USE OF CLUSTER REGRESSION FOR YIELD PREDICTION IN WINE GRAPE

Rodrigo A. Ortega and Luis E. Acosta
Departamento de Industrias Universidad Técnica Federico Santa María
Av. Santa María 6400, Vitacura, Santiago, Chile

Luis A. Jara
Neoag Agricultura de Precisión
Napoleón 3565, Of. 202, Las Condes, Santiago, Chile

ABSTRACT

Yield prediction is an essential component of the production chain of wineries. Accurately knowing, in advance, the amount of grapes being produced is crucial to establishing a proper logistic. Yield prediction models based on field and ancillary variables have been developed; predictions can be made by variety at the global or local (field) level. Segmenting the data sets into different groups and then running the corresponding regressions within each group may improve the quality of the predictions. The use of ancillary variables such as aerial or satellite imagery may facilitate data clustering. The present work had for objective to explore different mathematical models for early yield estimation of wine grape. Three-year data were used. Data consisted on the weight and number of bunches per meter row, taken at different times before harvest:> 90 days before harvest (DBH), 60-90 DBH, 30-60 DBH, and < 30 DBH. At each field, samples (15 to 20 per field) were collected in a systematic design, with three replications at each sampling point. Ancillary data consisted on a vegetation index (either PCD or NDVI) taken at veraison. Several mathematical models, using cluster regression as a base, were evaluated including: general (one variety at several farms), farm (one variety at each farm), and field (one variety at each field). Clusters were made using a hierarchical clustering algorithm. Results demonstrated that in general, local models performed better than the general ones and that the predictions were acceptable.   It is possible to predict yield as early as > 90 DBH.

Key words: yield prediction, cluster regression, wine grape, vegetation indices

Introduction

Yield prediction is an essential component of the production chain of wineries. Accurately knowing, in advance, the amount of grapes being produced is crucial to establishing a proper logistic. Ortega et al. (2007) and Ortega et al. (2008) have developed simple models based on field sampling and vegetation indices (VI) to predict tomato and wine grape yields, with good accuracy when the unit of prediction was a given field. The use of proper algorithms may improve the quality of the prediction; for example, the use of cluster regression (CR) has shown a very good potential for improving prediction results. Ríos (2010) working on the same data set as Ortega et al. (2008) showed that a CR algorithm improved the quality of yield prediction at the field level; even more, he demonstrated that using CR with a proper number of clusters would allow a good prediction of wine grape yield directly from a VI used as an ancillary variable; on the other hand, Quinteros (2011) working on a data set that related corn yield to soil fertility and N rate, found a large improvement on yield prediction when using the same CR algorithm. The CR procedure basically consists on segmenting the data sets into different groups and then running the corresponding regressions within each group. Ancillary variables, easy and inexpensive to determine, are key to delineate clusters.

The present work had for objective to explore different mathematical models for early yield estimation of wine grape using a CR algorithm.

MATERIALS AND METHODS

Three-year data were used (2007/2008, 2008/2009, and 2009/2010 growing seasons). During each season, data was collected according to the procedures described in Ortega et al. (2008), which have been followed up to today. Data consisted on the weight and number of bunches per meter row, taken at different times before harvest:> 90 days before harvest (DBH), 60-90 DBH, 30-60 DBH, and < 30 DBH. At each field, samples (15 to 20 per field) were collected in a systematic design, with three replications at each sampling point. Ancillary data consisted on a vegetation index (either PCD or NDVI) taken at veraison during summer 2008. Yield was estimated at each point and sampling date, obtaining a data set as the one given in table 1 as an example.

Table 1.Example of a data set for one farm and variety[1].

Farm	Variety	Field	Year	Observed yield	>90DBH	60-90DHB	30-60DBH	<30DBH
				Y	x_1	x_2	x_3	x_4
				------------------------------kg/ha-----------------------------				
Buin	Cabernet Sauvignon	620-1	2008	8029	4966	8281	10222	9985
Buin	Cabernet Sauvignon	620-1	2009	7159		4712	8119	8241
Buin	Cabernet Sauvignon	620-1	2010	2941	1419	1790	3416	2824

[1]only part of the data set is shown

Regressions between observed grape yield and those obtained at different sampling dates were performed in the software Lingo, using the algorithm

"Classification and Regression via Integer Optimization" (CRIO), proposed by Bertsimas and Shioda (2007).

