Combining target sampling with route-optimization to optimise yield estimation in viticulture

B. Oger[1-2], P. Vismara[2-3] and B. Tisseyre[1]
[1] *ITAP, Univ. Montpellier, Montpellier SupAgro, Irstea, France*
[2] *MISTEA, Univ. Montpellier, Montpellier SupAgro, INRA, France*
[3] *LIRMM, Univ. Montpellier, Montpellier SupAgro, CNRS, France*
baptiste.oger@supagro.fr

This PDF file is a pre-print version of the final publication which can be found at:
https://doi.org/10.3920/978-90-8686-888-9_20

Abstract

This paper describes a new approach for yield sampling in viticulture. It combines approaches based on auxiliary information and path optimization to offer more consistent sampling strategies, integrating statistical approaches with computer methods. To achieve this, groups of potential sampling points, comparable according to their auxiliary data values are created. Then, an optimal path connecting several points, one from each group of potential sampling points and minimizing the route distance is constituted. This part is performed using constraint programming, a programming paradigm offering tools to deal efficiently with combinatorial problems. The paper presents the formalization of the problem, as well as the tests performed on real fields. Results show that combining target sampling and path optimization can reduce by 45% the average sampling circuit length compared to previous methods based on auxiliary data while being almost equivalent in yield prediction error.

Keywords: sampling optimization, yield estimation, model sampling, NDVI, constraint programming.

Introduction

In order to optimize harvest organization, prepare the winemaking process and establish commercial strategies, the wine industry needs to know the yield of each vine field. Ideally, yield has to be estimated a few days before harvest with a relative error of less than 10 %. Although models have been developed to forecast the yield at the regional level (Cristofolini and Gottardini 2000), their results were not precise enough to manage logistic issues linked to harvest operations at the farm or at the winery level. Therefore, precise estimation of vine field yield always requires fruit sampling and counting. This estimation must be carried out quickly (few minutes per field) at a time when the workload at harvest or for the preparation of the harvest is critical. Practical constraints, like the time available to visit all the fields before harvest, limit the number of sampled sites per field. Therefore, yield estimation is based on a low number of sampling sites (4/5 per field) where yield components (number of clusters, number of berries per cluster, mean berry weight) are manually measured by a practitioner. Due to these practical constraints and the high within-field variability of grape yield usually observed, the small number of observation results in high errors in yield estimation (generally around 20 to 30%).

Recent works (Carillo et al. 2016) have shown the interest of integrating auxiliary data to improve sampling strategies and yield estimation for perennial crops. Among possible auxiliary information, vegetation indices derived from multispectral airborne images is of great interest since they can be used to characterise the spatial variability of several fields; in one acquisition, with a high spatial resolution (< 1 m.) and at an optimal date. In viticulture, Carrillo et al., (2016) showed the potential of normalised difference vegetation index (NDVI) to drive target sampling of the main grape yield components (e.g. bunch number, berry weight) to improve yield estimation. They demonstrated the value of using NDVI information to determine relevant within-field sampling sites selection based on the distribution of NDVI values.

Although interesting, the methodology proposed by Carrillo et al. (2016) presents a significant drawback. Indeed, it does do not take into consideration the relative position of the sites to be sampled, and the fact that vine fields are structured in rows. This peculiarity implies that rows cannot be crossed, leading to sampling plans optimized in terms of prediction but potentially unrealistic in terms of sampling routes and resulting travelled distance (and time) for the operator.

This paper proposes a new approach to optimally design within-field sampling routes which take into account the spatial organisation of the crop (rows) and spatial location of sampling sites. The originality of the approach, called constraint sampling, is to combine statistical and computer methods. It can be decomposed into two steps. In the first step, potential sampling sites are sorted into different groups according to their auxiliary data value in a similar way to traditional targeted sampling. The second step finds an optimal route that passes through one sampling site from each group. A Constraint Programming solver is used to build an optimal route in terms of travelled distance. This kind of solver has already been used in precision viticulture to solve the differential harvest problem (Briot et al. 2016).

Materials and methods

Sampling sites and selection principles

The purpose of *constraint sampling* is to select N sampling sites constituting a sampling route in the vine field. Accounting for classical sampling practices in viticulture, N will vary between 5 and 10. It is assumed that there is a finite number of sites on the plot where sampling can be carried out, these sites are called potential sampling sites. For instance, in the data presented in this article, a potential sampling site is defined every 15m. The sampling sites are then chosen from the list of potential sampling sites. In order to be able to apply selection methods based on auxiliary data, an NDVI value must be associated with each potential sampling point. For each potential site, the method assumes that: i) the co-ordinate of the potential site, ii) the row that the potential site belongs to and iii) the corresponding NDVI values are known.

