Is the optimal strategy to decide on sampling route always the same from field to field using the same sampling method to estimate yield?

Baptiste Oger, Cecile Laurent, Philippe Vismara, Bruno Tisseyre

Received: 14 April 2020 | Accepted: 23 December 2020 | Published: 5 February 2020
DOI:10.20870/oeno-one.2021.55.1.3334

OENO One
VINE AND WINE OPEN ACCESS JOURNAL

Is the optimal strategy to decide on sampling route always the same from field to field using the same sampling method to estimate yield?

Baptiste Oger[1,2,*], Cécile Laurent[1,3], Philippe Vismara[2,4] and Bruno Tisseyre[1]

[1] ITAP, Université de Montpellier, Montpellier SupAgro, INRAE, France
[2] MISTEA, Université de Montpellier, Montpellier SupAgro, INRAE, France
[3] Fruition sciences, France
[4] LIRMM, Université de Montpellier, CNRS, France

[*]corresponding author: baptiste.oger@hotmail.com

ABSTRACT

Aim: This short communication aims at providing insights to verify whether common yield sampling protocols (*i.e.*, one round trip within the fields over two representative rows) are optimal whatever the considered fields. In addition, it aims to show how factors like the spatial organisation of the within-field yield variability, the length of the rows, the erratic variance, etc. may affect the optimal sampling route and the error of the yield estimation.
Methods and Material: A new algorithm based on constraint programming and stochastic approaches was used to provide optimal sampling routes for vineyards. This algorithm guarantees the representativeness of the measurement sites and a minimization of the walking distance. Practical constraints (trellised structure, starting point, etc.) are considered by the algorithm to optimise the walking distance and the resulting sampling route. The algorithm has been applied to 60 simulated vineyards with known yield variability. Characteristics like yield spatial structure, row length and proportion of erratic variance were controlled during the simulation process and were used to study how they affect the optimal sampling route derived from the algorithm.
Results: The row length as well as the spatial organization of the within-field yield variability are the main factors that determine the optimality of a sampling route. Spatial organisation of the yield happens to have a strong incidence; fields with small yield patterns (Range of the semi-variogram = 25 m) showed a yield estimation error of less than 2 % with an optimal sampling route of three minutes with 7 sampling sites, whereas it takes more than 5 minutes (with 9 sampling sites) to achieve the same estimation error for fields with larger spatial patterns (range > 50 m). Results also highlight the relevance of original sampling routes which intend to sample only the beginnings of rows or mixed approaches based on a round trip in two inter-rows and complementary samples on the beginnings of one or more rows.
Conclusions: This study shows that an optimal sampling route strongly depends on the field characteristics. The optimal sampling route should therefore be tailored to each field. This approach is a first step which shows how this methodology could be used to identify other factors of influence. It could also apply to real fields to optimise other logistic operations in viticulture.
Significance and Impact of the Study: This short communication demonstrates the necessity to tailor sampling strategy to characteristics of each field to provide both an optimised sampling route (minimum walking distance with minimum samples) and the best possible estimate. It also proposes an original approach based on field simulations and an optimal sampling route generation algorithm. This approach makes it possible to produce new insights (and also to validate empirical practices) that can help the wine industry to better manage the logistics at harvest. This paper also gives considerations when it comes to the choice of a sampling route for a given field.

KEYWORDS

sampling, yield, viticulture, sapling protocol, grape yield, yield estimation

INTRODUCTION

Precise knowledge of field yields is critical for the wine industry, mainly for the logistical organisation of the harvest among other reasons (Clingeleffer *et al.,* 2001). Field yield estimation is often carried out by sampling. Observations of yield components are then made on a limited number of vines (Carrillo *et al.,* 2016; Arnó *et al.,* 2017; Wolpert and Vilas, 1992) and the mean value is generally used to provide an estimation of the field yield. To have a relevant estimation of the final yield, sampling is often carried out a few days before harvest, at a critical period in terms of workload. As a result, the implementation of a yield estimation protocol is often the result of a balance between the accepted error on the estimation and the time required to carry out the observations (Wolpert and Vilas, 1992). Vine fields (even small) present a high yield variance (Taylor *et al.,* 2005); the average coefficient of variation was found to be around 40 %. This variability is mostly explained by the spatial variation of environmental factors (soil, water availability, fertility, etc.) but also to other biotic factors (disease, weed or inter-vine competition, etc.).

This within-field variability affects the quality of estimates resulting from sampling. Indeed, when estimating yield by sampling, the error of estimation (and resulting confidence associated with the estimate) is a function of the number "N" of observations and the variance of the sample (observed variables). For a given field (with a given yield variance), the higher the "N", the more confident the estimate is, but the longer the time required to carry out the sampling. Note that the sampling time may present high variations depending on the location of the observations over the field. Indeed, the sampling time is directly related to "N", but also to the time needed to travel from one observation site to another. Optimising sampling time, therefore, requires to optimise both the number "N" of observations according to the field variability and the location of the observation sites to limit the travel time.

In the scientific literature, there are very few papers which have focused on optimizing the location of sampling sites for yield estimation in viticulture. Most of the studies focused on: I) the number "N" of observations to be considered to reach a reliable estimation and to minimise the error of estimation (Wolpert and Vilas, 1992; Carrillo *et al.,* 2016), II) the type of observations and sensing systems to limit measurement time and/or measurement errors (Diago *et al.,* 2012; Reis *et al.,* 2012; Nuske *et al.,* 2011; Serrano *et al.,* 2005; Dunn and Martin, 2004), III) the optimization of the representativeness of the observation sites by targeting samples based on an auxiliary variable (*i.e.*, vegetative index) available with a high spatial resolution (Wulfsohn *et al.,* 2012; Meyers *et al.,* 2011, Carrillo *et al.,* 2016).

