Model and auxiliary data for an accurate estimate of the field mean: Appendix

Baptiste Oger, Sébastien Roux, Gilles Le Moguedec, Bruno Tisseyre

Model and auxiliary data for an accurate estimate of the field mean: Appendix

B. Oger[1-2], S. Roux[2], G. Le Moguedec[3] 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, France
baptiste.oger@supagro.fr

This document supplements the proposed paper '*Model and auxiliary data for an accurate estimate of the field mean*' for the 13th European Conference on Precision Agriculture - ECPA 2021 (Oger et al. 2021). This conference paper proposes a discussion on inference methods for crop production estimation following previous studies (Carillo et al., 2016; Araya-Alma et al. 2019). This appendix includes the full demonstration of the theoretical results. For context and interpretation, please refer to the conference paper.

Notations and hypothesis

For a given plot, K is defined as the within-plot set of sites, numbered form 1 to k. For a given site $i \in K$, X_i and Y_i are respectively the auxiliary variable and the variable of interest measured on that site. An auxiliary data is assumed available at low cost and at a high spatial resolution. The auxiliary variable X_i is known for every site but the variable of interest Y_i only for a set N of sampled sites (with $Cardinal(N) = n$). It is assumed that Y_i can be predicted everywhere form X_i via a simple linear model:

$$\mathrm{Y_i} = \beta_0 + \beta_1 \mathrm{X_i} + \varepsilon_i \qquad with \qquad \varepsilon_i \sim N(0, \sigma^2) \tag{1}$$

Let $\overline{Y}$ be the average value of Y for the whole plot:

$$\overline{Y} = \frac{1}{k} \times \sum_{i \in K} Y_i \tag{2}$$

To focus on estimator related error, two hypotheses are made here. First, the calculations are made under the assumption that measurement sites are determined (the N set is fixed). Second, model residuals are considered independent, i.e. unaffected by the spatial structure of the plot.

Mean-based estimator

Estimator

A first estimator $\widehat{\overline{Y}_1}$ of $\overline{Y}$ based on the mean inference. $\widehat{\overline{Y}_1}$ is defined in Equation 3 as the average of the values Y_i taken on the n sampled sites ($i \in N$):

$$\boldsymbol{\widehat{\overline{Y}_1} = \frac{1}{n} \times \sum_{i \in N} Y_i} \tag{3}$$

Expectation

Main statistical properties of this estimator are given by its expected value (for the bias) and the variance (for the precision). The $\widehat{\overline{Y}_1}$ expectation (*E*) is written:

$$E\left[\widehat{\overline{Y}_1}\right] = E\left[\frac{1}{n} \times \sum_{i \in N} Y_i\right]$$

Although the mean estimator is not based on auxiliary data, it is assumed that such data exist to characterize its properties. By taking the linear relationship between Y_i and X_i presented in Equation 1, it is possible to replace Y_i by its expression from X_i:

$$\mathrm{E}\left[\widehat{\bar{Y}_1}\right] = \mathrm{E}\left[\frac{1}{n} \times \sum_{i \in N} (\beta_0 + \beta_1 X_i + \varepsilon_i)\right]$$

$$\mathrm{E}\left[\widehat{\bar{Y}_1}\right] = \mathrm{E}\left[\frac{1}{n} \times \sum_{i \in N} \beta_0\right] + \mathrm{E}\left[\frac{1}{n} \times \sum_{i \in N} (\beta_1 X_i)\right] + \mathrm{E}\left[\frac{1}{n} \times \sum_{i \in N} \varepsilon_i\right]$$

$$\mathrm{E}\left[\widehat{\bar{Y}_1}\right] = \beta_0 + \beta_1 \overline{X_{\imath \in N}} + 0$$

$$\mathrm{E}\left[\widehat{\bar{Y}_1}\right] = \beta_0 + \beta_1 (\overline{X_{\imath \in N}} + \bar{X} - \bar{X})$$

$$\mathrm{E}\left[\widehat{\bar{Y}_1}\right] = \beta_0 + \beta_1 \bar{X} + \beta_1 (\overline{X_{\imath \in N}} - \bar{X})$$

And finally:

$$\boldsymbol{E\left[\widehat{\bar{Y}_1}\right] = \bar{Y} + \beta_1 (\overline{X_N} - \bar{X})} \quad (4)$$

$\widehat{\bar{Y}_1}$ presents a bias which depend on β_1,the slope parameter, $\overline{X_N}$, the auxiliary data mean value for the sample and $\bar{X}$, the auxiliary data mean value for the whole plot. The difference between $\overline{X_N}$ and $\bar{X}$ reflects the influence of sample representativeness, ideally the sample should have properties close to those of the population. Since representativeness is expressed in terms of auxiliary data, its influence depends on β_1 which characterizes the strength of the correlation between the variables Y_i and X_i.

