



HAL
open science

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

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

► **To cite this version:**

Baptiste Oger, Gilles Le Moguédec, Philippe Vismara, Bruno Tisseyre. A new criterion based on estimator variance for model sampling in precision agriculture. *Computers and Electronics in Agriculture*, 2022, 200, pp.107184. 10.1016/j.compag.2022.107184 . hal-03744279

HAL Id: hal-03744279

<https://hal.inrae.fr/hal-03744279>

Submitted on 23 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

B. Oger¹⁻², G. Le Moguédec³, P. Vismara²⁻⁴ and B. Tisseyre¹

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

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

36 Introduction

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

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

49 Despite the development of these new data collection methods, some decisions still require sampling
50 on the field as some measurements are still inaccessible using the current sensors. However, the
51 available new sources of information are valuable because they allow, when accessible with a high
52 spatial resolution, to characterize the variability and the spatial structure of the fields (Kitchen et al.
53 2020, Damian et al. 2020). Moreover, even when the desired measurement variables are not directly
54 accessible, the observations from the sensors can be more or less related to the variable of interest.
55 This is the case, for example, between yield and NDVI vigour observations obtained by remote sensing
56 in viticulture (Carillo et al, 2016) or between soil parameters and soil electrical conductivity (Corwin et
57 al. 2003). In this context, new sampling approaches based on these sources of information have
58 emerged. For example, *stratified sampling* and *target sampling* approaches use high spatial resolution
59 observations to drive the choice of measurement sites on the field (Miranda et al., 2018;
60 Uribeetxebarria et al., 2019; Arnó et al., 2017). Other methods propose to go further by also mobilizing
61 these observations when inferring the estimation of the variable of interest. The estimator is then built
62 on the basis of a model linking the sampled quantity to the available auxiliary high spatial resolution
63 information. These approaches, described as *model sampling*, have shown promising results in
64 agriculture (Murthy et al., 1995; Araya-Alman et al., 2019).

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

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

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

79 Material and method

80 Hypotheses and notations

81 In this section, bold notations represent matrices and vectors.

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

$$91 \quad \mathbf{Y}_N | \mathbf{X}_N = \beta_0 \mathbf{1}_n + \beta_1 \mathbf{X}_N + \boldsymbol{\varepsilon}_N \quad \text{Eq. 1}$$

92 With:

$$93 \quad \boldsymbol{\varepsilon}_N \sim N(\mathbf{0}_n, \sigma^2 \mathbf{I}_n) \quad \text{Eq. 2}$$

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

100 The set S , consisting of the sites selected in the sample, and the set R , consisting of the sites not
 101 selected in the sample, form a partition of the set N : $N = S \cup R$ and $S \cap R = \emptyset$. We can thus
 102 decompose the vectors \mathbf{Y}_N and \mathbf{X}_N as shown in Equations 3 and A7 in the appendix.

$$103 \quad \mathbf{Y}_N = \begin{bmatrix} \mathbf{Y}_S \\ \mathbf{Y}_R \end{bmatrix} \quad \text{and} \quad \mathbf{X}_N = \begin{bmatrix} \mathbf{X}_S \\ \mathbf{X}_R \end{bmatrix} \quad \text{Eq. 3}$$

104 Formalization of an estimator

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

107
$$T = \sum_{i \in N} Y_i \quad \text{Eq. 4}$$

108 Which can also be written:

109
$$T = \mathbf{1}_N^t \mathbf{Y}_N \quad \text{Eq. 5}$$

110
$$T = \mathbf{1}_S^t \mathbf{Y}_S + \mathbf{1}_R^t \mathbf{Y}_R$$

111 \hat{T} is defined as the estimator of T . The values of the vector \mathbf{Y}_S , which correspond to the measured
 112 values of the quantity of interest, being known, the problem is to estimate the values of \mathbf{Y}_R .
 113 $\mathbf{1}_R^t \mathbb{E}(\mathbf{Y}_R | \mathbf{Y}_S, \mathbf{X}_S, \mathbf{X}_R)$ is chosen as the estimator of $\mathbf{1}_R^t \mathbf{Y}_R$ because it minimizes the quadratic risk. The
 114 statistical work on the mathematical expression of this estimator is detailed in the appendix from Eq.
 115 A1 to Eq. A26.

116
$$\hat{T} = \mathbf{1}_S^t \mathbf{Y}_S + \mathbf{1}_R^t \mathbb{E}(\mathbf{Y}_R | \mathbf{Y}_S, \mathbf{X}_S, \mathbf{X}_R) \quad \text{Eq. 6}$$

117 Estimator and forecast properties

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

120
$$\mathbb{E}(\hat{T}) = \sum_{i \in N} Y_i \quad \text{Eq. 7}$$

121 This is an unbiased estimator with variance:

122
$$\mathbb{V}(\hat{T}) = (n - s)^2 \times \left(\frac{1}{s} + \frac{(\bar{X}_R - \bar{X}_S)^2}{\sum_{i \in S} (X_i - \bar{X}_S)^2} \right) \times \sigma^2 \quad \text{Eq. 8}$$

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

127
$$\mathbb{V}(\tilde{T}) = (n - s)^2 \times \left(\frac{1}{s} + \frac{1}{n - s} + \frac{(\bar{X}_R - \bar{X}_S)^2}{\sum_{i \in S} (X_i - \bar{X}_S)^2} \right) \times \sigma^2 \quad \text{Eq. 9}$$

128 The variance of the forecast thus depends on:

- 129
- 130 • n , the size of the set of potential sampling sites within the field (N);
 - 131 • s , the number of sampling sites or the size of the set S;
 - 132 • σ^2 , the variance of the residual of the model;
 - 133 • $X_{i \in S}$, the values taken individually by the measurement sites for the auxiliary data;
 - 134 • \bar{X}_S , the average value of the measurement sites for the auxiliary data;
 - 135 • \bar{X}_R , the average value of the non-selected sites for the auxiliary data.

