}$, and the spherical $\rho(u, v|a, \tau) = (1-\tau)e^{-d/a}\left[1 - \frac{3}{2}\frac{d}{a} + \frac{d^3}{2a^3}\right]$ if $d \leqslant a$ and $\rho(u, v|a, \tau) = 0$ if $d > a$. Here u is the number of sampling units between s_i and s'_i in the row direction, v the number of sampling units between s_i and s'_i in the column direction, and $d = \sqrt{u^2 + v^2}$. The model parameters are the range rate a and ratio $\tau = \tau_0/(\tau_0 + \tau_d)$. Here τ_0 is the nugget effect, i.e., the variation at or near the origin (independent of distance), τ_d is the partial sill (a function of distance between sampling points) and $(\tau_0 + \tau_d)$ is the sill, i.e. the maximum variation far from the origin.

It was demonstrated by Ambrosio et al. [32] how this approach should be followed to design systematic samples using a land-use map. Here, we propose instead to use a crop-type map [33] to estimate the parameters of the correlogram model. The estimated correlogram can be used to evaluate the anticipated variance [34] as a function of design variables for the set of sampling strategies used in practice [35,36].

2.2.4.2.1. Anticipated variance

We illustrate the proposed approach considering segments of size n_0 and a simple random sample of segments of size n. The variance of the sampling error is $V\hat{y}_N = N^2(1-n/N)S^2/n$, where $S^2 = 1/(N-1)\sum_{i=1}^{N}(y_i - \bar{y}_N)$ is the population variance. The anticipated variance is the model-based expected value of this design-based sampling variance $EV\hat{y}_N = N^2(1-n/N)ES^2/n$, where $ES^2 = \sigma^2\Psi(N, n_0|a, \tau)$ and $\Psi(N, n_0|a, \tau) = n_0(Nn_0 - 1)(N-1)^{-1}[1 - \Phi(N, n_0|a, \tau)] - Nn_0(n_0-1)(N-1)^{-1}[1 - \Phi(n_0|a, \tau)]$. Here, $\Phi(N, n_0|a, \tau)$ is the average correlation between pairs of observations over the $C^2_{Nn_0}$ pairs in the population. $\Phi(n_0|a, \tau)$ is the average correlation between pairs of observations over the $C^2_{n_0}$ pairs in a segment.

2.2.4.2.2. The optimization problem

The design variables are the segment and sample sizes, and we find optimal values of these variables by solving the following optimization problem: $\min_{\{(n,n_0)\}} AV\hat{Y} = \min_{\{(n,n_0)\}} N_h^2(1-n/N)\sigma^2/n\Psi(N, n_0|a, \tau)$, subject to $C_0 + \sum_{h=1}^{L} C_c n + \sum_{h=1}^{L} C_w n n_0 + \sum_{h=1}^{L} C_k\sqrt{An} \leqslant C$. Here, C_c is the cost of adding

a segment to the sample, excluding travel cost but including positioning cost (travel to the first segment visited from the interviewer home base and then back to that base from the last segment visited during the data-collection trip), and C_w is the observing cost, including the cost of locating the segment

The solution to this problem is the optimum segment size n_0 and optimum sample size of segments n. In addition to the budget, this optimum solution is conditioned to the correlogram model parameters (a, τ).

3. Data

We used field data collected by the NSO of Spain to illustrate our approach to the integration of continuous ground data in the sampling design using linear models, and field data collected by the NSO of Senegal to illustrate our approach to the integration of categorical ground data into the sampling design using multinomial models.

The two RS products traditionally used as auxiliary data for crop acreage and yield estimation are pixel classification by crop type (for the former) and a set of vegetation indices (for the latter). For training pixel classification models, ground georeferenced data of crop type is required. In many countries, data from the agricultural sector are collected using national household or farm surveys, and no geographic information at parcel level is available.

Azzari et al. [22] compared several ways of generating the data needed to train pixel classification models when no such parcel-level information is available. Focusing on integrating ground survey data of maize in Malawi and Ethiopia from national household surveys using the Sentinel-2 satellite, the authors evaluated the accuracy of pixel classification producing georeferenced data at parcel level, ranging from the full parcel boundary to only one point (centroid).

