Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA)
Available online at www.inia.es/sjar
doi: http://dx.doi.org/10.5424/sjar/20110904-531-10

Spanish Journal of Agricultural Research 2011 9(4), 1156-1167
ISSN: 1695-971-X
eISSN: 2171-9292

Predictive modelling in grape berry weight during maturation process: comparison of data mining, statistical and artificial intelligence techniques

R. Fernandez Martinez*, F. J. Martinez-de-Pison Ascacibar,
A. V. Pernia Espinoza and R. Lostado Lorza

*EDMANS Group, ETSII, Edificio Departamental D202, C/ Luís de Ulloa, 20,
Universidad de La Rioja, 26004 Logroño, Spain*

Abstract

Environmental and geographical factors are two of the key aspects conditioning the growth of any crop, in such a way that the ability to predict significant variables of grape maturation can be highly useful to vine-growers. Berry weight is one of the variables monitored during this period, and the wineries have called for the development of an accurate prediction model. This study compares various types of data mining (DM) and artificial intelligence (AI) algorithms for developing an efficient prediction model for determining the variations in weight of grape berries during the ripening process according to the environmental and geographical properties not only throughout the ripening period but throughout the plant's cycle. The final objective is the search for a model that is efficient for data for new years with different properties to those in the past. This model helps the grower to harvest the grapes on the most suitable date for producing the best possible wine.

Additional key words: crop growth; learning algorithms; models; ripening.

Resumen

Modelado predictivo del peso de la baya de uva durante el proceso de maduración: comparación de técnicas de minería de datos, estadísticas e inteligencia artificial

Los factores ambientales y geográficos constituyen uno de los elementos fundamentales que condicionan el desarrollo de cualquier cultivo, de tal manera que poder predecir variables significativas de la maduración de la uva a partir de estos factores puede ser de gran utilidad para el viticultor. Una de las variables observadas durante este periodo es el peso de la baya, y el desarrollo de un modelo preciso para su predicción es una de las necesidades demandadas por las bodegas. En el presente estudio, se muestra una comparativa realizada entre diversos tipos de algoritmos de minería de datos e inteligencia artificial para el desarrollo de un modelo de predicción eficiente, que permita determinar la variación del peso de las bayas de uva durante el periodo de maduración según las características ambientales y geográficas, a lo largo no solo del periodo de maduración sino también de todo el ciclo vegetativo. El objetivo final es la búsqueda de un modelo que sea eficiente para datos de nuevos años con características diferentes a los de los históricos. Con este modelo es posible ayudar al viticultor a vendimiar las uvas en la fecha más adecuada para posteriormente producir el mejor vino posible.

Palabras clave adicionales: algoritmos de aprendizaje; crecimiento de cultivos; maduración; modelos.

Introduction

In recent years, the vine-growing industry has focussed part of its efforts on controlling the grape maturing process, as this is a key aspect for improving the quality of wines. This control relies on the use of new technologies that permit the gathering of information on those factors that have an impact on vineyards, such

*Corresponding author: roberto.fernandez@unirioja.es
Received: 29-12-10. Accepted: 31-10-11

Abbreviations used: AI (Artificial Intelligence); DM (Data Mining); DOC (Qualified Designation of Origin); MAE (Mean Absolute Error); NPM (Non-Parametric Models); PAV (Pair-Adjacent Violators); PCA (Principal Component Analysis); PM (Parametric Models); RMSE (Root Mean Squared Error); SPM (Semi-Parametric Models).

as environmental conditions, which allows a more accurate evaluation to be made of the crop evolution.

Many factors affect the chemical and physical processes undergone by the berry during ripening, such as illnesses, fertilisation modes, cultivation modes, etc. One of the physical changes that vine-growers control during ripening is the variation in weight (Peynaud, 1989; Ollat *et al*., 2002). Berry growth of vine crops is influenced by myriad factors, such as location and climate factors (Buttrose *et al*., 1971; Greer and Weston, 2010). Crops that grow annually, including vines, record differences in the maturation of the berries depending on the weather conditions to which the crop has been exposed (Coombe, 1992). Furthermore, the climate changes recorded in recent years, above all the variation in temperatures and the different rainfall pattern, are having an ever greater impact on these kinds of crops.

The ability to predict certain variables beforehand may be of great use to vine-growers. In this case, knowing how berry weight is going to evolve over the coming days, depending on the performance of the environmental variables in the vineyard, may help to know how the fruit is maturing.

Environmental conditions do not just influence maturation but also the vine's entire development process (Ebadi *et al*., 1996; Girona *et al.*, 2009), although it is during this stage that a proper analysis of temperatures, humidity, rainfall and other factors can help to provide the most important information. During the study period, the influence of the analysed variables is clearer, as the use of some actions such as irrigation, are regulated by the DOC (qualified designation of origin) Rioja Regulatory Council and its use is forbidden prior to a specific date (BOE, 2003; APA, 2004).

Both the evolution of size and weight vary depending on the year and the weather conditions at the vineyard. Humidity and the amount of water present in the soil, depending mainly on precipitation, mean that the berry receives a continuous supply of water that is conducive to weight increase (Amerine, 1956; Huglin, 1998). Temperature is a factor that favours berry growth, being essential for good plant development and for ensuring the grape matures fully, although very high temperatures can have a depressive effect on growth (Amerine and Winkler, 1944; Ribéreau-Gayon *et al.*, 1982; Mareca, 1983). Exposure to sunlight and ambient temperature are related as they are both responsible for the berry's temperature (Bergqvist *et al.*, 2001).