As a result, existing studies rarely take into account the two contradictory components leading to an optimal sampling: the optimisation of the sampling effort (time which includes the measurement time and the travel time/distance) and the minimisation of estimation error. For the wine industry, these two components are very important to produce the best possible estimation in the shortest possible time. Without reliable references, the sampling protocols used are often based on rules of thumb and the same protocol is always applied whatever the field, *i.e.*, two representative rows are chosen corresponding to one round trip within the field and observations are more or less randomly carried out along these rows (Rousseau Jacques, pers. communication) or sometimes according to a grid previously defined by an expert. Note that the use of high-resolution spatial data like remote sensing or soil mapping to consider more sophisticated sampling process remains rare, at least in France since less than 2 % of the vineyard area is benefiting from this type of service in France (Lachia *et al.,* 2019). Whatever the protocol, the underlying rules of thumb used by the wine industry to perform yield sampling remain difficult to justify rigorously.

Recent papers (Meyers *et al.,* 2020; Oger *et al.,* 2020) proposed new sampling approaches which simultaneously takes into account the stochastic nature of the spatial variable and the sampling time to produce optimal sampling routes in viticulture. The approach proposed by Meyers *et al.* (2020) provides optimal samples in terms of NDVI values taken by the selected pixels. The length of sampling routes is controlled by the selection of neighbouring pixels, reducing sampling times. In this approach, the choice of the measurement sites within a pixel (90 m²) is left at the practitioner discretion. The Oger *et al.* (2020) approach combines stochastic methods with constraint optimization to minimize the practitioner's travel time. The approach ensures the representativeness of the measurement in the attribute space (values) by considering the distribution of a high spatial

resolution auxiliary data and the optimality of the path to go from one observation site to another. The Oger *et al.* (2020) approach presents the advantage to directly propose to the practitioner sample sites distributed over an optimal sampling route and does not limit sampling routes to successive pixels (squares of 90 m²) where the practitioner still has to randomly select sampling sites. Assuming that the yield is fully known and a value is available for each within-field site, the Oger *et al.* (2020) method is interesting because it allows one to reverse its application to verify what the optimal sampling route according to the characteristics of the field would be. The application of such an approach to fields with known yield values is interesting because it allows: (I) to check whether empirical sampling protocols like the two-row round trip are relevant and if they apply properly to all the fields, (II) to explore whether original and non-trivial routes can emerge from the application of this algorithm and (III) to see whether specific field characteristics promote particular sampling procedures. The objective of this study is, therefore, to apply the methodology of Oger *et al.* (2020) on theoretical yield data with known features to study how factors like the spatial organisation of the yield, the length of the rows, the erratic variance, etc., may affect the optimal sampling route and the error of the yield estimation. This work remains theoretical because it requires prior knowledge of the yield at any point in the field (which is not the case in reality). However, the aim of this study is to verify whether the same optimal sampling route patterns apply for all the fields or, on the contrary, whether specific sampling routes tailored to each field (or type of field) should be considered. The paper will briefly present the approach as well as the theoretical yield data and their characteristics. It will then present the results with a discussion that will focus on practical issues.

MATERIALS AND METHODS

1. Data

Theoretical yield data were generated through a simulation process. This simulation process, described in Oger *et al.* (2020), generates spatialised data by summing Gaussian fields to non-spatialised residual noise (erratic data). By setting the semi-variogram of Gaussian field and the noise proportion, it was thus possible to control parameters of the resulting theoretical data. This paper focused on the influence of three parameters on the optimal sampling: I) The range for the Gaussian field semi-variogram. This corresponds to the autocorrelation distance of the theoretical yield data. The range defines the minimum distance (in meters) that must separate two sites for them to be considered as spatially independent. It defines the average size of the yield spatial patterns within a field, II) The nugget effect, which is the proportion of erratic (non-spatialised) variance of the theoretical yield data. This measurement is expressed as a percentage of the total variance, III) Row length in relation to field width.

For the simulation processes, the magnitude of variation of values for the range and the nugget effect were determined from within field yield observations obtained from yield monitoring systems in precision viticulture (Taylor *et al.,* 2005; Bramley *et al.,* 2019). For row length, simple rectangular structures with areas of 1 hectare were tested. Table 1 summarizes the values of the parameters tested in this article.

Theoretical fields were generated by varying only one parameter at a time, with the other two parameters taking their default values. The initial resolution is 1 pixel/m². Yield values are then extracted on the rows assuming a trellised structure with a 2.5 m distance between rows and 1 m between vines on the row (4000 vine plants/ha). Simulations were run with a Gaussian yield distribution with an average yield around 1000 g/vine and a coefficient of variation at 30 %. For each combination, 10 different fields were simulated. The final theoretical dataset consists of 60 (6 × 10) simulated fields.

2. Sampling Route Optimisation

For each field, the optimal sampling route was obtained by applying the approach described in Oger *et al. (*2020). This approach uses constraint programming principles and stochastic methods to find the best sampling route according to

TABLE 1. Parameter values for the generation of theoretical fields (default values in bold)

Theoretical field yield parameters	Values		
Range (m)	25	**50**	75
Nugget effect (%)	**20**	50	
Row length (m) × field width (m)	50 × 200	**100 × 100**	200 × 50

defined constraints. Without going into too much computational detail, a first constraint ensures that the N selected measurement sites are separated by a minimum Euclidean distance to avoid autocorrelation and to make sure observed yield values are independent. This minimum Euclidean distance is defined by half of the yield data range (refer to the previous section). For each field, the second constraint aims at ensuring that the N measurement sites are representative of the yield value distribution of the field, one measurement site is selected in each of the intervals defined by the N yield quantiles as proposed by Carrillo *et al.* (2016). Finally, the selected sampling route must be optimal in terms of walking distance. Measurement sites are selected to fit the two first constraints while their position and the order in which they are visited must minimise the walking distance. Optimisation was performed using a solver, a software program which considered possible combinations to select the best one. While exploring possible combinations, the approach seeks to find a better solution than the best one found so far. For simple cases with small values of N, the real optimum is found in a short time. In the most complex cases, computations are stopped after ten hours to ensure the solution is close enough to the real optimum. This time is generally sufficient to find a value close enough to the optimum or the optimum itself but without being able to demonstrate it. The core of the sampling approach was written in Java using the Choco solver (Prud'homme *et al.,* 2016). Computation was made on a Linux server with Intel(R) Xeon(R) CPU X5690 3.47GHz.