The authors concluded that collecting full-parcel boundary data or GPS coordinates of the polygon defined by complete parcel corner points yields the best-quality information for model training; however, the use of mid-sized sample (3000–4000 parcels) plot centroids could perform similarly to full plot boundaries. The authors did not consider the statistical model required to achieve the integration of ground data with Sentinel-2 to improve the accuracy of crop acreage estimates, obtaining design-based consistency.

In Sen4Stat, the focus was on developing an open-source system that permits any user to generate from Sentinel 1 SCL Sentinel-2 L1C and/or L2A images the RS data required to improve agricultural statistics. In addition to vegetation indices, using cloudmask and cloud-free mosaics for optical sensors, and for transforming data from SAR sensors, the system allows the choice of several techniques of supervised classification (including random forest) for pixel classification and crop-type map generation.

Raw data in spectral bands can be directly integrated with ground data using multinomial models. Indeed, as shown by Hogland et al. [10], using spectral-band digital numbers in conjunction with multinomial models is a pixel classification method.

4. Results and discussion

4.1. Spain

4.1.1. Crop acreage estimates at national level

We consider the area covered by an image of size 100 $\times$ 100 km in Castilla y León (Spain), with X coordinates (in meters): (300000,400000) and Y coordinates (4600000,4700000). The sampling unit is a square segment of side 700 meters and sample size in the area is 419 segments.

Ground data are observed at parcel level. However, we aggregated data on crop acreage at segment level for computations, so that y_i represents ground data on acreage of the study crop in segment i. Further, the RS data are $x_i = [1\ x_i]$, where x_i is the number of pixels classified as belonging to the study crop in segment i. Model parameter estimates are $\hat{B}_{1\pi} = 0.046$ for the independent term and $\hat{B}_{2\pi} = 0.903$ for the angular coefficient of x_i. Results for barley data observed in 2018 are in Table 1.

These results show that using RS data, estimator accuracy improved considerably; the amplitude of the confidence interval decreased and the estimation error was reduced by half. The relative efficiency was high; using RS data, the current sample size could be reduced to less than one fifth without loss of accuracy. This is so because RS data are reliable for crop acreage. These results are design-based, which implies that the results change with the sample design. Thus, the reduced sample should be chosen using the same design as the NSO is currently using.

4.1.2. Crop yield estimates

Ground data on yield are observed at parcel level, with y_{ij} representing these data for the study crop in parcel j of segment i. The RS data are denoted

Table 1
Barley acreage estimates in a 100 × 100 km area of Castilla y León, 2018

Data	Acreage (hectare)	Uncertainty: 95% Confidence interval (hectares) Limits	Uncertainty: 95% Confidence interval (hectares) Amplitude	Uncertainty: Coeff. of variation (%)*	Relative efficiency of RS data
Ground	236165.4	Lw: 215951.7 Up: 256379.0	40427.2	4.37	–
Ground + RS	228550.1	Lw: 219699.8 Up: 237400.3	17700.5	1.98	5.2

*Quotient between root square of the sampling variance and estimate.

Table 2a
Barley yield estimates in a 100 × 100 km area of Castilla y León, 2018. NDVI

Data	Yield (kg/hectare)	Uncertainty: 95% Confidence interval (kg/hectare) Limits	Uncertainty: 95% Confidence interval (kg/hectare) Amplitude	Uncertainty: Coeff. of variation (%)	Relative efficiency of RS data
Ground	4213.6	Lw: 4093.9 Up: 4333.2	239.2	1.45	–
Ground + RS	4155.6	Lw: 4033.2 Up: 4278.1	244.9	1.50	0.95

Table 2b
Barley yield estimates in a 100 × 100 km area of Castilla y León, 2019. LAI

Data	Yield (kg/hectare)	Uncertainty: 95% Confidence interval (kg/hectare) Limits	Uncertainty: 95% Confidence interval (kg/hectare) Amplitude	Uncertainty: Coeff. of variation (%)	Relative efficiency of RS data
Ground	2352.0	Lw: 2227.4 Up: 2476.6	249.1	2.70	–
Ground + RS	2327.8	Lw: 2215.1 Up: 2440.6	225.4	2.47	1.22

