iVES OENO One vine & wine

DOI: https://doi.org/10.20870/oeno-one.2023.57.3.7404

OPEN ACCESS

ORIGINAL RESEARCH ARTICLE

How to better estimate bunch number at vineyard level?

Baptiste Oger[1]*, Cécile Laurent[2], Philippe Vismara[3] and Bruno Tisseyre[1]

[1] ITAP, Univ. Montpellier, L'institut Agro Montpellier, INRAE, France
[2] Fruition Sciences, MIBI, 672 Rue du Mas de Verchant, 34000 Montpellier, France
[3] MISTEA, Univ. Montpellier, L'institut Agro Montpellier, INRAE, France

*correspondence:
baptiste.oger@supagro.fr

Associate editor:
Franco Meggio

Received:
17 February 2023

Accepted:
15 May 2023

Published:
17 July 2023

This article is published under the **Creative Commons licence** (CC BY 4.0).

Use of all or part of the content of this article must mention the authors, the year of publication, the title, the name of the journal, the volume, the pages and the DOI in compliance with the information given above.

ABSTRACT

Despite the extensive use of sampling to estimate the average number of grape bunches per vine, there is no clearly established sampling protocol that can be used as a reference when performing these estimations. Each practitioner therefore has their own sampling protocol. This study characterised the effect of differences between sampling protocols in terms of estimation errors. The goal was to identify the most efficient practices that will improve the early estimation of an important yield component: average bunch number. First, the appropriateness of including non-productive vines (i.e., dead and missing vines) in the sampling protocol was tested; the objective was to determine whether it is relevant to estimate two yield components simultaneously. Second, sampling protocols with sampling sites of varying size were compared to determine how the spatial distribution of observations and potential spatial autocorrelation affect estimation error. Third, a new confidence interval for estimation error was determined to express expected error as a percentage. It aimed at designing a new tool for finding the best sample size in an operational context. Tests were performed on two vineyards in the South of France, in which the number of bunches per vine had been exhaustively determined on all the plants before flowering. The results show that the simultaneous estimation of number of bunches and proportion of dead and missing vines increased the estimation errors by a factor of 2. Despite the low spatial autocorrelation of bunch number, the results show that the observation must be spread across at least 2 or 3 sampling sites to reduce estimation errors. Finally, the confidence intervals expressed as a percentage were validated and used to define an adequate sample size based on a compromise between the expected precision and the variability observed in the first measurements.

KEYWORDS: yield, sampling, cluster, missing vines, estimation error, confidence interval

INTRODUCTION

In viticulture, estimating yield at the vineyard scale early in the season is important for the planning of vineyard operations, investment and even marketing and commercialisation strategies (Laurent *et al*., 2021). Estimations are needed before veraison, when yield components are still developing and the berry mass is yet to be determined. Currently, estimations are mainly based on the observation of the number of bunches per vine, which is one of the first observable yield components. The number of bunches also often explains most of the mean vineyard yield variability (30 % to 70 % of total yield variability) compared to other yield components (Carrillo *et al*., 2016; Clingeleffer *et al*., 2001).

As it is not possible to manually count all the bunches present in a given vineyard, winegrowers follow sampling protocols to make estimations. New technologies based on embedded cameras and image recognition algorithms have been proposed in the literature to observe all the bunches in a vineyard (Millan *et al*., 2018; Nuske *et al*., 2011; Victorino *et al*., 2020). These technologies could potentially be used to estimate the mean number of bunches per vine or linear meter, as well as other yield components, such as the bunch volume or the number of berries per bunch. However, these methods are still under development and are currently seldom used in real commercial conditions. As a result, sampling remains by far the most common method for yield estimation in viticulture. To our knowledge, despite the extensive use of sampling for bunch estimation, no clear established sampling protocol is used as a reference among professionals. The literature in this area remains sparse and often limited to the application of classical statistics (Wolpert and Vilas, 1992; Clingeleffer *et al*., 2001). As a result, the wine industry is known to mostly apply a random bunch sampling approach. Sampling protocols differ greatly from one practitioner to another; in particular, large differences can be observed in terms of: *i)* total number of vines sampled per vineyard, ii) the arrangement of the sampled vines, which can be grouped within sampling sites of varying sizes (a varying number of consecutive vines sampled together along a row), and *iii)* the counting protocol, which can include or omit the missing vines. This diversity of practices raises a number of issues related to the design of sampling method using a yield component such as the number of bunches per vine.

Determining the number of vines to be sampled is directly related to variance in the vineyard and the expected precision of the estimate. The use of mean bunch number has been widely addressed by Wolpert and Vilas (1992) in the context of classical statistics which assume observations to be independent. While estimate precision is determined by user-defined operational constraints, yield variability in a vineyard can differ greatly from one vineyard to another (Taylor *et al*., 2005); moreover, it is generally not known before sampling is implemented. Therefore, it is necessary to find a way of defining the number of vines that need to be sampled to obtain the expected precision of yield estimation.

To our knowledge, sampling site size has not been thoroughly investigated in relation to mean bunch number estimation. In the scientific literature, sampling protocols are most often based on individual vines (Roessler and Amerine, 1958; Wulfsohn *et al*., 2012) or on sampling sites of 4 or 5 vines (Carillo *et al*., 2016; Araya-Alman *et al*., 2019). However, there is no objective reason for such designs. The size of the sampling sites and their spatial location raises the issue of the stochastic variance (i.e., bunch number variability from one vine to another) and the spatial autocorrelation of neighbouring vines. Increasing the size of a sampling site (i.e., its spatial footprint) to generate mean number of bunches per vine for several vines tends to reduce the incidence of stochastic variance, which is an advantage when analysing yield components in relation to other parameters (Carillo *et al*., 2016; Bramley, 2001). However, the yield components of the vines included in in these larger sampling sites can be more or less autocorrelated depending on the spatial arrangement of the vineyard yield components. Although spatial autocorrelation is usually low and largely determined by pruning operations (Taylor and Bates, 2013), the extent to which it affects the estimate accuracy is still unclear.

When missing vines are incorporated in the counting protocol (with number of bunches equal to 0), it follows that a second yield component is included in the estimate: number of missing vines per vineyard. From a practical point of view, it could be useful to estimate both of the yield components (i.e., number of bunches per vine and the number of unproductive or missing vines per vineyard) in a single survey. However, the number of bunches per vine and the number of unproductive or missing vines per vineyard are two different variables: one is continuous, while the other is categorical, and they can be independent of each other and have differing distributions within the same vineyard; this can result in the same sampling protocol giving rise to different estimation accuracies for each component. To the best of our knowledge, scientific studies on yield estimation have not investigated the effect of including or omitting missing vines in the counting protocol on the accuracy of the estimation of number of bunches per vine. In the absence of rigorous studies on this subject, the industry lacks information to be able to adapt its counting protocols in order to improve the accuracy of yield estimation.

In the light of the issues surrounding the fact that the wine industry follows different protocols for estimating number of grape bunches, the aim of this study was to investigate issues relating to the early estimation of mean bunch number at the vineyard level. These are related to i) the inclusion or omission of missing vines in the sampling protocol and the associated impact on estimation accuracy, ii) the impact of the number and size of the sampling sites on the estimate accuracy, and finally, iii) proposing an original approach that allows the optimal sample size to be defined as a trade-off

between the objectives in terms of estimation accuracy and the time available.

MATERIALS AND METHODS

1. Notations

For a given vineyard, m refers to the mean number of bunches per vine and σ to the standard deviation of the number of bunches per vine.

The objective was to sample a vineyard to obtain an estimation of m. The sampling site was defined as a set of consecutive vines in the same row. The size of a sampling site corresponded to the number of trunks/vines within it (Figure 1). The sample was the set of all the vines formed by selecting one or more sampling sites.