The method requires a distance matrix to be computed. This matrix gives the distance between each couple of potential sampling sites. This distance must take into account the structure of the vineyard. It corresponds to the shortest walking distance between two sites. If the points are in the same row, it corresponds to classical Euclidian distance. If they belong to different rows, the distance is computed considering that the practitioner has to leave the row, reach the desired rows passing by all extremities of intermediate rows and finally reach the targeted sampling site. As rows have two extremities, two different distances can be computed and the shortest one is kept.

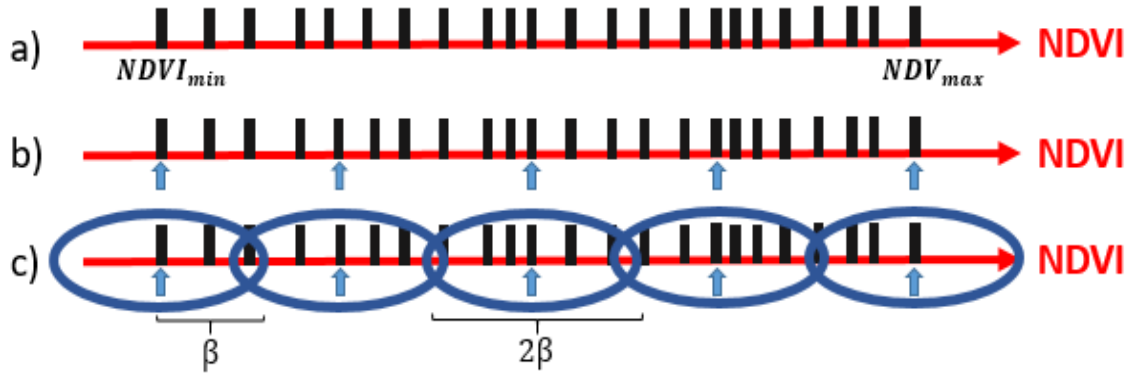

Figure 1: choice of potential sampling sites (for N = 5 sampling sites) based on NDVI values with method (iii) adapted from Kennard & Stone (1969); a) Distribution of observed NDVI values (each black dash represents one NDVI value). b) Selection of values corresponding to a potential sampling (arrows) and c) Groups of potential sampling sites built to account for the distribution of NDVI values, the width of groups is controlled by β

To consider auxiliary data efficiently, the idea of *model sampling* proposed by Carillo et al. (2016) was considered. The *model sampling* approach aims at calibrating a linear regression which relates the yield to auxiliary data (NDVI). This author also shown that yield components, especially berry weight, were linearly related to NDVI. Therefore, sampling sites can be selected representatively to build this linear model. This model is then used to estimate yield using all available high-resolution NDVI data.

The approach proposed in this paper relies on the following principle. Potential sampling sites are split into N homogeneous groups. Once groups are formed, one element from each group is selected in order to optimize the length of the route connecting all these points. It will ensure a good repartition of selected points by accounting for auxiliary data distribution, which is a key element to build a linear model with very few points (Kennard and Stone 1969). Three different ways to create these groups were tested:

(i) The first method relies on quantiles. If K is the number of potential sampling points, then each group has $\frac{K}{N}$ elements. A first group will contain the $\frac{K}{N}$ potential sampling points with the lowest NDVI values. A second one the $\frac{K}{N}$ potential sampling sites with NDVI values just above those of the first group and the last group contains the $\frac{K}{N}$ highest NDVI values.

(ii) The second method uses the K-means algorithm. This clustering algorithm is efficient for partitioning K elements into N groups. The main principle of K-means is to minimize the difference between points in the same groups and thus maximise the difference between a group's mean.

(iii) The last method is derived from the Kennard & Stone (1969) approach. As described by these authors, this approach selects elements by iteratively choosing a new sitethat is the furthest from the sitesalready selected in terms of auxiliary data (Figure 1b). The approach was adapted to create groups, one centred around each value selected by Kennard & Stone approach. A parameter called β, expressed as a percentage of the NDVI range, set the width of the groups (Figure 1c). Using this method, groups may overlap with each other or, on the contrary, some sites may not belong to any group. This

depends on the number of groups (N) and their length (β). Figure 1 illustrates the method with potential sampling sites projected on an NDVI axis.

