A new criterion based on estimator variance for model sampling in precision agriculture

Baptiste Oger, Gilles Le Moguédec, Philippe Vismara, Bruno Tisseyre

A New Criterion based on Estimator Variance for Model Sampling in Precision Agriculture

B. Oger[1-2], G. Le Moguédec[3], P. Vismara[2-4] and B. Tisseyre[1]
[1]*ITAP, Univ. Montpellier, Montpellier SupAgro, INRAE, France*
[2]*MISTEA, Univ. Montpellier, Montpellier SupAgro, INRAE, France*
[3]*AMAP, Univ. Montpellier, INRAE, CIRAD, CNRS, IRD France*
[4]*LIRMM, Univ. Montpellier, CNRS, France*

Corresponding author: baptiste.oger@supagro.fr

Abstract

Model sampling has proven to be an interesting approach to optimize the sampling of an agronomic variable of interest at the field level. The use of a model improves the quality of the estimates by making it possible to integrate the information provided by one or more auxiliary data. It has been shown that such an approach gives better estimations compared to more traditional approaches.

Through a statistical work describing the properties of model sampling variance, this paper details how the different factors either related to sample characteristics or to the correlation between the auxiliary data and the variable of interest, affect estimation error. The resulting equations show that the use of samples with a mean close to the field mean and with a substantial dispersion reduces the estimation variance. On the basis of these statistical considerations, a variance criterion is defined to compare sample properties. The lower the value of the criterion of a sample, the lower the variance of the estimate and the expected errors. These theoretical insights were applied to real commercial vine fields in order to validate the demonstration.

Nine vine fields were considered with the objective to provide the best yield estimation. High resolution vegetative index derived from airborne multispectral image was used to drive the sampling

Abbreviations:

C_S	variance criterion
N	set of potential sampling sites
n	size of the set N; $n = Card(N)$
NDVI	normalized difference vegetation index
R	set of sites not selected in the sample
RMSE	root mean square error
RRMSE	relative root mean square error
S	set of sampled sites
s	size of the set S; $s = Card(S)$
T	field yield
$\hat{T}$	field yield estimation
$\tilde{T}$	field yield forecast (accounting for T variance)
X_i	auxiliary data (NDVI) for site i
Y_i	variable of interest (yield) for site i
β_0 & β_1	linear model parameters
σ^2	variance of the residual of the model

and the estimation. The theoretical considerations were verified on the nine fields; as the observed estimation errors correspond quite well to the values predicted by the equations. The selection of a large number of random samples from these fields confirms that samples associated with higher values of the chosen criterion result, on average, in larger yield estimation errors. Samples with the highest criterion values are associated with mean estimation errors up to two times larger than those of average samples. Random sampling is also compared to two target sampling approaches (Clustering based on quantiles or on k-means algorithm) commonly considered in the literature, whose characteristics improve the value of the proposed criterion. It is shown that these sampling strategies produce samples associated with criterion values up to 100 times smaller than random sampling. The use of these easy-to-implement methods thus guarantees to reduce the variance of the estimation and the estimation errors.

Introduction

In crop production, sampling is a common practice used to estimate the agronomic variable of interests for a given field, whether it is related to crops, soil, diseases, etc. The state of the production system is strengthened by the estimation resulting from the sample and allows farmers to adjust their decision-making. During the estimation process, a sample of observations is made at a limited number of measurement sites within the field. The number of measurement sites is generally fixed by operational constraints such as available time. The quantity of interest is then characterized from this sample of observations by inference techniques based on an estimator.

New methods granting fast acquisition of field data have developed with the information and communication technology in agriculture. In particular, remote sensing methods are increasingly used to characterize canopy vigour through vegetation indices (Liaghat & Balasundram, 2010; Venkataratnam, 2001; Barnes & Baker, 2000), but also allowing a wide variety of data to be collected directly from fields (Rehman et al, 2014).

Despite the development of these new data collection methods, some decisions still require sampling on the field as some measurements are still inaccessible using the current sensors. However, the available new sources of information are valuable because they allow, when accessible with a high spatial resolution, to characterize the variability and the spatial structure of the fields (Kitchen et al. 2020, Damian et al. 2020). Moreover, even when the desired measurement variables are not directly accessible, the observations from the sensors can be more or less related to the variable of interest. This is the case, for example, between yield and NDVI vigour observations obtained by remote sensing in viticulture (Carillo et al, 2016) or between soil parameters and soil electrical conductivity (Corwin et al. 2003). In this context, new sampling approaches based on these sources of information have emerged. For example, *stratified* sampling and *target sampling* approaches use high spatial resolution observations to drive the choice of measurement sites on the field (Miranda et al., 2018; Uribeetxebarria et al., 2019; Arnó et al., 2017). Other methods propose to go further by also mobilizing these observations when inferring the estimation of the variable of interest. The estimator is then built on the basis of a model linking the sampled quantity to the available auxiliary high spatial resolution information. These approaches, described as *model sampling*, have shown promising results in agriculture (Murthy et al., 1995; Araya-Alman et al., 2019).

However, the methods used by the *model sampling* and *target sampling* approaches to guide the choice of measurement sites remain rather empirical. Considering that the number of sample is determined by operational constraints, this article proposes a more in-depth reflection on the choice of a fixed number of measurement sites when using a model. The study focuses on the estimation of

an expectation (field mean) or a cumulative value over the entire field. It is assumed that the quantity of interest is more or less strongly linearly related to an available auxiliary data (i.e. NDVI, soil apparent conductivity, etc.). The statistical properties of an estimator based on a linear model are then described using a matrix formalism.

To support this reflection, this article proposes a rigorous formalism to describe the uncertainty associated with an estimate made with model sampling. The purpose of this statistical study is to define a criterion which can relate how the sampling site selection affects the final estimation for a given sample size. This work is supplemented by a validation case study about yield estimation in viticulture based on NDVI auxiliary data in order to evaluate the robustness of the approach and to compare different sampling methods.

Material and method

Hypotheses and notations

In this section, bold notations represent matrices and vectors.

For a given field, the objective is to estimate the total production. This field is divided in elementary sites so that the total production is the sum of production of each site. Only a limited number of these sites can be sampled in order to build an estimator of the total production. These sites are chosen from the set N of potential measurement sites. For each potential measurement site ($i \in N$), numbered from 1 to n, there is a value for the quantity of interest noted Y_i. This value is only known for the s sampled sites ($i \in S$). A second variable, noted X_i, corresponding to an auxiliary data is available for each potential measurement site ($i \in N$). It is assumed that a linear relationship relates the quantity of interest to the auxiliary data. It is then possible to write the values of Y_i knowing X_i as shown in equation 1.

$$\boldsymbol{Y_N}|\boldsymbol{X_N} = \beta_0 \mathbf{1_n} + \beta_1 \boldsymbol{X_N} + \boldsymbol{\varepsilon_N} \qquad Eq.1$$

With:

$$\boldsymbol{\varepsilon_N} \sim N(\mathbf{0_n}, \sigma^2 \boldsymbol{I_n}) \qquad Eq.2$$

Where $\boldsymbol{Y_N}$ and $\boldsymbol{X_N}$ are two vectors of length n containing respectively the values of the quantity of interest and the auxiliary data. It should be noted that in the standard writing of the linear model in matrix form, $\boldsymbol{X_N}$ represents an incidence matrix, here $\boldsymbol{X_N}$ represents a vector because there is only one auxiliary data. The $\mathbf{0_n}$ **and** $\mathbf{1_n}$ vectors of length n contain respectively only 0 and only 1. The matrix $\boldsymbol{I_n}$ the identity matrix of dimension $n \times n$. Finally, β_0, β_1 and σ^2 represents the model parameters relating $\boldsymbol{Y_N}$ to $\boldsymbol{X_N}$.