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

137 \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
 138 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.

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

146 Variance criterion for the selection of measurement sites

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

$$149 \quad C_S = \frac{(\overline{X_R} - \overline{X_S})^2}{\sum_{i \in S} (X_i - \overline{X_S})^2} \quad Eq. 10$$

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

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

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

161 General method for the case study

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

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

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

170 Sampling methods

171 The first method implemented for selecting the s measurement sites is *random sampling* (Wulfsohn,
172 2010). In this approach, the set of S sampled sites is drawn from the set of N available sites by a random
173 draw.

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

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

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

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

183 For all three methods (random sampling, quantile and k-means), 1000 samples of size n ranging from
184 4 to 15 are drawn for each field (see next sections for the presentation of the fields).

185 Measurement of the quality of the estimate

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

$$189 \quad \text{Error (\%)} = \frac{|\tilde{T} - T|}{T} \quad \text{Eq. 11}$$

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

$$192 \quad \text{RMSE} = \sqrt{\sum_{i \in \text{Samples}} \frac{(\tilde{T}_i - T)^2}{\text{Cardinal}(\text{Samples})}} \quad \text{Eq. 12}$$

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

$$194 \quad \text{RMSE} = \sqrt{(\mathbb{E}(\tilde{T}) - T)^2 + \mathbb{V}(\tilde{T})} \quad \text{Eq. 13}$$

195 And as bias is nul (Eq. 14):

$$196 \quad \text{RMSE} = \sqrt{\mathbb{V}(\tilde{T})} \quad \text{Eq. 14}$$

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

$$199 \quad \text{RRMSE (\%)} = \frac{\text{RMSE}}{T} \times 100 \quad \text{Eq. 15}$$

200 Data

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

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



207
 208 *Figure 1: Representation of the plots on the INRAE Pech-Rouge domain. Field colour represent local NDVI from low (red) to*
 209 *high (green) computed with Avion Jaune multispectral images. P104 is further north. Background: Google maps.*

210 NDVI values were derived from a multispectral image with a resolution of 1 pixel = 1m² taken on August
 211 31, 2008 by Avion Jaune (Narbonne, Hérault, France). The spectral regions captured in the images
 212 were: blue (445-520 nm), green (510-600 nm), red (632-695 nm) and near infrared (757-853 nm). From
 213 this image, the aggregation method described by Acevedo-Opazo et al. (2008) was used to obtain 9 m²
 214 image pixels, reducing the effect of canopy and bare ground discontinuity on the measured values.
 215 NDVI was finally calculated from the processed images according to Rouse et al. (1973). Mechanical or
 216 chemical weed control was performed on the row spacing; therefore, weed control had extremely
 217 small effect on NDVI values.

218 *Table 1 : Characteristics of the experimental fields*

219	220	221	222	223	224	225	226	227
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%	

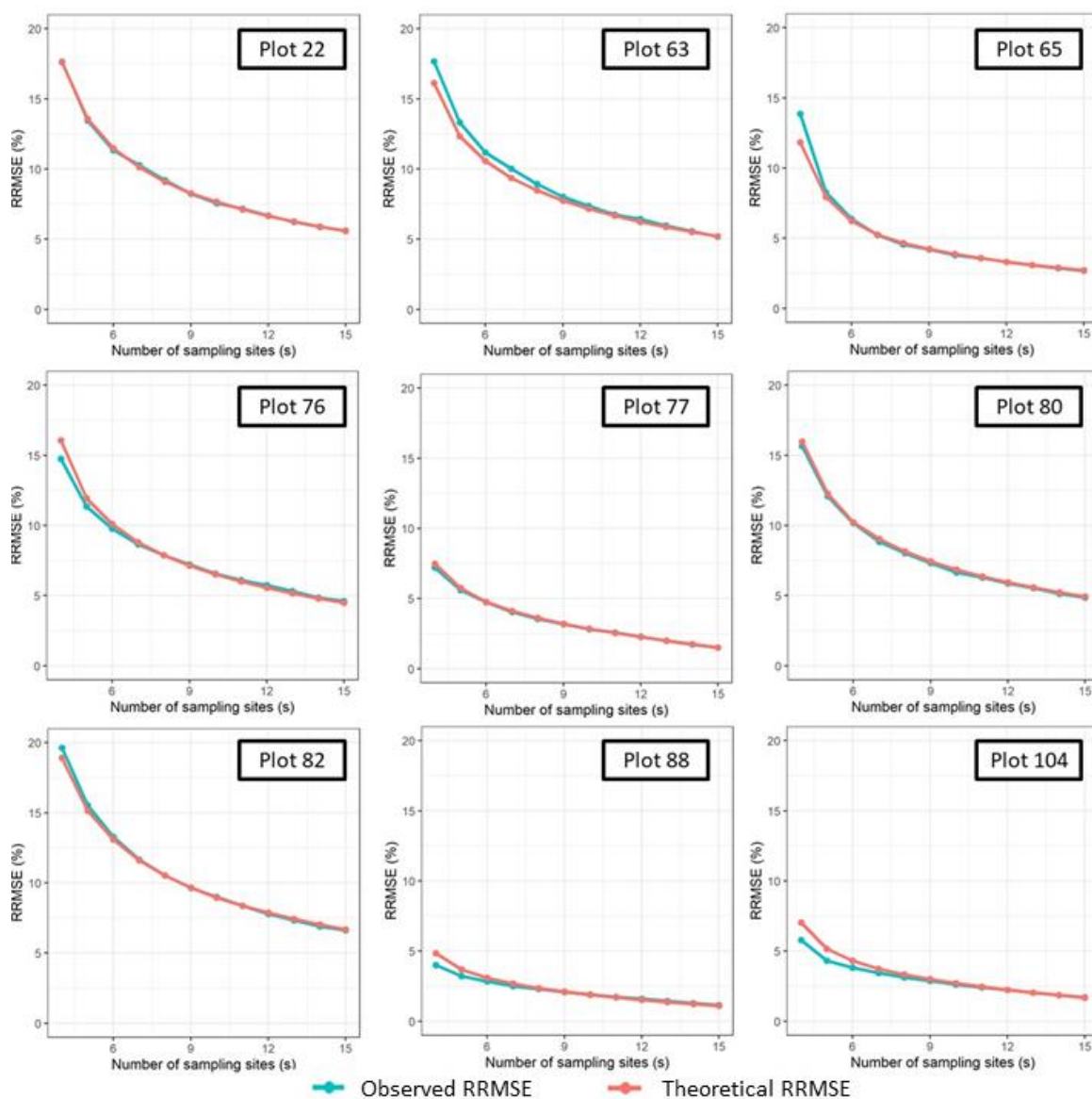
228 Local yield measurements on the fields were made locally on the nodes of a 15x15 m sampling grid. At
 229 each grid node, yield was measured on 5 consecutive vines along the row and the average yield was
 230 assigned to the coordinates of the grid node. The final database consisted of a set of 313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