The second step of the approach consists in selecting one sampling site per group. These N sampling sites must be all different and have to form the shortest possible sampling circuit. There are numerous possible choices to select these sampling sites and many ways to order them to form a circuit. It is therefore a highly combinatorial optimization problem. Constraint Programming is one of the programming paradigms able to deal with such problems. It aims at solving a problem expressed as a set of variables and a set of constraints on these variables. Such a problem is called a Constraint Satisfaction Problem (CSP). A Constraint Solver is used to find a solution to the problem that satisfies all the constraints. The efficiency of these solvers relies on the implementation of many methods such as filtering, which allows quick detection of combinations of values that do not lead to an optimal solution. The interest of constraint solvers lies in their ability to address many types of constraints.

Without going into small detail, let $S = \{1, \dots, K\}$ be the set of potential sampling sites and $\{G_i\}_{i\in(1,\dots,N)}$ the set of groups covering S, formed in the previous step. For decomposition (i) and (ii) all groups are disjoint and $\{G_i\}_{i\in(1,\dots,N)}$is a partition of S. P_i is defined as the selected site for group G_i. The first constraint imposes that all P_i must be different (this constraint is immediately satisfied in the case of methods (i) and (ii)). P_0 represents the point of departure and arrival; it is a fixed parameter representing the initial position of the practitioner. The length of the optimum route passing through all the $P_{i\in(0,\dots,N)}$ must be a minimum. This is a particular case of the vehicle routing problem (VRP) where the goal is not to find a Hamiltonian tour (visiting once every site) but a tour covering only a subset of sites. Recent work about the *WeightedSubCircuit* constraint (Vismara et al. 2018) has proposed a filtering algorithm that is well adapted to address this type of situation. All these constraints and variables constitute the constraint satisfaction problem. An instance of this problem is built from each dataset and solved with the solver in order to get an ordered set of sites that form a sampling circuit. The program returns the list of sampling sites, the order in which they are visited and the associated distance.

Yield estimation

The aim is to estimate Y, the average grape weight (GW) per vine. For each site selected by the sampling method ($s \in selected$), $GW(s)$, the observed grape weight per vine value, is available. A linear model linking the NDVI to GW is built from these sites (Eq. 1):

$$\widehat{GW}(s) = a \times NDVI(s) + b \qquad (1)$$

For a given site s, $\widehat{GW(s)}$ represents an estimate of $GW(s)$. The parameters a and b are obtained from a linear regression on the N sites selected by the sampling method.

With $S = \{1, \dots, K\}$ being the full set of potential sampling sites available, $\widehat{\overline{GW}}$, the estimate of $\overline{GW}$, can be computed from the model using all these potential sites (Eq. 2):

$$\hat{Y} = mean_{s\in S}(\widehat{GW}(s)) \qquad (2)$$

Estimation error

The estimation error is a deviation from the actual yield value (Y), expressed as a percentage (Eq. 3).

$$Error\ (\%) = \frac{|Y - \hat{Y}|}{Y} \quad (3)$$

Reference methods

The method is compared to two references:

- A conventional *random sampling* where the N sampling sites are randomly selected. $\widehat{\overline{GW}}$ is directly estimated from the mean of observed GW values. Here, $selected$ represent the N sampling sites chosen randomly (Eq. 4). *Random sampling* represents what is generally done in practice in terms of yield estimation

$$\hat{Y} = mean_{s \in selected}(GW(s)) \quad (4)$$

- A *model sampling* whose method principles have been described by Carillo et al. (2016). Sampling sites are chosen according to NDVI values. One site is randomly selected for each of the N NDVI quantiles. *Model sampling* uses a model based on the NDVI/yield relationship, as described in Eq. 1. The main difference is that *model sampling* does not consider the spatial position of the selected sampling sites.

To compare the length of the routes between the different methods, the optimal route between the selected sites must be computed for the reference methods. This is done with the Concorde TSP solver. As for *constraint sampling*, the route includes the starting point of the practitioner (P_0).

Experimental data

The data used to test the method came from INRA Pech-Rouge (Narbonne, France). The experiment and the database were detailed by Carrillo et al., (2016). It is briefly summarised hereafter. NDVI values from 9 different vine fields were considered. All of them are non-irrigated and exposed to Mediterranean climate with precipitation occurring during spring and with hot and dry summers. The characteristics of each plot are shown in Table 1.