The set S, consisting of the sites selected in the sample, and the set R, consisting of the sites not selected in the sample, form a partition of the set N : $N = S \cup R$ and $S \cap R = \emptyset$. We can thus decompose the vectors $\boldsymbol{Y_N}$ and $\boldsymbol{X_N}$ as shown in Equations 3 and A7 in the appendix.

$$\boldsymbol{Y_N} = \begin{bmatrix} \boldsymbol{Y_S} \\ \boldsymbol{Y_R} \end{bmatrix} \qquad and \qquad \boldsymbol{X_N} = \begin{bmatrix} \boldsymbol{X_S} \\ \boldsymbol{X_R} \end{bmatrix} \qquad Eq.3$$

Formalization of an estimator

The objective is to estimate T, the sum of local yield values (Y_i) on the field. By separating the values for which an observation is available (S), from the unobserved values (R) as defined in Eq. 3:

$$T = \sum_{i \in N} Y_i \qquad Eq.4$$

Which can also be written:

$$T = \mathbf{1}_{\boldsymbol{N}}^{t} \boldsymbol{Y_N} \qquad Eq.5$$

$$T = \mathbf{1}_{\boldsymbol{S}}^{t} \boldsymbol{Y_S} + \mathbf{1}_{\boldsymbol{R}}^{t} \boldsymbol{Y_R}$$

$\hat{T}$ is defined as the estimator of T. The values of the vector $\boldsymbol{Y_S}$, which correspond to the measured values of the quantity of interest, being known, the problem is to estimate the values of $\boldsymbol{Y_R}$. $\mathbf{1}_{\boldsymbol{R}}^{t}\, \mathbb{E}(\boldsymbol{Y_R} | \boldsymbol{Y_S}, \boldsymbol{X_S}, \boldsymbol{X_S})$ is chosen as the estimator of $\mathbf{1}_{\boldsymbol{R}}^{t} \boldsymbol{Y_R}$ because it minimizes the quadratic risk. The statistical work on the mathematical expression of this estimator is detailed in the appendix from Eq. A1 to Eq. A26.

$$\hat{T} = \mathbf{1}_{\boldsymbol{S}}^{t} \boldsymbol{Y_S} + \mathbf{1}_{\boldsymbol{R}}^{t}\, \mathbb{E}(\boldsymbol{Y_R} | \boldsymbol{Y_S}, \boldsymbol{X_S}, \boldsymbol{X_R}) \qquad Eq.6$$

Estimator and forecast properties

For this estimator, we are interested in classical indicators such as the first and second order moments of the estimator in order to characterize its bias and the distribution around this bias:

$$\mathbb{E}(\hat{T}) = \sum_{i \in N} Y_i \qquad Eq.7$$

This is an unbiased estimator with variance:

$$\mathbb{V}(\hat{T}) = (n-s)^2 \times \left(\frac{1}{s} + \frac{(\overline{X_R} - \overline{X_S})^2}{\sum_{i \in S} (X_i - \overline{X_S})^2} \right) \times \sigma^2 \qquad Eq.8$$

The reasoning held here led to the construction of an estimator of the expectation of T. If a forecast is to be made, in the same way as for a linear regression prediction, the individual variance ε_i for each of the unobserved Y_i ($i \in R$) must be considered as $\mathbb{V}(\hat{T})$ only represents the variance of the expectation estimator. The forecast $\tilde{T}$ of a single value of the quantity of interest has for variance:

$$\mathbb{V}(\tilde{T}) = (n-s)^2 \times \left(\frac{1}{s} + \frac{1}{n-s} + \frac{(\overline{X_R} - \overline{X_S})^2}{\sum_{i \in S} (X_i - \overline{X_S})^2} \right) \times \sigma^2 \qquad Eq.9$$

The variance of the forecast thus depends on:

- n, the size of the set of potential sampling sites within the field (N);
- s, the number of sampling sites or the size of the set S;
- σ^2, the variance of the residual of the model;
- $X_{i \epsilon S}$, the values taken individually by the measurement sites for the auxiliary data;
- $\overline{X_S}$, the average value of the measurement sites for the auxiliary data;
- $\overline{X_R}$, the average value of the non-selected sites for the auxiliary data.

This variance logically tends towards 0 when s tends towards n.

$\tilde{T}$ is a forecast of T, the sum of Y_i. The previous reasoning is applicable to $\frac{\tilde{T}}{n}$ which is an estimator of the expectation of $Y_{i \in N}$. The variance of $\frac{\tilde{T}}{n}$ is of the formula $\frac{\mathbb{V}(\tilde{T})}{n^2}$ and has similar properties.

This result allows to characterize the uncertainty associated with $\tilde{T}$ in relation to the size s of the sample (S) and the size n of the set of potential sampling sites (N), the values of the auxiliary data for the whole field, which are known, and the quality of the relationship between the data of interest and the auxiliary variable (σ). The values of n and σ are fixed and only depend on the field characteristics. The value of s is chosen by the practitioner and is also fixed depending on the available time and the expected quality for the estimation. Finally, the values X_S, $\overline{X_S}$ and potentially $\overline{X_R}$ which have an incidence on the variance can direct the choice of sampling sites. The following section will therefore focus on the part of the variance that depends on the auxiliary data chosen for the sample.

Variance criterion for the selection of measurement sites

The variance criterion C_S is defined as the part of the variance of the estimator (Eq. 8) or the prediction (Eq. 9) associated with the auxiliary data values of the measurement sites:

$$C_S = \frac{(\overline{X_R} - \overline{X_S})^2}{\sum_{i \in S}(X_i - \overline{X_S})^2} \qquad Eq. 10$$

For a given sample size s, the variance criterion defines the fraction of variance that depends on the choice of measurement sites. In a situation where s is fixed by operational constraints (available time, destructive measurements ...), the sampling plan leading to the lowest estimation variance will be the one with the lowest value of C_S.

In the numerator, $(\overline{X_R} - \overline{X_S})^2$, is the quadratic difference between the sample mean and the mean of the whole population. This can be understood as the representativeness of the auxiliary values on the sample sites. For a given sampling size, the closer the mean value of sample sites to the mean of the field the lower the C_s value.

In the denominator $\sum_{i \in S}(X_i - \overline{X_S})^2$, is the sum of the squared deviations between the measurement sites and their own mean, it represents the dispersion of the sample values. Indeed, the higher variability of sample values around their mean, the lower the C_S value.

General method for the case study

The first objective is to verify the relevance of the assumptions made (linear model, independent measurement sites) on a real dataset. To do so, experimental errors are compared with expected errors derived from the theoretical variance.

The second objective is to validate, through experimentation, the relevance of the variance criterion C_S. The idea is to establish a link between the value of the variance criterion (C_S) and the quality of the estimate produced. To this aim, the C_S value is computed and compared with the quality of the estimate produced for a large number of samples.

Three sampling methods are tested and compared, two of them mobilizing the auxiliary data.

Sampling methods

The first method implemented for selecting the s measurement sites is *random sampling* (Wulfsohn, 2010). In this approach, the set of S sampled sites is drawn from the set of N available sites by a random draw.

The second method is based on the principle of *target sampling*. This partitions the set N into s subsets according to the values for the auxiliary data (defined as variable X). A single measurement site is then

randomly selected in each of the s subsets (Carillo et al., 2016; Oger et al. 2019). Two partitioning methods are tested:

- The quantile method where the set N is cut according to the percentiles in order to obtain s subsets of the same size.
- The k-means algorithm (MacQueen, 1967).

These approaches naturally tend to favour a dispersion of the sampled values and thus to minimize the variance criterion C_S.

For all three methods (random sampling, quantile and k-means), 1000 samples of size n ranging from 4 to 15 are drawn for each field (see next sections for the presentation of the fields).

Measurement of the quality of the estimate

The quality of the estimation is measured by the estimation error. This is defined as the absolute value of the relative difference between the value taken by the estimator and the estimated quantity. Its value is expressed as a percentage of the estimated quantity:

$$Error\ (\%) = \frac{\left|\tilde{T} - T\right|}{T} \qquad Eq.11$$

The root mean square error (RMSE) is a measure of the quality of an estimate over a large number of estimates. Defining Samples as a set of samples, it is calculated as follows:

$$RMSE = \sqrt{\sum_{i \in Samples} \frac{\left(\tilde{T}_i - T\right)^2}{Cardinal(Samples)}} \qquad Eq.12$$

In theory, RMSE is also defined as the sum of the squared bias and the variance (Wasserman, 2004):

$$RMSE = \sqrt{\left(\mathbb{E}\left(\tilde{T}\right) - T\right)^2 + \mathbb{V}\left(\tilde{T}\right)} \qquad Eq.13$$

And as bias is nul (Eq. 14):

$$RMSE = \sqrt{\mathbb{V}\left(\tilde{T}\right)} \qquad Eq.14$$

For standardisation purpose, the Relative Root Mean Square Error (RRMSE) is computed from experimental and theoretical RMSE as Eq. 15:

$$RRMSE\ (\%) = \frac{RMSE}{T} \times 100 \qquad Eq.15$$

Data

The fields used to test the method belong to INRAE Pech-Rouge (Narbonne, France - co-ordinates: E:709800, N:6226840, RGF93 datum, Lambert93) (Figure 1). The experiment and the resulting data are detailed in Carrillo et al. (2016). They are briefly summarized hereafter. The auxiliary data corresponds to a vegetation index: the NDVI. Nine fields were represented in this dataset. All were non-irrigated

and exposed to the Mediterranean climate with precipitation occurring in spring and a hot and dry summer.

Figure 1: Representation of the plots on the INRAE Pech-Rouge domain. Field colour represent local NDVI from low (red) to high (green) computed with Avion Jaune multispectral images. P104 is further north. Background: Google maps.