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

233 Results

234 Validation

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



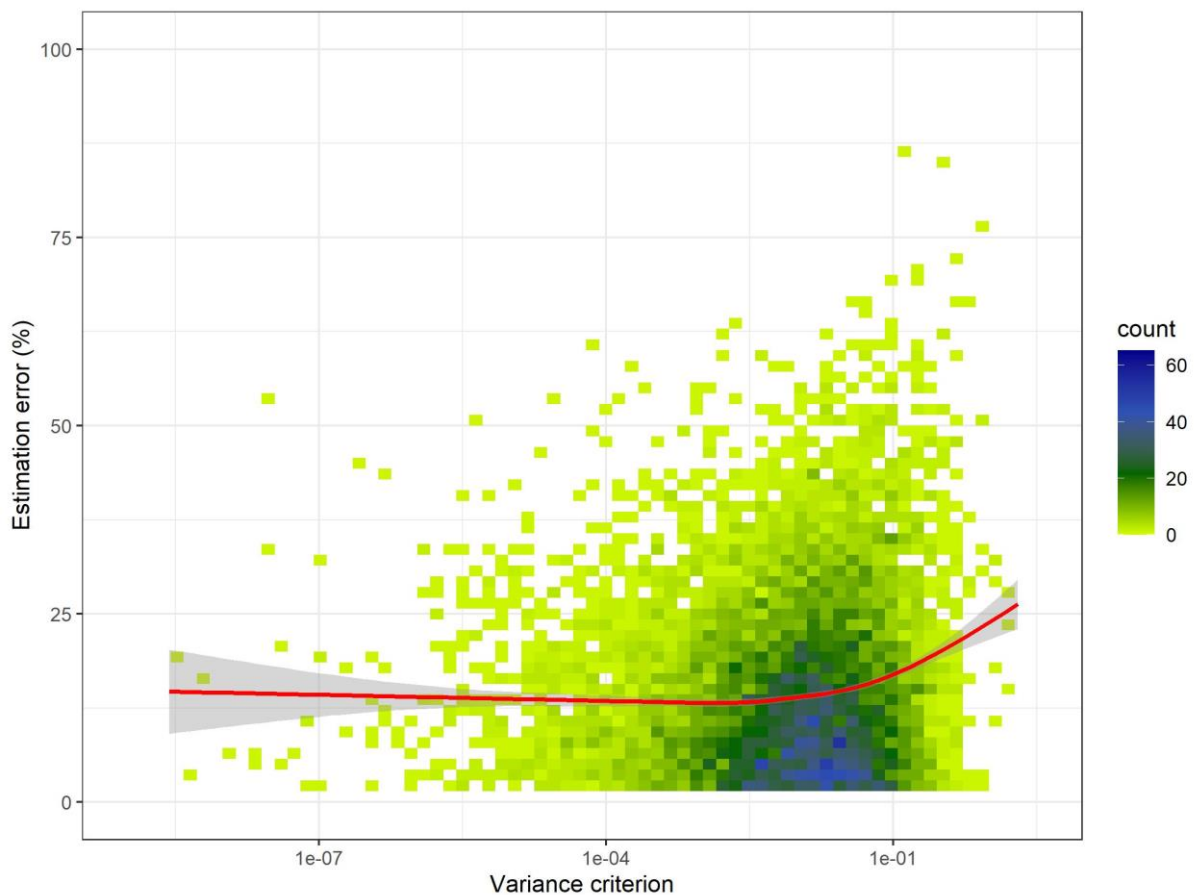
241

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

247 Variance criterion and random sampling

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

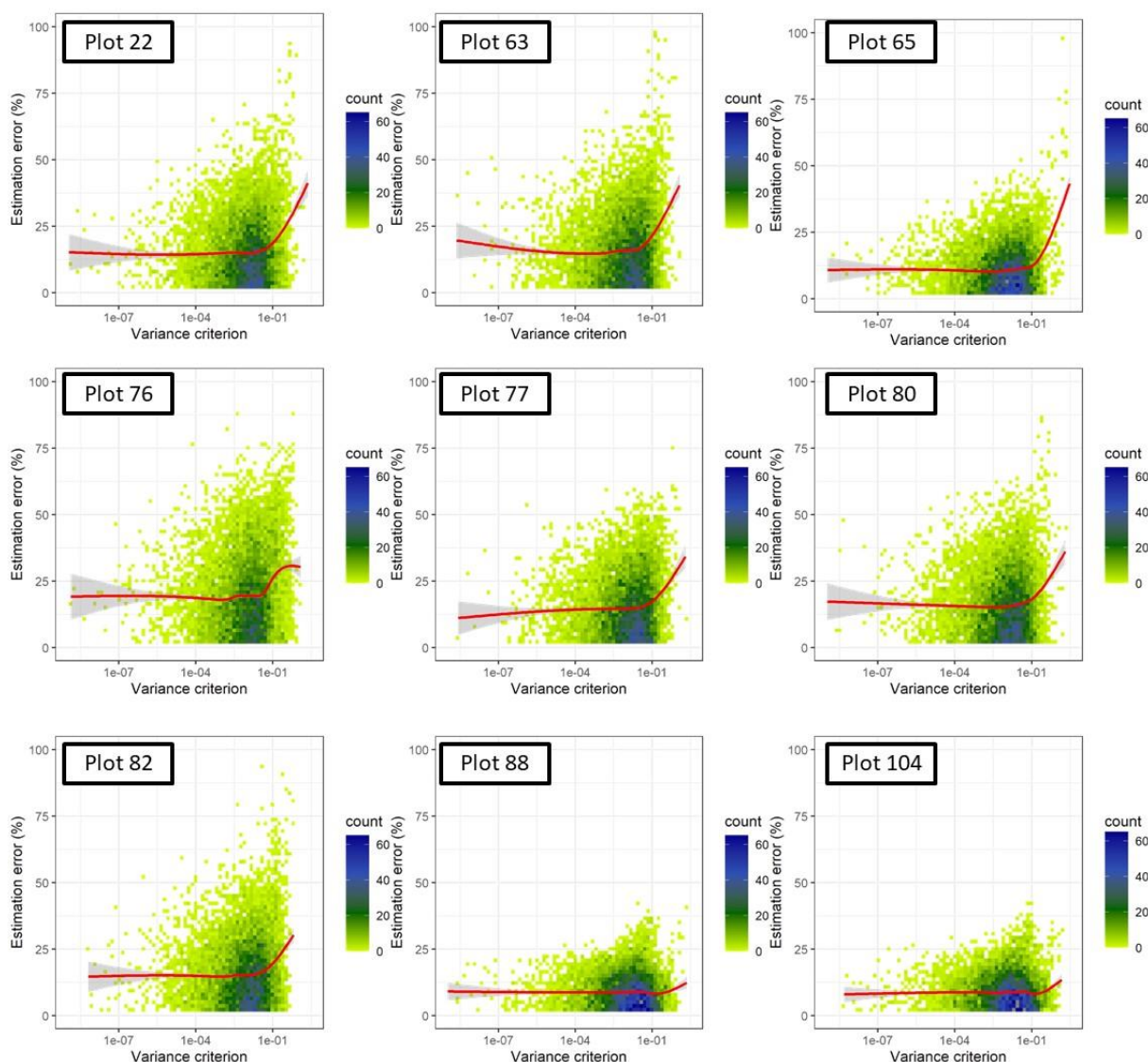
254 The values of the variance criterion taken for these random samples are concentrated around the
255 median (0.012) with 45% of the values between 10^{-2} and 10^{-1} and a dispersion ranging from 10^{-10}
256 to 10^1 . The red curve shows a local regression (Jacoby 2000) of the evolution of the mean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



261
262 *Figure 3: Relationship between variance criterion (C_s) and estimation error. The average estimation error (in red) increases*
263 *when the estimates are made with a sample that has a high variance criterion.*

264 For these fields, an increase in the estimation error as a function of the variance criterion is observed.
265 This increase is slow at first and then accelerates. This observation is consistent with the theoretical
266 equation for the variance of the estimate (Eq. 9). Indeed, in equation 9, the variance criterion is added
267 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
268 remains very small compared to the sum of terms $\frac{1}{s}$ and $\frac{1}{n-s}$ and variation of variance criterion then
269 have almost no impact on the variance of the estimate. When the variance criterion reaches values of

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



273
 274 *Figure 4: Evolution of the mean estimation error as a function of the variance criterion for all nine fields.*