Any sample can hence be described by all of the following:

- the number of sampling sites, k
- the size of the sampling sites, s
- the size of the sample, noted n, which is equal to the number of vines within the sample $n = k \times s$
- The mean number of bunches per vine over the n vines of the sample, $\overline{X}$, which is used to estimate m
- the standard deviation of the number of bunches per vine over the n vines of the sample, σ_{Sample}

The coefficient of variation (CV) (Eq. 1) derived from the last two parameters represents the dispersion of the sample expressed as a percentage:

$$CV = \frac{\sigma_{Sample}}{\overline{X}}$$

The estimation error associated with a sample is calculated afterwards by comparing its mean to the actual vineyard mean . A relative error (%) is computed as described in Eq. 2:

$$Error(\%) = \frac{|\overline{X} - m|}{m}$$

The estimated mean number of bunches per vine is often reported by wine growers as a number of bunches per hectare based on the number of vines within the vineyard (Eq. 3).

$$\frac{bunches}{hectare} = \frac{plantedvines}{hectare} \times (1 - \%dead \wedge missing\ vines) \times \frac{bunches}{vine}$$

Usually, the number of bunches per vine simply corresponds to the mean number of bunches observed on productive vines. In this case, the mean number of bunches per vine in a vineyard m is noted as m_1 (Eq. 4). Since the proportion of missing and dead vines is not always known when estimating the number of bunches per vine, it can seem easier to estimate $[(1 - \%dead \wedge missing\ vines) \times \frac{bunches}{vine}]$ at once by counting the mean number of bunches per planted vine. In this case, dead and missing vines are included and counted as vines with 0 bunches. Therefore two yield components are estimated simultaneously: number of bunches per productive vine $[\frac{bunches}{vine}]$ and proportion of dead and missing vines $[\%dead \wedge missing\ vines]$. In this case, m is noted as m_2 and corresponds to the mean number of bunches per planted vine (Eq. 5).

$$m_1 = \frac{bunches}{vine}$$

$$m_2 = (1 - \%dead \wedge missing\ vines) \times \frac{bunches}{vine}$$

2. Sampling protocols

Two sampling protocols were rigorously studied to determine the accuracy of the resulting estimations: Complete Random Sampling and Productive Random Sampling (Figure 1).

Complete Random Sampling (CRS) comprises random sampling of all the vines planted in the vineyard. Dead, missing and productive vines can have the same probability of being sampled. When a missing vine is selected it is counted as a plant with 0 bunches. CRS provides an estimate of m_2.

In Productive Random Sampling (PRS), sampling sites are selected randomly with the same probability of being sampled. Only productive vines are considered, and dead and missing vines are ignored. When a sampling site is selected, all of its vines are taken into account in the sample.

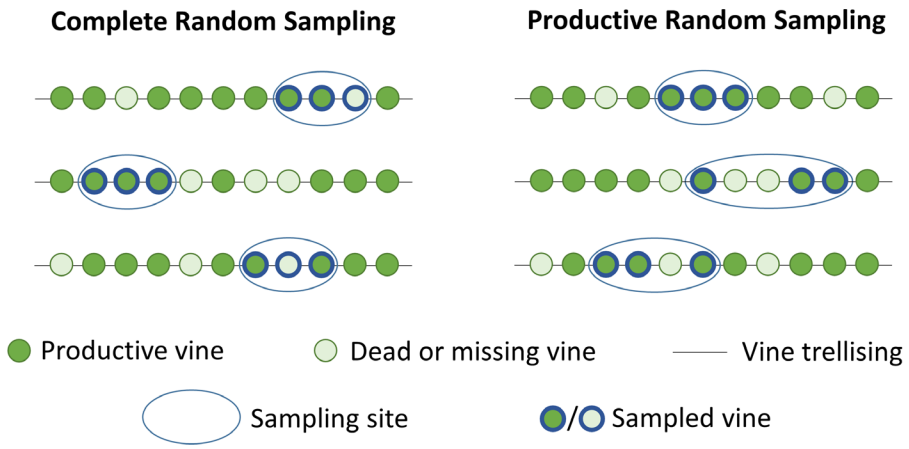


FIGURE 1. Representation of the two proposed random sampling protocols: complete and productive, with k = 3 sampling sites of sizes = 3 for a total of n = 9 vines as an example.

Every other sampling site that has a vine in common with selected sites is excluded for the rest of the sampling protocol to ensure that a vine never appears twice in a sample. PRS provides an estimate of m_1. An estimate of m_2 can be deduced from m_1 when an estimate of the proportion of dead and missing vines ($\%dead \wedge missing\ vines$) is available (Eq. 4 & Eq. 5).

3. Expected estimation errors

This section describes the method used to obtain information about the accuracy of the estimation derived from a given sample. To derive information about the expected estimation errors associated with a given sample, the following classical assumptions are made: i) the number of bunches per vine in the vineyard has a normal distribution $N(m, \sigma)$, and ii) the selected vines are independent of each other.

Using frequentist (i.e., classical) statistics (Smithson, 2000), it is possible to compute a confidence interval for m or $(\overline{X} - m)$ based on sample properties; this confidence interval is only expressed as number of bunches per vine. However, to address the operational issues of yield estimation, it is more appropriate to express the errors as a percentage (Eq. 2) in order for the estimation error incidence on total yield to be better understood. Since intervals expressed as a percentage are difficult to obtain using frequentist statistics, Bayesian statistics were used to compute the confidence interval of relative error.

A probability distribution for m and σ was computed from the observations. As they represent the parameters of a normal distribution, a normal-inverse-gamma (NIG) distribution was chosen to represent m and σ. To ensure that the approach would be applicable to all the vineyards, a weak Bayesian prior was chosen so that the posterior probability distribution would only depend on the sample properties (O'Hagan, 2010):

$(m, \sigma)NIG$

With the Bayesian a posteriori () parameters:

$m * \overline{X}v * \frac{1}{n}$

$a * \frac{n-1}{2}$

$b * \frac{(n-1) \times \sigma_{Sample}}{2}$

For each sample, 10,000 possible values of m were obtained from its normal-inverse-gamma distribution (Eq. 6). From each given value of m a value of $\frac{|\overline{X} - m|}{m}$ was calculated. This set of values was used to build an empiric distribution of relative error (Eq. 2).

For representation purposes, this density was converted into a credible interval (O'Hagan, 2010). Credible intervals are the Bayesian equivalent of confidence intervals in frequentist statistics. In this case, it can be interpreted as a confidence interval (Hespanhol *et al.*, 2019). For simplicity, credible intervals will hereafter be referred to as confidence intervals.

The confidence interval with the desired confidence level was derived from the distribution of $\frac{|\overline{X} - m|}{m}$ computed with Eq. 6 and Eq. 2 using percentiles. For example, the 90 % confidence interval corresponds to the 90th percentile of the observed distribution of 10 000 $\frac{|\overline{X} - m|}{m}$ values (Figure 2). This confidence interval is only computed from the properties of a given sample: $\overline{X}$, σ_{Sample} and n.

A validation process was carried out to ensure that the confidence interval was correct; this comprised checking if the effective estimation error obtained with the real m value was within the confidence interval. For a large number of samples and confidence intervals (10,000), the proportion of cases in which the estimation error was included in the confidence interval was computed. If the assumptions are correct, this proportion will be equal to the confidence level of this interval. The different steps used to compute and validate the error confidence interval of a sample are summarised in Figure 3.

4. Dataset

Experimental dataset

The experimental dataset is composed of two vineyards located in the Occitanie region in the South of France (Vineyard 1: 43.547417, 3.8414769; Vineyard 2: 43.144570, 3.131338, WGS84).

Both vineyards belong to a commercial vineyard and are rain-fed and grown under a Mediterranean climate. They both have an inter-row distance of 2.5 m and a between-vines distance of 1.2 m (Vineyard 1) and 1m (Vineyard 2). All of the bunches on each vine in each of the two vineyards were counted manually just before flowering (Figure 4). Missing or

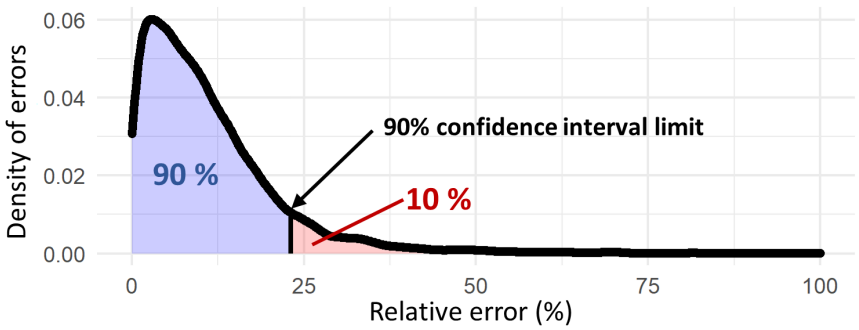


FIGURE 2. The confidence interval limit is deduced from the density of the errors computed from the normal-inverse-gamma distribution.