Current data mining (DM) and artificial intelligence (AI) techniques allow prediction models to be designed based on past data that support the decision-making process.

Several authors (Behera and Panda, 2009; Bojacá *et al.*, 2009) have developed and used models that explain the effect meteorological variables have on the growth of different kinds of crops. Knowing how these variables impact on the crop, monitoring either their natural or artificial development, constitutes a major step forward towards better product control. There are likewise some models capable of simulating grapevine systems that indicate their impacts on grape production (Due *et al.*, 1993; Valdés-Gómez *et al.*, 2009).

This paper considers a comparative study with multiple DM and AI techniques developing an overall dynamic model that predicts the weight of the berry in vineyards according to several influential variables during ripening.

The aim is to develop overall models that learn from the past but which are capable of continuing to be efficient when presented with new conditions in the future. This method uses not only basic regression models, but also considers models based on artificial intelligence, such as neural networks, Gaussian functions, etc.

Material and methods

Study area

The DOC Rioja grows a range of different grape varieties, including Tempranillo, Garnacha, Mazuelo, Graciano and Viura. This research and all the samples taken involve the Tempranillo variety of red grape. It was chosen as it is the most widespread throughout the region, which accounts for over 60% of the cultivated vineyards (Martínez de Toda and Sancha, 1995).

Soil properties, latitude and altitude all play a role, above all when growing a crop in a specific location, although environmental factors are amongst the main aspects that condition any type of crop. The vine, like any plant, has an ecological window and finds its preferred habitat in specific microenvironments. An ideal habitat for vine growing is the region of La Rioja, which produces crops of the highest quality, making Rioja one of the world's most esteemed wine labels.

The vine-growing region of the DOC Rioja (Fig. 1) covers an area of 60,905 ha in the heart of the Ebro

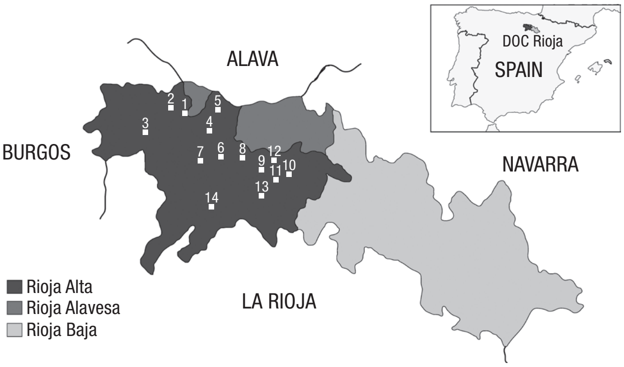


Figure 1. DOC Rioja region and the 14 locations used in this research.

Valley in Spain. The northern growing area, Rioja Alta, with an Atlantic climate of long hours of sunlight and stony, mostly limestone soils, has better qualities for producing aged wines than the other areas of the DOC Rioja (Pascual and Cabrerizo, 1995).

The research has been conducted within this area, where 14 sites, with the Tempranillo variety of grape growing at different altitudes and on vine stock that has been in the ground for different lengths of time, were chosen. Table 1 shows the pertinent details for each one of the chosen locations.

Field measurements

From 2002 to 2008 sampling was conducted in several selected vineyards, which were sampled weekly during the grape maturing process.

The readings in the vineyards were always taken by the same person in each case and using the following procedure: the berries were picked within a radius of approximately 20 m in each vineyard and collected, from around 20 plants in order to gather 100. Two berries were taken from the arms on each plant, two berries from the middle and one berry from the tip, with the orientation and location of the bunch on the plant being alternated. Once the berries had been picked, they were weighed together in the laboratory using precision scales.

The climate data was provided every 15 minutes by the weather stations the Government of La Rioja has within the observation area (eleven in total). These stations provide data on air temperature, rainfall, relative humidity and wind strength and direction. Table 2 provides a summary of the main variables measured during the years considered for the modelling.

Modelling process stages

The modelling process has been undertaken in three stages:

1. Initially, identification was made of those climate variables that have the greatest impact on berry growth (Ribèreau-Gayon *et al.*, 2006; Jackson, 2008) and a database was created based on past records of the maturing process in prior campaigns. Once all this information is stored, a feature transformation in order to make a variables reduction based in Principal Component Analysis (PCA) techniques is performed. This

Table 1. Information on the different locations used in this research. Locations, coordinates, elevations and year of plantation

	Site	Latitude (º)	Longitude (º)	Altitude (m)	Variety	Year of plantation	Data points
1	Haro - Vicuana	42°34’34.47”N	2°50’06.55”O	438	Tempranillo	1993	41
2	Haro - El Cuervo	42°34’59.89”N	2°51’45.77”O	465	Tempranillo	1997	40
3	Cihuri - La Arena	42°34’19.85”N	2°55’14.39”O	500	Tempranillo	1997	40
4	San Vicente de la Sonsierra - La Liende	42°32’57.33”N	2°46’09.78”O	440	Tempranillo	1987	45
5	San Vicente de la Sonsierra - Santamaría	42°34’27.22”N	2°45’13.78”O	590	Tempranillo	2000	42
6	San Asensio - Camino Carrera	42°30’20.73”N	2°44’34.66”O	457	Tempranillo	1985	38
7	San Asensio - El Roble	42°29’53.18”N	2°45’53.51”O	580	Tempranillo	1979	40
8	Cenicero - Las Quince	42°29’44.14”N	2°40’05.93”O	434	Tempranillo	2000	34
9	Cenicero - Carril	42°28’11.09”N	2°37’54.91”O	560	Tempranillo	1986	35
10	Fuenmayor - Los Llanos	42°26’49.70”N	2°32’08.88”O	520	Tempranillo	1986	40
11	Fuenmayor - Los Llanos	42°27’26.43”N	2°33’34.23”O	428	Tempranillo	2000	39
12	Fuenmayor - El Cuadro	42°28’42.65”N	2°34’05.92”O	430	Tempranillo	1996	38
13	Sotés - Palomar	42°24’03.35”N	2°36’39.32”O	650	Tempranillo	1996	40
14	Alesanco - Ajas	42°23’41.06”N	2°48’42.11”O	635	Tempranillo	1990	41