NDVI values were derived from a multispectral image with a resolution of 1 pixel = 1m² taken on August 31, 2008 by Avion Jaune (Narbonne, Hérault, France). The spectral regions captured in the images were: blue (445-520 nm), green (510-600 nm), red (632-695 nm) and near infrared (757-853 nm). From this image, the aggregation method described by Acevedo-Opazo et al. (2008) was used to obtain 9 m² image pixels, reducing the effect of canopy and bare ground discontinuity on the measured values. NDVI was finally calculated from the processed images according to Rouse et al. (1973). Mechanical or chemical weed control was performed on the row spacing; therefore, weed control had extremely small effect on NDVI values.

Table 1 : Characteristics of the experimental fields

Field	Area (ha)	Variety	Total Number of Sites (n)	Pearson correlation coefficient (NDVI/yield)	Average field yield (g/vine)	Field yield standard deviation (g/vine)	Yield coefficient of variation
P22	1.72	Syrah	45	0.13	1766	992.6	56.21%
P63	1.33	Syrah	42	0.28	1132	692.4	61,17%
P65	0.69	Syrah	33	0.86	1183	949.6	80,27%
P76	1.14	Carignan	37	0.39	824	661.2	80,24%
P77	1.24	Syrah	19	0.48	1427	1025.7	71,88%
P80	0.54	Syrah	40	0.63	1147	878.9	76,63%
P82	1.15	Syrah	53	0.47	968	613.7	63,40%
P88	0.85	Syrah	21	-0.04	2321	831.2	35,81%
P104	0.81	Carignan	23	0.18	2366	1091.6	46,14%

Local yield measurements on the fields were made locally on the nodes of a 15x15 m sampling grid. At each grid node, yield was measured on 5 consecutive vines along the row and the average yield was assigned to the coordinates of the grid node. The final database consisted of a set of 313 sites

distributed over the 9 different fields. For each site, an NDVI value was assigned as the mean of the 4 nearest pixels. The characteristics of each field are presented in Table 1.

Results

Validation

Figure 2 compares the theoretical and observed RRMSEs of yield estimates as a function of the number of measurement sites (*s*) for each of the nine fields considered (Table 1). The number of measurement sites varies from 4 to 15 for each field. The blue curve corresponds to the observed RRMSE (Eq. 12 & 15). Each point represents the averaged RRMSE over the 1000 samples. The red curve gives the average of the theoretical RRMSEs calculated with the theoretical variance equation of the forecast as proposed (Eq. 9, 14 & 15).

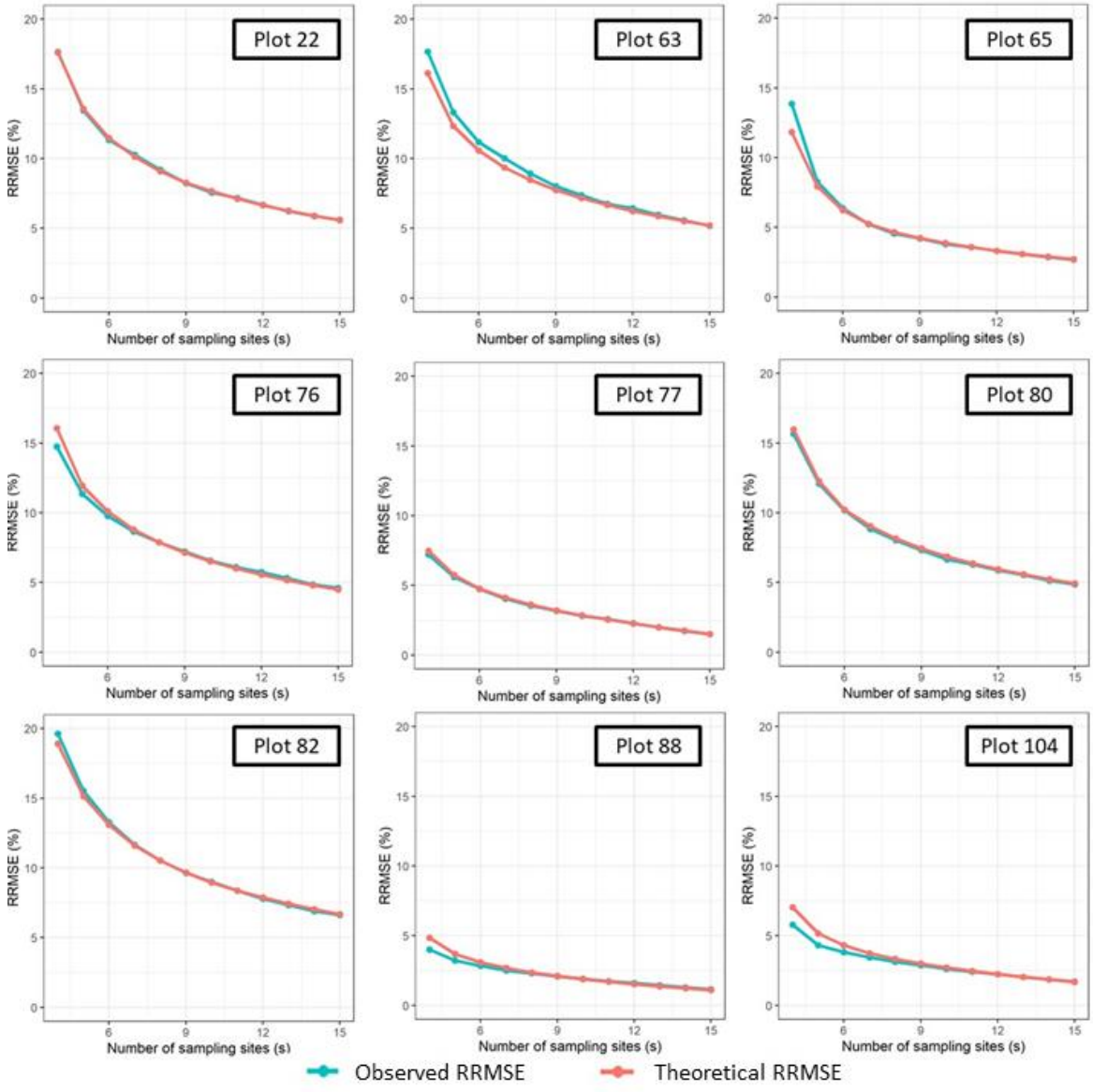

Figure 2: Observed (blue) and theoretical (red) RRMSEs; averaged for 9 vineyard fields (from left to right and top to bottom: P22, P63, P65, P76, P77, P80, P82, P88, P104) with a variable number of sample sites. Observed RRMSEs are computed from Eq. 12 and Eq. 15 and correspond to the relative error between field yields and sampling estimation using a model-based estimator (Eq. 6 and Eq. A26) with random samples. Theoretical RRMSEs are deduced from Eq. 14, Eq. 15 with the NDVI values of the sampled sites.

Variance criterion and random sampling

Figure 3 shows the result of 9,000 *random samplings* on the available data, all fields combined (1,000 *random samplings* per field). Each random sample is composed of 8 measurement sites ($s = 8$) and is associated to a yield estimate based on the model estimator (Eq. 6). The estimation error results for each of the 9000 samples are represented in Figure 3 as a function of the value of the variance criterion. The coloured areas represent the sample density according to their estimation error and variance criteria values.

The values of the variance criterion taken for these random samples are concentrated around the median (0.012) with 45% of the values between 10^{-2} and 10^{-1} and a dispersion ranging from 10^{-10} to 10^{1}. The red curve shows a local regression (Jacoby 2000) of the evolution of the mean estimation error as a function of the observed variance criterion. The 95% confidence interval of the curve is represented by a gray shading. For low values of variance criterion, the estimation error corresponds to a plateau with error values close to 15%, and then the estimation error starts to increase when the variance criterion exceeds 10^{-1}.