Table 1. Description of the 9 fields

Field (Id)	Area (ha)	Variety	Row Spacing (m)	Vine Spacing (m)	Potential Sampling Sites
P22	1.72	Syrah	2.5	1	45
P63	1.33	Syrah	2.5	1	42
P65	0.69	Syrah	2.5	1	33
P76	1.14	Carignan	2.25	1.5	37
P77	1.24	Syrah	2.5	1	19
P80	0.54	Syrah	2.5	1	40
P82	1.15	Syrah	2.5	1	53
P88	0.85	Syrah	2.25	1.5	21
P104	0.81	Carignan	2.25	1.5	23

NDVI values were derived from a 1 m. resolution multi-spectral image taken the 31th of August 2008 by Avion Jaune (Montpellier, Hérault, France). The spectral regions captured in the images were: blue (445–520 nm), green (510–600 nm), red (632–695 nm) and near-infrared (757–853 nm). From these, 1 m square image pixels, aggregation method described in Carillo et al. (2016) was used to obtain 9m square image pixels reducing the effect of canopy discontinuity and bare soil on measured values. NDVI was finally computed from processed images. Mechanical or chemical weeding was performed over the inter-row spacing; therefore, row cover crop did not affect NDVI values.

Sampling sites were selected regularly over the fields with measurement made on each node of a 15m² width sampling grid. At each node, yield components [bunch number per vine (BuN) and bunch weight (BuW)] were measured in 2009. Each site was considered as 5 consecutive vines in the row. BuW was estimated at harvest by weighing 10 bunches (2 bunches per vine) also randomly taken from the same 5 consecutive vines. BuN was determined by counting all bunches of the 5 consecutive vines of each sampling point. Grape weight per vine (GW) was then calculated from BuW and BuN. The distance between vines along the row was 1m or 1.5m. Data were associated with the spatial co-ordinates of the central vine. The final data base was a set of 313 sites over the 9 different fields. The number of sites per field varied from 19 to 45 sampling sites. Each site was then characterized by GW as field parameter and NDVI values. NDVI value was assigned to each site as the mean of 4 pixels corresponding to a square of 36m².

Actual yield values measured at harvest are not available. For each field, the average of all available measured GW values is then used as the reference yield value (Y) when computing estimation error (Eq. 3).

Implementation

The core of the approach was written in java and used the Choco solver (Prud'homme et al. 2016). The calculations to obtain the distance matrix, groups of individuals classified according to their NDVI value, estimation errors and route distance were made with R.

As explained in the description of the constraints, the approach presented here takes into account the starting point of the practitioner (P_0) which is included in the sampling circuit. Varying the starting point thus changes the sampling route. In order to increase the number of situations tested, this starting point was positioned on different ends of row across the vineyards. The approach was then applied to 86 situations instead of 9.

Results and discussion

Evaluation of sampling strategies

Figure 2a and 2b show the results of estimation errors and sampling route distance observed for the different sampling methods. Remember that "i-quantile", "ii-kmeans", "iii-kennard β=10" and "iii-kennard β=15" refer to the *constraint sampling* methods (i.e. methods that account simultaneously for auxiliary data distribution and distance between sampling points) while *model sampling* and *random sampling* refers to methods that account only on auxiliary data distribution. "iii-kennard β=10" and "iii-kennard β=15" are based on the same approach with different group widths (β=10% & β=15%). Results for the different starting points are averaged for each field and then all together to give the same weight to each field.

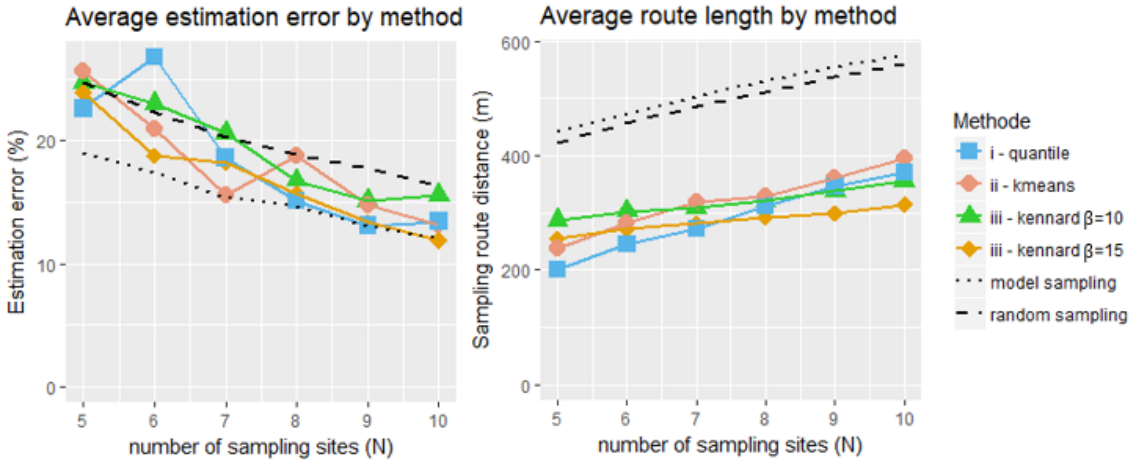

Figure 2a. (Estimation error %) and 2b. (Sampling route distance): Results and comparison to reference samplings