Distances were expressed as walking time (min.). Walking times do not consider additional constraints specific to a given field that could alter the walking speed (grass, slope, soil surface conditions, etc.). They only take into account vineyard specificities associated with the trellised structure. It is not possible to move between two rows while being in the field. Going from one inter-row to another implies having to reach one of the field edges. Each measurement site can be accessed from two different inter-rows. This distance also takes into account a starting point where the sampling route must begin and end. It is positioned in the southwest corner of each field. The distance optimized by the solver corresponds to this walking distance that passes through each measurement site and returns to the starting point. This promotes the choice of measurement sites close to the starting point. Common starting points enable a simple comparison of the sampling routes obtained.

3. Sampling route characterisation

To clarify the presentation of the results, two types of sampling routes were considered. The first one corresponded to what is assumed to be most commonly performed by practitioners; this consists in an empirical sampling protocol where measurements are carried out following one round trip within the fields across two, or more, representative rows. Rows are therefore walked from one end to the other, forming a sampling route joining the two sides of the field. This type of route is called thereafter *row-based sampling route* (RBSR). The second type of sampling route never reaches both sides of the field *i.e.*, no row is walked entirely. This type of sampling route corresponds in reality to a large diversity of cases, but a common feature is to focus on the field edge close to the starting point. All the sampling routes presenting these features were considered as *edge-based sampling rout*e (EBSR).

Sampling routes obtained with the solver were characterised using three criteria: I) The type of sampling route: RBSR or EBSR. II) The walking time required to get from one observation site to another, regardless of the protocol chosen to carry out the measurements and the time associated with these measurements. The time required to make observations (number of clusters, average cluster weight, etc.) at a sampling site may vary depending on the protocol used. However, it was assumed in this work that, for a given situation, the measurement protocol was the same for each sampling site. As a result, for the same number of observation sites, the sampling time was only influenced by the travel distance between the observation sites. Therefore, the walking time depends only on the distance to be covered and the walking speed of the practitioner, which is assumed here to be constant at 0.9 m/s. III) The estimation error corresponds to the difference between the value predicted from measurement sites along the sampling route and the actual average yield of the field. The predicted value ($\hat{\bar{Y}}$) is constructed as the average of the N yield observations made during sampling. The actual yield value ($\bar{Y}$) corresponds to the average of all the simulated yield values of the field. The calculation of the estimation error, expressed as a percentage, is defined by Equation 1.

$$Estimation\ Error\ (\%) = \frac{\left|\hat{\bar{Y}} - \bar{Y}\right|}{\bar{Y}} \times 100 \quad (Eq.1)$$

RESULTS

Figures 1.A, 1.B and 1.C shows the results of optimal sampling routes, either EBSR or RBSR, expressed as estimation errors and walking times for the different field characteristics. Figure 1.D shows results obtained with a simple random sampling which was based on the selection of sites randomly chosen among all the measurement sites without any path optimisation. All the curves share the same logical trends; the estimation errors decrease with an increase in the number N of samples. However, improving the quality of the estimation has a cost since the sampling effort estimated by the "walking time" increases with the number of measurements. A comparison between Figures 1.A and 1.D shows the value of the optimal sampling approach, as proposed in this study, compared to a simple random sampling approach. Sampling optimisation simultaneously improves the estimation error by 5 % to 9 % and reduces the running time by half for the examples considered. Only results obtained for simulated fields with different range were presented (Figure 1.D) for random sampling, but very similar results (results not shown) were obtained for the other simulated fields (row length, nugget effect).

Figure 1 shows that the characteristics of the fields do not affect the optimal sampling route in the same way. The range (Figure 1.A) and, to a lesser extent the row length (Figure 1.B), significantly affect the optimal sampling route, while the proportion of erratic variance in the total yield variance of the field (nugget effect) has a small effect on the optimal sampling route (Figure 1.C). For clarity, Figure 1 does not show the variability resulting from the ten simulations, in the average standard deviation of the results is 1.7 % and 0.6 minutes for the error and the walking time, respectively.