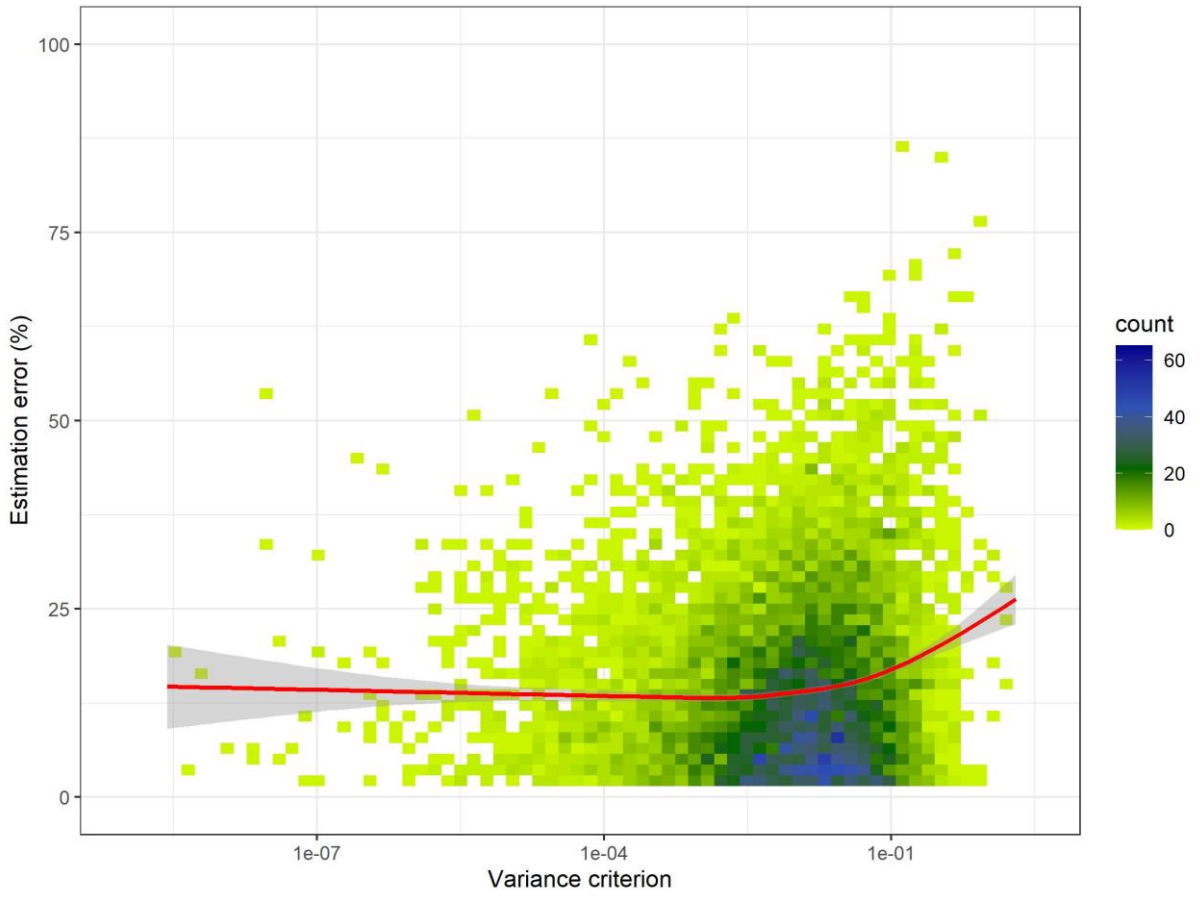

Figure 3: Relationship between variance criterion (C_S) and estimation error. The average estimation error (in red) increases when the estimates are made with a sample that has a high variance criterion.

For these fields, an increase in the estimation error as a function of the variance criterion is observed. This increase is slow at first and then accelerates. This observation is consistent with the theoretical equation for the variance of the estimate (Eq. 9). Indeed, in equation 9, the variance criterion is added to the terms $\frac{1}{s}$ and $\frac{1}{n-s}$. For the lowest values (less than 10^{-2}), the value of the variance criterion remains very small compared to the sum of terms $\frac{1}{s}$ and $\frac{1}{n-s}$ and variation of variance criterion then have almost no impact on the variance of the estimate. When the variance criterion reaches values of

the order of $\frac{1}{s}$, its variations significantly affect the variance of the estimate. An increase in the variance criterion then has an impact on the variance of the estimate, which increases the estimation error.

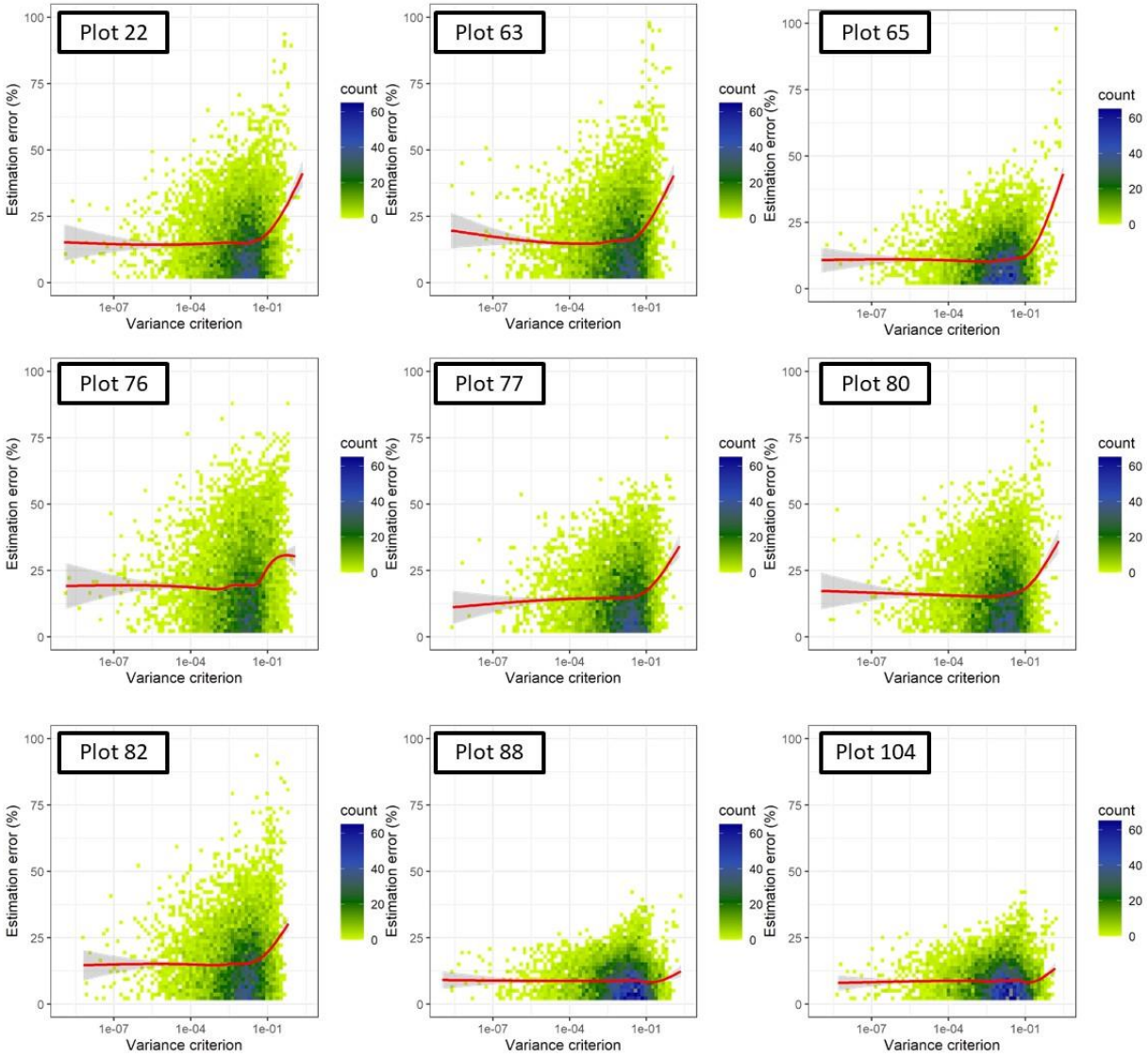

Figure 4: Evolution of the mean estimation error as a function of the variance criterion for all nine fields.

Using a similar procedure as in Figure 3, the nine graphs in Figure 4 show the individual results obtained on all fields. The results for each field are very similar to those presented in Figure 3: a high proportion of samples with a variance criterion value between 10^{-2} and 10^{-1} and an increase in estimation error for samples with a variance criterion exceeding 10^{-1}.

However, the plots have different error profiles represented by the flattening of the density of estimation errors and the value of the plateau of the red curve. These differences can be partly interpreted using the properties of the fields (Table 1). Fields with similar properties such as P88 and P104 (low) or P82 and P63 have similar error profiles. In particular, fields P88 and P104 correspond to the lowest errors of estimation compared to other fields. This can be explained by i) their low values of n which tends to minimize the difference $(n - s)$ in the expression of the variance ii) their low coefficients of variation (low within field yield variability) due to the very high average yields observed on these fields. It should be noted that these two fields show low correlations between NDVI and yield, but that this does not counterbalance the effect of the other factors.

The effect of the correlation between NDVI and yield can be deduced from fields P65 and P76 which have very similar CVs and similar sizes (n values). Field P65, which shows a very good correlation between NDVI and yield, gives better results than field P82. Field 80 ($n = 40$) also hints the importance of the correlation since it presents similar results to those of field 77, although the value of n is twice as small ($n = 19$).

Additional simulations (result not shown) tend to confirm that i) a decrease of n (by only considering part of the fields) where reducing estimation error while ii) increasing the yield variance and iii) decreasing the correlation between yield and NDVI (by adding a random noise to either NDVI or yield) was increasing estimation errors. However, these effects vary substantially from one plot to another.

