Advances in Animal Biosciences: Precision Agriculture (ECPA) 2017, (2017), **8:2**, pp 515–519 © The Animal Consortium 2017
doi:10.1017/S2040470017000656

advances in animal biosciences

Using ancillary yield data to improve sampling and grape yield estimation of the current season

M. Araya-Alman[1,3†], C. Acevedo-Opazo[1], S. Guillaume[3], H. Valdés-Gómez[2], N. Verdugo-Vásquez[1,3], Y. Moreno[4] and B. Tisseyre[3]

[1]*CITRA, Universidad de Talca, 2 Norte, 685 Talca, Chile;* [2]*Pontificia Universidad Católica, Av. Vicuña Mackenna 4860, Santiago, Chile;* [3]*UMR ITAP, Montpellier SupAgro, Irstea, 2 Place Pierre Viala, 34060 Montpellier, France;* [4]*CTVV, Universidad de Talca, 2 Norte, 685 Talca, Chile*

*This paper proposes a methodology aiming at using historical yield data to improve yield sampling and yield estimation. The sampling method is based on a collaboration between historical data (at least three years) and yield measurements of the year performed on some sites within the field. It assumes a temporal stability of within field yield spatial patterns over the years. The first factor of a principal component analysis (PCA) is used to summarize the stable temporal patterns of within field yield data and it represents a large part of the variability of the different years assuming yield temporal stability and a high positive correlation between this factor and the yield. This main factor is then used to choose the best sites to sample (target sampling). Yield measurements are then used to calibrate a model that relates yield values to coordinates on the first factor of the PCA. This sampling method was tested on three vine fields (*Vitis vinifera L.*) in Chile and France with different varieties (Chardonnay, Cabernet Sauvignon and Syrah). For each of these fields, yield data of several years were available at the within field level. After temporal stability of yield patterns was verified for almost all the fields, the proposed sampling method was applied. Results were compared to those of a classical random sampling method showing that the use of historical yield data allows sampling sites selection to be optimized. Errors in yield estimations were reduced by more than 10% in all the cases, except when yield stable patterns are affected by specific events, i.e. early frost occurring on Chardonnay field.*

Keywords: historical yield data, principal component analysis, sampling optimization

Introduction

To plan harvest, optimize quality management and limit operation costs, the wine industry needs to know the yield of each vine field a few days before harvest with a relative error less than 10%. Precise estimation of vine field yield always requires intensive fruit sampling and counting. Yield estimation must be carried out quickly at harvest time when the workload is important. Practical constraints, like the time available to visit all the fields before harvest, limit the number of sampling sites. Regarding the significant grape yield variability usually observed at the field level (Taylor *et al.*, 2005), errors resulting from this estimation are usually higher than expected.

Recent works propose the use of auxiliary data like NDVI (Carrillo *et al.*, 2016) to optimize sample locations. They have shown that target sampling based on spatial patterns highlighted by high spatial resolution auxiliary data improves yield estimation. Considering the development of yield monitoring systems (Taylor *et al.*, 2016) and temporal stability of yield patterns usually observed at within-field level for grapes (perennial cultivation) (Tisseyre *et al.*, 2008), yield data of previous years may constitute a relevant source of information to improve sampling and yield estimation.

This work aims at investigating the relevance of historical yield maps to improve yield prediction of the current year. It proposes an approach based on the collaboration between historical data and yield measurements performed at some sites within the field. The main objective of this paper is to present a method that accounts for historical yield maps (multi-dimensional aspect of the problem) to define an optimised target sampling. The method was applied on three vine fields from Chile and France. To demonstrate the relevance of the proposed approach, yield estimation results are compared with those arising from a conventional method based on a random drawing of the sampling sites.

Materials and methods

Sampling method and yield estimation

The proposed method assumes a temporal stability of yield spatial distribution over the years, i.e. sites having higher or lower yield compared to the mean yield for a given year are

† E-mail: miaraya@utalca.cl

always the same in other years. In addition, the proposed approach assumes that each field is managed under homogeneous climatic conditions, cultivar, and any other practices that could affect growth and vine yield.

To remove year effect on yield and to make grape yield for different years comparable, yield data were standardized as stated in Eq. (1):

$$\tilde{Y}_j(i) = \frac{Y_j(i) - \overline{Y}_j}{\sigma_j} \quad (1)$$

where, $\tilde{Y}_j(i)$ is the standardized yield of each site *i*, $Y_j(i)$ is the historical yield value in year *j* for site *i*, $\overline{Y}_j$ is the mean yield in year *j*, σj is the standard deviation in year *j*.