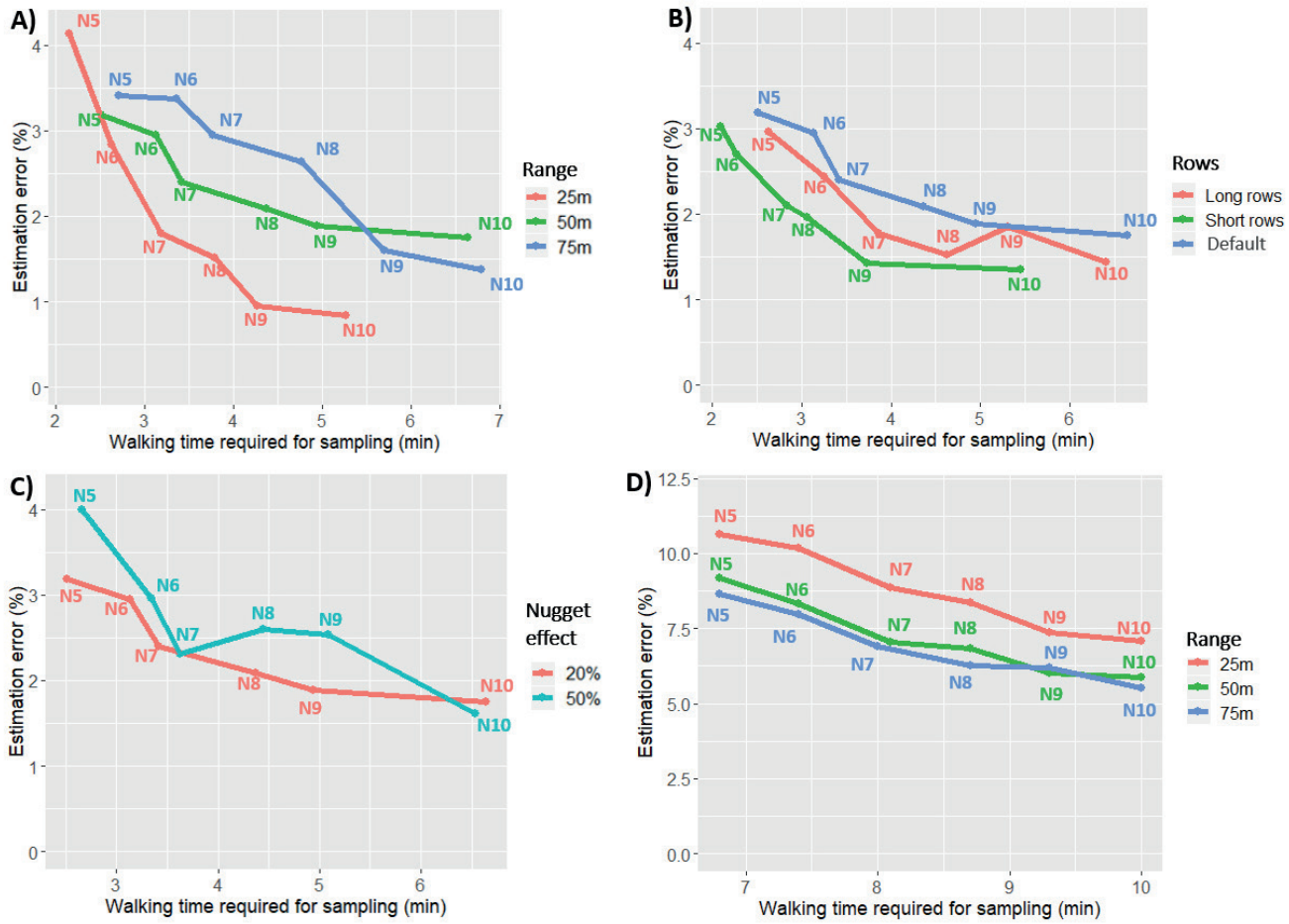

FIGURE 1. Average estimation error and walking times depending on field characteristics: A) the field range, B) row length, C) percentage of random variability (nugget effect). Results are the mean value over ten simulations. Each curve is made of 6 points corresponding to sampling routes with N = {5,6,7,8,9,10} sampling sites. D) gives the same result as A) but for random sampling and results are the mean value over ten simulations and 100 repetitions per simulation.

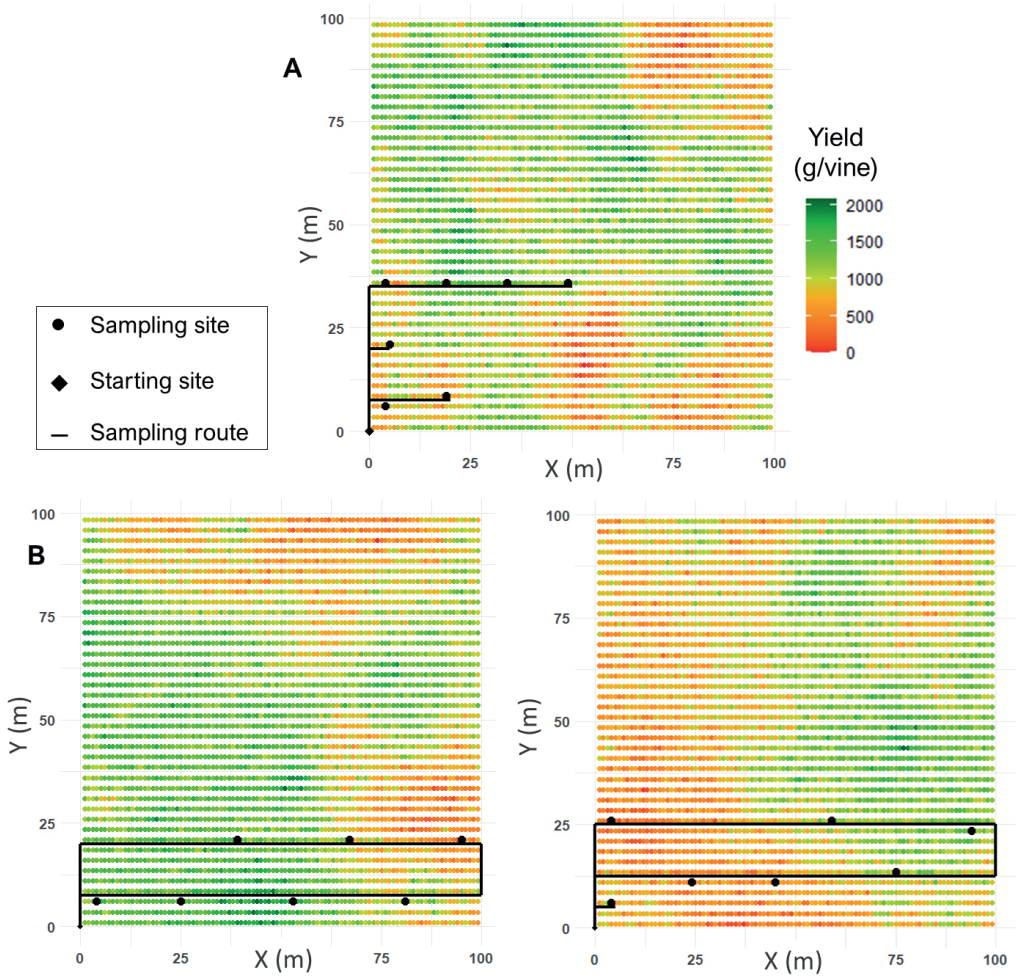

FIGURE 2. Illustration of sampling routes for three typical fields with different range with N = 7
Field A: Range = 25 m, Nugget effect = 20 %, Row length = 100 m
Field B: Range = 50 m, Nugget effect = 20 %, Row length = 100 m
Field C: Range = 75 m, Nugget effect = 20 %, Row length = 100 m