Table 2. Magnitude of weather variables such as temperature, relative humidity, wind speed and rainfall during the vegetative growth periods

Variable		2002	2003	2004	2005	2006	2007	2008
Temperature (°C)	Max	36.90	40.00	36.80	38.10	37.00	37.00	35.00
	Min	–2.40	–3.80	–3.10	–9.00	–5.10	–5.70	–4.20
	Mean	12.86	13.88	12.69	12.64	13.65	12.67	12.78
Relative humidity (%)	Max	95.00	93.16	94.00	92.80	92.48	89.11	90.97
	Min	24.96	30.17	40.77	17.88	25.11	37.71	45.49
	Mean	68.95	65.50	67.53	64.52	65.48	66.50	68.82
Wind speed (km h^{-1})	Mean	7.60	4.60	7.11	7.36	7.20	6.55	5.62
Precipitation (mm)	Total	409.50	409.30	433.00	427.40	419.30	366.50	603.40

process likewise included the development of a stratified sampling that allows existing processes to be standardised in order to increase the degree of reliability of the models created.

2. A subsequent validation is made of a battery of different techniques in DM and AI with a view to singling out those that generate the best predictive models.

3. Finally, the models created are tested with new data to identify the degree of generalisation of the models created.

Data analysis and pre-processing

The model's design and testing requires a database with all the meteorological variables, vineyard data and berry weight. A total of 14 data groups are used, one for each location, with 7 years of field measurements (2002-2008). The model is calibrated by choosing 6 years at random and separating the 14 locations of the remaining year for testing.

The design of the regression model is shown in Fig. 2. The purpose of this model is to predict the weight of the grape berries during ripening taking into account the data provided by the vineyard and the meteorological variables to which the latter has been exposed. The variables used are shown in Table 3.

Prior to the development of the models, several techniques are used for detecting spurious data (Castejon-Limas *et al.*, 2004) and the final data obtained is analysed. This involves the use of several display techniques (Fig. 3), such as histograms, scatter diagrams, etc., which enable us to observe the structure of the data.

The raft of available variables makes the process of developing the models more complex and more liable to generate an erroneous output, so the first step is to compress the number of variables. Methods were studied for selecting the most influential variables, although the decision was taken to use PCA methods for the compression in model inputs keeping the maximum of information.

This projection technique compresses the number of correlated variables to provide a smaller number of uncorrelated variables by means of an orthogonal linear transformation of the data into a new system of coordinates, which means the least amount of information possible is lost (Gorban *et al.*, 2007). The new principal components or factors are a linear combination of the original variables that are independent of each other. The variables to which the PCAs are applied are the ones with the highest correlation between each

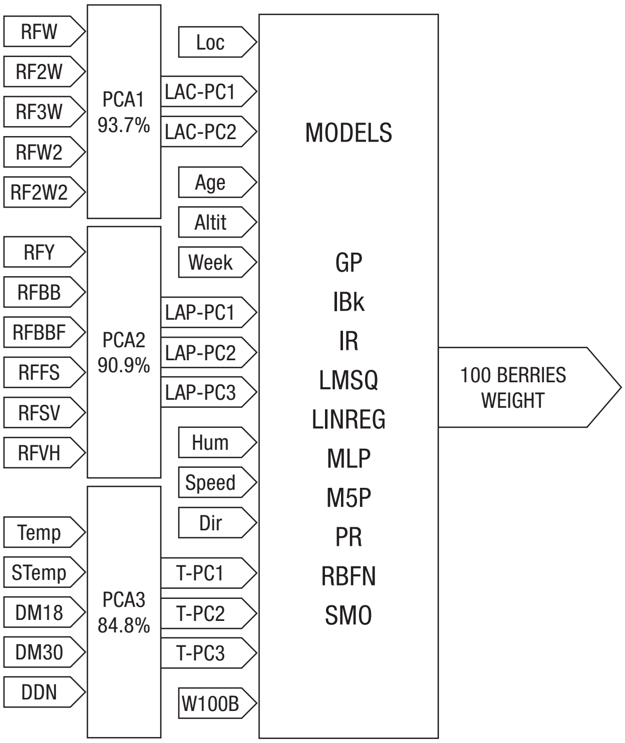


Figure 2. Design of the regression model. All variables are defined in Table 3.