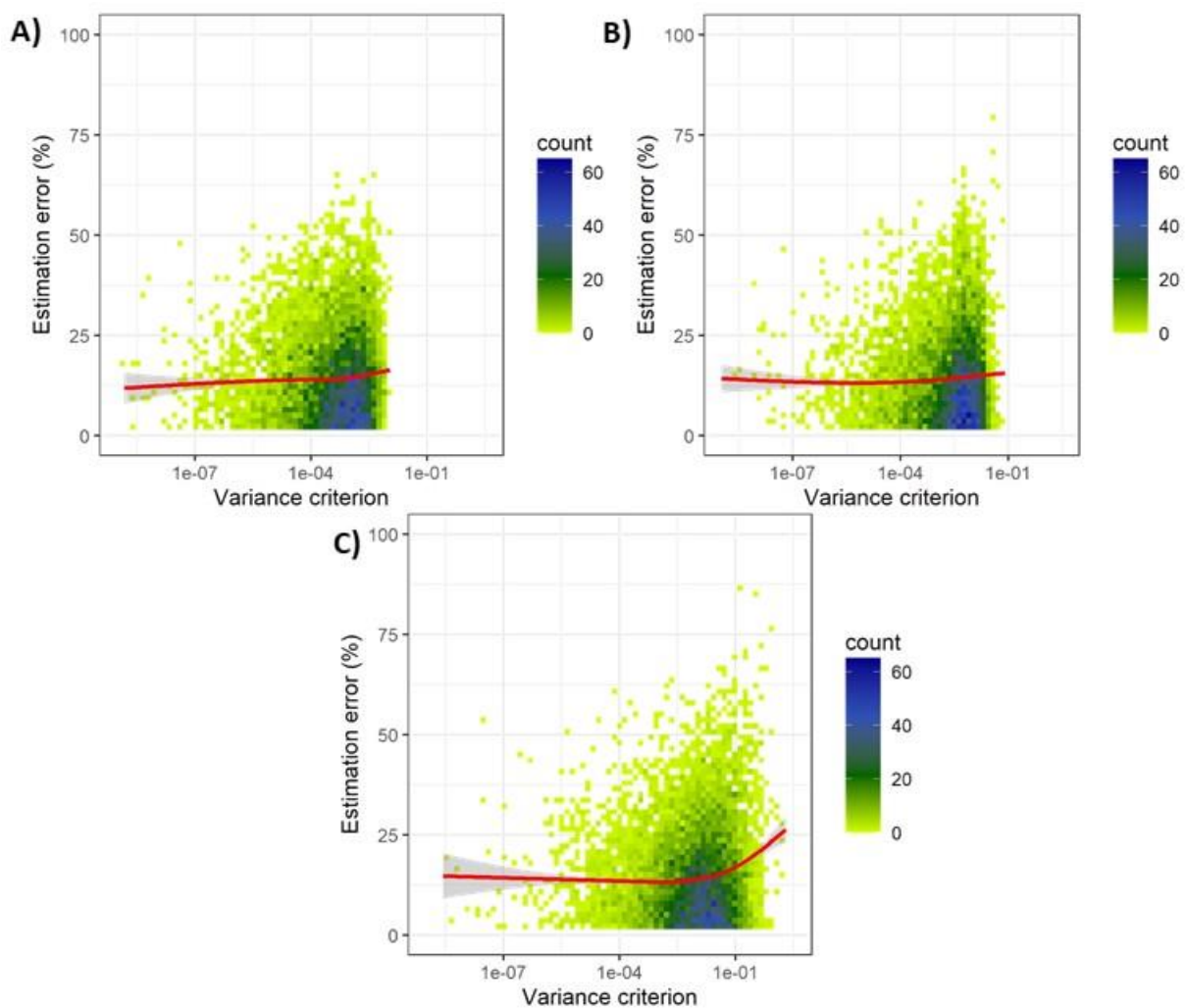
275 Using a similar procedure as in Figure 3, the nine graphs in Figure 4 show the individual results obtained
 276 on all fields. The results for each field are very similar to those presented in Figure 3: a high proportion
 277 of samples with a variance criterion value between 10^{-2} and 10^{-1}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

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

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

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

297 Variance criterion and targeted sampling



298

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

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

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

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

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

319 This result explains from a theoretical point of view, the interest of approaches implemented more or
320 less empirically in the existing literature (Carillo et al., 2016; Araya-Alman et al., 2017; Meyers et al.,
321 2020; Oger et al. 2020). These later propose sampling methods based on auxiliary data which aimed
322 at driving the selection of measurement sites such as quantile intervals. Indeed, by constraining the
323 attribute values of the measurement sites taking into account the distribution values of auxiliary
324 variable, these approaches tend to (i) reduce the difference between the sample mean and the
325 population mean, which is the numerator of the variance criterion $(\bar{X}_R - \bar{X}_S)^2$, and (ii) increase the
326 dispersion of sample values, which is the denominator of the variance criterion $\sum_{i \in S} (X_i - \bar{X}_S)^2$. These
327 two associated phenomena limit the values of the variance criterion and thus the variance of the
328 estimate.

329 Further thought

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

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

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

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

368 Conclusion

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

382 Acknowledgements

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

385

386 Appendix

387 Abbreviations

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

403

404 Hypotheses and notations

405 Bold notations represent matrices and vectors.

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

$$415 \quad \mathbf{Y}_N | \mathbf{X}_N = \beta_0 \mathbf{I}_N + \beta_1 \mathbf{X}_N + \boldsymbol{\varepsilon}_N \quad \text{Eq. A1}$$

416 With:

$$417 \quad \boldsymbol{\varepsilon}_N \sim N(\mathbf{0}_N, \sigma^2 \mathbf{I}_N) \quad \text{Eq. A2}$$

418 Where \mathbf{Y}_N and \mathbf{X}_N are two vectors of length n containing respectively the values of the quantity of
419 interest and the auxiliary data. It should be noted that in the standard writing of the linear model in
420 matrix form, \mathbf{X}_N represents an incidence matrix, here \mathbf{X}_N represents a vector because there is only
421 one auxiliary data. The vector $\mathbf{0}_n$ and $\mathbf{1}_n$ vectors of length n containing respectively only 0 and only
422 1. The matrix \mathbf{I}_n the identity matrix of dimension $n \times n$. Finally, β_0 , β_1 and σ^2 represents the model
423 parameters relating \mathbf{Y}_N to \mathbf{X}_N .