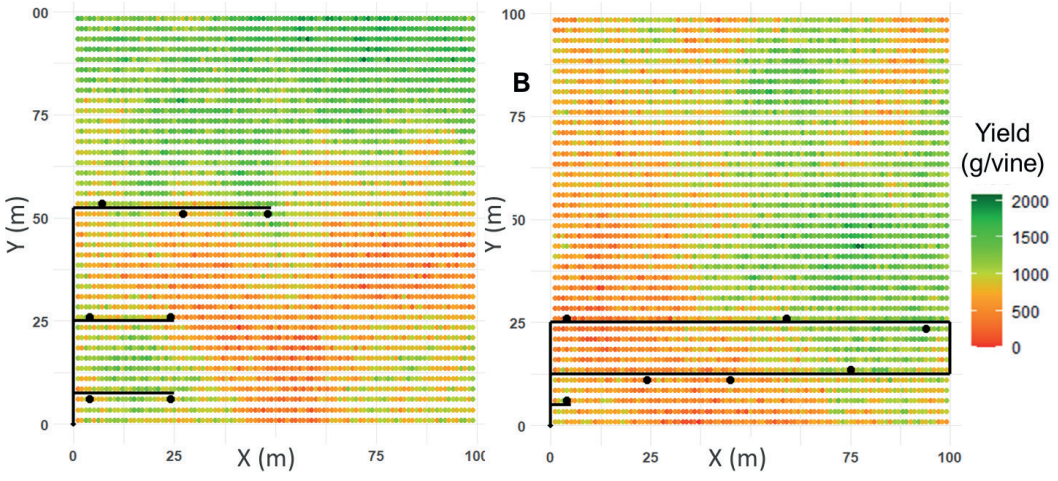

FIGURE 3. Illustration of sampling routes for two high range fields with opposite gradient orientation and N = 7
Field A & B: Range = 75 m, Nugget effect = 20 %, Row length = 100 m

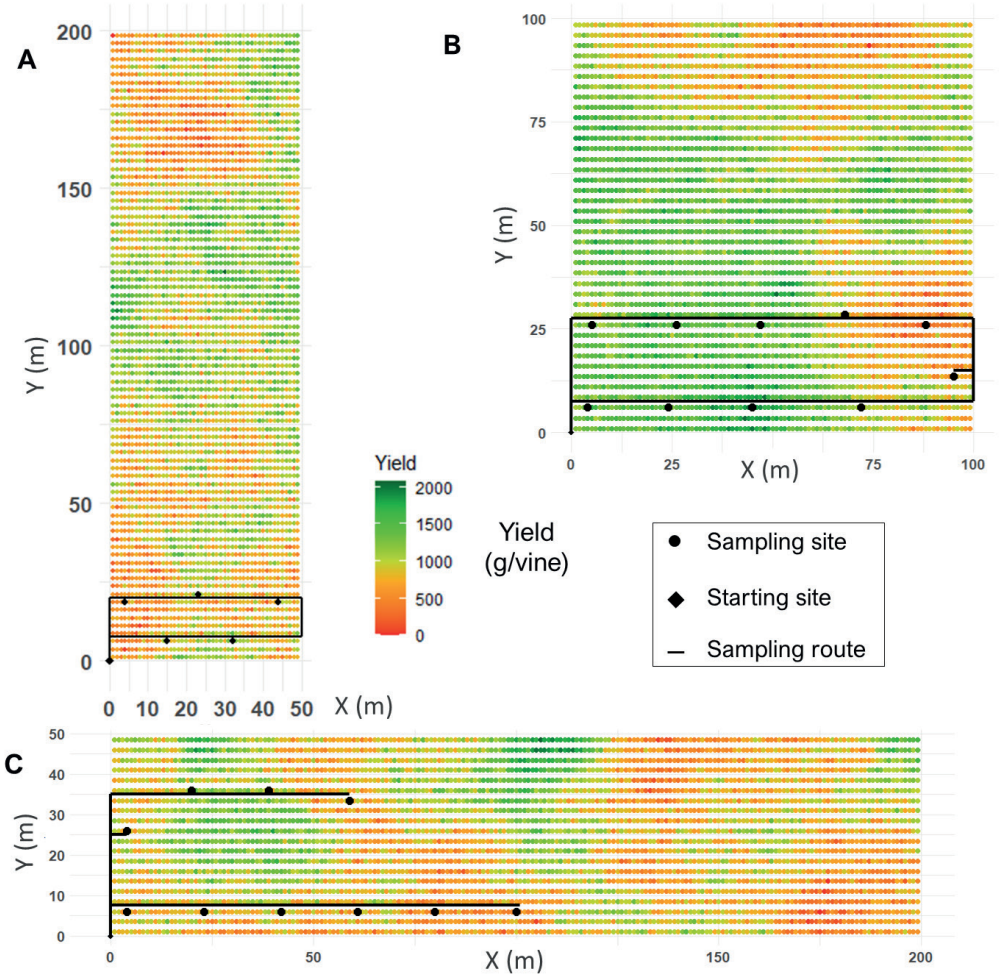

FIGURE 4. Illustration of sampling routes for three different row length
Field A: Range = 50 m, Nugget effect = 20 %, Row length = 50 m, N = 5;
Field B: Range = 50 m, Nugget effect = 20 %, Row length = 100 m, N = 10;
Field C: Range = 50 m, Nugget effect = 20 %, Row length = 200 m, N = 10.

Regarding the range (Figure 1.A), fields with lower ranges (25 m) show lower estimation errors for a given "walking time". On average, for a range of 25 m, it is possible to achieve an estimation error less than 2 % with an optimal three minutes sampling route with 7 sampling sites, whereas it takes more than 5 minutes (with 9 sampling sites) to achieve the same estimation error for fields with more extensive ranges (50 m and 75 m). In general, the lower the range, the shorter the sampling route and the walking time. Focusing on the length of the rows (Figure 1.B), it is also a factor which affects an optimal sampling route. For short rows, lower estimation errors are achieved with less sampling effort. The effect of nugget effect (Figure 1.C) is less obvious although, larger nugget effects (50 %) are associated with slightly larger estimation errors compared to fields with a low nugget effect (20 %).

Focusing on the range effect, Figure 2 shows examples of sampling routes for three fields with different range values. The three examples share some common features; sampling routes are optimized from the starting point located in the southwest corner of the field (coordinates X = 0, Y = 0). It is clear that the sampling points (and the resulting sampling route) intend to minimize the distance to this starting point for each field. Figure 2 also shows the two types of sampling routes described previously (EBSR or RBSR). The field with the shorter range is associated with an edge-based sampling route (EBSR), while the fields with longer ranges (50 m and 75 m) are associated with a row-based sampling route (RBSR). In this example, for the same number of sampling sites, the optimal route changes with the range.