Variance:

Looking at its variance:

$$Var\left[\widehat{\bar{Y}_1}\right] = Var\left[\frac{1}{n} \times \sum_{i \in N} Y_i\right]$$

Using Equation 1:

$$Var\left[\widehat{\bar{Y}_1}\right] = Var\left[\frac{1}{n} \times \sum_{i \in N} (\beta_0 + \beta_1 X_i + \varepsilon_i)\right]$$

As β_0, β_1 and X_i correspond to numerical values, their variances are equal to 0. For X_i, this is true under the hypothesis that measurement sites are determined.

$$Var\left[\widehat{\bar{Y}_1}\right] = Var\left[\frac{1}{n} \times \sum_{i \in N} \varepsilon_i\right]$$

With $\mathrm{V}ar(aX) = a^2 Var(X)$:

$$Var\left[\widehat{\bar{Y}_1}\right] = \frac{1}{n^2} \times Var\left[\sum_{i \in N} \varepsilon_i\right]$$

And finally:

$$Var\left[\widehat{\bar{Y}_1}\right] = \frac{1}{n^2} \times n\sigma^2$$

$$\boldsymbol{Var\left[\widehat{\bar{Y}_1}\right] = \frac{\sigma^2}{n}} \quad (5)$$

The variance can be described with σ, the residual variance of the linear model and n, the number of sampling sites. This result is logical, the higher the variability of the variable of interest, the more variable the estimate will be. Conversely, a larger sample size improves confidence in the estimate.

Model-based estimator

Estimator

Let $\hat{Y}_i$ be the estimate for Y_i based on the linear model calibrated using the values of Y_i on the n sampling sites. A second estimator, $\widehat{\bar{Y}_2}$, can be described from the set of values predicted by the model:

$$\widehat{\bar{Y}_2} = \frac{1}{k} \times \sum_{i=1}^{k} \hat{Y}_i \quad (6)$$

which can be rewritten:

$$\widehat{\bar{Y}_2} = \frac{1}{k} \times \sum_{i=1}^{k} \left(\widehat{\beta_0} + \widehat{\beta_1} X_i\right)$$

$$\widehat{\bar{Y}_2} = \widehat{\beta_0} + \widehat{\beta_1} \bar{X}$$

$\widehat{\beta_0}$ and $\widehat{\beta_1}$ are the respective estimators of β_0 and β_1. In the simple linear model, these estimators are well known (Wasserman, 2004):

$$\widehat{\beta_1} = \frac{\sum_{i \in N}(X_i - \overline{X_N})(Y_i - \overline{Y_N})}{\sum_{i \in N}(X_i - \overline{X_N})^2}$$

$$\widehat{\beta_0} = \overline{Y_N} - \widehat{\beta_1} \times \overline{X_N}$$

And where $\overline{X_N}$ represents the mean of X_i for $i \in N$ and $\overline{Y_N}$ represents the mean Y_i for $i \in N$.

Expectation

As for $\widehat{\bar{Y}_1}$, the expectation expression is derived for $\widehat{\bar{Y}_2}$:

$$\mathrm{E}\left[\widehat{\bar{Y}_2}\right] = E\left[\widehat{\beta_0} + \widehat{\beta_1} \bar{X}\right]$$

$$\mathrm{E}\left[\widehat{\bar{Y}_2}\right] = E\left[\widehat{\beta_0}\right] + E\left[\widehat{\beta_1}\right] \bar{X}$$

One of the properties of the linear model is that the estimators $\widehat{\beta_0}$ and $\widehat{\beta_1}$ for β_0 and β_1 are unbiased (Wasserman, 2004). In other words, that $E\left[\widehat{\beta_0}\right] = \widehat{\beta_0}$ and $E\left[\widehat{\beta_1}\right] = \widehat{\beta_1}$. As a result:

$$\mathrm{E}\left[\widehat{\bar{Y}_2}\right] = \widehat{\beta_0} + \widehat{\beta_1} \bar{X}$$

And finally:

$$\mathbf{E}\left[\widehat{\bar{Y}_2}\right] = \bar{Y} \quad (7)$$

This result shows that the estimation based on model and auxiliary data inference is unbiased.