424 \mathbf{X}_N and $\boldsymbol{\varepsilon}_N$ are assumed to be multinormal vectors and independent. In particular \mathbf{X}_N follows a
425 multinormal distribution of expectation $\boldsymbol{\mu}_N$ and of variance \mathbf{V}_N . It is possible to write the expectation
426 and variance of the conditional distribution of the observations of $\mathbf{Y}_N | \mathbf{X}_N$:

$$427 \quad \mathbb{E}(\mathbf{Y}_N | \mathbf{X}_N) = \beta_0 \mathbf{I}_N + \beta_1 \mathbf{X}_N \quad \text{Eq. A3}$$

$$428 \quad \mathbb{V}(\mathbf{Y}_N | \mathbf{X}_N) = \sigma^2 \mathbf{I}_N \quad \text{Eq. A4}$$

429 Therefore, the deconditioned vector \mathbf{Y}_N , follows a multinormal distribution of expectation and
 430 variance:

$$431 \quad \mathbb{E}(\mathbf{Y}_N) = \mathbb{E}(\mathbb{E}(\mathbf{Y}_N|\mathbf{X}_N)) = \mathbb{E}(\beta_0\mathbf{I}_N + \beta_1\mathbf{X}_N)$$

$$432 \quad \mathbb{E}(\mathbf{Y}_N) = \beta_0\mathbf{I}_N + \beta_1\boldsymbol{\mu}_N \quad \text{Eq. A5}$$

433 And:

$$434 \quad \mathbb{V}(\mathbf{Y}_N) = \mathbb{V}(\mathbb{E}(\mathbf{Y}_N|\mathbf{X}_N)) + \mathbb{E}(\mathbb{V}(\mathbf{Y}_N|\mathbf{X}_N)) = \mathbb{V}(\mathbb{E}(\beta_0\mathbf{I}_N + \beta_1\mathbf{X}_N)) + \mathbb{E}(\mathbb{V}(\boldsymbol{\varepsilon}_N))$$

$$435 \quad \mathbb{V}(\mathbf{Y}_N) = \beta_1^2\mathbf{V}_N + \sigma^2\mathbf{I}_N \quad \text{Eq. A6}$$

436 The set S , consisting of the sites selected in the sample, and the set R , consisting of the sites not
 437 selected in the sample, form a partition of the set N : $N = S \cup R$ and $S \cap R = \emptyset$. We can thus
 438 decompose the vectors \mathbf{Y}_N and \mathbf{X}_N as shown in Equations A7, A8 and A9.

$$439 \quad \mathbf{Y}_N = \begin{bmatrix} \mathbf{Y}_S \\ \mathbf{Y}_R \end{bmatrix} \quad \text{and} \quad \mathbf{X}_N = \begin{bmatrix} \mathbf{X}_S \\ \mathbf{X}_R \end{bmatrix} \quad \text{Eq. A7}$$

440 We can also decompose the parameters of the multi-normal distribution of \mathbf{X}_N :

$$441 \quad \boldsymbol{\mu}_N = \begin{bmatrix} \boldsymbol{\mu}_S \\ \boldsymbol{\mu}_R \end{bmatrix} \quad \text{Eq. A8}$$

$$442 \quad \mathbf{V}_N = \begin{bmatrix} \mathbf{V}_{SS} & \mathbf{V}_{SR} \\ \mathbf{V}_{RS} & \mathbf{V}_{RR} \end{bmatrix} \quad \text{Eq. A9}$$

443 Estimation of the regression parameters from the sample

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

$$447 \quad \mathbf{Y}_S|\mathbf{X}_S = \beta_0 + \beta_1\mathbf{X}_S + \boldsymbol{\varepsilon}_S \quad \text{with} \quad \boldsymbol{\varepsilon}_S \sim N(\mathbf{0}_S, \sigma^2\mathbf{I}_S)$$

$$448 \quad \mathbf{Y}_S|\mathbf{X}_S = [\mathbf{1}_S \quad \mathbf{X}_S][\boldsymbol{\beta}] + \boldsymbol{\varepsilon}_S \quad \text{with} \quad [\boldsymbol{\beta}] = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \text{Eq. A10}$$

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

$$450 \quad \hat{\boldsymbol{\beta}} = ([\mathbf{1}_S \quad \mathbf{X}_S]^t [\mathbf{1}_S \quad \mathbf{X}_S])^{-1} \cdot [\mathbf{1}_S \quad \mathbf{X}_S]^t \mathbf{Y}_S \quad \text{Eq. A11}$$

451 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
 452 expression of $\hat{\boldsymbol{\beta}}$ as follow (Equation A12) :

$$453 \quad \hat{\boldsymbol{\beta}} = \frac{s}{\sum_{i \in S} (X_i - \overline{X}_S)^2} \begin{bmatrix} \frac{1}{s} \times \sum_{i \in S} (X_i - \overline{X}_S)^2 + \overline{X}_S^2 & -\overline{X}_S \\ -\overline{X}_S & 1 \end{bmatrix} \begin{bmatrix} \overline{Y}_S \\ \overline{X_S Y_S} \end{bmatrix} \quad \text{Eq. A12}$$

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

$$456 \quad \mathbb{E}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \text{Eq. A13}$$

457
$$\mathbb{V}(\hat{\boldsymbol{\beta}}) = \frac{\sigma^2}{\sum_{i \in S} (X_i - \bar{X}_S)^2} \begin{bmatrix} \frac{1}{s} \times \sum_{i \in S} (X_i - \bar{X}_S)^2 + \bar{X}_S^2 & -\bar{X}_S \\ -\bar{X}_S & 1 \end{bmatrix} \quad \text{Eq. A14}$$