However, Figure 3 shows that for large ranges, EBSR may also be promoted as an optimal sampling route. In this case, a large range (compared to the dimension of the field) affects the spatial variability of yield which tends to follow a trend (gradient). In practice, this type of spatial distribution may be observed when the yield is driven by an isotropic factor such as the slope, soil depth gradient, water access, etc. In this case, the optimal sampling route is dependent on the relative direction of the rows with the yield gradient. When the yield gradient and the rows present more or less the same direction (Figure 3.B), RBSR is promoted. Indeed, the yield gradient on Figure 3B follows the direction of the row: *i.e.*, yield increases from left to right, and rows are oriented from left to right. Conversely, when the gradient is perpendicular to the row direction (Figure 3.A), EBSR is promoted by the algorithm. Indeed, the gradient on Figure 3A follows a direction perpendicular to the rows' orientation: *i.e.*, yield increases from bottom to top, and rows are oriented from left to right.

Figure 4 shows the effect of row length on optimal sampling routes across three examples. Fields with short rows are associated with RBSR even with a limited number of sampling sites (N = 5) (Figure 4.A). Conversely, long rows promote EBSR where entire rows are never explored (Figure 4.B and 4.C).

For different field parameters, Figure 5 gives the proportion of sampling strategies corresponding to RBSR against EBSR in the function of the number of sampling sites. Each point of the figure corresponds to the average results over ten simulated fields.

Figure 5.A shows clearly that for fields with short ranges, the optimal sampling route is an EBSR in a large majority. As already seen before, the range has a significant effect on the choice of the optimal sampling strategy and this result is confirmed here over several fields. However, for fields with ranges of 50 m and 75 m, the effect is lessened and the proportion of full row sampling routes (RBSR)

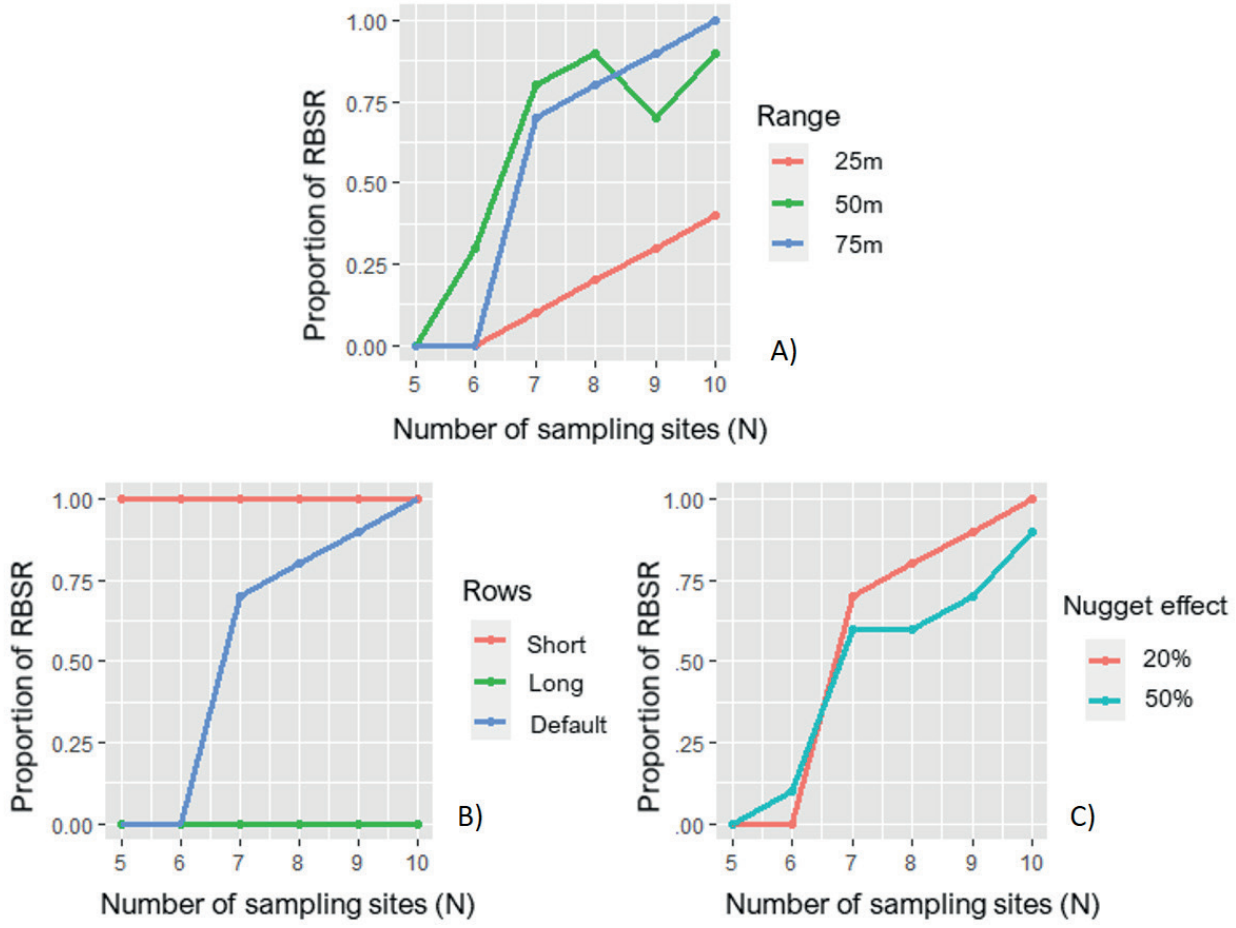

FIGURE 5. Proportion of RBSR sampling strategies sampling route depending on the number of sampling sites (N) and field type (A) range, B)row length of the field and C) nugget effect).
Results are frequency over ten simulations.

reaches a limit; the proportion curve associated with high ranges (75 m) never reaches 100 % of RBSR. This result is explained by simulated fields whose yield gradient is more or less perpendicular to the row direction which promotes EBSR over RBSR (Figure 3).

Figure 5.B also shows clearly the incidence of the length of the row on the best possible sampling route. Long rows always promote EBSR while short row fields always promote RBSR. This result verifies that of Figure 4: when the rows get longer, the optimal sampling strategy always avoids going all along the rows. Exactly the opposite is true for short rows, which is why RBSR is systematically proposed for short rows in this case.