As a consequence, each yearly within-field yield data has the same mean and variance. Thus, every year is comparable to others.

A principal component analysis (PCA) was used to summarize the information of all available years for a given field. The assumption is that the first factor explains a significant amount of the variance, and thus reflects a temporal pattern over the years (all years are positively correlated with the principal factor). Once this assumption has been checked, the scores of the principal factor can be used to rank all the sites according to the yield value observed at these sites. Sites with a negative (resp. positive) score regarding Factor 1 have usually a yield value lower (resp. higher) than the mean yield of the field (for all the years). Then, the scores on the axis associated to Factor 1 are stratified to select the most representative sampling sites.

The score distribution is divided into *n* quantiles (*n* corresponding to the number of sampling sites). In each quantile, the closest site to the Factor 1, i.e. that with the lowest absolute score regarding Factor 2, is selected as a sampling site (Fig. 1).

Then to provide a yield estimate for the new year ($j+1$), the following steps must be considered:

i. Make a measurement at the *n* sampling sites for the current year $j+1$;
ii. Estimate the coefficients *a* and *b* of the linear model (Eq. 2) that determines the relationship between the yield in $j+1$ and the stable temporal structure.

$$Y_n^{j+1} = a * F1(n) + b \quad (2)$$

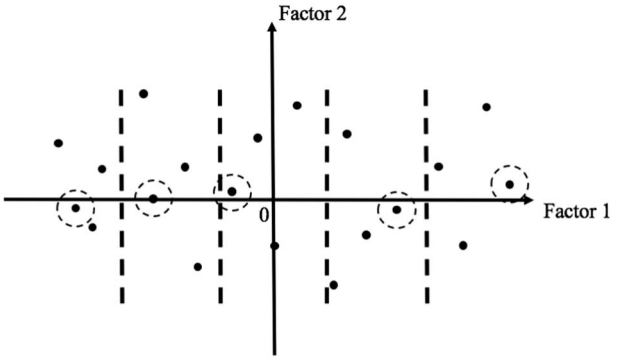

Figure 1 Selection of the most representatives sampling sites (n = 5) using the PCA analysis.

where Y_n^{j+1} is the yield measured at the *n* sites in year $j+1$, $F1(n)$ are the scores of the *n* sites on factor F1 and, *a* and *b* are the coefficients of the linear model.

iii. Estimate the yield across all sites *i* for the incoming year, $\hat{Y}_i^{j+1}$, using the linear model defined in Eq. 2.
iv. Finally, the mean field yield ($\widehat{\overline{Y}}^{j+1}$) is estimated with the mean of the yield estimates at each site *i*, with *I* being the total number of sites in the field (Eq. 3).

$$\widehat{\overline{Y}}^{j+1} = \frac{\sum_i \hat{Y}_i^{j+1}}{I} \quad (3)$$

Tests were performed with a number of sites varying from $n=2$ to $n=10$.

Evaluation of quality estimation

For a given plot, a given year *j* and a given number of samples *n*, the estimation error is calculated according to equation (4).

$$\text{Error} = \frac{\sqrt{\left(\hat{Y}_J - Y_j\right)^2}}{Y_j} \quad (4)$$

where $\hat{Y}_j$ is the yield estimate in year *j* using *J* historical yield maps except that related to year *j*, Y_j is the true yield computed with the mean of all available yield values of the year *j*.

If *J* years were available in the database, it was therefore possible to compute *J* errors values leaving out one year at a time. The average of all the observed errors was then computed. Results of the sampling approach were compared to those of a commonly used sampling method based on a random selection of *n* sites among all the available sites. Mean field yield is then computed from the yield observations of the *n* sites. To avoid a very good or very bad yield estimation eventually caused by the random choice of the sampling sites, a bootstrap method was applied (Efron, 1979). More specifically, *n* sampling sites were selected randomly and the estimated mean grape yield ($\hat{Y}_b$) corresponding to the bootstrap sample *b* was calculated, *b* being composed of the *n* sampling sites. This process was repeated *B* times, which provided *B* bootstrap samples. Bootstrapping was implemented with *B = 1000*. The estimated mean field yield was then computed as indicated in Eq. 5.

$$\widehat{\overline{Y}}_j = \frac{1}{B}\sum_{b=1}^{B} \hat{Y}_b^J \quad (5)$$

The estimated variance of the considered sampling method was defined as indicated in Eq. 6.