$x_{ij} = \underset{1 \leqslant l \leqslant L}{row}(x_{ijl})$, a row vector with 1 in the first position and a set of vegetation indices in the remaining positions. The latter include the normalized difference vegetation index (NDVI) and leaf area index (LAI) (sum of LAI simulated by a Savitsky-Golay interpolation, fitting all LAI observations in the growing season) during (i) sprouting, (ii) flowering, and (iii) ripeness, plus (iv) maximum value of the LAI S-G interpolation and yield simulated by the simple algorithm for yield (SAFY) crop-growth model. Results for barley data observed in 2018 are in Table 2a for NDVI only and in Table 2b for LAI and yield simulated by SAFY.

The relationship between crop yield and the vegetation indices is statistically significant. However, their correlation is not as strong as for crop acreage and as a result, using RS data, yield estimator accuracy improves little. Using RS data, the estimation error is of the same order of magnitude as using only ground data. The relative efficiency is nearly 1; using RS data, the current sample size could be reduced little without loss of accuracy. This is because RS data reliability for crop yield is not as great as for crop acreage. Additional research is required to make the production estimates more reliable through an improved estimate for yield from RS; physically based models are more reliable and can be integrated with RS data.

4.1.3. Crop production estimates

We estimate crop production as the product of the crop acreage and yield estimates. Results are in Table 3.

Thanks to improvement in the crop acreage estimator using RS data, the production estimator accuracy increased considerably; the estimation error decreased by half, even if the RS data failed to improve the yield estimate.

4.1.4. Crop acreage estimates at province level

To illustrate provincial estimation using areal sampling, we consider barley acreage estimates at the provincial level within the 100 × 100 km area of Castilla y León (Spain). Results are in Table 4.

The estimate accuracy at provincial level is, as expected, less than at the national level, but the RS relative efficiency is of the same order of magnitude. In provinces where the sampling error using only ground data is very high (León), the RS contribution is qual-

Table 3
Barley production estimates in a 100 × 100 km area of Castilla y León, 2018

Data	Production (tons) (1000 kg)	Uncertainty		
		95% confidence interval (tons: 1000 kg)		Coeff. of variation (%)
		Limits	Amplitude	
Ground	987841	Lw: 903640 Up: 1072042	168402	4.35
Ground + RS	965733	Lw: 915194 Up: 1016272	101078	2.67

Table 4
Barley acreage estimates at provincial level

Province	Using only ground data		Using ground & RS data		Relative efficiency of RS data
	Acreage (has.)	Error (CV%)	Acreage (has.)	Error (CV%)	
León	6853.2	24.15	6834.5	16.20	2.3
Palencia	88602.0	7.36	90535.3	3.33	4.7
Valladolid	128209.5	5.57	119707.4	2.66	5.1
Zamora	12324.2	17.71	10948.2	8.16	6.4
Total area	235989.1	4.37	228028.5	1.98	5.2

Table 5
Barley acreage estimates at municipality level

Municipality	Sample size (segments)	Acreage	
		Hectares	Coeff. of variation (%)
Belver de los Montes	1	212.96	29.1
Castroverde	3	2914.22	8.0
Pinilla de Toro	4	963.30	10.0
Quintanilla del Monte	1	466.65	20.3
Toro	1	615.91	14.0
Vezdemarbán	3	1358.22	12.6
Villalpando	2	560.05	39.1
Villamayor de Campos	1	1056.23	11.1
Villanueva del Campo	1	784.03	13.2
Villar de Fallaves	1	844.16	11.0
Villardondiego	1	516.40	11.5
Villavendimio	1	656.07	10.4
Total Zamora	20	10948.20	8.2

itative in the sense that it allows the estimates to be labeled as official statistics by reducing the coefficient of variation below the standard limits (20%) considered acceptable in official statistics.

4.1.5. Crop acreage estimates at the municipality level based on linear mixed models

To illustrate small-area estimation using areal sampling, we consider barley acreage estimates at municipality level in that part of Zamora province within the 100 × 100 km area of Castilla y León (Spain). Results are in Table 5.