Variance:

To be able to fully compare the two estimators, it is nevertheless necessary to look at the $\widehat{\bar{Y}_2}$ variance:

$$Var\left[\widehat{\bar{Y}_2}\right] = Var\left[\widehat{\beta_0} + \widehat{\beta_1}\bar{X}\right]$$

The variance of a sum can be decomposed as follows: : $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$. As a result:

$$Var\left[\widehat{\bar{Y}_2}\right] = Var\left[\widehat{\beta_0}\right] + Var\left[\widehat{\beta_1}\bar{X}\right] + 2Cov\left[\widehat{\beta_0}, \widehat{\beta_1}\bar{X}\right]$$

$$Var\left[\widehat{\bar{Y}_2}\right] = \sigma^2_{\widehat{\beta_0}} + \sigma^2_{\widehat{\beta_1}}\bar{X}^2 + 2\bar{X}Cov\left[\widehat{\beta_0}, \widehat{\beta_1}\right]$$

However, in the simple linear model, the expressions of the components $\sigma^2_{\widehat{\beta_0}}$, $\sigma^2_{\widehat{\beta_1}}$ and $Cov\left[\widehat{\beta_0}, \widehat{\beta_1}\right]$ are known and expressed (Wasserman, 2004):

$$\sigma^2_{\widehat{\beta_1}} = \frac{\sigma^2}{\sum_{i\in N}(X_i - \overline{X_N})^2}$$

$$\sigma^2_{\widehat{\beta_0}} = \sigma^2 \times \left(\frac{1}{n} + \frac{\overline{X_N}^2}{\sum_{i\in N}(X_i - \overline{X_N})^2}\right)$$

$$Cov\left[\widehat{\beta_0}, \widehat{\beta_1}\right] = -\frac{\overline{X_N}}{\sum_{i\in N}(X_i - \overline{X_N})^2} \times \sigma^2$$

By replacing these components by their expressions in the expression of $\widehat{\bar{Y}_2}$ variance:

$$Var\left[\widehat{\bar{Y}_2}\right] = \sigma^2 \times \left(\frac{1}{n} + \frac{\overline{X_N}^2}{\sum_{i\in N}(X_i - \overline{X_N})^2}\right) + \frac{\sigma^2}{\sum_{i\in N}(X_i - \overline{X_N})^2}\bar{X}^2 - \frac{\overline{X_N}}{\sum_{i\in N}(X_i - \overline{X_N})^2} \times \sigma^2 2\bar{X}$$

$$Var\left[\widehat{\bar{Y}_2}\right] = \frac{\sigma^2}{n} + \frac{\overline{X_N}^2}{\sum_{i\in N}(X_i - \overline{X_N})^2}\sigma^2 + \frac{\sigma^2}{\sum_{i\in N}(X_i - \overline{X_N})^2}\bar{X}^2 - \frac{\overline{X_N}}{\sum_{i\in N}(X_i - \overline{X_N})^2} \times \sigma^2 2\bar{X}$$

It is then factorized by $\frac{\sigma^2}{\sum_{i\in N}(X_i - \overline{X_N})^2}$:

$$Var\left[\widehat{\bar{Y}_2}\right] = \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum_{i\in N}(X_i - \overline{X_N})^2}\left(\overline{X_N}^2 + \bar{X}^2 - 2\bar{X} \times \overline{X_N}\right)$$

$$\boldsymbol{Var\left[\widehat{\bar{Y}_2}\right] = \frac{\sigma^2}{n} + \frac{(\overline{X_N} - \bar{X})^2}{\sum_{i\in N}(X_i - \overline{X_N})^2} \times \sigma^2} \qquad (8)$$

As for $\widehat{\bar{Y}_1}$, the $\widehat{\bar{Y}_2}$ variance depends on σ, the residual variance of the linear model and n, the number of sampling sites. It also depends on the sampled sites and their auxiliary data. To reduce its variance: the sample must i) be representative, so its mean ($\overline{X_N}$) is close to the field mean ($\bar{X}$) and ii) present a strong dispersion so the samples sites are far apart from their own mean $\sum_{i\in N}(X_i - \overline{X_N})^2$. This is due to the fact that a dispersed sample will allow a better estimation of the model parameters.