Finally, Figure 4.C shows that a higher proportion of erratic variance (nugget effect) tends to promote EBSR when sampling sites greater than 6.

DISCUSSION

In the wine industry, a tendency to adopt the same sampling route for all fields is commonly encountered. However, based on a posteriori knowledge of yield distribution, results exposed in this paper show that the optimal strategy to design a sampling route for grape yield estimation may vary from one field to another in the function of field characteristics. The optimal route sampling seeks to minimize the effort to find sites that are representative of the distribution of yield values. Logically, the lower range of yield reduces the minimum distance to be covered to find two spatially independent sites. Therefore, low ranges make it possible to find a higher variability of yield values in the direct vicinity of the starting point which explains why EBSR is promoted in this case. This also explains why the travel distance decreases with the yield range (Figure 1), EBSR being generally shorter as it does not require to travel twice the length of the rows to find relevant observation sites. The extreme case would be a field with no spatial autocorrelation of yield values (*i.e.*, yield values are perfectly random with no range), in which choosing N independent sampling sites might result in selecting N contiguous vines on the same row. For large range yield, when the yield gradient and the rows present more or less the same direction (Figure 3.A), RBSR is logically promoted. Conversely, when the gradient is perpendicular to the row direction (Figure 3.B), EBSR is promoted by the algorithm. This is consistent considering that RBSR allows a larger diversity of yield values to be explored more quickly when the variability is organised along the rows, whereas EBSR is more efficient to explore the diversity of yield values by travelling through different rows when yield gradient is perpendicular to the rows. Similarly, short rows always provide more flexibility to find short sampling routes. Indeed, they bring the possibility to access a large diversity of yield values quickly in a limited time (Figure 4.A) which always promote RBSR. For long rows, RBSR becomes time-consuming with no added information in exploring entire rows (Figure 4.C) which justify EBSR in this case. For the same reason, although this conclusion is not that consistent with the results, an increasing nugget effect may result in more heterogeneous yield values in the surroundings of the starting point, which logically promotes EBSR.

Note that operational hybrid sampling routes do exist for fields corresponding to more complex configuration. In this case, sampling route is largely based on RBSR which consists in a one-way round trip across two rows with one or more measurement sites coming from a third incompletely covered row added (Figure 2.C or Figure 3.B; which are the same). In contrast, Figure 3.A shows three incompletely-covered rows.

It should be kept in mind that the results of this study are based on simulated data, which represent a simplified version of reality. The errors of estimation exposed here are not indicative of what can be found in practice, the context here is a purely theoretical framework where the spatial distribution of the yield is fully known. For example, it was assumed that for each measurement site, the yield was fully known, as if all bunches of the plant had been weighed. Such a destructive approach is not realistic in a commercial situation because of measurement time and yield loss. In practice, the estimation of the yield on a site is itself the result of a sampling of one or two bunches chosen and weighed by the operator. The result of this process is an error in estimating the yield at each site and a resulting error in estimating the average yield of the field which is necessarily higher than that reported in this work. In this study, uncertainty in the representativeness of the sampling sites is taken into account by the nugget effect which corresponds to erratic variance caused among other things by the error in estimating the yield at each site. However, it remains unclear whether the range of variation chosen for the nugget effect in this work represents the impact of the diversity of yield estimation methods at the level of the measurement site.

Considerations discussed in this study are based on simple field characteristics. This simplified framework enables to identify the impacts of different parameters affecting sampling route. However, the characteristics of a real field are often more complex. For example, rows can have irregular length, fields can have irregular shapes, different sizes etc. Other elements can also affect travel time such as slopes or the presence of a discontinuity in the row structure allowing the practitioner to pass from one row to the other without having to walk all along it. Logistical issues may also count in the sampling route design. The intention of this paper is therefore not to give settled values to be respected but rather guidelines to consider to optimize yield sampling at a lower cost, and effort when information about the yield spatial structure is available. It is thus to be noted that simple and quick field observations such as the row length can be used to instruct the choice towards an EBSR or RBSR strategy. The row length is simple and available information that can be considered without additional cost. This can moreover be achieved without interfering with the decision on the trade-off between estimation error and sampling time, which is left to the practitioner's discretion. The starting point corresponds here to the fixed entry point for a given field. Its position has an influence on the distance to be covered to reach certain sites. When possible, adjusting its position could reduce the total sampling time. Note that the total sampling time also depends on the measurement time, which is not discussed here as this work focuses on minimising walking time. A proper sampling strategy should consider both walking time minimisation and suited measurement protocol.

Thus, based on the study of yield spatial structure, results shed light on some generic considerations when sampling for grape yield estimation. However, yield spatial structure is generally not known before sampling. Ancillary data *i.e.*, data that are correlated to yield can then be used for this purpose. These data are often chosen because they are readily available at higher resolution and at a lower cost than yield data. Vegetation indices such as NDVI (Carillo *et al.,* 2016) measured by satellite, UAV or aerial imagery and historical yield data (Araya-Alman *et al.,* 2017) are examples of auxiliary data which are already being considered for grape yield estimation. However, as the correlations between yield and ancillary data are specific to each field (Carillo *et al.,* 2016), the use of these data must be made based on field knowledge and local calibration as far as possible. Temporality must be considered as it might affect the correlation between variables. When fully known, an ancillary can then be used to directly drive measurement site selection according to the same considerations exposed in this paper and thus help in yield estimation before harvest.

The results obtained in this study are dependent on the sampling strategy used. In this case, this later aimed at selecting measurement sites that are representative of the yield distribution. The choice of this approach may explain why the proposed optimal routes are strongly influenced by the spatial structure of the yield and its organisation with respect to row orientation. However, most targeted sampling methods aim to consider the distribution of the variable to be estimated and may well lead to similar results. This study does not allow us to demonstrate this, however, the proposed methodology may well be used to evaluate sampling methods by simultaneously taking into account: the quality of the estimate made and the sampling effort.

Finally, these considerations on optimal routes for yield sampling may be applied to other variables of interest such as fruit maturation (Meyers *et al.,* 2020), Brix degree (Kasimatis and Vilas, 1985) or water status (Herrero-Langreo *et al.,* 2018). This study could also be extended to other crops associated with distance constrained by a trellised structure.