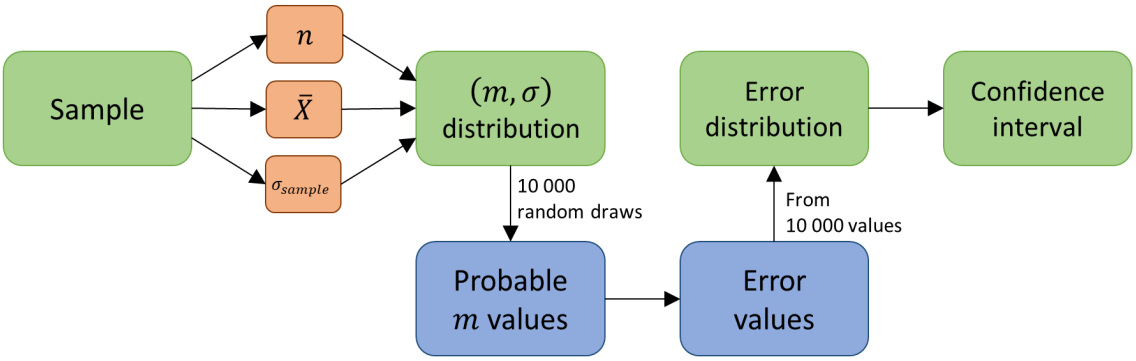


FIGURE 3. Workflow of the computation of the error confidence interval of a sample.

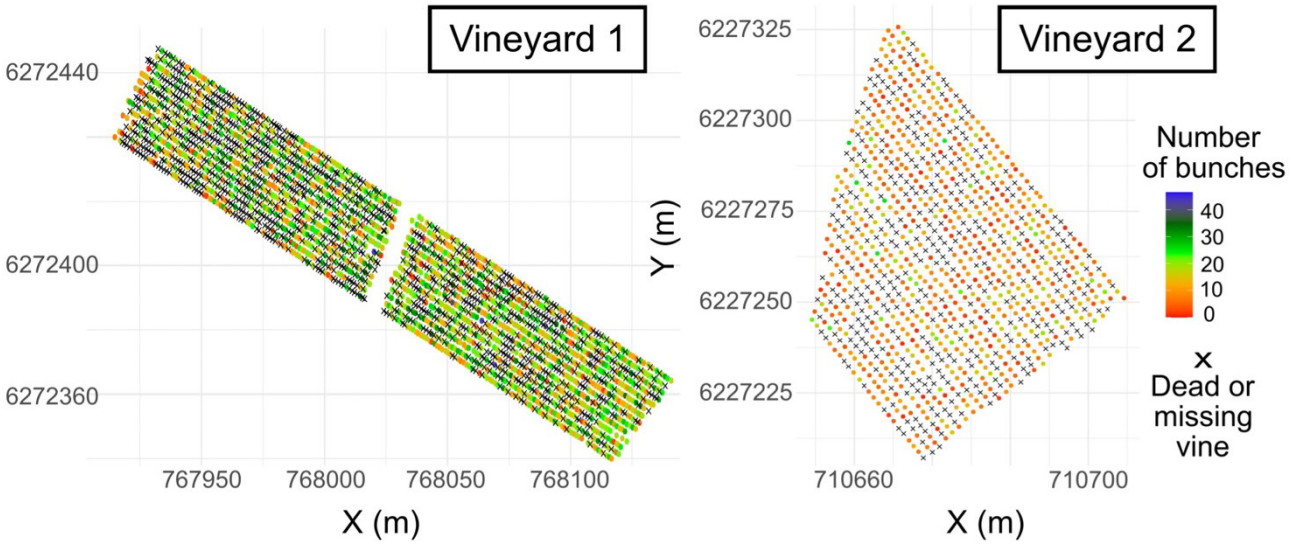


FIGURE 4. Maps of the number of bunches per vine for the two experimental vineyards. Coordinates are in Lambert 93.

dead vines were also counted and georeferenced at the same time. Counting was carried out in May 2022 in Vineyard 1, and in May 2014 in Vineyard 2. The coordinates of the vines in Vineyard 1 were acquired using a RTK GNSS (Real Time Kinematic Global Positionning Satellite System) receiver (Ancelin *et al*., 2022), with a centimeter correction giving a positioning accuracy of +/- 10 mm. The RTK GNSS rover receiver was connected to a smartphone to record and share observations through the Mergin Maps application (LTD, 2022). In Vineyard 2, the coordinates were retrieved from the coordinates of the row edges and the vine number. The characteristics of both vineyards in terms of bunch number and missing vines are summarised in Table 1.

Simulated dataset

The independence or non-independence of the observations that constitute a sample is known to affect the estimation error (Smithson, 2000). In viticulture, this independence can be affected by the spatial structure of the vineyards:vines that are near each other tend to show very similar number of bunch (spatial autocorrelation). In precision viticulture, yield has been found to be spatially structured; i.e., grape yield showed high to moderate spatial autocorrelation (Taylor *et al*., 2005). To account for different levels of spatial autocorrelation in the number of bunches, the experimental dataset (Vineyard 1 and Vineyard 2) was completed with new simulated vineyards to test the different sampling protocols. These datasets were used to determine the effect of spatial autocorrelation, ensuring that all other vineyard parameters remained constant.

Four 100 m × 100 m (1 ha) vineyards were simulated (Figure 5). The inter-row distance was set at 2.5 m and the within-row inter-vine distance at 1 m, resulting in a plant density of 4,000 vines/ha. In order to be consistent with a real situation, the vineyard simulations were based on the characteristics of Vineyard 1 , with a mean bunch number per vine of 18.5 and a standard deviation of 7.3. The objective

TABLE 1. General information about vineyard properties, bunch number and missing vines.

Vineyard properties					Bunch per productive vine statistics		Bunch per planted vine (including dead and missing vines) statistics	
Vineyard ID	Variety (Rootstock)	Area (ha)	Number of productive vines	Number of missing (or dead) vines	Average value m_1	Standard deviation	Average value m_2	Standard deviation
Vineyard 1	Syrah (Sélection Oppenheim n°4)	0.8	1474	1096	18.5	7.3	10.6	10.7
Vineyard 2	Syrah (Ruggeri 140)	0.5	675	355	8.9	5.2	5.8	6.0

was to generate four levels of spatial auto-correlation (0 %, 10 %, 20 % and 30 %) with different semi-variogram sills and nugget effects. The range was set at 25 m. As proposed by Oger *et al*. (2021), bunch numbers per vine were generated using a two-step approach. First, Gaussian vineyards with no nugget effect were simulated using the "gstat" package (Gräler *et al*., 2016). Their sill was respectively set at i) 0 %; ii) 10 %; iii) 20 % and iv) 30 % of the total variance (for the four levels of spatial autocorrelation). These Gaussian fields represented the spatialised part of the simulated vineyards. Second, to obtain the unstructured variability of the simulated vineyards, a random nugget effect was then added to each Gaussian field. The nugget effects were added by using a simple centred normal distribution of variance: i) 100 %; ii) 90 %; iii) 80 % and iv) 70 % of the total variance (σ). The four final vineyards had equivalent variances (and sills), but showed differing nugget effects.

Simulations and analyses were performed using the open source statistical software R (R Core Team, 2022).