Variance criterion and targeted sampling

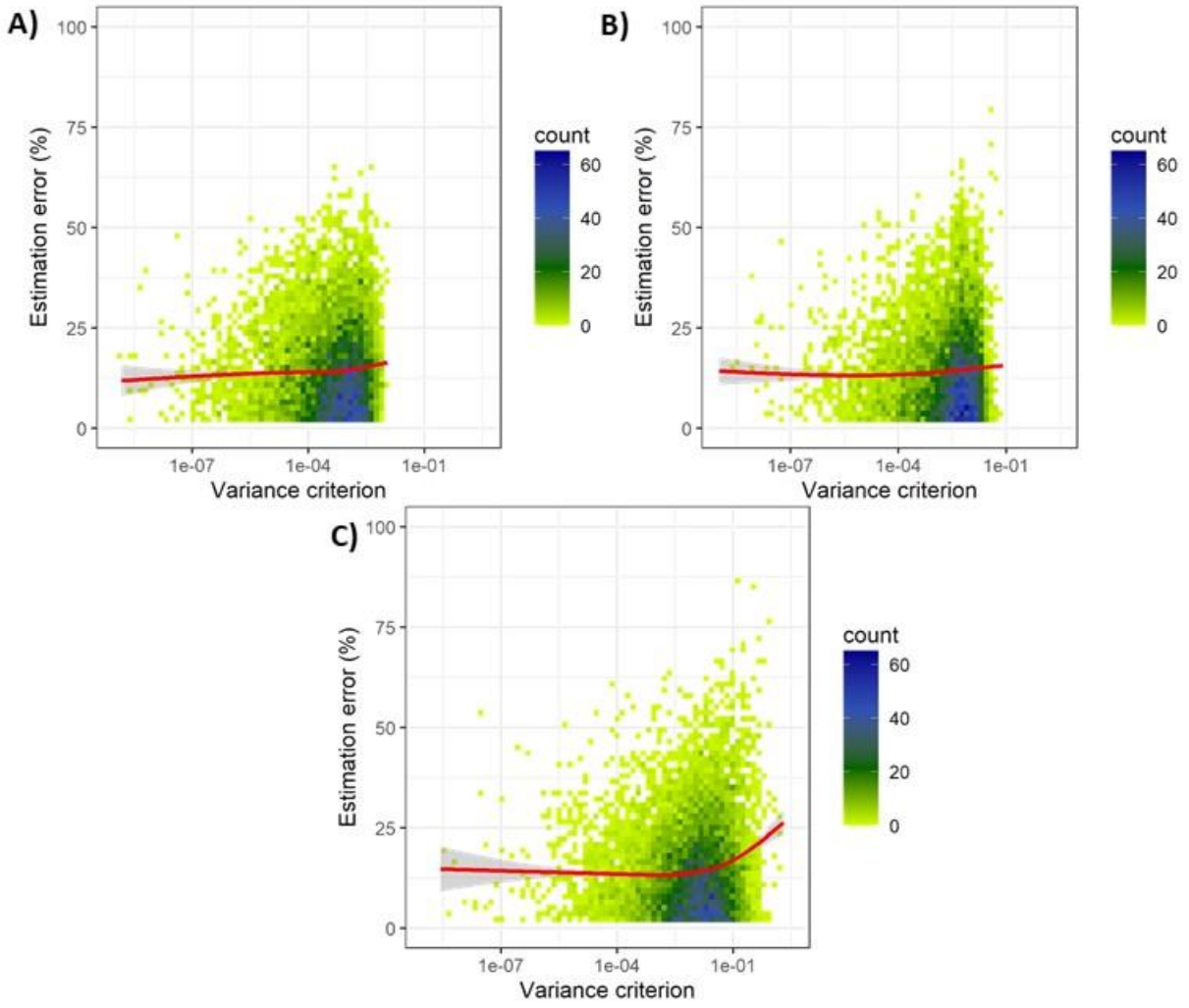

Figure 5: The target sampling approaches are associated to smaller variance criterion values, thus limiting the estimation error. The figure compares target sampling based on the quantile approach (4A) and the k-means approach (4B) to random sampling (4C).

Figure 5 highlights the value of *target sampling* approaches. For the record, these sampling strategies forced the samples to be taken from several classes representing the distribution of auxiliary values which aims at favouring the dispersion of sample values. Figure 5A presents the results obtained by the quantile method while the Figure 5B presents the results obtained by the *k-means method*. Both

results are obtained with 9,000 *target samplings* (1,000 per field) with samples from 8 measurement sites. For comparison, Figure 5C reproduces the results of Figure 3 obtained by *random sampling*.

The comparison of figures 5A and 5B with figure 5C shows that the estimation errors with the *target sampling* approaches are lower than those obtained with the *random sampling*. Contrary to *random sampling*, the regression between the mean estimation error and the observed variance criterion does not present a minimum at which the estimation errors increase rapidly. For both approaches, the average estimation error is around 13% and don't depend on the values taken for the variance criterion.

This result can be explained by the way target sampling approaches constrain the values taken by the variance criterion. For quantile-*based target sampling*, these values ranged from 10^{-10} to 10^{-2} , and from 10^{-9} to 10^{-1} for the *k-means approach*. For both, the maximum values of the variance criterion (C_S) remain low enough to avoid high-variance estimations. This is illustrated by the red curve which only presents a slight increase for these two approaches compared to random sampling.

This result explains from a theoretical point of view, the interest of approaches implemented more or less empirically in the existing literature (Carillo et al., 2016; Araya-Alman et al., 2017; Meyers et al., 2020; Oger et al. 2020). These later propose sampling methods based on auxiliary data which aimed at driving the selection of measurement sites such as quantile intervals. Indeed, by constraining the attribute values of the measurement sites taking into account the distribution values of auxiliary variable, these approaches tend to (i) reduce the difference between the sample mean and the population mean, which is the numerator of the variance criterion $(\overline{X_R} - \overline{X_S})^2$, and (ii) increase the dispersion of sample values, which is the denominator of the variance criterion $\sum_{i \in S}(X_i - \overline{X_S})^2$. These two associated phenomena limit the values of the variance criterion and thus the variance of the estimate.

Further thought

The results presented in figure 3 show that, for a fixed number of sampling size, the estimation errors can be related to the variance criterion in the case of a linear model sampling. The choice of measurement sites according to their auxiliary data values thus appears to be a suitable tool to control a large proportion of the estimation error. Figure 4 shows that field properties – such as field size, yield variability or its correlation to auxiliary data – affect estimation error. Figure 5 shows that the selection of the measurement sites should be performed using *target sampling* approaches with quantile or k-means clustering. Also, new sampling approaches seeking to directly minimize the variance criterion could be promising.

The variance criterion defined in this paper makes it possible to compare two samples of the same size even before the measurements have been made or the estimate has been inferred. On the studied fields, up to 9 measurement sites are necessary to guarantee an estimation error lower than 10%. This number could be a little larger in real conditions as it is assumed here that there is no measurement error. Further work could be performed to try to characterize the interactions between the variance criterion and the number of sampling sites. However, these interactions would be field specific as they also depend on the size of the field and the correlation between the auxiliary data and the variable of interest. For a given sampling size, the direct use of the variance criterion equation allows to estimate the expected precision of the estimation from the value of auxiliary variable of the sample. The confidence that can be placed in an estimate is thus made quantifiable. This is a major issue in sampling problems in plant production. This information could be used to support the professional in defining the number of samples based on available sampling time and the expected quality of estimate to

achieve better trade-offs between operational constraints (time) and accuracy in yield estimation. However, the characterization of this variability remains dependent on the knowledge of the standard deviation of the model's residuals. This standard deviation may be specific to local conditions, the considered auxiliary information and its relation with the variable of interest. It may therefore be difficult to estimate, depending on the crops and the variables considered. The establishment of references to know the expected values for such model parameters in crop production represents a challenge for the development of model sampling approaches.

The proposed criterion is based on relatively simple hypotheses which, even if they are not always fully verified on real data, ensure that its use is applicable to real fields. The tests presented on a limited number of fields corresponding to different conditions confirms the relevance of the proposed formalisation and the potentiality for its practical use. However, the robustness of the method and the validity of the hypotheses on which it is based need to be tested in a wider range of situations and case studies. In particular, the linear model is based on the assumption of independence of the residuals, which means that the spatial structure of the variable of interest is entirely explained by the auxiliary data. This work could be extended to a more general framework adapting the expression of the variance of the residual of the model integrating a spatial structure. Furthermore, the approach and the theoretical considerations could also be extended to other types of models or to higher dimensional data to make it more adaptable to the diversity of plant production systems.