$$V(\hat{Y}_J) = \frac{1}{B}\sum_{b=1}^{B} (\hat{Y}_b^J - \widehat{\overline{Y}}_J)^2 \quad (6)$$

The error in % was derived from the estimated variance $V(\hat{Y}_j)$ and the estimated mean field yield (Eq. 7).

$$\text{Error}\,(\%) = \frac{\sigma(\hat{Y}_j)}{\widehat{\overline{Y}}_j} \times 100 \text{ with } \sigma(\hat{Y}_J) = \sqrt{V(\hat{Y}_J)} \quad (7)$$

Both methods were performed with a number of sites varying from $n=2$ to $n=10$.

Table 1 *Principal characteristics of the fields under study*

Cultivar	Area (ha)	Sampling sites	Date of plantation	Mean yield (kg.vine^{-1})	Number of years
Syrah	1.2	30	1991	1.8	7
Chardonnay	1.7	19	1994	3.0	5
Cabernet Sauvignon	1.6	18	1998	2.2	4

Data used to validate the method

The study was carried out in commercial vineyards located in Chile and France. In the southern part of France (Gruissan-Aude, RGF93 datum, Lambert93, E:709800, N:6226840) the experiment was conducted on *cv.* Syrah during 7 years (1999 to 2005). In Chile, the experiment was carried out in the Maule Valley (Talca, WGS84 datum, 35°22.2'S, 71°35.39'W). The first field was planted with a *cv.* Chardonnay and was monitored during 5 seasons (2011 to 2016), while the second one (*cv.* Cabernet Sauvignon) was monitored during 4 seasons (2009 to 2013). The *cv.* Chardonnay was affected by a severe frozen event that just modified the spatial structure of the yield. For all the experimental fields, yield per vine was manually measured on each sampling site (4 to 5 vines per site). Table 1 show the principal characteristics of each experimental field.

Tools

Data mapping was performed with the 3D-Field software (version 2.9.0.0., Copyrigtht 1998–2007, Vladimir Galouchko, Russia). Two classes were considered for each yield map: low (0.4–1.69 kg per vine) and high yield (1.7–3.5 kg per vine), each class gathering 50% of the observations (two quantiles). The proposed method was implemented and executed in MATLAB R2014a (The MathWorks, Inc., Natick, Massachusetts, USA).

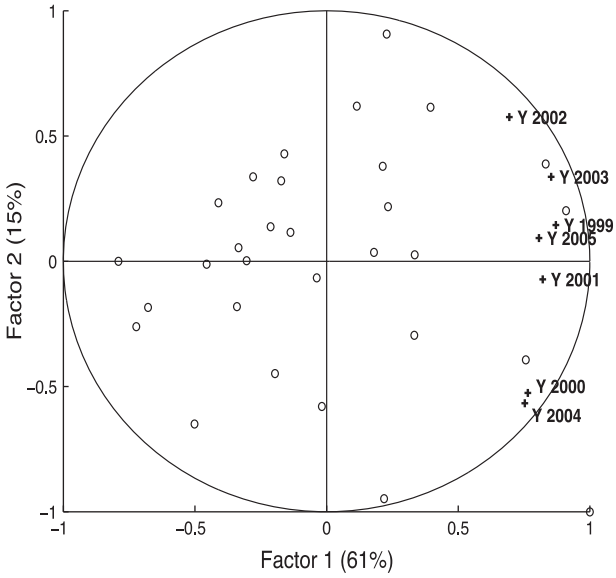

Figure 2 Scatter plot and correlation coefficients of the principal component analysis (first two factors) with data centred and reduced according to a per field basis cv. Syrah. White circles represent the sampling sites.

Results and discussion

Temporal stability of spatial yield patterns

Figure 2 reports the results of the PCA for the *cv.* Syrah only. Factors 1 and 2 account the 61% and 15% of the yield variation, respectively. Factor 1 is strongly correlated to the yield observations for most of the seasons. This factor effectively explains a significant proportion of the yield spatial variability for all the seasons. As a result, the distribution of individuals (sampling sites) on Factor 1 represents the average distribution of yield values for each of the 7 seasons. Sites that consistently exhibit low yields are located on the left side of Factor 1 (negative coordinates), while sites with consistent high yield values can be found on the right side of Factor 1 (positive coordinates). Similar results (not shown) were observed for the other fields under study (Cabernet Sauvignon and Chardonnay).