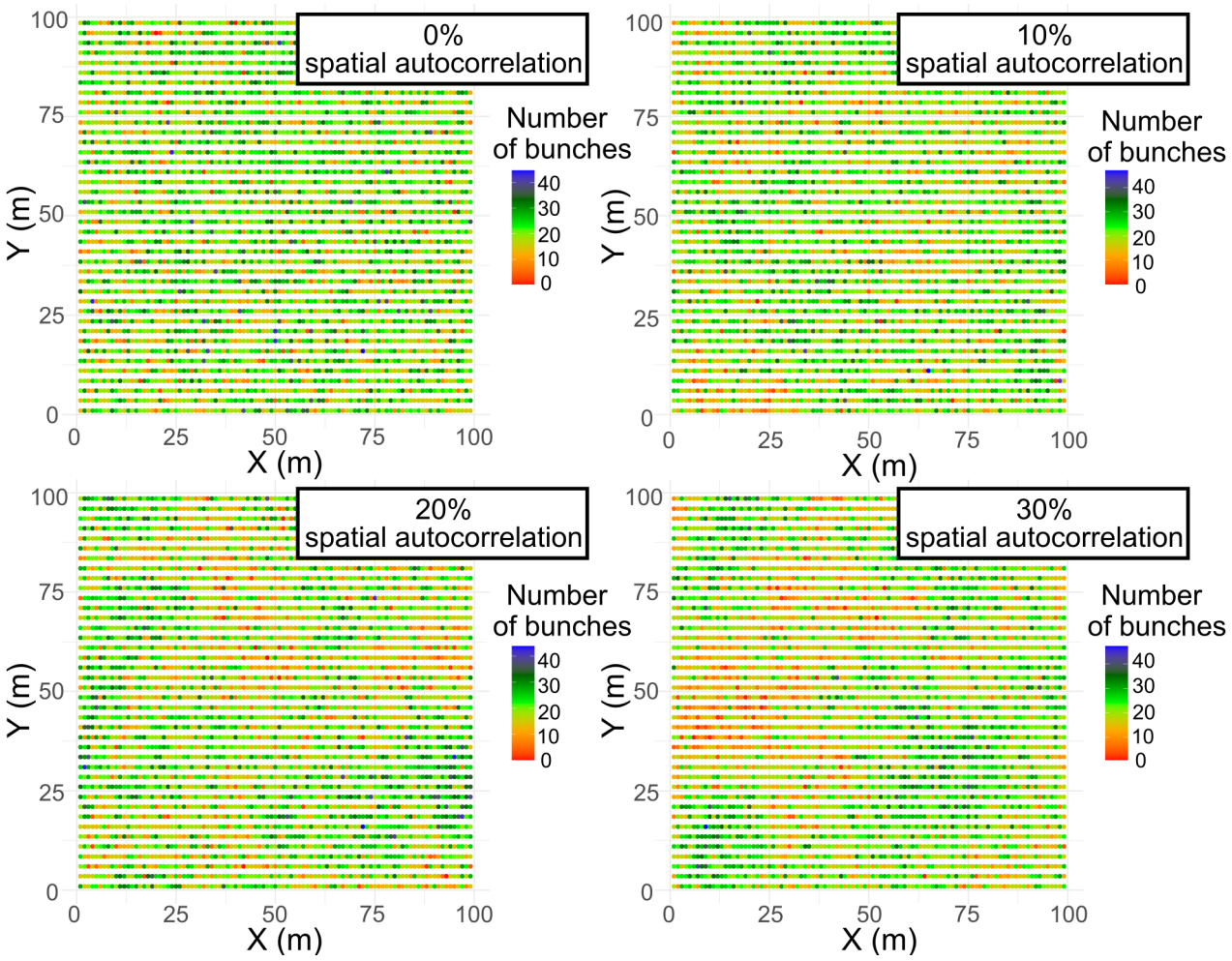


FIGURE 5. Representation of number of bunches per vine for the four simulated vineyards with spatial autocorrelation varying from 0 % to 30 %.

RESULTS

1. Should dead and missing vines be included in bunch sampling?

This first part of the study focused on the estimation of two yield components: the proportion of dead and missing vines and the number of bunches per vine. As the proportion of dead and missing vines affects the number of bunches per vineyard, this part aimed to identify whether it was appropriate to sample both yield components simultaneously (CRS protocol) or not (PRS protocol).

Figure 6 shows the bunch number estimation errors obtained for both of the real study vineyards after complete random sampling (CRS) and productive random sampling (PRS). Number of sampling sites, k, ranged from 1 to 15. Sampling site size was $s = 1$, thus the sample size is $n = k$. The red curves represent the estimation error compared to m_2 when dead and missing vines were included and counted as vines with zero bunches when sampled (CRS). The continuous blue curves represent the estimation error compared to m_2 when dead and missing vines are known with no error (PRS 0 %) while dashed lines represent errors observed with PRS when an error of 15 %, 30 % and 45 % is considered on the dead and missing vine estimation. The values were derived from 10,000 samples for each sample size. For both vineyards, the error logically decreased as the sample size increased.

For each vineyard, the mean errors for CRS were double those obtained for PRS 0 % (no error in the estimation of dead and missing vines): for a vine sample size of $n = 5$, the observed error was 37 % and 18 % for CRS and PRS respectively. The difference in mean estimation error between the two sampling protocols remained the same (i.e., CRS values twice as high as PRS values) with increasing sample size. Such differences in estimation error are to be expected, since with CRS two yield components were estimated simultaneously (number of bunches per vine and proportion of missing vines), while proportion of missing vines was known in PRS 0 %. When the estimation errors for dead and missing vines were added to PSR (blue dashed curves), the estimation errors were logically higher, but they were mostly always lower than those observed in CRS. The accuracy of CRS was only higher when the percentage of missing vines was high (Vineyard 1), the error of the missing vine estimate was very large (30 % or 45 %) and the sample size was big. To detail this phenomenon, Figure 7 shows the estimation error of dead and missing vines obtained in CRS.

Regarding the estimation of the proportion of dead and missing vines in the vineyard, Figure 7 shows that 15 sampling sites of size $s = 1$ were necessary to reach a mean error of between 25 and 30 %. This error can increase to 40 % or 50 % with sampling variability (coloured area). The same amount of error was obtained with fewer observations when sampling for number of bunches per vine. Estimating the proportion of dead and missing vines require larger sample sizes to be relevant. Therefore, it may be counterproductive to estimate both yield components simultaneously, since the sample size would need to be increased at a time when vineyard workload is already heavy. This result was of course dependent on the examples considered here, as well as the proportion of dead or missing plants; this point will be discussed later in the article.

For the estimation of bunch number per vine, the number of dead and missing vines is considered as known (i.e., estimated using another appropriate method) in the rest of the article. Therefore, the following results only focus on number of bunches on productive vines (PRS), and the

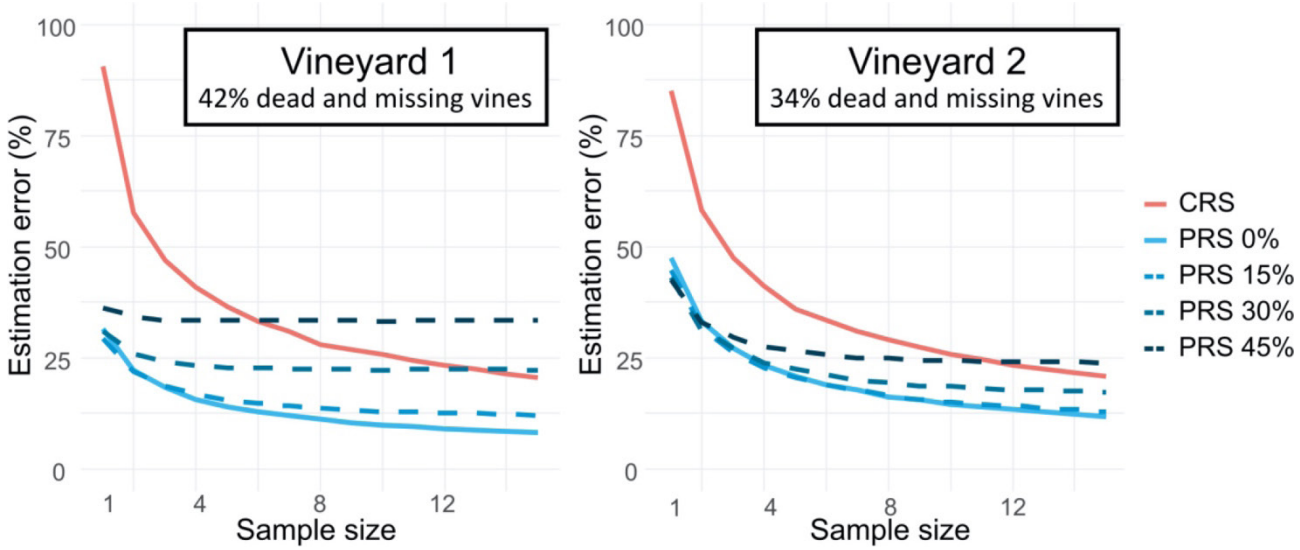


FIGURE 6. Estimation errors (%) related to the number of bunches per vine with complete random sampling (CRS) and productive random sampling (PRS) in each of the two study vineyards. For PRS, four scenarios, corresponding to different estimation errors of the proportion of missing and dead vines per vineyard (0 %, 15 %, 30 % and 45 %), were compared.