Conclusion

This paper proposes a statistical formalization of uncertainty for sampling methods based on auxiliary data and a linear model. It is shown that the quality of the estimates resulting from these methods depends on external factors but also on the choice of the measurement sites. The article thus proposes a criterion based on the selected measurement sites in order to control the expected quality of the estimation. A such criterion seems relevant to compare samples or sampling methods. This work shows that for a fixed number of measurements, samples with the best representativeness and the best dispersion allow to reach lower estimation variance. In practice, it is therefore interesting to balance the measurement sites between sites for which rather low values are expected and others for which rather high values are expected. It also shows that *target sampling* approaches based on classification algorithms as proposed in the literature tend to select samples with interesting properties with respect to this criterion and are therefore more likely to produce limited estimation errors. This work opens up new perspectives for sampling approaches based on auxiliary data such as variables obtained by remote sensing.

Acknowledgements

This work was supported by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004 (#Digitag).

Appendix

Abbreviations

C_S	variance criterion
N	set of potential sampling sites
n	size of the set N; $n = Card(N)$
NDVI	normalized difference vegetation index
R	set of sites not selected in the sample
RMSE	root mean square error
S	set of sampled sites
s	size of the set S; $s = Card(S)$
T	field yield
$\hat{T}$	field yield estimation
$\tilde{T}$	field yield forecast (accounting for T variance)
X_i	auxiliary data (NDVI) for site i
Y_i	variable of interest (yield) for site i
β_0 & β_1	linear model parameters
σ^2	variance of the residual of the model

Hypotheses and notations

Bold notations represent matrices and vectors.

For a given field, the objective is to estimate the total production. This field is divided in elementary sites so that the total production is the sum of production of each site. Only a limited number of these sites can be sampled in order to build an estimator of the total production. These sites are chosen from the set N of potential measurement sites. For each potential measurement site ($i \in N$), numbered from 1 to n, there is a value for the quantity of interest noted Y_i. This value is only known for the s sampled sites ($i \in S$). A second variable, noted X_i, corresponding to an auxiliary data which is available for each potential measurement site ($i \in N$). It is assumed that a linear relationship relates the quantity of interest to the auxiliary data. It is then possible to write the values of X_i knowing Y_i as shown in equation A1.

$$\boldsymbol{Y_N}|\boldsymbol{X_N} = \beta_0 \boldsymbol{I_N} + \beta_1 \boldsymbol{X_N} + \boldsymbol{\varepsilon_N} \qquad Eq.A1$$

With:

$$\boldsymbol{\varepsilon_N} \sim N(\boldsymbol{0_N}, \sigma^2 \boldsymbol{I_N}) \qquad Eq.A2$$

Where $\boldsymbol{Y_N}$ and $\boldsymbol{X_N}$ are two vectors of length n containing respectively the values of the quantity of interest and the auxiliary data. It should be noted that in the standard writing of the linear model in matrix form, $\boldsymbol{X_N}$ represents an incidence matrix, here $\boldsymbol{X_N}$ represents a vector because there is only one auxiliary data. The vector $\boldsymbol{0_n}$ **and** $\boldsymbol{1_n}$ vectors of length n containing respectively only 0 and only 1.The matrix $\boldsymbol{I_n}$ the identity matrix of dimension $n \times n$. Finally, β_0, β_1 and σ^2 represents the model parameters relating $\boldsymbol{Y_N}$ to $\boldsymbol{X_N}$.

$\boldsymbol{X_N}$ and $\boldsymbol{\varepsilon_N}$ are assumed to be multinormal vectors and independent. In particular $\boldsymbol{X_N}$ follows a multinormal distribution of expectation $\boldsymbol{\mu_N}$ and of variance $\boldsymbol{V_N}$. It is possible to write the expectation and variance of the conditional distribution of the observations of $\boldsymbol{Y_N}|\boldsymbol{X}_N$:

$$\mathbb{E}(\boldsymbol{Y_N}|\boldsymbol{X_N}) = \beta_0 \boldsymbol{I_N} + \beta_1 \boldsymbol{X_N} \qquad Eq.A3$$

$$\mathbb{V}(\boldsymbol{Y_N}|\boldsymbol{X_N}) = \sigma^2 \boldsymbol{I_N} \qquad Eq.A4$$

Therefore, the deconditioned vector $\boldsymbol{Y_N}$, follows a multinormal distribution of expectation and variance:

$$\mathbb{E}(\boldsymbol{Y_N}) = \mathbb{E}\big(\mathbb{E}(\boldsymbol{Y_N}|\boldsymbol{X_N})\big) = \mathbb{E}(\beta_0 \boldsymbol{I_N} + \beta_1 \boldsymbol{X_N})$$

$$\mathbb{E}(\boldsymbol{Y_N}) = \beta_0 \boldsymbol{I_N} + \beta_1 \boldsymbol{\mu_N} \qquad Eq.A5$$

And:

$$\mathbb{V}(\boldsymbol{Y_N}) = \mathbb{V}\big(\mathbb{E}(\boldsymbol{Y_N}|\boldsymbol{X_N})\big) + \mathbb{E}\big(\mathbb{V}(\boldsymbol{Y_N}|\boldsymbol{X_N})\big) = \mathbb{V}\big(\mathbb{E}(\beta_0 \boldsymbol{I_N} + \beta_1 \boldsymbol{X_N})\big) + \mathbb{E}\big(\mathbb{V}(\boldsymbol{\varepsilon_N})\big)$$

$$\mathbb{V}(\boldsymbol{Y_N}) = {\beta_1}^2 \boldsymbol{V_N} + \sigma^2 \boldsymbol{I_N} \qquad Eq.A6$$

The set S, consisting of the sites selected in the sample, and the set R, consisting of the sites not selected in the sample, form a partition of the set N : $N = S \cup R$ and $S \cap R = \emptyset$. We can thus decompose the vectors $\boldsymbol{Y_N}$ and $\boldsymbol{X_N}$ as shown in Equations A7, A8 and A9.

$$\boldsymbol{Y_N} = \begin{bmatrix} \boldsymbol{Y_S} \\ \boldsymbol{Y_R} \end{bmatrix} \qquad and \qquad \boldsymbol{X_N} = \begin{bmatrix} \boldsymbol{X_S} \\ \boldsymbol{X_R} \end{bmatrix} \qquad Eq.A7$$

We can also decompose the parameters of the multi-normal distribution of $\boldsymbol{X_N}$:

$$\boldsymbol{\mu_N} = \begin{bmatrix} \boldsymbol{\mu_S} \\ \boldsymbol{\mu_R} \end{bmatrix} \qquad Eq.A8$$

$$\boldsymbol{V_N} = \begin{bmatrix} \boldsymbol{V_{SS}} & \boldsymbol{V_{SR}} \\ \boldsymbol{V_{RS}} & \boldsymbol{V_{RR}} \end{bmatrix} \qquad Eq.A9$$

Estimation of the regression parameters from the sample

The regression is constructed from the observations of the variables X and Y, that are chosen for sampling, these being contained in the vectors $\boldsymbol{Y_S}$ et $\boldsymbol{X}_S$. The following equation repeats Eq. A1 for the set S:

$$\boldsymbol{Y_S}|\boldsymbol{X}_S = \beta_0 + \beta_1 \boldsymbol{X_S} + \boldsymbol{\varepsilon_S} \quad with \quad \boldsymbol{\varepsilon_S} \sim N(\boldsymbol{0_S}, \sigma^2 \boldsymbol{I_S})$$

$$\boldsymbol{Y_S}|\boldsymbol{X}_S = [\boldsymbol{1_S} \quad \boldsymbol{X_S}][\boldsymbol{\beta}] + \boldsymbol{\varepsilon_S} \quad with \quad [\boldsymbol{\beta}] = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad Eq.A10$$

The estimation of $\boldsymbol{\beta}$ from the set S by least squares leads to the following estimator:

$$\widehat{\boldsymbol{\beta}} = ([\boldsymbol{1_S} \quad \boldsymbol{X_S}]^t [\boldsymbol{1_S} \quad \boldsymbol{X_S}])^{-1} . [\boldsymbol{1_S} \quad \boldsymbol{X_S}]^t \boldsymbol{Y_S} \qquad Eq.A11$$

By defining $\overline{X_S} = \sum_{i \in S} \frac{X_i}{s}$, $\overline{Y_S} = \sum_{i \in S} \frac{Y_i}{s}$ and $\overline{X_S Y_S} = \sum_{i \in S} \frac{X_i \times Y_i}{s}$, it becomes possible to rewrite the expression of $\widehat{\boldsymbol{\beta}}$ as follow (Equation A12) :

$$\widehat{\boldsymbol{\beta}} = \frac{s}{\sum_{i \in S} (X_i - \overline{X_S})^2} \begin{bmatrix} \frac{1}{s} \times \sum_{i \in S} (X_i - \overline{X_S})^2 + \overline{X_S}^2 & -\overline{X_S} \\ -\overline{X_S} & 1 \end{bmatrix} \begin{bmatrix} \overline{Y_S} \\ \overline{X_S Y_S} \end{bmatrix} \qquad Eq.A12$$