Table 3. Variables used in the model

Vineyard variables	
Location	Loc
Vineyard age (year)	Age
Altitude (m)	Altit
Environmental variables related to the amount of rainfall	
Total rainfall over the preceding week (mm)	RFW
Total rainfall over the preceding two weeks (mm)	RF2W
Total rainfall over the preceding three weeks (mm)	RF3W
Total rainfall since the beginning of the year (mm)	RFY
Total rainfall since bud break (mm)	RFBB
Total rainfall during the penultimate week (mm)	RFW2
Total rainfall during the penultimate and antepenultimate week (mm)	RF2W2
Total rainfall between bud break and flowering (mm)	RFBBF
Total rainfall between flowering and setting (mm)	RFFS
Total rainfall between setting and véraison (mm)	RFSV
Total rainfall between véraison and harvest (mm)	RFVH
Environmental variables related to wind and humidity	
Prevailing wind direction over the preceding week (N,S,E,W)	Dir
Average relative humidity over the preceding week (%)	Hum
Average wind speed over the preceding week (km h^{-1})	Speed
Environmental variables related to temperature	
Average temperature over the preceding week (ºC)	Temp
Aggregate of average daily temperatures since the beginning of the year (ºC)	STemp
Days with average temperatures above 18ºC during maturation	DM18
Days with maximum temperatures above 30ºC during maturation	DM30
Average differences between maximum and minimum daily temperature during maturation (ºC)	DDN
Weight variables in preceding weeks	
Weight of 100 berries in preceding week (g)	W100B

other, such as all those related to temperature or all those related to the amount of rainfall.

Once the spurious values have been discarded and a selection made of the definitive variables that will provide the input for the models, the various families of models can be trained.

Data mining techniques

In order to find models that generate a low prediction error a battery of algorithms are used. These can be divided into three large groups: parametric models (PM), semi-parametric models (SPM) and non-parametric models (NPM). Models that range from the most classic, based on parametric statistics, to non-parametric models that work best with variables with unknown probability functions, high noise, many empty values, dependent variables, etc.

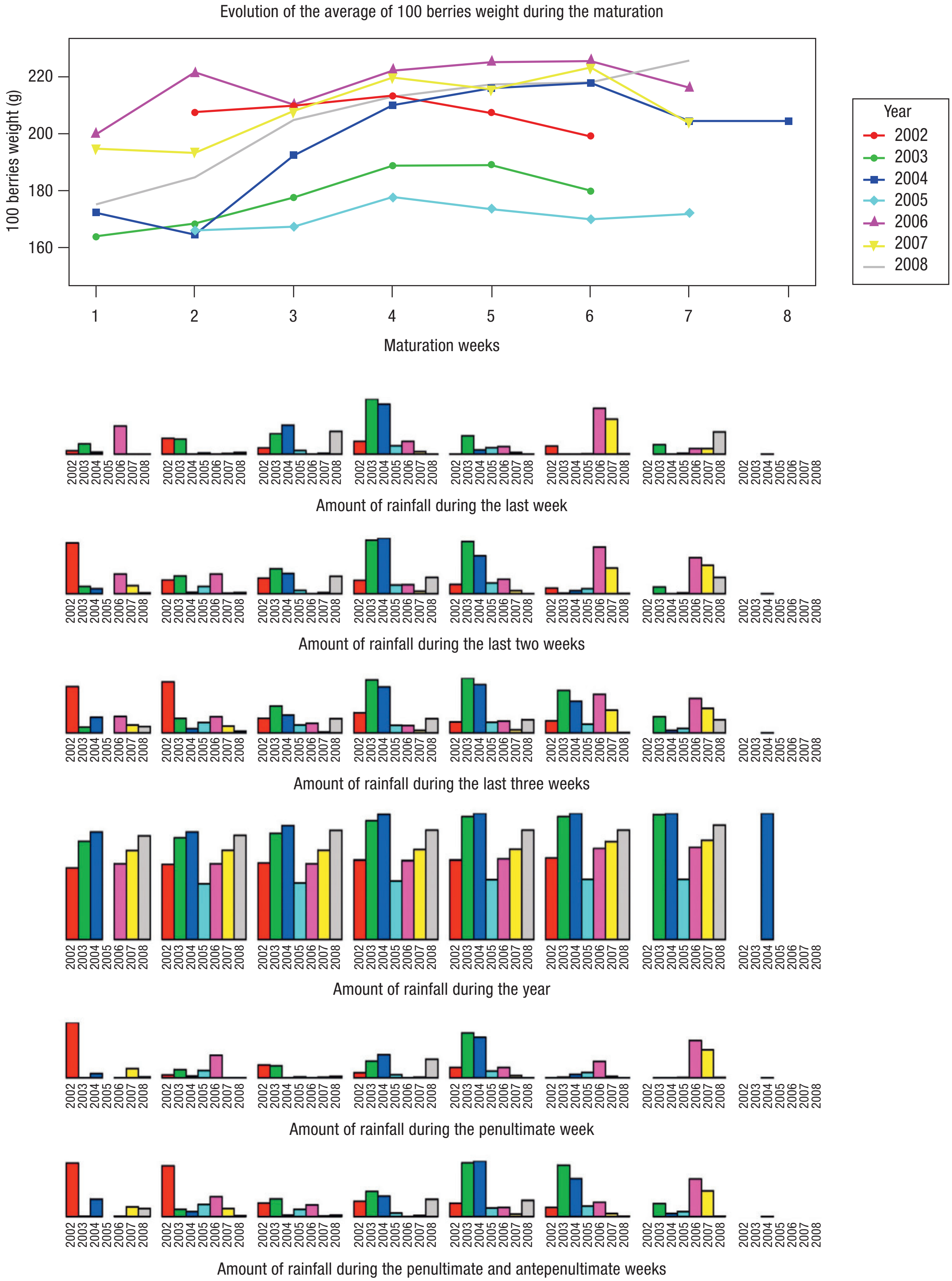
The parametric methods used are:

— LeastMedSq (LMSQ) (Portnoy and Koenker, 1997): It is an implementation of least median squared linear regression that minimizes the median squared error. Linear regression algorithms are used to form predictions.