Root mean squared error

By looking at Equations 4 and 7, $\hat{\bar{Y}}_2$ appears biased while $\hat{\bar{Y}}_2$ is unbiased. But at the same time, $\hat{\bar{Y}}_2$ has higher variance than $\hat{\bar{Y}}_1$ (Equations 5 and 8).To be able to compare them, theoretical Mean Square Error (MSE) and Root Mean Square Error (RMSE) are computed for both estimators from bias and variance as stated in Equation 9:

$$MSE\left(\hat{\bar{Y}}\right) = Bias\left(\hat{\bar{Y}}\right)^2 + Variance\left(\hat{\bar{Y}}\right) \tag{9}$$

$$RMSE\left(\hat{\bar{Y}}\right) = \sqrt{MSE\left(\hat{\bar{Y}}\right)} \tag{10}$$

Based on Equation 10, RMSE of $\hat{\bar{Y}}_1$ and $\hat{\bar{Y}}_2$ are given is given respectively in Equation 11 and Equation 12:

$$RMSE\left(\hat{\bar{Y}}_1\right) = \sqrt{\frac{\sigma^2}{n} + {\beta_1}^2 \times (\overline{X_{\iota \in N}} - \bar{X})^2} \tag{11}$$

$$RMSE\left(\hat{\bar{Y}}_2\right) = \sqrt{\frac{\sigma^2}{n} + \frac{(\overline{X_N} - \bar{X})^2}{\sum_{i \in N}(X_i - \overline{X_N})^2} \times \sigma^2} \tag{12}$$

Derive the expression of $RMSE\left(\hat{\bar{Y}}_2\right) < RMSE\left(\hat{\bar{Y}}_1\right)$ allows to identify which condition estimator $\hat{\bar{Y}}_2$ is provides lower RMSE than $\hat{\bar{Y}}_1$ and thus when estimation by the linear model should be preferred.

$$RMSE\left(\hat{\bar{Y}}_2\right) < RMSE\left(\hat{\bar{Y}}_1\right) \Rightarrow MSE\left(\hat{\bar{Y}}_2\right) < MSE\left(\hat{\bar{Y}}_1\right)$$

$$MSE\left(\hat{\bar{Y}}_2\right) < MSE\left(\hat{\bar{Y}}_1\right) \Rightarrow \left[\frac{\sigma^2}{n} + \frac{(\overline{X_N} - \bar{X})^2}{\sum_{i \in N}(X_i - \overline{X_N})^2} \times \sigma^2\right] < \left[\frac{\sigma^2}{n} + {\beta_1}^2 \times (\overline{X_{\iota \in N}} - \bar{X})^2\right]$$

$$\Rightarrow \frac{(\overline{X_N} - \bar{X})^2}{\sum_{i \in N}(X_i - \overline{X_N})^2} \times \sigma^2 - \beta_1^2 \times (\overline{X_{\iota \in N}} - \bar{X})^2 < 0$$

$$\Rightarrow (\overline{X_N} - \bar{X})^2 \times \left[\frac{\sigma^2}{\sum_{i \in N}(X_i - \overline{X_N})^2} - {\beta_1}^2\right] < 0$$

From these expressions, $RMSE\left(\hat{\bar{Y}}_2\right) < RMSE\left(\hat{\bar{Y}}_1\right)$ if and only if :

$$\boldsymbol{\frac{\sigma^2}{{\beta_1}^2} < \sum_{i \in N}(X_i - \overline{X_N})^2} \tag{13}$$

Following this result, the choice of the inference strategy must be done according to i) σ^2, the residual variance of the linear model; ii) β_1, the regression slope; iii) $\sum_{i \in N}(X_i - \overline{X_N})^2$, the dispersion of the sample in terms of auxiliary data. The $\frac{{\beta_1}^2}{\sigma^2}$ ratio reflects the existing correlation between the estimated and the auxiliary data. If the regression slope is small in front of the residuals variance, the correlation is low and it favours the use of a mean for estimator. Conversely, if the slope is strong in front of σ and if the sample values are dispersed (for a better estimate of the model parameters), this favours the use of an estimator based on the linear model.

References:

Araya-Alman, M., Leroux, C., Acevedo-Opazo, C., Guillaume, S., Valdés-Gómez, H., Verdugo-Vásquez, N., Pañitrur-De la Fuente, C., Tisseyre, B. (2019). A new localized sampling method to improve grape yield estimation of the current season using yield historical data. *Precision Agriculture*.

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

Oger, B., Roux, S., Le Moguédec, G., Tisseyre, B., (2020). Model and auxiliary data for an accurate estimate of the field mean. 13th European Conference on Precision Agriculture (ECPA 2021).

Wasserman, L. (2004). All of statistics: A concise course in statistical inference. *New York: Springer.*