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

460
$$\widehat{\sigma^2} = \frac{(\mathbf{Y}_S - [\mathbf{1}_S \ \mathbf{X}_S] \cdot \hat{\boldsymbol{\beta}})^t (\mathbf{Y}_S - [\mathbf{1}_S \ \mathbf{X}_S] \cdot \hat{\boldsymbol{\beta}})}{s - 2} \quad \text{Eq. A15}$$

461 Conditional law

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

464
$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_S \\ \mathbf{X}_R \\ \mathbf{Y}_S \\ \mathbf{Y}_R \end{bmatrix} \quad \text{Eq. A16}$$

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

466
$$\mathbb{E} \begin{bmatrix} \mathbf{X}_S \\ \mathbf{X}_R \\ \mathbf{Y}_S \\ \mathbf{Y}_R \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_S \\ \boldsymbol{\mu}_R \\ \beta_0 \mathbf{1}_S + \beta_1 \boldsymbol{\mu}_S \\ \beta_0 \mathbf{1}_R + \beta_1 \boldsymbol{\mu}_R \end{bmatrix} \quad \text{Eq. A17}$$

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

468
$$\mathbb{V} \begin{bmatrix} \mathbf{X}_S \\ \mathbf{X}_R \\ \mathbf{Y}_S \\ \mathbf{Y}_R \end{bmatrix} = \begin{bmatrix} \mathbf{V}_S & \mathbf{V}_{SR} & \beta_1 \mathbf{V}_S & \beta_1 \mathbf{V}_{SR} \\ \mathbf{V}_{RS} & \mathbf{V}_R & \beta_1 \mathbf{V}_{RS} & \beta_1 \mathbf{V}_R \\ \beta_1 \mathbf{V}_S & \beta_1 \mathbf{V}_{SR} & \beta_1^2 \mathbf{V}_S + \sigma^2 \mathbf{I}_S & \beta_1^2 \mathbf{V}_{SR} \\ \beta_1 \mathbf{V}_{RS} & \beta_1 \mathbf{V}_R & \beta_1^2 \mathbf{V}_{RS} & \beta_1^2 \mathbf{V}_R + \sigma^2 \mathbf{I}_R \end{bmatrix} \quad \text{Eq. A18}$$

469 It should be noted that the matrices \mathbf{V}_S and \mathbf{V}_R are symmetrical and that matrices \mathbf{V}_{SR} and \mathbf{V}_{RS} are
470 the transposed matrices of each other.

471 By distinguishing the values of \mathbf{X}_S , \mathbf{X}_R et \mathbf{Y}_S which are known (1) from those of \mathbf{Y}_R which are unknown
472 (2), the notations \mathbf{m}_1 , \mathbf{m}_2 , $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{12}$, $\boldsymbol{\Sigma}_{21}$ et $\boldsymbol{\Sigma}_{22}$ are introduced:

473
$$\mathbb{E} \begin{bmatrix} \mathbf{X}_S \\ \mathbf{X}_R \\ \mathbf{Y}_S \\ \mathbf{Y}_R \end{bmatrix} = \begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{bmatrix}$$

474 With:

475
$$\mathbf{m}_1 = \begin{bmatrix} \boldsymbol{\mu}_S \\ \boldsymbol{\mu}_R \\ \beta_0 \mathbf{1}_S + \beta_1 \boldsymbol{\mu}_S \end{bmatrix} \quad \text{and} \quad \mathbf{m}_2 = [\beta_0 \mathbf{1}_R + \beta_1 \boldsymbol{\mu}_R] \quad \text{Eq. A19}$$

476 And:

477
$$\mathbb{V} \begin{bmatrix} \mathbf{X}_S \\ \mathbf{X}_R \\ \mathbf{Y}_S \\ \mathbf{Y}_R \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

478 With:

479
$$\boldsymbol{\Sigma}_{11} = \begin{bmatrix} \mathbf{V}_S & \mathbf{V}_{SR} & \beta_1 \mathbf{V}_S \\ \mathbf{V}_{RS} & \mathbf{V}_R & \beta_1 \mathbf{V}_{RS} \\ \beta_1 \mathbf{V}_S & \beta_1 \mathbf{V}_{SR} & \beta_1^2 \mathbf{V}_S + \sigma^2 \mathbf{I}_S \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_{12} = \begin{bmatrix} \beta_1 \mathbf{V}_{SR} \\ \beta_1 \mathbf{V}_R \\ \beta_1^2 \mathbf{V}_{SR} \end{bmatrix} \quad \text{Eq. A20}$$

480
$$\boldsymbol{\Sigma}_{21} = [\beta_1 \mathbf{V}_{RS} \quad \beta_1 \mathbf{V}_R \quad \beta_1^2 \mathbf{V}_{RS}] \quad \text{and} \quad \boldsymbol{\Sigma}_{22} = [\beta_1^2 \mathbf{V}_R + \sigma^2 \mathbf{I}_R]$$

481 Formalization of an estimator

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

484
$$T = \sum_{i \in N} Y_i \quad \text{Eq. A21}$$

485 Which can also be written:

486
$$T = \mathbf{1}_N^t \mathbf{Y}_N \quad \text{Eq. A22}$$

487
$$T = \mathbf{1}_S^t \mathbf{Y}_S + \mathbf{1}_R^t \mathbf{Y}_R$$

488 \hat{T} is defined as the estimator of T . The values of the vector \mathbf{Y}_S , which correspond to the measured
489 values of the quantity of interest, being known, the problem is to estimate the values of \mathbf{Y}_R .
490 $\mathbf{1}_R^t \mathbb{E}(\mathbf{Y}_R | \mathbf{Y}_S, \mathbf{X}_S, \mathbf{X}_R)$ is chosen as the estimator of $\mathbf{1}_R^t \mathbf{Y}_R$ because it minimizes the quadratic risk.