Considering that the sample size at municipality level is small or null, the accuracy of municipality estimates is good, thanks to the RS contribution. In most municipalities, the estimates could be labeled as official statistics because the coefficient of variation is smaller than the standard limit (20%), even when the sample size is only one sampling unit

4.2. Senegal

4.2.1. Crop acreage estimates in Nioro Department

To demonstrate crop acreage estimation using point sampling and multinomial models, we consider the Nioro Department of Senegal, in the Kaolack Region. The ground data were observed in a sample of 345 points pixels selected using a list frame. The source of RS data is a crop-type map. In fact, we used a dual frame, since we complemented the NSO list frame with an area frame based on satellite images, in which agricultural areas were distinguished from non-agricultural areas and were stratified into crop types.

In this case, the NSO list frame alone is not sufficient to integrate field and RS data This is because the number of pixels in the agricultural areas is required for expanding the sample estimates to the entire population. This expansion factor is provided by the area frame.

We focus on the two main crops (millet and groundnut) observed in the field sample. The remaining crops (mainly maize) are included in a third category ($K = 3$) called other, whose probability estimate is one minus the estimate's probability of millet and groundnut. The RS data are coded according to the Earth Observation (EO) crop type into which pixels are classified: $x_i = [1\ 0\ 0\ 0]$ for any pixel i in the EO class maize, $x_i = [0\ 1\ 0\ 0]$ for any pixel i in the EO class millet, $x_i = [0\ 0\ 1\ 0]$ for any pixel i in the EO class groundnut, and $x_i = [0\ 0\ 0\ 1]$ for any pixel i in the EO class other crops. Model parameter estimates are in Table 6.

Table 6
Model parameter estimates ($\hat{B}_{\pi millet}$ and $\hat{B}_{\pi groundnut}$)

Crop	EO crop type map			
	Maize	Millet	Groundnut	Other crops
Millet	−0.010655	1.47958334	0.88931346	8.76504093
Groundnut	−0.659204	−0.59871700	2.89873948	1.21898514

Table 7
Crop acreage estimates, Nioro (Senegal)

Crop type	Hectare	Uncertainty				
		Standard error	Coefficient of variation (%)	Limits of 95% confidence interval		
				Lower	Upper	Amplitude
Millet	89215	3661.1	4.11	81978.8	96330.4	14351.5
Groundnut	78815	2923.9	3.71	73089.1	84550.9	11461.8

Table 8
Efficiency of RS data for crop acreage estimation Nioro (Senegal)

Crop type	Standard errors of proportion estimators		Relative efficiency of RS data
	Using only ground data	Using ground & RS data	
Millet	3.37	1.90	3.13
Groundnut	3.34	1.52	4.80

Table 9
Confusion matrix based on the sample

Ground data		Number of pixels in the EO class (%)*			
Crop	Number of points/pixels in the sample	Maize	Millet	Groundnut	Other
Maize	48	9 (18.8)	26 (54.2)	12 (25.0)	1 (2.0)
Millet	134	10 (7.4)	97 (72.4)	23 (17.2)	4 (3.0)
Groundnut	163	5 (3.1)	14 (8.6)	143 (87.7)	1 (0.6)
Total	345	24	137	178	6

*% Percentage of pixels correctly classified.

Crop acreage estimates of millet and groundnut based on ground and RS data are in Table 7.

We evaluate the RS data efficiency RE_{RSk} for the acreage estimation of millet and groundnut. Results are in Table 8.

The effect of integrating RS data in the ground sample data was a reduction in the sampling error of millet and groundnut and, as a result, in the confidence interval of these two crops, without loss of design-based consistency. In other words, using RS data, the cost of estimating millet and groundnut acreage could be reduced to less than a third of the current cost, without loss of accuracy. These results are design-based, so it is understood that the reduced sample size should be selected using the currently sampling design used by the NSO.

4.2.2. Estimation directly based on pixel classification

There is consensus in the official statistics community on using methods providing design-consistent estimates of the population characteristics under study and measures of estimator uncertainty, such as sampling error, coefficient of variation, or confidence intervals. The method proposed in this paper for integrating RS data in the NSO sampling design agrees with this consensus.