CONCLUSION

This work shows that to be optimal, a sampling route must be tailored to the characteristics of the field. The row length, as well as the spatial organization of the within-field yield variability, are factors that determine the optimality of a sampling route. This work opens up interesting perspectives. Indeed, the approach could be used to identify whether other factors affect the optimal definition of a sampling route (*e.g.*, by taking into account the slope and the resulting effort). Thus, this work could well be applied to real cases and propose optimal sampling routes by taking into account the actual length of the rows, the actual starting points corresponding to field access, the expected spatial organisation of the yield data based on previous knowledge (yield maps from previous years, multispectral images, soil electrical resistivity maps), etc. Beyond these practical aspects, this work also highlights the interest of spatial simulation in association with constraint optimisation, to provide insights for optimising the logistics of viticulture operations.

Constraint models could be adapted to fit real case studies. In particular, similar approaches could well be used to propose the optimization of machine routes to respond to economical as well as environmental issues such as the reduction of fossil fuel consumption.

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

REFERENCES

Araya-Alman, M., Acevedo-Opazo, C., Guillaume, S., Valdés-Gómez, H., Verdugo-Vásquez, N., Moreno, Y., Tisseyre, B. (2017). Using ancillary yield data to improve sampling and grape yield estimation of the current season. *Proceedings of the 11th European Conference on Precision Agriculture, Advances in Animal Biosciences 8(2),* 515-519. https://doi.org/10.1017/S2040470017000656

Arnó, J., Martínez-Casasnovas, J.A., Uribeetxebarria, A., Escolà, A., & Rosell-Polo, J.R. (2017). Comparing efficiency of different sampling schemes to estimate yield and quality parameters in fruit orchards. In: J.A. Taylor (Ed.) *Proceedings of the 11th European Conference on Precision Agriculture, Advances in Animal Biosciences 8(2)*, 471-476. https://doi.org/10.1017/S2040470017000978

Bramley, R. G. V., Ouzman, J., Trought, M. C. T., Neal, S. M., & Bennett, J. S. (2019). Spatio-temporal variability in vine vigour and yield in a Marlborough Sauvignon Blanc vineyard. *Australian Journal of Grape and Wine Research, 25*, 430-438. https://doi.org/10.1111/ajgw.12408

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

Clingeleffer, P.R., Martin, S., Krstic, M., & Dunn, G.M. (2001). Crop development, crop estimation and crop control to secure quality and production of major wine grape varieties. A National Approach: Final Report to Grape and Wine Research & Development Corporation. *Grape and Wine Research & Development Corporation* (Victoria, Australia: CSIRO and NRE).

Diago, M.P., Correa, C., Millàn, B., Barreiro, P., Valero, C., & Tardaguila, J. (2012). Grapevine Yield and Leaf Area Estimation Using Supervised Classification Methodology on RGB Images Taken under field conditions. *Sensors. 12*, 16988-17006. https://doi.org/10.3390/s121216988

Dunn, G., & Martin, S. (2004). Yield prediction from digital image analysis: A technique with potential for vineyard assessments prior to harvest. *Australian Journal of Grape and Wine Research*, *10*, 96-198. https://doi.org/10.1111/j.1755-0238.2004.tb00022.x

Herrero-Langreo, A., Tisseyre, B., Roger, J. M. & Scholasch, T. (2018). Test of sampling methods to optimize the calibration of vine water status spatial models. *Precision Agriculture volume 19*, 365–378. https://doi.org/10.1007/s11119-017-9523-8

Kasimatis, A. N. & Vilas, E. P. (1985) Sampling for Degrees Brix in Vineyard Plots. *American Journal of Enology and Viticulture 36*: 207-213. https://www.ajevonline.org/content/36/3/207.short

Lachia, N., Pichon, L., & Tisseyre, B., (2019). A collective framework to assess the adoption of precision agriculture in France: description and preliminary results after two years. *In: Precision agriculture'19., ED. John v. Stafford, Ampthill, UK, Wageningen Academic Publishers*. 851-857. https://doi.org/10.3920/978-90-8686-888-9_105

Meyers, J.M., Sacks, G.L., van Es, H.M., & Vanden Heuvel, J.E. (2011). Improving vineyard sampling efficiency via dynamic spatially-explicit optimisation. *Australian Journal of Grape and Wine Research, 17,* 306–315. https://doi.org/10.1111/j.1755-0238.2011.00152.x

Meyers, J.M., Dokoozlian, N., Ryan C., Bioni, C. & Vanden Heuvel, J.E. (2020). A New, Satellite NDVI-Based Sampling Protocol for Grape Maturation Monitoring. *Remote Sens. 2020, 12*, 1159-1170. https://doi.org/10.3390/rs12071159

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

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

Prud'homme, C., Fages, J.G., & Lorca, X. (2016). *Choco Documentation*. TASC, INRIA Rennes, LINA CNRS UMR 6241, COSLING S.A.S, http://www.choco-solver.org

Reis, M.J.C.S., Morais, R., Peres, E., Pereira, C., Contente, O., & Soares, S. (2012). Automatic detection of bunches of grapes in natural environment from color images. *Journal of Applied Logic*, 10, 285–290. https://doi.org/10.1016/j.jal.2012.07.004

Serrano, E., Roussel, S., Gontier, L., & Dufourcq, T. (2005). Estimation précoce du rendement de la vigne : corrélation entre le volume de la grappe de *Vitis Vinifera* en cours de croissance et son poids à la récolte. In: H. Schultz (Ed.) *Proceeding of the Groupe Européen d'Etude des Systémes de Conduite de la Vigne*, (pp. 311-318). https://www.infowine.com/intranet/libretti/libretto4713-01-1.pdf

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

Wolpert, J.A., & Vilas, E.P. (1992). Estimating vineyard yields: Introduction to a simple, two-step method. *American Journal of Enology and Viticulture*, *43*, pp. 384-388. https://www.ajevonline.org/content/43/4/384.short

Wulfsohn, D., Aravena-Zamora, F., Potin-Téllez, C., Zamora, I., & García-Fiñana, M. (2012). Multilevel systematic sampling to estimate total fruit number for yield forecasts. *Precision Agriculture*, *13*(2), 256-275. https://doi.org/10.1007/s11119-011-9245-2