— LinearRegression (LINREG) (Wilkinson and Rogers, 1973): Although requiring a major restriction of the model's linearity, this algorithm is used to view the data behaviour using a linear model. Furthermore, it uses the Akaike criterion for model selection, and is able to deal with weighted instances.

The semi-parametric methods used are:

— M5P algorithm (M5P) (Quinlan, 1992): Implementation of base routines for generating M5Model


Evolution of the average of 100 berries weight during the maturation

Amount of rainfall during the last week

Amount of rainfall during the last two weeks

Amount of rainfall during the last three weeks

Amount of rainfall during the year

Amount of rainfall during the penultimate week

Amount of rainfall during the penultimate and antepenultimate weeks

Figure 3. Evolution of the average weight of 100 berries in 'Rioja Alta' in the 'Tempranillo' variety according to the rainfall over preceding weeks.

trees. A decision list for regression problems is generated using separate-and-conquer. It builds a model tree in each iteration using M5 algorithm and makes the 'best' leaf into a rule. Quinlan's M5P can learn such piece-wise linear models. M5P also generates a decision tree that indicates when to use which linear model.

— PaceRegression (PR) (Li and Shue, 2004): The basic idea of regression analysis is to fit a linear model to a set of data. The classical ordinary least squares estimator is simple, computationally cheap, and has well-established theoretical justification. Nevertheless, the models produced are often unsatisfactory. Page regression improves the classical ordinary least squares regression by evaluating the effect of each variable and using a clustering analysis to improve the statistical basis for estimating their contribution to the overall regressions. Under regularity conditions, pace regression is provably optimal when the number of coefficients tends to infinity. It consists of a group of estimators that are either overall optimal or optimal under certain conditions.

And finally, the non-parametric methods selected are:

— GaussianProcesses (GP) (Mackay, 1998; Williams, 1998): It implements a non-parametric Bayesian technique. Bayesian regression techniques assume a prior distribution over the function hypothesis space and calculate a posterior distribution using Bayes rule and the available learning data. Instead of assuming a prior over the parameter vectors, GP assume a prior over the target function itself.

— IBk (IBk) (Aha *et al.*, 1991): It is instance-based learning that works as a k-nearest-neighbour classifier. A variety of different search algorithms are used to speed up the task of finding the nearest neighbours.

— IsotonicRegression (IR) (Stout, 2008): A non-parametric method that is designed for applications where the expected value of a response variable (y) increases or decreases in one or more explanatory variables ($x_1, ..., x_p$). It implements the method for learning an isotonic regression function based on the pair-adjacent violators approach. It minimizes the squared error between the observed class probabilities and the resulting calibrated class probabilities. The basic Pair-Adjacent Violators (PAV) algorithm iteratively merges pairs of neighbouring data points that violate the monotonicity constraint by computing their weighted mean. The result is a function that increases monotonically in a stepwise fashion.

— MultilayerPerceptron (MLP) (Haykin, 1999): It is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. A supervised learning technique called back propagation is used for training the network connecting many simple perceptron-like models in a hierarchical structure, which can distinguish data that is not linearly separable.

— RBFNetwork (RBFN) (Haykin, 1999): The main characteristic of the radial basic function neural network is the use of a normalized distance between the input points and the hidden nodes to define the activation of each node. The closer the two points, the stronger the activation.

— SMOreg (SMO) (Smola and Scholkopf, 1998; Shevade *et al.*, 1999): A sequential minimal optimization (SMO) algorithm for training a support vector regression using polynomial or RBF kernels. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. This algorithm solves large quadratic programming (QP) optimization problems, widely used for the training of support vector machines. SMO breaks up large QP problems into a series of smallest possible QP problems, which are then solved analytically.

The models are trained using cross-validation, as their calculation times are not very high and allow the entire training dataset, with 483 entries, to be used to create and validate models. This method involves dividing the initial database into 10 subsets. To calculate this error, 9 subsets were chosen to train the model, with the one subset omitted from the training being used to calculate the error of the partial sample. This procedure is repeated ten times, each time using a different test subset. Finally, the error is calculated as the arithmetic mean of the ten errors of the partial samples.

The purpose of this paper is to determine the algorithm or group of algorithms that provide the best prediction or, in other words, the algorithm that yields the lowest Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) between model simulation and observed data which are not used for model construction. RMSE and MAE are calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y(k) - \hat{y}(k))^2} \qquad [1]$$

and

$$MAE = \frac{1}{n}\sum_{k=1}^{n}\left|y(k) - \hat{y}(k)\right| \qquad [2]$$

where y and $\hat{y}$ are, respectively, the measured and predicted outputs and n is the number of points of the database used to validate the models.