Although the proposed approach allows for any form of RS data, including raw reflectance data in the form of pixel digital numbers, we focus on the use of croptype maps as auxiliary information. In this context, a natural question that arises is: Why not directly use the number of pixels in each EO croptype class to estimate crop acreage instead of using it as auxiliary data in a statistical model?

We follow the suggestion of a referee to clarify this question, comparing the results of our design-based approach (shown in Table 7) with estimates directly based on the croptype map. The comparison is necessarily limited to the crop acreage estimates because the usual algorithms used for croptype map generation do not provide measures of uncertainty comparable to those of Table 7.

Using a random forest classifier, four EO crop types were considered in the croptype map of Nioro: maize,

Table 10
Probability that the cover of a pixel of the EO class j is the crop k

EO class (j)	Millet ($k=1$)	Groundnut ($k=2$)	Maize & Other_crops ($k=3$)
Maize ($j=1$)	0.395	0.206	0.399
Millet ($j=2$)	0.739	0.093	0.168
Groundnut ($j=3$)	0.113	0.841	0.046
Other ($j=4$)	0.997	0.001	0.002

Table 11
Estimates number of pixels, $\hat{N}_k$

EO class		Estimates of the number of pixels covered by crop k in each EO crop type class, $\hat{N}_{kj} = N_j \times \hat{\mu}_{kj}$		
Crop type (j)	Number of pixels N_j	Millet ($k=1$)	Groundnut ($k=2$)	Maize & Other ($k=3$)
Maize ($j=1$)	1032628	407587	213088	411953
Millet ($j=2$)	9470572	7000329	876040	1594203
Groundnut ($j=3$)	8076448	910538	6791735	374175
Other ($j=4$)	604853	603038	605	1210
Estimates of the total number of pixels by crop $\hat{N}_k = \sum_{j=1}^{J} \hat{N}_{kj}$	19184501	8921492	7881468	2381541

Table 12
Crop acreage estimates at the district (arrondissement) level, Nioro (Senegal)

Arrondissement	Millet		Groundnut	
	Acreage (has.)	Error (CV%)	Acreage (has.)	Error (CV%)
Medina Sabakh	20067.2	8.6	19765.3	7.3
Paoskoto	38316.0	5.3	35018.2	4.0
Wack Ngouna	30831.7	11.9	24030.7	10.7
Total Nioro	89215,0	4.1	78815,0	3.7

millet, groundnut, and others. The remaining pixels were classified in a non-agricultural landuse class. The number of pixels in each EO croptype class is in Table 11.

As seen in Table 9, there is uncertainty in the pixels classification and any estimate based on it are subject to this uncertainty. Most pixels of millet in the sample (72.4%) are correctly classified in the EO class millet, but a non-negligible percentage of them were confounded with groundnut (17.2%), maize (7.4%), and others (3.0%). For groundnut, the percentage of pixels that were correctly classified is higher (87.7%) than for millet, but the confusion with millet (8.6%) and maize (3.1%) is non-negligible.

In the proposed approach, the multinomial logit model is used for estimating the probability μ_{kj} that the crop covering a pixel classified in the EO class j is actually crop k. The model parameter estimates are in Table 6 and the probability estimates are in Table 10.

To estimate the number of pixels N_{kj} in EO class j that are actually covered by crop k, we multiplied the total number N_j of pixels in EO class j by the estimator of the aforementioned probability $\hat{\mu}_{kj}$: $\hat{N}_{kj} = N_j \times \hat{\mu}_{kj}$. The estimator of the total number of pixels covered by crop k in Nioro, $\hat{N}_k = \sum_{j=1}^{J} \hat{N}_{kj}$ is the sum of the estimators in each EO class. This estimator is design-consistent and the estimates based on it are in Table 11.

Both the number of pixels in EO crop type millet (9470572, i.e., 94705.72 hectares, as a pixel represents 100 m^2) and in EO crop type groundnut (8076448, i.e., 80764.48 hectares) are larger than the design-based number estimates: 8921492 (89214.92 hectares) for millet and 7881468 (78814.68 hectares) for groundnut. For millet, the difference was 5490.8 hectares (6.2%) and for groundnut 1949.8 hectares (2.5%). For the former, the difference is greater than the design-based standard error (3661.10 hectares) and the coefficient of variation (4.11%). For the latter the difference is smaller than the design-based standard error (2923.94 hectares) and the coefficient of variation (3.71%). These differences are not statistically significant, since the estimates directly based on the EO croptype classes are within the design-based confidence limits, namely, [81978.88, 96330.4] hectares for millet and [73089.15, 84550.98] hectares for groundnut.