In a classical regression setting there are *n* data points ($\boldsymbol{x_i}$, y_i), $\boldsymbol{x_i} \in \Re^d$, $y_i \in \Re^d$, and i= 1,…, n. We wish to find a linear relationship between $\boldsymbol{x_i}$ and y_i, i.e, $y_i \approx \boldsymbol{\beta}' \boldsymbol{x}_i$ for all i, where the coefficients $\boldsymbol{\beta} \in \Re^d$ are found minimizing $\sum_{i=1}^{n}(y_i - \beta' x_i)^2$ or $\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}' \boldsymbol{x}_i\right|$. The CRIO algorithm seeks finding *k* disjoint regions, where $P_k \subset \Re^d$ and corresponding coefficients $\boldsymbol{\beta_\kappa} \in \Re^d$, k= 1,…, k, such that if $\boldsymbol{x_0} \in \boldsymbol{P_k}$ the prediction for y_0 will be $\hat{y}_0 = \boldsymbol{\beta}_k' \boldsymbol{x}_0$.

Several regression models were evaluated at different levels of detail, including: general (one variety at several farms), farm (one variety at each farm), and field (one variety at each field). The following models were tested at each level of detail:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

$$y = \beta_0 + \beta_2 x_2 + \varepsilon$$

$$y = \beta_0 + \beta_3 x_3 + \varepsilon$$

$$y = \beta_0 + \beta_4 x_4 + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

In each case all the regression assumptions including collineality were tested. The best model was selected by its R^2, obtained by regressing observed yields on estimated ones.

Clusters were made using an ancillary variable (x_5) corresponding to the vegetation index (VI). The hierarchical clustering method by nearest neighbor and Euclidean distance squared was used.

Models were constructed only when there were more than five observations per cluster.

Results and discussion

Some examples of general and farm prediction models are presented.

General models

Table 2 shows the R^2's for all the models per variety, when including all farms and fields, without clustering. In general, better predictions are obtained when models included samples taken closer to harvest.

Table 2. Overall models per variety across farms and fields.

Variety	x1	n	x2	n	x3	n	x4	n	x1 + x2	n	x1 + x2 + x3	n	x1 + x2 + x3 + x4	n
Cabernet Sauvignon	0.51	39	0.54	46	0.69	44	0.64	27	0.54	35	0.75	35	0.81	19
Carmenere	0.62	21	0.80	21	0.74	22	0.58	15	0.91	17	0.91	17	0.91	11
Chardonnay	0.76	16	0.84	19	0.93	21			0.56	14	0.78	14		
Merlot	0.60	23	0.45	33	0.81	34	0.80	19	0.61	23	0.85	21	0.99	9
Sauvignon Blanc			0.15	14	0.82	19	0.89	17						
Syrah	0.17	4	0.32	4	0.30	6	0.86	4	0.80	4				

n=number of observations.

Table 3 presents the overall models (including all farms and fields) per each variety with sampling times x1 and x2, with clustering. It is observed that, in general, there was a significant improvement in the R^2 when clustering. The best results were obtained for the Chardonnay variety, with three clusters with an $R^2 > 0.93$. On the other hand, the variety Merlot presented the lowest R^2, probably because the VI does not vary as widely as with the other varieties, given its lower vigor.

Table 3. Overall models per variety across farms and fields with clustering [1].

Variedad	Two clusters		Three clusters	
	R^2	n	R^2	n
Cabernet Sauvignon	0.81	44	0.82	43
Chardonnay	0.92	19	0.93	19
Merlot	0.55	23		

[1] Based on sampling times x1 and x2

Farm models

In farms where there were enough data points, it was possible to develop local models by variety. Figure 1 presents the effects of sampling date on prediction when two clusters were considered at the Buin Location. It can be seen that good predictions can be reached when sampling as early as > 90 days DBH (x1). The R^2 of prediction varied from 0.77 to 0.99, when using samples from 30 to 90 DBH (x2), and those from x1, x2, and x3 (30 to 60 DBH) samples, respectively. This means that accurate yield prediction can be obtained early in the season, which will improve, as sampling time gets closer to harvest.