We can then establish that the vector $\widehat{\boldsymbol{\beta}}$ follows a bi-normal distribution of expectation (Equation A13) and variance (Equation A14) :

$$\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad Eq.A13$$

$$\mathbb{V}(\widehat{\boldsymbol{\beta}}) = \frac{\sigma^2}{\sum_{i\in S}(X_i - \overline{X_S})^2}\begin{bmatrix} \frac{1}{s}\times\sum_{i\in S}(X_i - \overline{X_S})^2 + \overline{X_S}^2 & -\overline{X_S} \\ -\overline{X_S} & 1 \end{bmatrix} \qquad Eq.A14$$

Finally, we are interested in the estimator of σ^2, the last parameter of the linear model. This estimation is done with $s-2$ degrees of freedom:

$$\widehat{\sigma^2} = \frac{\left(\boldsymbol{Y_S} - [\mathbf{1_S} \quad \boldsymbol{X_S}].\widehat{\boldsymbol{\beta}}\right)^t\left(\boldsymbol{Y_S} - [\mathbf{1_S} \quad \boldsymbol{X_S}].\widehat{\boldsymbol{\beta}}\right)}{s-2} \qquad Eq.A15$$

Conditional law

In this part, we are interested in the joint vector $\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{bmatrix}$ which we wish to decompose using the notations presented in Eq. A7. We then obtain:

$$\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X_S}\\ \boldsymbol{X_R}\\ \boldsymbol{Y_S}\\ \boldsymbol{Y_R}\end{bmatrix} \qquad Eq.A16$$

From Eq. A5 and A8, it is possible to describe the expectation of the joint distribution:

$$\mathbb{E}\begin{bmatrix}\boldsymbol{X_S}\\ \boldsymbol{X_R}\\ \boldsymbol{Y_S}\\ \boldsymbol{Y_R}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\mu_S}\\ \boldsymbol{\mu_R}\\ \beta_0\mathbf{1_S} + \beta_1\boldsymbol{\mu_S}\\ \beta_0\mathbf{1_R} + \beta_1\boldsymbol{\mu_R}\end{bmatrix} \qquad Eq.A17$$

Similarly, from Eq. A6 and A9, it is possible to describe the variance of the joint distribution:

$$\mathbb{V}\begin{bmatrix}\boldsymbol{X_S}\\ \boldsymbol{X_R}\\ \boldsymbol{Y_S}\\ \boldsymbol{Y_R}\end{bmatrix} = \begin{bmatrix} \boldsymbol{V_S} & \boldsymbol{V_{SR}} & \beta_1\boldsymbol{V_S} & \beta_1\boldsymbol{V_{SR}} \\ \boldsymbol{V_{RS}} & \boldsymbol{V_R} & \beta_1\boldsymbol{V_{RS}} & \beta_1\boldsymbol{V_R} \\ \beta_1\boldsymbol{V_S} & \beta_1\boldsymbol{V_{SR}} & \beta_1{}^2\boldsymbol{V_S} + \sigma^2\boldsymbol{I_S} & \beta_1{}^2\boldsymbol{V_{SR}} \\ \beta_1\boldsymbol{V_{RS}} & \beta_1\boldsymbol{V_R} & \beta_1{}^2\boldsymbol{V_{RS}} & \beta_1{}^2\boldsymbol{V_R} + \sigma^2\boldsymbol{I_R} \end{bmatrix} \qquad Eq.A18$$

It should be noted that the matrices $\boldsymbol{V_S}$ and $\boldsymbol{V_R}$ are symmetrical and that matrices $\boldsymbol{V_{SR}}$ and $\boldsymbol{V_{RS}}$ are the transposed matrices of each other.

By distinguishing the values of $\boldsymbol{X_S}$, $\boldsymbol{X_R}$ et $\boldsymbol{Y_S}$ which are known (1) from those of $\boldsymbol{Y_R}$ which are unknown (2), the notations $\boldsymbol{m_1}$, $\boldsymbol{m_2}$, $\boldsymbol{\Sigma_{11}}$, $\boldsymbol{\Sigma_{12}}$, $\boldsymbol{\Sigma_{21}}$ et $\boldsymbol{\Sigma_{22}}$ are introduced:

$$\mathbb{E}\begin{bmatrix}\boldsymbol{X_S}\\ \boldsymbol{X_R}\\ \boldsymbol{Y}_S\\ \boldsymbol{Y_R}\end{bmatrix} = \begin{bmatrix}\boldsymbol{m_1}\\ \boldsymbol{m_2}\end{bmatrix}$$

With:

$$\boldsymbol{m_1} = \begin{bmatrix}\boldsymbol{\mu_S}\\ \boldsymbol{\mu_R}\\ \beta_0\mathbf{1_S} + \beta_1\boldsymbol{\mu_S}\end{bmatrix} \quad and \quad \boldsymbol{m_2} = [\beta_0\mathbf{1_R} + \beta_1\boldsymbol{\mu_R}] \qquad Eq.A19$$

And:

$$\mathbb{V}\begin{bmatrix}\boldsymbol{X_S}\\ \boldsymbol{X_R}\\ \boldsymbol{Y_S}\\ \boldsymbol{Y_R}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}}\\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}}\end{bmatrix}$$

With:

$$\boldsymbol{\Sigma_{11}} = \begin{bmatrix}\boldsymbol{V_S} & \boldsymbol{V_{SR}} & \beta_1\boldsymbol{V_S}\\ \boldsymbol{V_{RS}} & \boldsymbol{V_R} & \beta_1\boldsymbol{V_{RS}}\\ \beta_1\boldsymbol{V_S} & \beta_1\boldsymbol{V_{SR}} & {\beta_1}^2\boldsymbol{V_S} + \sigma^2\boldsymbol{I_S}\end{bmatrix} \quad and \quad \boldsymbol{\Sigma_{12}} = \begin{bmatrix}\beta_1\boldsymbol{V_{SR}}\\ \beta_1\boldsymbol{V_R}\\ {\beta_1}^2\boldsymbol{V_{SR}}\end{bmatrix} \qquad Eq.A20$$

$$\boldsymbol{\Sigma_{21}} = [\beta_1\boldsymbol{V_{RS}} \quad \beta_1\boldsymbol{V_R} \quad {\beta_1}^2\boldsymbol{V_{RS}}] \quad and \quad \boldsymbol{\Sigma_{22}} = [{\beta_1}^2\boldsymbol{V_R} + \sigma^2\boldsymbol{I_R}]$$

Formalization of an estimator

The objective is to estimate T, the sum of local yield values (Y_i) on the field. By separating the values for which an observation is available (S), from the unobserved values (R) as defined in Eq. A7:

$$T = \sum_{i \in N} Y_i \qquad Eq.A21$$

Which can also be written:

$$T = \mathbf{1}_N^t \boldsymbol{Y_N} \qquad Eq.A22$$

$$T = \mathbf{1}_S^t \boldsymbol{Y_S} + \mathbf{1}_R^t \boldsymbol{Y_R}$$

$\hat{T}$ is defined as the estimator of T. The values of the vector $\boldsymbol{Y_S}$, which correspond to the measured values of the quantity of interest, being known, the problem is to estimate the values of $\boldsymbol{Y_R}$. $\mathbf{1}_R^t\ \mathbb{E}(\boldsymbol{Y_R}|\boldsymbol{Y_S}, \boldsymbol{X_S}, \boldsymbol{X_S})$ is chosen as the estimator of $\mathbf{1}_R^t \boldsymbol{Y_R}$ because it minimizes the quadratic risk.

$$\hat{T} = \mathbf{1}_S^t \boldsymbol{Y_S} + \mathbf{1}_R^t\ \mathbb{E}(\boldsymbol{Y_R}|\boldsymbol{Y_S}, \boldsymbol{X_S}, \boldsymbol{X_R}) \qquad Eq.A23$$

By decomposing $\mathbb{E}(\boldsymbol{Y_R}|\boldsymbol{Y_S}, \boldsymbol{X_S}, \boldsymbol{X_S})$ using the conditional distribution of a multinormal distribution and the notations introduced in the previous subsection, $\hat{T}$ can be derived as expressed in equation A24.

$$\hat{T} = \mathbf{1}_S^t \boldsymbol{Y_S} + \mathbf{1}_R^t \left(\boldsymbol{m_2} + \boldsymbol{\Sigma_{21}}.\boldsymbol{\Sigma_{11}^{-1}}.\begin{bmatrix}\boldsymbol{X_S} - \boldsymbol{\mu_S}\\ \boldsymbol{X_R} - \boldsymbol{\mu_R}\\ \boldsymbol{Y_S} - \beta_0\boldsymbol{I_S} - \beta_1\boldsymbol{\mu_S}\end{bmatrix}\right) \qquad Eq.A24$$

It is possible to rewrite the expression for $\hat{T}$ as in equation A25 to make the size s of the sample (S) and the size n of the of potential measurement sites (N) appear.