However, the comparison must focus not on the results, which are always uncertain, but on the methods. The methods make the major difference; whereas the design-based estimators are in the mainstream of official statistics, the estimators based directly on croptype maps are not. The proposed approach provides design-consistent estimates together with the usual measures of uncertainty (sampling error, coefficient of variation, and confidence intervals). The two main objections to estimators based directly on the croptype map are that (i) they are not design-consistent and (ii) the usual uncertainty measures are not available.

4.2.3. Crop acreage estimates at district level in Nioro Department

To illustrate crop acreage estimation at district level using point sampling and multinomial models, we considered the three districts (arrondissements) of Nioro, Medina Sabakh, Paoskoto, and Wack Ngouna. The estimates are in Table 12.

The estimate accuracy at district (arrondissement) level is, as expected, lower than at the national level. However, the sampling error is within the standard limit (coefficient of variation $< 20\%$) for labeling as official statistics.

The differences between the number of pixels in the EO crop type millet and the design-based estimate of the number of pixels covered by millet are smaller than the sampling error in Medina Sabakh (0.7%) and Paoskoto (4.1), and larger than the design-based estimate in Wack Ngouna (12.4%). The differences between the number of pixels in the EO crop type groundnut and the design-based estimate of the number of pixels covered by groundnut are smaller than sampling error in the three districts: Medina Sabakh (3.4%), Paoskoto (2.5%), and Wack Ngouna (1.2%).

5. Concluding remarks

A result well established in the literature is that RS data improve the cost-efficiency of design-consistent crop acreage estimators based on linear models and simple area samples of georeferenced polygons (segments). The strong RS reliability for crop acreage estimation facilitates a notable improvement of crop production estimates, even if RS data are not sufficiently reliable for yields.

However, many countries use complex list samples of non-georeferenced households or farms for agricultural statistics. We have demonstrated in this paper that with the use of multinomial models, a point georeferenced by parcel in these complex samples is sufficient to enhance the cost efficiency of design-consistent, national-level crop acreage estimators using RS data.

The open-source Sen4Stat system is being designed in such a way that any NSO can use it to improve the cost efficiency of the procedure currently used to obtain crop acreage and production estimates, at both national and minor administrative area levels (province and municipality). The only inputs required by the system are the observed data in the sample currently designed by the NSO for field data collection. The system allows for both a segment simple sample, based on area frames, and a point complex sample, based on list frames of households/farms. In both cases, the system generates from Sentinel images a crop-type map and integrates it with ground data, using linear models in the segment case and multinomial models in the point case.

We have extended the method developed in this paper to two additional RS applications, multi-season estimates and optimization of the sample design. Additional work is required to obtain the data required to test these two additional RS applications.

Acknowledgments

This research was carried out in the framework of a consortium integrating the Université Catholique de Louvain (UCLouvain) as contractor and the Universidad Politécnica de Madrid (UPM), CS ROMANIA S.A. (C-S RO) and Systèmes d'Information à Référence Spatial (SIRS/CLS) as sub-contractors. The national statistical offices of Ecuador, Malawi, Senegal, Spain, and Tanzania participated in the work as partners and provided the required ground data. SIRS/CLS elaborated the ground databases, C-S RO developed the system for image classification, and UPM developed statistical prototypes for the integration of ground and RS data. UCLouvain coordinated the project and contributed mainly to the specification of products and services, system design RS database elaboration for integration with the ground database, and the full-scale demonstration plan.

We thank the referees and the subject editor for their very helpful comments and suggestions.

Funding

This work was supported by the European Space Agency in the Sen4Stat project framework [Contract

No. 4000127181/19/I-NS]. This agency contributed to the study design, together with a steering committee made up of members of international organizations such as the World Bank and FAO, with the aim of achieving synergies between this project and those launched by these organizations to improve agricultural statistics.