Model comparison

When comparing general versus local models in terms of prediction quality it was found that the latter performed better than the former ones (figure 2). This means that for properly predicting yield of a given variety, local data must be available in a reasonable number in order to apply the CRIO procedure.

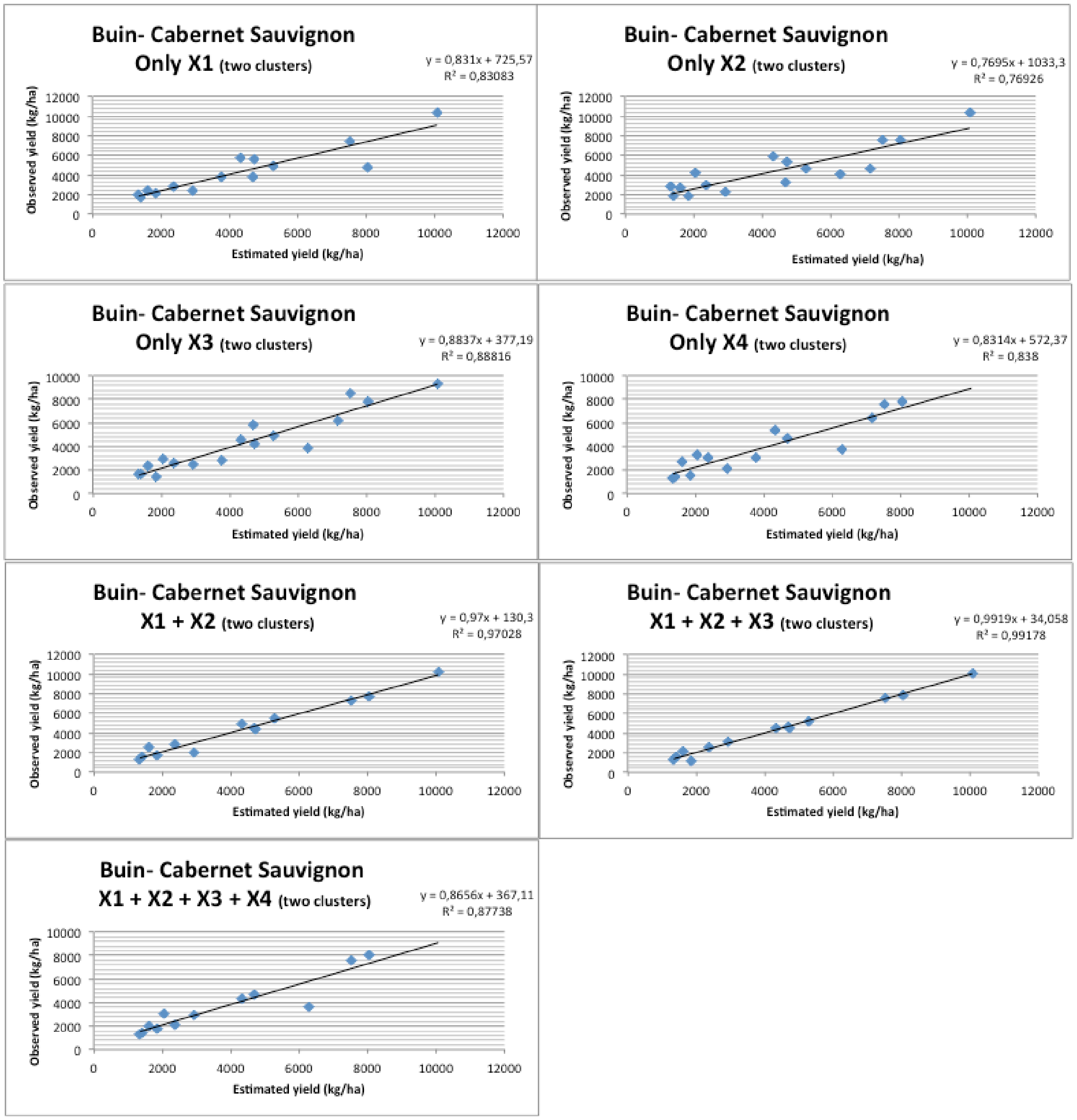

Figure 1. Local yield prediction at the Buin location for the variety Cabernet Sauvingnon, when using two clusters and different sampling dates.