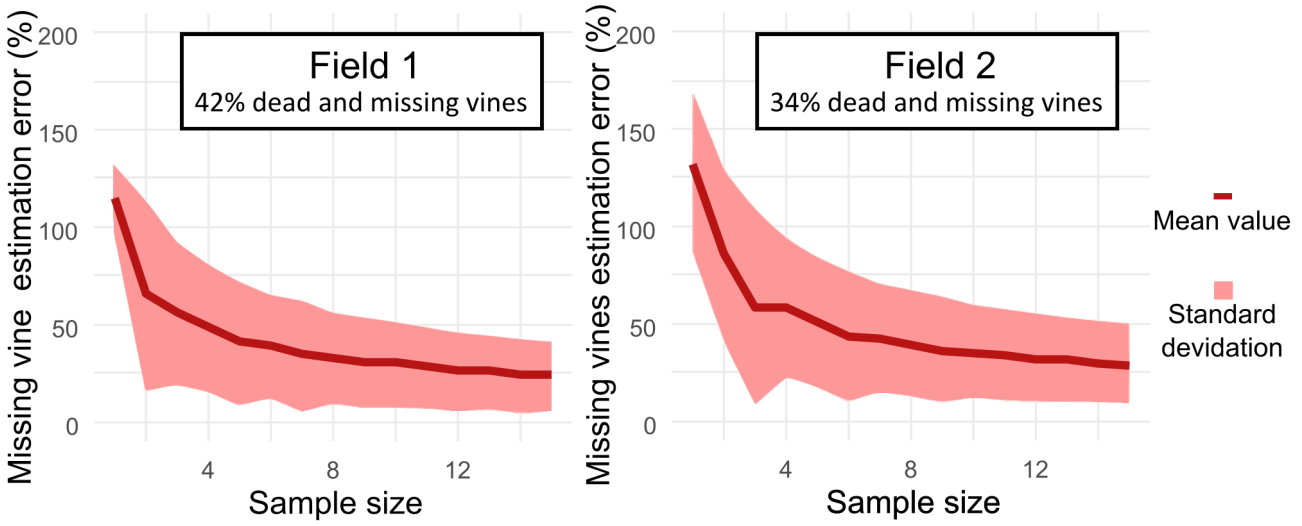


FIGURE 7. Dead and missing vines estimation errors when performed with complete random sampling (simultaneously with the estimation of number of bunches per vine).

estimation errors were computed from m_1, the mean number of bunches per vine.

2. How to select sampling sites to count bunches?

Depending on the vineyard, sampling protocols can include varying number of sampling sites (k) and varying sizes of sampling sites (s). The objective here was to characterise and determine how the spatial structure of a fixed number of observations within sampling sites of arbitrary size can affect the estimation.

Figure 8 represents the estimation errors in the simulated vineyards of increasing spatial autocorrelation (0 %, 10 %, 20 % and 30 %).

For the simulated vineyard with no spatial autocorrelation (Figure 8, top left), the estimation error was always constant regardless of sampling site size. In this case, the median error was 7 %, with a first quartile at 4 % and a third quartile at 11 %. The estimation errors increased with increasing spatial autocorrelation when the sampled vines were grouped in a reduced number of sampling sites. In the most extreme cases, when 12 sampled vines were grouped within a single sampling site, the median error increased from 7 % for the vineyard with no spatial autocorrelation to 13 % for the simulated vineyard with 30 % spatial autocorrelation (Figure 8, bottom right).

Regarding the two real vineyards used in the study, the changes in estimation errors with an increasing number of sampling sites were very similar to those observed in the simulated vineyards (Figure 9). The sampling process was the same: 10,000 samples comprising $n = 12$ vines with varying sampling sites of $s = 1,2,3,4,6 \vee 12$ consecutive vines. However, while both vineyards had a small level of spatial autocorrelation of the number of bunches, an increase in estimation errors was observed for larger sampling sites. This trend was very slight in Vineyard 1, with a median error that only increased from 7 % to 9 %; it was more noticeable in Vineyard 2, with the median error increasing from 12 % to 18 %. Vineyard 1 had a lower spatial autocorrelation (3.3 % of the vineyard variance), and was therefore more similar to the simulated vineyard with 0 % spatial autocorrelation (Figure 8, top left), which explains why the errors were almost constant regardless of the different designs of the sampling sites. It should be noted, however, that a single large sampling site with 12 vines was not optimal and led to 2 % additional error compared to other sampling designs. Vineyard 2 had a higher spatial autocorrelation (9.6 %) and was similar to the simulated vineyard with 10 % spatial autocorrelation (Figure 8, top right). It showed the same trend in error estimation from 12 sampling sites to 1 large and unique sampling site.

3. How many vines should be sampled?

Based on the observed parameters (n, $\overline{X}$ and σ_{Sample}) of a sample this part of the study aimed to validate the possibility of refining the sampling strategy during the estimation process to reach a desired error of estimation. The Bayesian formalism based on the normal inverse gamma law described in Eq. 6 was used to compute the confidence interval of the relative error that were associated with samples of different sizes (n). Table 2 shows the proportion of samples of 10,000 random samples with $s = 1$ whose error fell within the 50 % (blue) and 90 % (red) confidence intervals derived from the use of the Bayesian approach on both real vineyards.

For both vineyards, between 49.88 % and 52.70 % of the observed estimation errors were within the 50 % confidence interval, and between 89.97 % and 94.57 % of the estimation errors lay within the 90 % interval. The confidence intervals for small sample sizes ($n = 3$) had a slight tendency to overestimate the variability of the errors, as these intervals contained slightly more than 50 % or 90 % of the estimation errors. Overall, the estimation errors correctly followed the computed confidence intervals. Table 2 validates the