Table 4. Results of the modelling process. Training errors for each model configuration (ordered by mean of RMSE[1]). This table presents the mean, maximum (max), minimum (min) and standard deviation (SD) of RMSE[1] and MAE[2] training errors for twenty models of each type of algorithm configuration. The last column shows the time required for creating the twenty models and obtaining the cross-validation errors

Algorithm	Group	RMSE[1] mean	RMSE[1] max	RMSE[1] min	RMSE[1] SD	MAE[2] mean	MAE[2] max	MAE[2] min	MAE[2] SD	TIME (s)
GP	NPM	0.0939	0.0951	0.0926	0.0005	0.0748	0.0755	0.0744	0.0003	224.88
SMO	NPM	0.0957	0.0970	0.0944	0.0006	0.0748	0.0758	0.0739	0.0005	78.79
LMSQ	PM	0.0958	0.0979	0.0941	0.0011	0.0752	0.0765	0.0735	0.0007	670.02
LINREG	PM	0.0961	0.0970	0.0949	0.0006	0.0756	0.0761	0.0746	0.0004	2.12
M5P	SPM	0.0962	0.0979	0.0949	0.0008	0.0757	0.0767	0.0746	0.0005	103.38
PR	SPM	0.0964	0.0976	0.0951	0.0006	0.0758	0.0764	0.0747	0.0005	1.00
MLP	NPM	0.1066	0.1139	0.0993	0.0030	0.0843	0.0906	0.0785	0.0028	76.98
IR	NPM	0.1079	0.1094	0.1055	0.0011	0.0848	0.0856	0.0831	0.0005	334.94
IBk	NPM	0.1146	0.1184	0.1127	0.0016	0.0905	0.0931	0.0887	0.0013	0.00
RBFN	NPM	0.1502	0.1551	0.1448	0.0027	0.1209	0.1238	0.1166	0.0023	34.08

[1] RMSE: root mean squared error. [2] MAE: mean absolute error.

Accordingly, twenty models of each type of algorithm configuration are trained with 86% of the data from the training database and the remaining data (14%) are used to validate each model.

WEKA suite (Witten and Frank, 2005) is used to develop the different models.

Results and discussion

Feature reduction

Feature transformation techniques have obtained a variable reduction and a modelling process optimization. PCA technique compresses the number of correlated variables providing a smaller number of uncorrelated variables. Three PCA analyses were conducted separately of meteorological variables in terms of temperature and rainfall. The first PCA (PCA1) for the total amount of rainfall over the preceding weeks compresses the 5 related variables to 2 principle components (PCs), which explains 93.7% of the data; the second PCA concerning the 6 variables of the total amount of rainfall throughout the full growth cycle reduces the original variables to 3 PCs, which explains 90.9% of the data. The third PCA also identifies 3 PCs, which explains 84.8% of the data, for the 5 variables in relation to temperature. Overall, the 16 variables related to temperature and rainfall, were reduced to 8 by the PCA (Fig. 4).

Model calibration

The models are obtained from a training dataset (483 instances) with 16 final variables as Fig. 2 shows. The

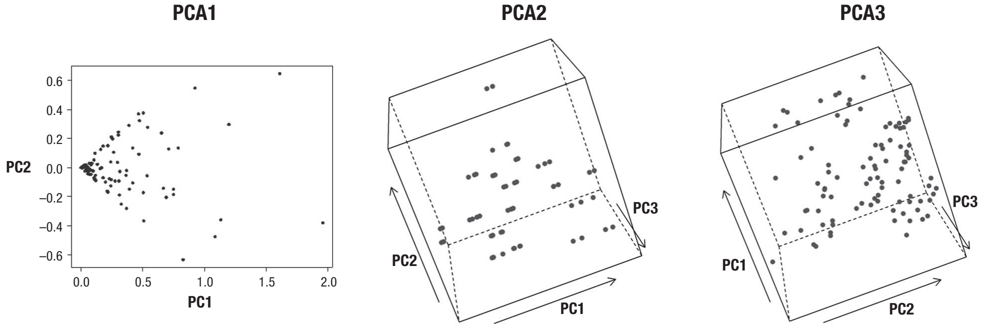


Figure 4. Principal component analysis projections: PCA1 of amounts of rainfall during preceding weeks, PCA2 of amounts of rainfall during the growth cycle, and PCA3 of temperature variables.