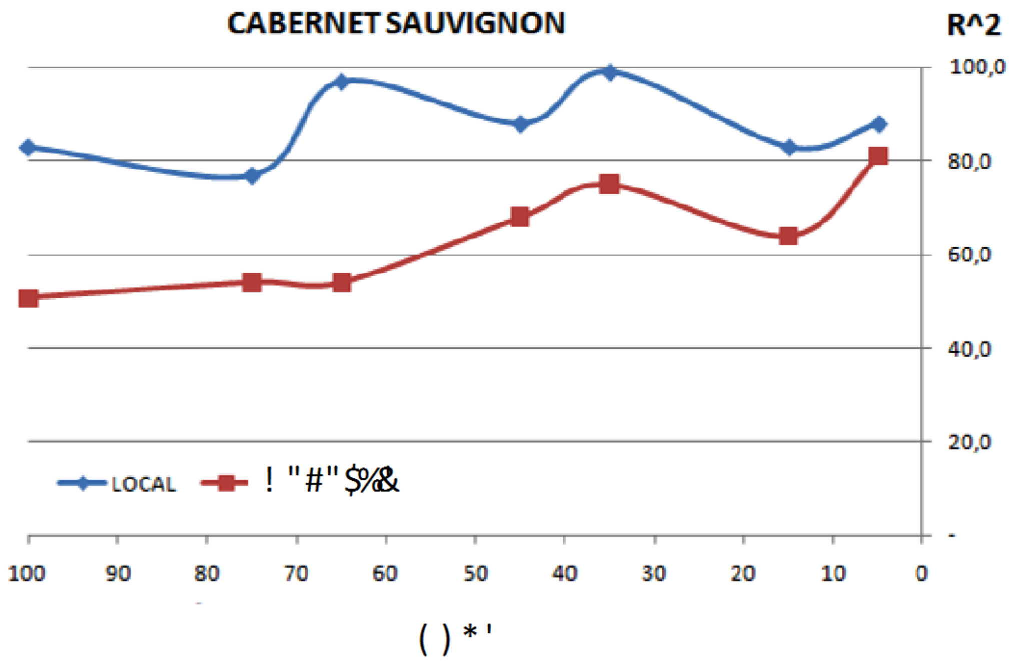

Figure 2. R^2s of prediction for the Cabernet Sauvignon variety using local and general models.

CONCLUSIONS

Estimating grape yield with a good accuracy is possible using an optimization algorithm such as CRIO; however, the result will be directly proportional to the quality and quantity of data.

The incorporation of multispectral images, and from them VIs, to spatialize information, determine the proper sampling size, or define sample location to enhance its representativeness, will generate a considerable improvement in early estimate (>90 days) of yield at harvest.

The local models are "better" than the general ones, because there is a spatial variability to consider. That is the effect of soil, climate and management, which are reflected in the results of local models.

REFERENCES

Bertsimas, D. and R. Shioda. 2007. Classification and Regression via Integer Optimization Operations Research 55(2):252–271.

Ortega, R.A., L.A. Jara, A.A. Esser, and A.A. Inostroza. 2008. Using multispectral imagery and directed sampling to estimate wine grape yield. Proceedings of the 9th International Conference on Precision Agriculture (ICPA), Denver, CO, USA. July 20–23, 2008 (CD rom).

Ortega, R., Esser, A., Inostroza, A., and Jara, L. 2007. Tomato yield and quality prediction by using a calibrated, satellite-based, green vegetation index (GVI). In: J. Stafford (ed.) Precision Agriculture ’07. Wageningen Academic Publishers. pp 573-579.

Quinteros, P. 2011. Modelo para predecir el rendimiento de maíz en function de las propiedades del suelo. Memoria de Título Ingeniero Civil Industrial. Universidad Técnica Federico Santa María. Santiago, Chile.

Ríos, F. 2010. Una propuesta del uso de la programación entera para la estimación de la producción de uva vinífera. Memoria de Título Ingeniero Civil Industrial. Universidad Técnica Federico Santa María. Santiago, Chile.