References

[1] Fonteneau F, Delincé J. Surveying Farms in the 21st Century. In: FAO editor. Handbook on the Agricultural Integrated Survey (AGRIS); Global Strategy to Improve Agricultural and Rural Statistics. Rome: FAO; 2018; pp. 1-6.

[2] Zezza A, Gourlay S, Molini V. Closing the data gap in agriculture through sustainable investment in the data value chain: Realizing the vision of the 50x2030 Initiative. Stat J IAOS. 2022; 38: 57-62. doi: 10.3233/SJI-220933.

[3] Hanuschak GA, Allen RD, Wigton WH. Integration of Landsat data into the crop estimation program of USDA's Statistical Reporting Service (1972-1982). Invited paper at: 8th International Symposium on Machine Processing of Remote Sensed Data. 1982 July 7–9; West Lafayette, Indiana: Purdue University. Available from: https//www.lars.purdue.edu/home/references/sym_1982/sym_1982.html.

[4] Allen JD. A look at the Remote Sensing Applications Program of the National Agricultural Statistics Service. J Off Stat. 1990; 6(4): 393-409.

[5] Ambrosio L, Alonso R, Villa A. Estimación de superficies cultivadas por muestreo de áreas y teledetección. Precisión relativa. Estad Esp. 1993; 35(132): 91-103.

[6] Delincé J. Cost-Effectiveness of Remote Sensing for Agricultural Statistics in Developing and Emerging Economies. In: FAO editor. GSARS Technical Report GO09-2015. Rome: FAO; 2015. doi: 10.13140/RG.2.2.25985.45927.

[7] Gallego FJ. Remote sensing and land cover area estimation. Int J Remote Sens. 2004; 25(15): 3019-47. doi: 10.1080/01431160310001619607.

[8] Firth D, Bennett KE. Robust models on probability sampling. J. R. Stat. Soc. Series B Stat. Methodol. 1998; 60:3-21. doi: 10.1111/1467-9868.00105.

[9] Agresti A. Categorical data analysis. New York: Wiley; 2002; 710 p.

[10] Hogland J, Billor N, Anderson N. Comparison of standard maximum likelihood classification and polytomous logistic regression used in remote sensing. Eur J Remote Sens. 2013; 46(1): 623-40. doi: 10.5721/EuJRS20134637.

[11] FAO. Handbook on Master Sampling Frames for Agricultural Statistics. Frame Development, Sample Design and Estimation. Improving Agricultural and Rural Statistics. Global Strategy. Rome: FAO; 2015 Dec. 170 p.

[12] Särndal CE, Swensson B, Wretman, J. Model Assisted Survey Sampling. New York: Springer; 1992; 695 p.

[13] Fuller WA. Sampling statistics. New York: Wiley; 2009; 472 p. doi: 10.1002/9780470523551.

[14] Cotter J, Nealon J. Area frame design for agricultural surveys. Research and Applications Division. National Agricultural Statistics Service. United States Department of Agriculture. Washington, D.C., 1987.

[15] FAO. Multiple frame agricultural surveys. Volume 1: Current surveys based on area and list sampling methods. Volume 2: Agricultural survey programmes based on area frame or dual frame (area and list) sample designs. Statistical development series 7 and 10. Rome: FAO; 1996, 1998.

[16] Ministry of Agriculture, Livestock and Fisheries. 2014/15 Annual Agricultural Sample Survey Report. Zanzibar (TZ); 2016; 72 p. Available from: http//hdl.handle.net/20.500.12018/2905.

[17] Yates F. Sampling Methods for Census and Surveys. London: Griffin. 1949; 318 p.

[18] Jinguji I. Dot sampling method for area estimation. In: Crop Monitoring for Improved Food Security. Srivastava M.K. editor, Bangkok (TH): FAO and ADB; 2015 RAP PUBLICATION 2014/28: pp. 27-48.

[19] Gay C, Porchier JC. Land cover and land use classification using TER-UTI. In: Holland TE and van den Broecke MPR, editors. Proceedings of Agricultural Statistics 2000, an International Conference on Agricultural Statistics 1998 Mar 18-20: pp. 93-201. Available from: https//www.isi-web.org/isi.cbs.nl/iamamember/Books/agric2000/page-193.pdf.