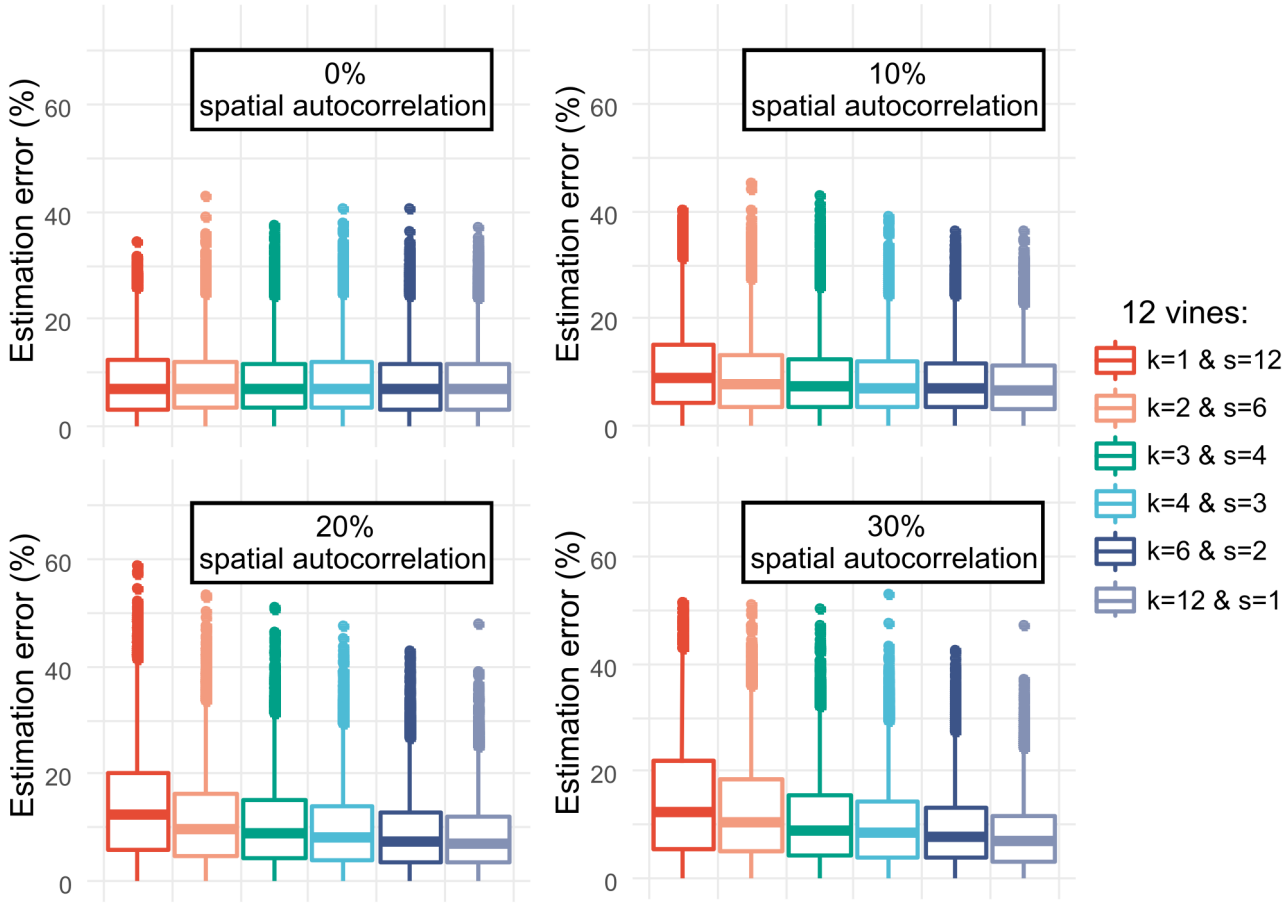


FIGURE 8. Estimation errors of the number of bunches per vine in simulated vineyards with different levels of spatial autocorrelation and different sampling site sizes. Boxplots with median, first and third quartile of the estimation error of the bunch number per vine using 10,000 samples comprising $n = 12$ vines. The estimation was based on productive random sampling and sampling sites of varying sizes: from $s = 12$ (red) to $s = 1$ (grey) consecutive vines.

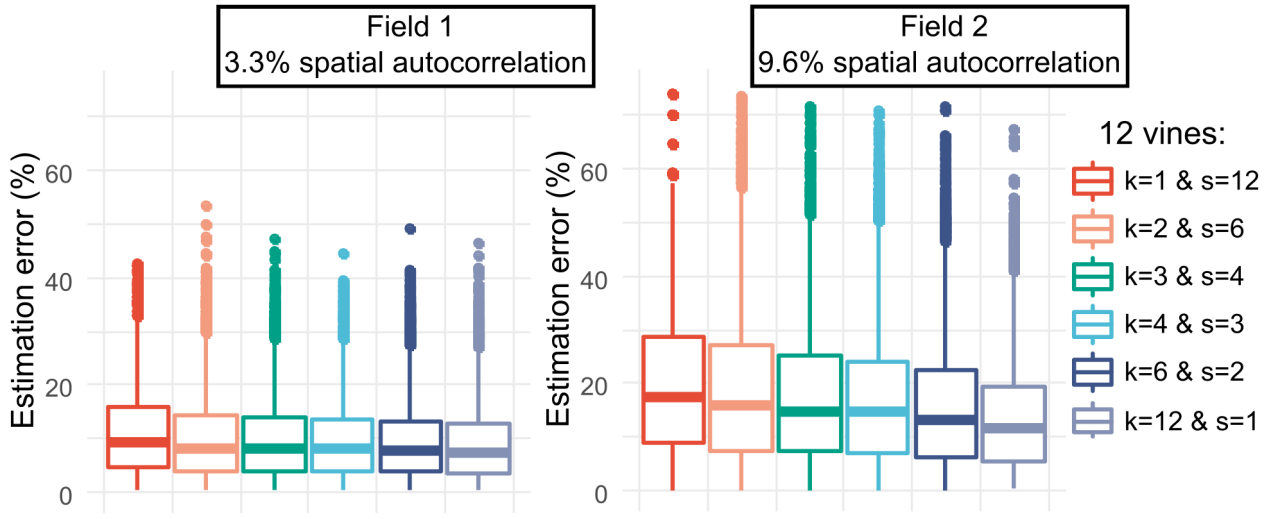


FIGURE 9. Estimation errors of number of bunches per vine observed in two real vineyards (Figure 4) with different of sampling site sizes. Boxplots with median, first and third quartile of bunch estimation error of 10,000 samples comprising $n = 12$ vines resulting from productive random sampling with varying sizes of sampling sites: from $s = 12$ (red) to $s = 1$ (grey) consecutive vines.

TABLE 2. Validation of confidence interval estimated using a Bayesian approach: proportion of samples whose error fell within the 50 % (blue) and 90 % (red) confidence intervals calculated from their size, mean and standard deviation (Eq. 6).

Sample size (n)	3	6	9	12	15
Vineyard 1	51.98 % 92.33 %	50.71 % 90.16 %	51.30 % 90.93 %	49.99 % 89.87 %	50.14 % 90.12 %
Vineyard 2	52.70 % 94.57 %	51.58 % 91.20 %	49.88 % 91.37 %	50.80 % 90.84 %	50.07 % 91.21 %

TABLE 3. Sample size required to obtain relative estimation errors lower than 10 % with a 90 % confidence interval for the number of bunches.

n value required to obtain relative error <10 % with 90 % probability		Sample standard deviation (σ_{Sample})				
		2	4	6	8	10
Sample mean ($\overline{X}$)	8	19	70	> 100	> 100	> 100
	12	10	33	70	> 100	> 100
	16	7	19	41	70	> 100
	20	5	13	27	46	70
	24	4	10	19	33	49

The sample size depends on the sample mean $\overline{X}$ and the sample standard deviation σ_{Sample}.

relevance of the working hypotheses (normal distribution, independence of the samples and negligible compared to vineyard size) to define confidence intervals. It should be noted that the independence of the observed vines depends on the spatial autocorrelation phenomenon discussed in the previous section. As seen previously, sampling sites of size $s = 1$ randomly distributed within the vineyards guaranteed the independence of the observations.

Table 3 shows how the methodology allowed the sample size (n) to be defined using the confidence intervals derived from Eq. 6 and validated in Table 2. Table 3 shows the sample size required to reach estimation errors lower than 10 % and with a 90 % confidence interval from the sample properties (sample mean $\overline{X}$ and sample standard deviation σ_{Sample}). In other words, this table represents the total number of vines that must be sampled in order to have more than a 90 % chance that the error is lower than 10 %. The values presented in Table 3 only depend on the sample and are valid regardless of the vineyard sampled. Sample mean ($\overline{X}$) values and sample standard deviation (σ_{Sample}) values were chosen based on the values in Table 2. Table 3 can be used in order to determine the total number of vines that need to be sampled to ensure that there is at least a 90 % chance of the estimation error being less than 10 %, knowing that a small sample of mean $\overline{X}$ and standard deviation σ_{Sample} is already available For example, if for a first sample of $n = 4$ vines the observed mean = 20 bunches per vine with a standard deviation σ_{Sample} of 4, Table 3 indicates that it would be necessary to sample $n = 13$ vines to obtain an estimate with a 10 % error in 90 % of the cases; since 4 vines have already been sampled, another 9 vines would be required to obtain a 10 % error level with 90 % confidence. By supposing the calculation can be done at the actual time of sampling, the number of vines to sample to reach 10 % error in 90 % of the cases can be re-evaluated during each measurement based on the observed mean and standard deviation of the sample.