Note that yield observations for some years are more correlated to Factor 1 than others (Fig. 2). Yield observations for the years 1999, 2001, 2005 and to a lesser extent 2003, exhibit a high correlation with Factor 1. For these seasons, yield distribution is similar and close to the mean distribution. These years can be considered as 'typical'. Other seasons, especially 2000, 2002 and 2004, although highly correlated with Factor 1, deviates from this typical behaviour. This may be due to environmental factors that could have strongly influenced on yield patterns. These years can be considered as 'less typical'. The identification of the environmental factors that might explain this deviation is out of the scope of this study. Note however that these factors may have had the same effect on yield distribution in 2000 and 2004 since these two years are highly correlated. A very similar trend was observed for *cv.* Chardonnay during 2014. In this case, the observed deviation was more significant due to a strong frost event that affected most of the central zone in Chile. This frost causes a change in the 'typical' spatial organisation of within field yield.

These results highlighted the relevancy of Factor 1 to model the 'typical', i.e. temporal stability of the spatial yield variability for a particular field. Factor 1 could then be used to choose the best sites to sample (target sampling) and to calibrate a historical yield-based model.

Spatial validation of the method

Figure 3 shows the relevancy of the proposed approach to estimate yield over the entire plot using specific sampling

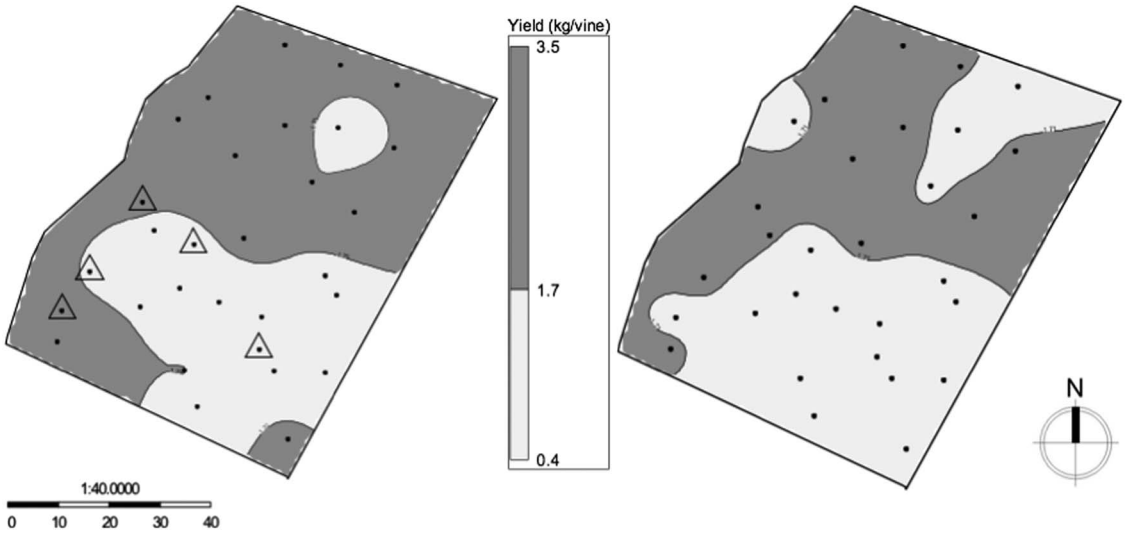

Figure 3 Observed (left) and estimated (right) yield maps of cv. Syrah (year 2001). Estimated map was derived from historical yield data of years 1999, 2000, 2002, 2003, 2004 and 2005, and yield data of 5 sites in 2001. Triangles represent the specific sampling sites.

Table 2 *Yield estimation errors (%) associated to the proposed method and the random sampling based methodology for different number of sampling sites (n = 2 to n = 10)*

		Errors (%)								
Cultivar	Method of estimation	n = 2	n = 3	n = 4	n = 5	n = 6	n = 7	n = 8	n = 9	n = 10
Syrah	Random	26.4	24.0	22.2	20.8	19.8	18.8	18.0	17.3	16.7
	Proposed method	18.4	17.1	19.2	9.4	8.8	10.1	8.8	9.5	5.8
Chardonnay	Random	23.6	21.3	19.7	18.4	17.5	16.6	15.9	15.3	14.8
	Proposed method	21.7	12.8	12.2	9.7	14.9	8.3	13.1	10.1	7.6
Cabernet Sauvignon	Random	24.9	22.5	20.8	19.6	18.6	17.7	16.9	16.3	15.7
	Proposed method	21.7	22.7	10.9	7.5	6.4	4.3	4.3	1.7	3.0

sites and a historical database of yield observations. The estimated yield map was created with five sites according to the previously described methodology. These five sites were then used to provide yield estimates across all the sites. Both maps (observed and estimated) exhibit similar magnitude of yield variation as well as very similar spatial patterns. Indeed, the correlation between the observed yield and the estimated yield by the proposed approach was found high ($r = 0.83$). The proposed approach was proved relevant because a high degree of consistency was reached with regard to the observed yield values while using a low number of measurement sites.