[20] Nusser SM, Goebel JJ. The National Resources Inventory: a long-term multi-resource monitoring programme. Environ Ecol Stat. 1997; 4(3): 181-204. doi: 10.1023/A1018574412308.

[21] Nusser SM, Goebel JJ, Fuller WA. Design and estimation for investigating the dynamics of natural resources. Ecol Appl. 1998; 8(2): 234-45. doi: 10.1890/1051-0761(1998)008[0234:DAEFIT]2.0.CO;2.

[22] Azzari G, Jain S, Jeffries G, Kilic T, Murray S. Understanding the requirements for surveys to support satellite-based crop type mapping: evidence from Sub-Saharan Africa. Remote Sens. 2021; 13(23), 4749. doi: 10.3390/rs13234749.

[23] Direction de l'Analyse de la Prévision et des Statistiques Agricoles. Ministère de l'Agriculture et de l'Equipement Rural. Senegal – Enquete Agricole Annuelle. Methodologie et plan de sondage de l'enquete agricole. Senegal. 2017. http//anads.ansd.sn/index.php/catalog/218/download/1846.

[24] MDcCulloch CE, Searle SR. Generalized, linear, and mixed models. New York: Wiley; 2001; 325 p.

[25] Ambrosio L, Iglesias L. Land cover estimation in small areas using ground surveys and remote sensing. Remote Sens. Environ. 2000; 74: 240-8. doi: 10.1016/S0034-4257(00)00114-0.

[26] Saei A, Chambers R. Small area estimation under linear and generalized linear mixed models with time and area effects. Methodology Working Paper M03/15. Southampton (UK): Southampton Statistical Sciences Research Institute, 2003: 31 p. http//eprints.soton.ac.uk/id/eprint/8165.

[27] Molina I, Saei A, Lombardia MJ. Small area estimates of labour force participation under a multinomial logit mixed model. J. R. Stat. Soc. Ser. A Stat. Soc. 2007; 170(4): 975-1000. doi: 10.1111/j.1467-985X.2007.00493.x.

[28] López-Vizcaíno E, Lombardía MJ, Morales D. Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. J. R. Stat. Soc. Ser. A Stat. Soc. 2015; 178(3): 535-65. doi: 10.1111/rssa.12085.

[29] Fuller WA. Environmental surveys over time. J Agric Biol Environ Stat. 1999; 4(4): 331-45. doi: 10.2307/1400493.

[30] Fuller WA, Breidt FJ. Estimation for supplemented panels. The Indian Journal of Statistics, Serie B. 1999; 61(1): 58-70. https//www.jstor.org/stable/25053068.

[31] Cochran WG. Relative accuracy of systematic and stratified random samples for a certain class of populations. Annals of Mathematical Statistics. 1946; 17(2): 164-77.

[32] Ambrosio L, Iglesias L, Marin C. Systematic sample design for the estimation of spatial means. Environmetrics. 2003; 14(1): 45-61. doi: 10.1002/env.564.

[33] Defourny P, Bontemps S, Bellemans N, Cara C, Dedieu G,

Guzzonato E, el al. Near real-time agriculture monitoring at national scale at parcel resolution: performance assessment of the Sen2-Agri automated system in various cropping systems around the world. Remote Sens. Environ. 2019; 221: 551-68. doi: 10.1016/j.rse.2018.11.007.

[34] Fuller WA, Isaki CT. Survey design under superpolutation models. In: Current Topics in Survey Sampling. Krewski D, Rao JNK and Platek R, editors, New York: Academic Press; 1981; pp. 199-226.

[35] Ambrosio L, Iglesias L. Identifying the most appropriate sampling frame for specific landscape types. Technical Report Series. GO-01-2014. FAO. https//www.fao.org/3/ca6436en/ca6436en.pdf.

[36] Ambrosio L, Iglesias L, Marín C. A model-assisted approach to identify a cost-efficient spatial sampling strategy. Paper presented at: Eighth International Conference on Agricultural Statistics (ICAS-VIII); 2019 Nov 18-21; New Delhi (IN).