$$\hat{T} = s\overline{Y_S} + (n-s)\beta_0 + \beta_1 \mathbf{1}_R^t \boldsymbol{X_R} \qquad Eq.A25$$

This formulation involves the coefficients β_0 and β_1. In practice, these are not known and replaced by their respective estimators:

$$\hat{T} = s\overline{Y_S} + (n-s)\widehat{\beta_0} + \widehat{\beta_1}\mathbf{1}_R^t \boldsymbol{X_R} \qquad Eq.A26$$

Estimator properties

For this estimator, we are interested in classical indicators such as the first and second order moments of the estimator in order to characterize its bias and the distribution around this bias:

$$\mathbb{E}\left(\hat{T}\right) = \mathbb{E}\left(s\overline{Y_S} + (n-s)\widehat{\beta_0} + \widehat{\beta_1}\mathbf{1}_R^t \boldsymbol{X_R}\right) = s\overline{Y_S} + (n-s)\beta_0 + \beta_1 \mathbf{1}_R^t \boldsymbol{X_R} \qquad Eq.A27$$

505 $$\mathbb{E}(\hat{T}) = s\overline{Y_S} + (n-s)\overline{Y_R} = n\overline{Y_N}$$

506 $$\mathbb{E}(\hat{T}) = \sum_{i \in N} Y_i$$

507 This is an unbiased estimator with variance:

508 $$\mathbb{V}(\hat{T}) = \mathbb{V}\left(s\overline{Y_S} + (n-s)\widehat{\beta_0} + \widehat{\beta_1}\mathbf{1}_{\boldsymbol{R}}^{\boldsymbol{t}}\boldsymbol{X}_{\boldsymbol{R}}\right)$$

509 $$\mathbb{V}(\hat{T}) = \left[(n-s) \quad \sum_{i \in R} X_i\right] . \mathbb{V}(\hat{\beta}) . \left[\sum_{i \in R}^{(n-s)} X_i\right] \qquad Eq.A28$$

510 This variance can be written:

511 $$\mathbb{V}(\hat{T}) = (n-s)^2 \times \left(\frac{1}{s} + \frac{(\overline{X_R} - \overline{X_S})^2}{\sum_{i \in S}(X_i - \overline{X_S})^2}\right) \times \sigma^2 \qquad Eq.A29$$

512 The variance of the estimator thus depends on:

513 - n, the size of the set of potential sampling sites within the field (N);
514 - s, the number of sampling sites or the size of the set S;
515 - σ^2, the variance of the residual of the model;
516 - $X_{i \epsilon S}$, the values taken individually by the measurement sites for the auxiliary data;
517 - $\overline{X_S}$, the average value of the measurement sites for the auxiliary data;
518 - $\overline{X_R}$, the average value of the non-selected sites for the auxiliary data.

519 This variance logically tends towards 0 when s tends towards n.

520 The reasoning held here led to the construction of an estimator of the expectation of T. If a prediction
521 is to be made, in the same way as for a linear regression prediction, the individual variance ε_i for each
522 of the unobserved Y_i ($i \in R$) must be considered. If $\tilde{T}$ is the forecast, it has for variance:

523 $$\mathbb{V}(\tilde{T}) = \mathbb{V}(\hat{T}) + (n-s).\mathbb{V}(\mathbf{1}_{\boldsymbol{R}}^{\boldsymbol{t}}\varepsilon_R) = \mathbb{V}(\hat{T}) + (n-s) \times \sigma^2$$

524 $$\mathbb{V}(\tilde{T}) = (n-s)^2 \times \left(\frac{1}{s} + \frac{1}{n-s} + \frac{(\overline{X_R} - \overline{X_S})^2}{\sum_{i \in S}(X_i - \overline{X_S})^2}\right) \times \sigma^2 \qquad Eq.A30$$

525 $\tilde{T}$ is a forecast of T, the sum of Y_i. The previous reasoning is applicable to $\frac{\tilde{T}}{n}$ which is an estimator of
526 the expectation of $Y_{i \in N}$. The variance of $\frac{\tilde{T}}{n}$ is of the formula $\frac{\mathbb{V}(\tilde{T})}{n^2}$ and has similar properties.

527

528 ## References:

529 M. Araya-Alman, C. Leroux, C. Acevedo-Opazo, S. Guillaume, H. Valdés-Gómez, N. Verdugo-Vásquez,
530 C. Pañitrur-De la Fuente & B. Tisseyre. *A new localized sampling method to improve grape yield*
531 *estimation of the current season using yield historical data.* Precision Agriculture volume 20, pages
532 445–459(2019) doi: 10.1007/s11119-019-09644-y

J. Arnó, J.A. Martínez-Casasnovas, A. Uribeetxebarria, & J. R. Rosell-Polo. (2017*). Comparing efficiency of different sampling schemes to estimate yield and quality parameters in fruit orchards.* Advances in Animal Biosciences, 8(2), 471-476.

E.M. Barnes & M.G. Baker, 2000*. Multispectral data for mapping soil texture: Possibilities and limitations.* Applied Engineering in Agriculture. VOL. 16(6): 731-741. doi: 10.13031/2013.5370

E. Carrillo, A. Matese, J. Rousseau, & B. Tisseyre, 2016. *Use of multi-spectral airborne imagery to improve yield sampling in viticulture.* Precision Agriculture, 17(1), 74-92. doi: 10.1007/s11119-015-9407-8

Corwin, D. L., Lesch, S. M., Shouse, P. J., Soppe, R., & Ayars, J. E., 2003. *Identifying soil properties that influence cotton yield using soil sampling directed by apparent soil electrical conductivity*. Agronomy Journal, 95(2), 352-364.

J. M. Damian, O.H D. C. Pias, M. R. Cherubin, A. Z. D. Fonseca, E. Z. Fornari & A. L. Santi, 2020. Applying the NDVI from satellite images in delimiting management zones for annual crops. Scientia Agricola, 77(1).

W. G. Jacoby (2000). *Loess: a nonparametric, graphical tool for depicting relationships between variables.* Electoral Studies Volume 19, Issue 4, December 2000, Pages 577-613. doi:10.1016/s0261-3794(99)00028-1

N. R. Kitchen, K. A., Sudduth, & S. T. Drummond, 1999. *Soil electrical conductivity as a crop productivity measure for claypan soils*. Journal of Production Agriculture, 12(4), 607-617.

S. Liaghat & S.K. Balasundram, 2010. *A Review: The Role of Remote Sensing in Precision Agriculture.* American Journal of Agricultural and Biological Sciences 5 (1): 50-55, 2010

MacQueen, J. B. (1967). *Some Methods for classification and Analysis of Multivariate Observations*. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press. pp. 281–297.

C. Miranda, L. G. Santesteban, J. Urrestarazu, M. Loidi & J. B. Royo, 2018. *Sampling Stratification Using Aerial Imagery to Estimate Fruit Load in Peach Tree Orchards.* Agriculture 2018, 8(6), 78; https://doi.org/10.3390/agriculture8060078

C. S. Murthy, S. Thiruvengadachari, P. V. Raju & S. Jonna. 1995. *Improved ground sampling and crop yield estimation using satellite data*. International Journal of Remote Sensing. Volume 17, 1996 - Issue 5 Pages 945-956. https://doi.org/10.1080/01431169608949057

B. Oger, P. Vismara & B. Tisseyre. 2020. *Combining target sampling with within field route-optimization to optimise on field yield estimation in viticulture*. Precision Agriculture (2020) doi: 10.1007/s11119-020-09744-0

A. Rehman, A. Z. Abbasi, N. Islam, Z. A. Shaikh. 2014. *A review of wireless sensors and networks' applications in agriculture*. Computer Standards & Interfaces. Volume 36, Issue 2, February 2014, Pages 263-270 https://doi.org/10.1016/j.csi.2011.03.004

A. Uribeetxebarria, J. A. Martínez-Casasnovas, B. Tisseyre, S. Guillaume, A. Escolà, J. R. Rosell-Polo, & J. Arnó. (2019). *Assessing ranked set sampling and ancillary data to improve fruit load estimates in peach orchards.* Computers and Electronics in Agriculture, 164, No. 104931.

572 L. Venkataratnam, 2001. Remote sensing and GIS inagricultural resources management. Proceedings
573 of the 1st National Conference on Agro-Informatics, June 3-4, Dharwad, India, pp: 20-29.
574 http://www.insait.org/abstracts.pdf

575 L. Wasserman. 2004. All of statistics: A concise course in statistical inference. New York: Springer.

576 D. Wulfsohn, 2010. Sampling techniques for plants and soil. In S. K. Upadhyaya, D. K. Giles, S.
577 Haneklaus, & E. Schnug (Eds.), Advanced engineering systems for specialty crops: A review of precision
578 agriculture for water, chemical, and nutrient application, and yield monitoring. Landbauforschung
579 Völkenrode, Special Issue 340, 3–30.