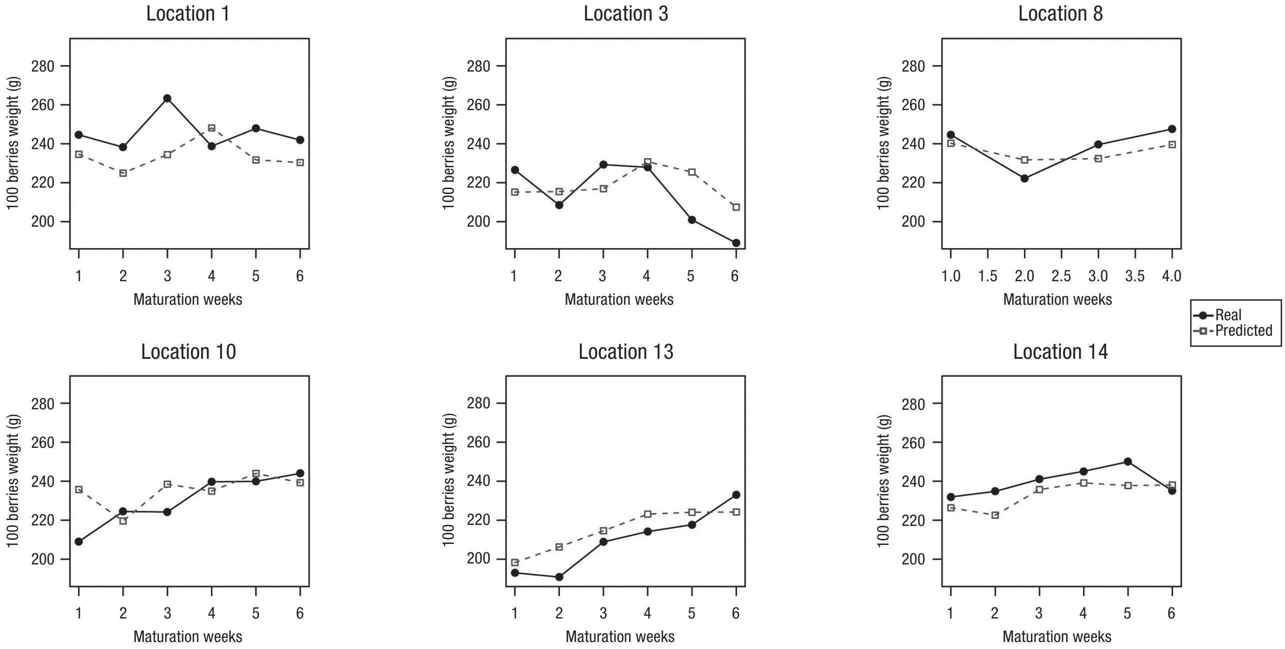


Figure 5. Comparison of the real value of the weight of 100 berries with the predicted value working with the test database.

result of the calibration process is shown in Table 4. In this case, the model's calibration dataset correspond to all the years but one, with the validation dataset being those for the discarded year (14% of the database, 70 instances). This table provides a summary of errors obtained from cross-validation training processes ordered by the lower RMSE corresponding to twenty trained models for each type of configuration and algorithm.

It can be seen that, in general, the linear models perform better than all the others within the battery of models used. Models such as LMSQ, LINREG and PR perform better than others such as MLP and RBFN. Nevertheless, the model with the best performance is GP, as it generates the smallest RMSE, as well as the lowest mean of RMSEs. By contrast, the models with the worst performance are those based on k-nearest neighbours (IBk) and the radial basis function networks (RBFNs). The differences recorded range between 6 and 9%. The models requiring the longest training time are the LMSQ with a high number of random samples used to generate the function.

Of the wide range of methods used, it was found that in the prediction of variables relating to crop growth, such as in the case of the weight of wine grape berries, the non-parametric models behave best. The models with the best behaviour obtained for this case are GP (RMSE=0.0939; MAE=0.0748) and SMO (RMSE=0.0957; MAE=0.0748). Not all the models of this type behaved in the same way and some in this group gave worse results. It is also significant that the models in the parametric techniques group also gave good results: LMSQ (RMSE=0.0958; MAE=0.0752) and LINREG (RMSE=0.0961; MAE=0.0756).

The GP model is based on Gaussian Processes for regression without hyper parameter-tuning using the kernel RBF (Radial Basis Function). Different values of σ are used for the calibration of this model. This parameter controls the width of the kernel and thereby the amount of generalization used by the GP. A σ of 0.7 gives the best result. It generates the smallest RMSE, as well as the lowest RMSE mean.

Model validation

Once the best model has been chosen (GP with a value for σ of 0.7), a comparison is made between the values obtained with that model and the real values in the validation database. Thus, a RMSE = 0.0842 and a MAE = 0.0708 is obtained with the best model obtained at the validation stage. It can be seen that even the errors are relatively lower than those of cross-validation proving the model's good performance.

Table 5 shows the validation errors in a function of experimental location.

Fig. 5 and 6 shows the results of using the model for predicting weight, compared with the real weight of the same 100 berries. The data used for the figure are those corresponding to the validation database.

As can be seen, the model's performance is reasonably close to the real performance of weight evolution during ripening and accurately detects changes in the weight trend. When the trajectory of the curve undergoes, a sudden change is detected and the model minimizes the error.

The LMSQ model is another that gives good results. Their errors are close to the GP's, albeit with the characteristic that being a linear regression model, it provides information on how each variable informs the final model (Table 6). These coefficients show which are the more representative variables and the value of their influence.

Table 5. Validation errors for the model chosen as the best model in the training process divided according to the locations

Location	Algorithm	RMSE[1]	MAE[2]
1	GP(σ=0.7)	0.1202	0.1098
2	GP(σ=0.7)	0.0549	0.0438
3	GP(σ=0.7)	0.1078	0.0940
4	GP(σ=0.7)	0.0697	0.0649
5	GP(σ=0.7)	0.0727	0.0654
6	GP(σ=0.7)	0.0656	0.0542
7	GP(σ=0.7)	0.0632	0.0551
8	GP(σ=0.7)	0.1368	0.1269
9	GP(σ=0.7)	0.0557	0.0539
10	GP(σ=0.7)	0.0557	0.0539
11	GP(σ=0.7)	0.0953	0.0731
12	GP(σ=0.7)	0.0663	0.0587
13	GP(σ=0.7)	0.0675	0.0625
14	GP(σ=0.7)	0.0601	0.0540

[1] RMSE: root mean squared error. [2] MAE: mean absolute error.

Conclusions

The maturation of the grape is influenced by a host of factors, such as location and weather conditions. The possibility of predicting the short-term weight of the berry helps vine-growers considerably when taking decisions. This allows the wineries to know how the grape is evolving during the ripening process and can evaluate different management strategies under various environments.