Figure 2a shows that all the methods follow the same trend with a decreasing error as the number of sampled sites increases. This result is logical, and consistent with the literature. As Carrillo et al (2016) have already shown, taking into account auxiliary data (*model sampling*) slightly improves the quality of yield estimation compared to a *random sampling*. Despite higher variability in the observed error, the integration of constraints does not increase estimation errors, the methods (i), (ii), (iii) allow, in most cases, to maintain lower errors than *random sampling*. Kennard and Stone decomposition with β=15 may be the best option when creating the N groups, the results could match those of *model sampling* on most of the cases. Note however that observed errors with constraint methods are higher than for *model sampling* (i.e. without constraints). This result may be logical considering that the addition of the constraints may lead to the choice of less optimal sites within the groups of potential sites. Also, the irregularity of the curves associated with *constraint sampling* can be explained by a smaller number of experiments. These curves are based on 86 results compared to the 1,000 repetitions considered for reference methods, resulting in a higher variability.

Figure 2b clearly illustrates the gains brought by *constraint sampling* in terms of travel distance across the vineyard. Logically, the travelled distance within the plot increases linearly with N, the number of sampling sites visited. The four curves representing *constraint sampling* are at the same level, with a reduced distance of about 45% compared to *model sampling* and *random sampling*. Overall, this method offers a good compromise between the quality of the estimate and the travel constraint on the plot.

Applying the approach to new data could consolidate the results presented here. It would also be interesting to test the method with plots of vines having different characteristics (shape, size), under different cultivation practices (weed management between rows) or with different auxiliary data available (e.g.: historical yield).

This is a first model that can still be improved. Increasing the number of usable auxiliary variables or allowing the method to adjust directly as the first sites are selected for instance, could improve the accuracy and quality of the results. From an efficiency point of view, improvements in the Constraint Programming model could reduce computation times.

Computation times

Computation times increase with the number of possible combinations. The higher K and N (the number of potential sampling sites and the number of sites to be selected respectively), the longer it will take. The way groups are created also affects the computational time. For instance, when using the Kennard and Stone approach to build a group, an increase of the β parameter (group width) can consistently increase computation times. In general, for plots with N<8, the computation times are in the order of a second. It took about a few hours in the most complex cases.

Conclusions

The methodology presented in this paper described a new approach for yield sampling in viticulture. The originality of the approach comes from the association of a previously published method based on auxiliary data and optimisation algorithms to propose relevant sampling routes in term of estimation error and travelled distance. While the *model sampling* principle guides sampling choice considering auxiliary information, optimisation through constraint programming ensures the relevancy of the chosen route in term of walking distance for the practitioner. Results presented here are of course preliminary results.

As available time is often the principal constraint for growers, they tend to rely on random samplings limited to a small part of the vineyard. Integrating spatial aspect accounting for travelling constraints is a key element to propose new methods that are relevant for field application. Further tests should be considered to confirm these first results and identify the limitations of the approach.

Acknowledgements

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

References

Briot, N., Bessiere, C., Tisseyre, B. & Vismara, P. (2015). Integration of Operational Constraints to Optimize Differential Harvest in Viticulture. Proceed. 10th European Conference on Precision Agriculture (ECPA 2015), 487–494.

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.

Cristofolini, F. & Gottardini, E. (2000). Concentration of airborne pollen of Vitisvinifera L. and yield forecast: a case study at S.Michele all'Adige, Trento, Italy. Aerobiologia, 16(1), 125–129.

Kennard W., and Stone L. A. (1969). Computer Aided Design of Experiments. Technometrics, 11, 137-148.

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/

Vismara P. & Briot N. (2018). A Circuit Constraint for Multiple Tours Problems. Proceed. 24th International Conference on Principles and Practice of Constraint Programming (CP 2018). Lecture Notes in Computer Science. Vol.11008,389–402.