As expected, the higher the standard deviation of the sample, the larger the sample size that is needed to obtain the same level of confidence in the estimation. Similarly, the higher the mean, the higher the confidence in the estimation. This last characteristic can be easily understood, since the error is relative to the mean, and when it increases, the relative error logically decreases (Eq. 2).

Table 3 also shows that samples with the same $\frac{\sigma_{sample}}{\overline{X}}$ ratio require the same sample size to reach 10 % error. For instance, a sample with a mean of 12 and standard deviation of 4 requires the same sample size as a sample with a mean of 24 and standard deviation of $8(2 \times 4)$ to obtain the same error with an equivalent confidence level. This highlights that confidence in the estimation is directly related to the coefficient of variation of the sample (Eq. 1). This is not surprising since the difference between estimation and reality $(\overline{X} - m)$ is often proportional to the variability, represented by σ_{Sample}. The relative errors $\frac{|\overline{X} - m|}{m}$ can therefore be associated with the coefficient of variation. Similar results can be obtained with other confidence levels. The higher the desired confidence level, the larger the sample sizes should be.

Figure 10 complements Table 3 by representing the 90 % confidence interval depending on the sample properties: its

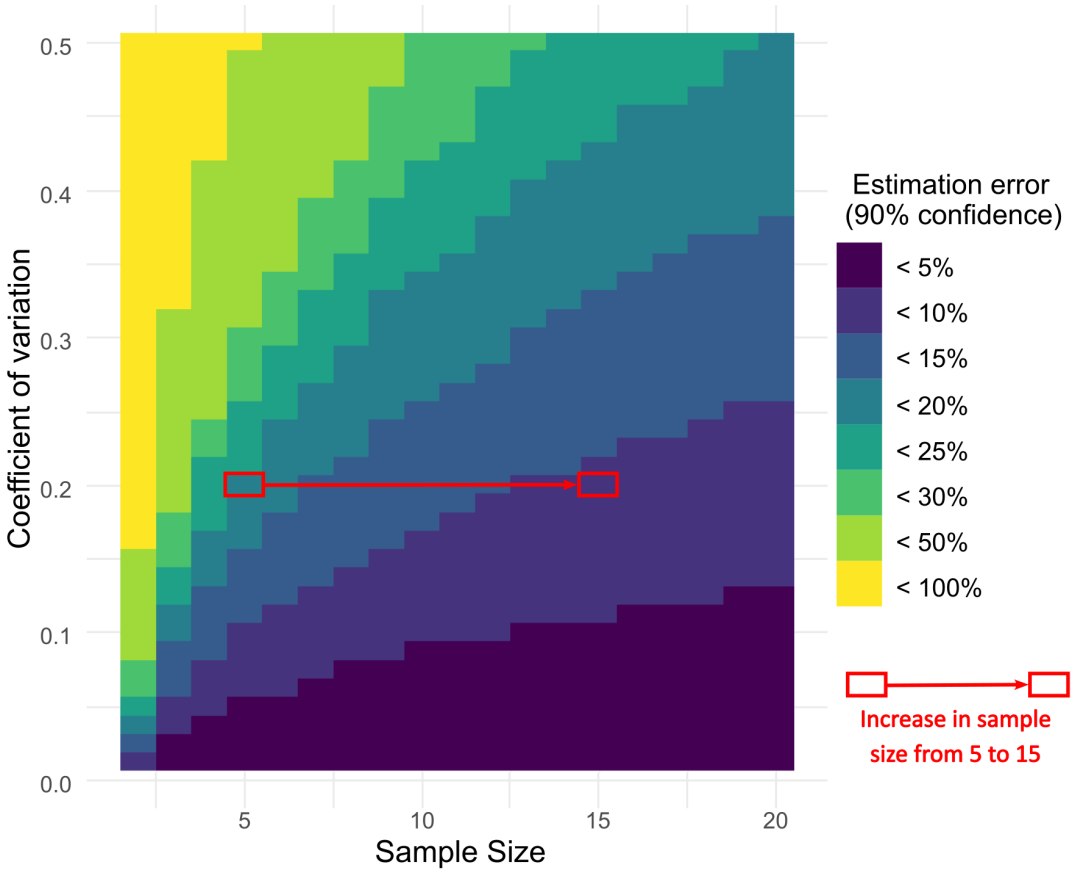


FIGURE 10. Expected estimation error (%) depends on the sample size (n) and its coefficient of variation.

size and its coefficient of variation. According to Figure 10, a sample of size $n = 5$ with a coefficient of variation of 20 % will give estimation errors lower than 20 % with a 90 % probability (red square on the left). For the same coefficient of variation, a sample size of $n = 10$ should result in estimation errors lower than 15 %, and a sample size of 15 will lead to estimation errors lower than 10 % (red square on the right). Using Figure 10, it is possible to quantify the uncertainty associated with a sample based only on its size and its coefficient of variation (Eq. 1) for any vineyard.

DISCUSSION

By addressing the three main objectives of the study described in the introduction, this study aimed to contribute to improving the design of sampling schemes for an important yield component: number of bunches per vine.

First, in a commercial context, the proportion of dead and missing vines is often unknown at the time that the number of bunches is being estimated. Although it is tempting to sample for these two components simultaneously, the results of this study show that this can be hazardous and may not be effective, as these two yield components may not have the same variability. The number of observations required for estimating the proportion of dead and missing vines with the same level of error is often higher than for the number of bunch. Therefore, at least in this case, in order to obtain the same level of error when estimating these two yield components simultaneously, observations must be carried out on a larger number of vines. From a practical point of view, this can be more time consuming during the flowering period, especially if the dead and missing vines can be sampled during a less specific time period. For these reasons, a specific sampling approach for each of these yield components should be preferred. In both study vineyards, the proportion of dead and missing vines was known and relatively high (42 % and 34 %). Therefore, it was even more important to estimate the proportion of dead and missing vines, since it had a significant impact on the final yield estimation. In the (unrealistic) case of a vineyard with no dead or missing plants, errors obtained with the PRS and CRS would be exactly the same. However, when the number of dead and missing vines is unknown at the time of estimating the number of bunches, it is important to note that this could drastically impact the number of bunches that need to be counted (and the duration of sampling) to reach an expected level of error; indeed, the higher the proportion of dead and missing vines, the higher the impact. A specific study on the impact of the proportion of missing vines on the obtained estimation errors could shed light on this issue. New approaches are being developed to specifically estimate the proportion of missing vines by aerial imagery (Chanussot *et al*., 2005; Di Gennaro and Matese,

2020; Tang *et al*., 2016), thus making independent estimation of different yield components even more relevant.

Second, distributing the sampled vines within a few large sampling sites resulted in higher estimation errors compared to smaller and more numerous sampling sites when number of bunches per vine showed spatial autocorrelation. The values of the vines that are close to each other are more similar, and the probability of overestimation or underestimation was higher when the majority of sampled vines was located in the same zone of the vineyard (i.e., in a low yield zone or a high yield zone). The higher the spatial autocorrelation, the higher the occurrence of this phenomenon. On the other hand, when there is no autocorrelation, the location of the vines within the vineyard is not expected to have any influence on the estimated number of bunches, and the arrangement of the sampling sites within the vineyard will not have any effect on the accuracy of the estimation. In previous studies, number of bunches per vine has often been found to have a low spatial autocorrelation, because variations due to environmental factors were controlled by pruning operations (Carrillo *et al*., 2016; Taylor and Bates, 2013). Accordingly, in the present study, both real vineyards showed low spatial autocorrelation (3.3 % and 9.6 %). However, scientific literature dealing with spatial autocorrelation of bunch numbers remains scarce. In practice, therefore, in order to minimise any potential influence of autocorrelation on estimation errors, bunch number estimation should not be based on a single large sampling site. Although there are operational constraints involved (e.g., cost of travel between sampling sites), at least two or three sampling sites should be sampled.