491
$$\hat{T} = \mathbf{1}_S^t \mathbf{Y}_S + \mathbf{1}_R^t \mathbb{E}(\mathbf{Y}_R | \mathbf{Y}_S, \mathbf{X}_S, \mathbf{X}_R) \quad \text{Eq. A23}$$

492 By decomposing $\mathbb{E}(\mathbf{Y}_R | \mathbf{Y}_S, \mathbf{X}_S, \mathbf{X}_R)$ using the conditional distribution of a multinormal distribution and
493 the notations introduced in the previous subsection, \hat{T} can be derived as expressed in equation A24.

494
$$\hat{T} = \mathbf{1}_S^t \mathbf{Y}_S + \mathbf{1}_R^t \left(\mathbf{m}_2 + \boldsymbol{\Sigma}_{21} \cdot \boldsymbol{\Sigma}_{11}^{-1} \cdot \begin{bmatrix} \mathbf{X}_S - \boldsymbol{\mu}_S \\ \mathbf{X}_R - \boldsymbol{\mu}_R \\ \mathbf{Y}_S - \beta_0 \mathbf{I}_S - \beta_1 \boldsymbol{\mu}_S \end{bmatrix} \right) \quad \text{Eq. A24}$$

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

497
$$\hat{T} = s \bar{Y}_S + (n - s) \beta_0 + \beta_1 \mathbf{1}_R^t \mathbf{X}_R \quad \text{Eq. A25}$$

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

500
$$\hat{T} = s \bar{Y}_S + (n - s) \widehat{\beta}_0 + \widehat{\beta}_1 \mathbf{1}_R^t \mathbf{X}_R \quad \text{Eq. A26}$$

501 Estimator properties

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

504
$$\mathbb{E}(\hat{T}) = \mathbb{E}(s \bar{Y}_S + (n - s) \widehat{\beta}_0 + \widehat{\beta}_1 \mathbf{1}_R^t \mathbf{X}_R) = s \bar{Y}_S + (n - s) \beta_0 + \beta_1 \mathbf{1}_R^t \mathbf{X}_R \quad \text{Eq. A27}$$

505
$$\mathbb{E}(\hat{T}) = s\bar{Y}_S + (n - s)\bar{Y}_R = n\bar{Y}_N$$

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

507 This is an unbiased estimator with variance:

508
$$\mathbb{V}(\hat{T}) = \mathbb{V}(s\bar{Y}_S + (n - s)\widehat{\beta}_0 + \widehat{\beta}_1 \mathbf{1}_R^t \mathbf{X}_R)$$

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

510 This variance can be written:

511
$$\mathbb{V}(\hat{T}) = (n - s)^2 \times \left(\frac{1}{s} + \frac{(\bar{X}_R - \bar{X}_S)^2}{\sum_{i \in S} (X_i - \bar{X}_S)^2} \right) \times \sigma^2 \quad \text{Eq. A29}$$

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 • $X_{i \in S}$, the values taken individually by the measurement sites for the auxiliary data;
 517 • \bar{X}_S , the average value of the measurement sites for the auxiliary data;
 518 • \bar{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
$$\mathbb{V}(\tilde{T}) = \mathbb{V}(\hat{T}) + (n - s) \cdot \mathbb{V}(\mathbf{1}_R^t \varepsilon_R) = \mathbb{V}(\hat{T}) + (n - s) \times \sigma^2$$

524
$$\mathbb{V}(\tilde{T}) = (n - s)^2 \times \left(\frac{1}{s} + \frac{1}{n - s} + \frac{(\bar{X}_R - \bar{X}_S)^2}{\sum_{i \in S} (X_i - \bar{X}_S)^2} \right) \times \sigma^2 \quad \text{Eq. A30}$$

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. Pañitru-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

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

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

538 E. Carrillo, A. Matese, J. Rousseau, & B. Tisseyre, 2016. *Use of multi-spectral airborne imagery to*
539 *improve yield sampling in viticulture*. *Precision Agriculture*, 17(1), 74-92. doi: 10.1007/s11119-015-
540 9407-8

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

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

547 W. G. Jacoby (2000). *Loess: a nonparametric, graphical tool for depicting relationships between*
548 *variables*. *Electoral Studies* Volume 19, Issue 4, December 2000, Pages 577-613. doi:10.1016/s0261-
549 3794(99)00028-1

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

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

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

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

560 C. S. Murthy, S. Thiruvengadachari, P. V. Raju & S. Jonna. 1995. *Improved ground sampling and crop*
561 *yield estimation using satellite data*. *International Journal of Remote Sensing*. Volume 17, 1996 - Issue
562 5 Pages 945-956. <https://doi.org/10.1080/01431169608949057>

563 B. Oger, P. Vismara & B. Tisseyre. 2020. *Combining target sampling with within field route-optimization*
564 *to optimise on field yield estimation in viticulture*. *Precision Agriculture* (2020) doi: 10.1007/s11119-
565 020-09744-0

566 A. Rehman, A. Z. Abbasi, N. Islam, Z. A. Shaikh. 2014. *A review of wireless sensors and networks'*
567 *applications in agriculture*. *Computer Standards & Interfaces*. Volume 36, Issue 2, February 2014, Pages
568 263-270 <https://doi.org/10.1016/j.csi.2011.03.004>

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

- 572 L. Venkataratnam, 2001. Remote sensing and GIS in agricultural 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.