Standardization make grape yield for different years comparable, therefore was possible to estimate the yield of an intermediate year (2001) using data from others years, including the last years (Fig. 3). These results show that yield data of previous years may constitute a relevant source of information to improve sampling and yield estimation. This proves the usefulness of historical yield maps for yield prediction.

Quality of the estimates with the model

Table 2 reports the accuracy of both methods, i.e. the proposed method and the random-based method, regarding their ability to properly estimate vine yield values. Using between two and ten sampling sites, results show that the proposed approach helps reduce yield estimation error by more than 10% in average. Regarding the *cv.* Syrah, it was possible to reach an estimation error close to 9% with five sampling sites ($n = 5$). It was not possible to obtain such an accurate estimate using the random-based method, even with ten sampling sites ($n = 10$).

For the *cv.* Chardonnay, the proposed approach also generated better yield estimates than the random-based method. However, in this case, yield estimation errors are much more erratic. It does not exhibit a particular trend with an increasing number of sampling sites. This result may be due to the severe frozen event that took place in 2013 affecting the within field yield spatial structure.

Yield for the *cv.* Cabernet Sauvignon was the best estimated by the proposed methodology. This corresponds to the field for which the highest differences in estimation errors were observed between both approaches. Indeed, the proposed approach generated estimates errors lower than 10% when $n \geq 5$ sampling sites were used, while the random-based method generated estimation errors near 20%.

However, results also show that when an event modifies the spatial structure of yield (i.e. frost event on Chardonnay field), the proposed sampling method becomes less relevant. This aspect constitutes one of the main limitations of the proposed approach. Indeed, any factor that induces a change in the spatial structure of yield can significantly alter results. In our database, frost event that occurred on *cv.* Chardonnay is a relevant example. Other factors of the environment (hail, wild life damage, etc.) may result in a similar limit, as well as management factors (variable rate fertilisation, variable rate irrigation, etc.). A future refinement of the method could be in detecting atypical years applying principal component analysis.

Conclusion

The methodology presented in this study demonstrated how historical yield datasets could help improve vine yield estimations at the within field level. This is a consequence of the optimization of sampling sites selection, which significantly reduced yield estimation errors compared to random vine sampling. The approach was proved to be very efficient on several vine fields grown under different climate conditions, management systems and cultivars. Currently, this methodology is not used by the vine growers. However, this approach could be considered in a near future due to the reduction in, (i) cost and time regarding yield sampling and, (ii) yield estimation errors. This methodology might also help add value to high resolution yield data provided by grape harvesting machines. Finally, the methodology could be tested on other variables such as NDVI or soil apparent electrical conductivity. This would imply important contributions to the wine industry.

Acknowledgements

This study was financially supported by National CONICYT Doctoral Fellowship 2015 N° 21151630, Chile, and by the experimental unite Pech Rouge, France. The authors also would like to Corentin Leroux for his contribution to this work.

References

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

Efron B 1979. Computers and the theory of statistics: thinking the unthinkable. SIAM review 21 (4), 460–480.

Taylor JA, Tisseyre B, Bramley RGV and Reid A 2005. A comparison of the spatial variability of vineyard yield in European and Australian production systems. In: Precision Agriculture '05. Proceedings of 5th ECPA, Uppsala, Sweden, June 8–11. JV Stafford (ed). Wageningen Academic Publishers pp. 907–914.

Taylor JA, Sánchez L, Sams B, Haggerty L, Jakubowski R, Djafour S and Bates TR 2016. Evaluation of a commercial grape yield monitor for use mid-season and at-harvest. OENO One 50 (2), 57–63.

Tisseyre B, Mazzoni C and Fonta H 2008. Within-field temporal stability of some parameters in viticulture: Potential toward a site specific management. Journal International des Sciences de la Vigne et du Vin 42 (1), 27–39.