Third, from just a few observations it was possible to obtain information about the expected error and to determine the sample size necessary to reach the desired accuracy and confidence for the estimation. As the survey progresses and the sample increases, the relative error confidence interval can be easily updated. This means that it is possible to adjust the sampling protocol in real time during the estimation process. In the more unfavourable cases, the use of these confidence intervals can help to identify a situation in which the sample variability is too high to obtain a relevant estimation within a reasonable time period. Based on this information, the practitioner can choose to invest available time in a vineyard of which the benefits in terms of accuracy will be higher than in another vineyard. This study shows how the Bayesian approach is relevant when computing confidence intervals for relative estimation errors (i.e., as percentages). From an operational point of view, these intervals, expressed as error percentages, are more consistent with how wine growers understand and express estimation errors compared to conventional frequentist confidence interval errors that are expressed as number of bunches per vine. It was also easier to evaluate the influence of errors on the final yield estimation using percentages. The fact that the computation of confidence intervals was based on Bayesian statistics also opened up the possibility of integrating a priori available vineyard information into the process. Indeed, in this study, a fully uninformative Bayesian prior was chosen, but it is possible to use an a priori density, which reflects existing knowledge of a vineyard, to better define the confidence intervals.

In the scientific literature, several approaches have been proposed to select measurement sites in a vineyard for yield estimation (Carillo *et al*., 2016; Araya-Alman *et al*., 2019, Oger *et al*., 2021). While these studies offer guidance on how to best situate sampling sites, the size, number and type of sampling sites have until now been rarely addressed. The results of the present study therefore complement existing knowledge and provide new tools for defining appropriate sampling protocols. However, it should be noted that the results obtained from the two real vineyards are specific to the South of France. Despite the simulation of vineyards that encompassed a wide variety of conditions, the results may still not be representative of the diversity of viticultural production systems. Finally, the estimation of number of bunches per vine should be placed within a more general context of yield estimation: although bunch number is often the component that explains the largest part of total yield variability (Carrillo *et al*., 2016; Clingeleffer *et al*., 2001), it is part of two- or three-step estimation methods in which other components must also be estimated. Therefore, estimation errors must be associated with those obtained from the estimation of the proportion of dead and missing vines and other yield components, such as bunch weight, number of berries per bunch or berry weight. Measurement errors were also not considered here. Therefore, how all these errors propagate and affect the final yield estimation remains an open question.

CONCLUSION

This study addressed some practical aspects of sampling number of bunches. Even when the sampling protocol is random and not based on a priori information, estimation accuracy can be improved by applying appropriate practices. For grapevine yield estimation, it is recommended to use a specific sampling protocol for each yield component. In particular, when possible, the proportion of dead and missing vines should be estimated independently to avoid negatively impacting the estimate of number of bunches per vine. It was also shown that choice of appropriate sampling strategy must be based on observations spread over several measurement sites that are randomly distributed within the vineyard in order to limit the effect of spatial autocorrelation on yield estimate. Based on available vineyard data, 20 to 30 vines spread over two or three sites of ten vines were needed to estimate the number of bunches with an error lower than 10 %. Finally, the Bayesian confidence intervals used here can contribute to new methodology for evaluating errors associated with a sample. This method allowed the size of an ideal sample to be defined in relation to the desired estimation error expressed as a percentage and the variability found in the first observations. This work opens the way towards the adaptation of sampling protocols in real time, and generally provides new knowledge that can be appropriated by viticultural stakeholders for their sampling methods.

ACKNOWLEDGEMENTS

This work was financed by the Occitanie region.

The authors would like to thank Christophe Abraham for his help in Bayesian statistics, James Arnold Taylor for his proofreading and Célia Crouzet, Pauline Faure, Jean-Philippe Gras, Clémence Huck, Yoann Valloo, and Yulin Zhang for their help in the acquisition of field data.

REFERENCES

Ancelin, J., Poulain, S., & Peneau, S. (2022). *jancelin/centipede: 1.0*. Zenodo. https://doi.org/10.5281/zenodo.5814960

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 20, 445–459 (2019). https://doi.org/10.1007/s11119-019-09644-y

Bramley, R. G. V. (2001). Vineyard sampling for more precise, targeted management. *Proceedings of the First National Conference on Geospatial Information & Agriculture*, 417– 327.

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. https://doi.org/10.1007/s11119-015-9407-8

Chanussot, J., Bas, P., & Bombrun, L. (2005). Airborne remote sensing of vineyards for the detection of dead vine trees. *Proceedings. 2005 IEEE International Geoscience and Remote Sensing Symposium, 2005. IGARSS '05.*, *5*, 3090–3093. https://doi.org/10.1109/IGARSS.2005.1526490

Clingeleffer, P. R.; Martin, S. R.; Dunn, G. M.; Krstic, M. P. (2001). *Crop Development, Crop Estimation and Crop Control to Secure Quality and Production of Major Wine Grape Varieties: A National Approach*. Grape and Wine Resarch & Development Corporation.

Di Gennaro, S. F., & Matese, A. (2020). Evaluation of novel precision viticulture tool for canopy biomass estimation and missing plant detection based on 2.5D and 3D approaches using RGB images acquired by UAV platform. *Plant Methods*, *16*(1), 91. https://doi.org/10.1186/s13007-020-00632-2

Gräler, B., Pebesma, E., & Heuvelink, G. (2016). Spatio-Temporal Interpolation using gstat. *The R Journal*, *8*(1), 204–218. https://doi.org/10.32614/RJ-2016-014

Hespanhol, L., Vallio, C. S., Costa, L. M., & Saragiotto, B. T. (2019). Understanding and interpreting confidence and credible intervals around effect estimates. *Brazilian Journal of Physical Therapy*, *23*(4), 290–301. https://doi.org/10.1016/j.bjpt.2018.12.006

Laurent, C., Oger, B., Taylor, J. A., Scholasch, T., Metay, A., & Tisseyre, B. (2021). A review of the issues, methods and perspectives for yield estimation, prediction and forecasting in viticulture. *European Journal of Agronomy*, *130*, 126339. https://doi.org/10.1016/j.eja.2021.126339

LTD (2022). *Mergin Maps*. http://www.merginmaps.com

Millan, B., Velasco-Forero, S., Aquino, A., & Tardaguila, J. (2018). On-the-Go Grapevine Yield Estimation Using Image Analysis and Boolean Model. *Journal of Sensors*, *2018*, 1–14. https://doi.org/10.1155/2018/9634752

Nuske, S., Achar, S., Bates, T., Narasimhan, S., & Singh, S. (2011). Yield estimation in vineyards by visual grape detection. *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2352–2358. https://doi.org/10.1109/IROS.2011.6095069

O'Hagan, A. (2010). *Kendall's Advanced Theory of Statistic 2B*. John Wiley \& Sons.

Oger, B., Vismara, P., & Tisseyre, B. (2021). Combining target sampling with within field route-optimization to optimise on field yield estimation in viticulture. *Precision Agriculture*, *22*(2), 432–451. https://doi.org/10.1007/s11119-020-09744-0

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/

Smithson, M. (2000). *Statistics with Confidence*. SAGE Publications Ltd. https://doi.org/10.4135/9781446218525

Tang, J., Woods, M., Cossell, S., Liu, S., & Whitty, M. (2016). Non-Productive Vine Canopy Estimation through Proximal and Remote Sensing**This work was supported by Wine Australia. *IFAC-PapersOnLine*, *49*(16), 398–403. https://doi.org/10.1016/j.ifacol.2016.10.073

Taylor, J. A., & Bates, T. R. (2013). Temporal and spatial relationships of vine pruning mass in Concord grapes. *Australian Journal of Grape and Wine Research*, n/a-n/a. https://doi.org/10.1111/ajgw.12035

Taylor, J., Tisseyre, B., Bramley, R., Reid, A., Stafford, J., & others. (2005). A comparison of the spatial variability of vineyard yield in European and Australian production systems. *Proceedings of the 5th European Conference on Precision Agriculture*, *5*, 907–914.

Victorino, G. F., Braga, R., Santos-Victor, J., & Lopes, C. M. (2020). Yield components detection and image-based indicators for non-invasive grapevine yield prediction at different phenological phases. *OENO One*, *54*(4), 833–848. https://doi.org/10.20870/oeno-one.2020.54.4.3616

Wolpert, J. A., & Vilas, E. P. (1992). Estimating Vineyard Yields: Introduction to a Simple, Two-Step Method. *American Journal of Enology and Viticulture*, *43*(4), 384–388. https://doi.org/10.5344/ajev.1992.43.4.384.