This paper compares several classical and current techniques in DM and AI to design models for predicting grape growth. It has shown that these methodologies can be accurate and can be applied with confidence to vineyards other that the one for which the model has been trained, within the same variety and the same weather area conditions.

A wide range of parametric and non-parametric models have been developed showing that, the results of non-parametric models are better than those obtained with parametric and semiparametric techniques when predicting several variables related with crops growth such as grape maturation.

From amongst the numerous configurations used, the model based on the Gaussian Processes (GP) algorithm is deemed to be the one providing the best predic-

Table 6. Relative importance of variables in the case of one of the linear regression model used (LMSQ)

Coeff.	Name	Description	Abs	Ratio (%)
–0.0122	Week	Maturation week	0.0122	1.52
0.0903	Age	Vineyard age	0.0903	11.26
0.0046	Altit	Vineyard altitude	0.0046	0.57
0.0068	Hum	Mean humidity during preceding week	0.0068	0.85
–0.0002	Speed	Mean wind speed during preceding week	0.0002	0.02
–0.0123	Dir	Prevailing wind direction during preceding week	0.0123	1.53
0.0571	LAC-PC1	PC1 of PCA projection of rainfall amounts during preceding weeks	0.0571	7.12
–0.0584	LAC-PC2	PC2 of PCA projection of rainfall amounts during preceding weeks	0.0584	7.28
–0.0515	LAP-PC1	PC1 of PCA projection of rainfall amounts during growth cycle	0.0515	6.42
–0.0907	LAP-PC2	PC2 of PCA projection of rainfall amounts during growth cycle	0.0907	11.31
0.1231	LAP-PC3	PC3 of PCA projection of rainfall amounts during growth cycle	0.1231	15.35
–0.042	T-PC1	PC1 of PCA projection of temperatures	0.0420	5.24
–0.0236	T-PC2	PC2 of PCA projection of temperatures	0.0236	2.94
0.0967	T-PC3	PC3 of PCA projection of temperatures	0.0967	12.06
0.8019	W100B	Weight of 100 berries in preceding week	0.8019	100.00

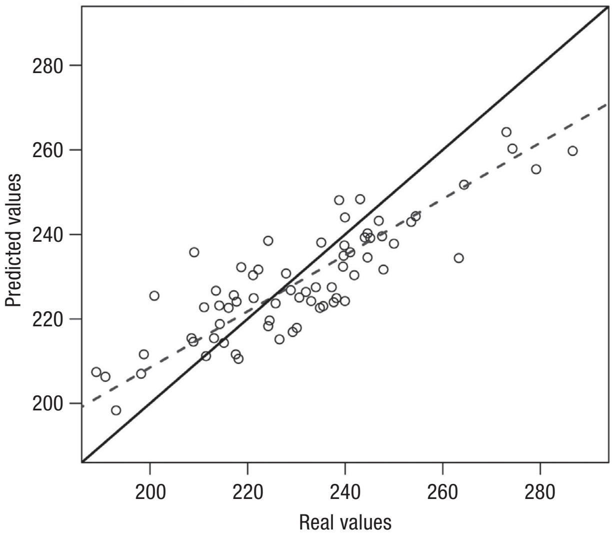


Figure 6. Scatter plots of a sample of forecasted and observed values corresponding to the GP model.

tion results. This model proves to be efficient at dealing with new data in the maturing process and with new weather conditions. Its use can help vine-growers to monitor the maturing process and establish how the evolution of the growth can be affected.

Given the models and to data, we can see that the weather has a bearing on the grape maturing process and that major changes during the cycle have an impact on the berry's end properties.

Acknowledgements

The authors thank for the financial support to the *Dirección General de Investigación* of the Spanish Ministry of Science and Innovation for the project DPI2007-61090, to the *Universidad de La Rioja* and *Banco Santander* for the project API11/13 and to the European Union for the project RFS-PR-06035. Finally, the authors also thank the Autonomous Government of La Rioja for its support through the 3rd *Plan Riojano de I+D+I* for the project FOMENTA 2010/13.

References

AHA D.W., KIBLER D., ALBERT M.K., 1991. Instance-based learning algorithms. Mach Learn 6, 37-66.

AMERINE M.A., 1956. The maturation of wine grapes. Wine and Vine 37, 27-30.

AMERINE M.A., WINKLER R.A.J., 1944. Composition and quality of musts and wines of California grapes. Hilgardia 15, 493-673.

APA, 2004. Order APA/3465/2004, of 20 October, approving the Regulation of the Qualified Designation of Origin 'Rioja' and its Regulatory Board. Boletín Oficial del Estado No. 259, 27/10/2004.

BEHERA S.K., PANDA R.K., 2009. Integrated management of irrigation water and fertilizers for wheat crop using field experiments and simulation modelling. Agr Water Manage 96(11), 1532-1540.

BERGQVIST J., DOKOOZLIAN N., EBISUDA N., 2001. Sunlight exposure and temperature effects on berry growth and composition of Cabernet Sauvignon and Grenache in the central San Joaquin Valley of California". Am J Enol Viticult 52(1), 1-7.

BOE, 2003. Legislation 24/2003, of 10 July, about vine and wine. Boletín Oficial del Estado No. 165, 11/7/2003.

BOJACÁ C.R., GIL R., COOMAN A., 2009. Use of geostatistical and crop growth modelling to assess the variability of greenhouse tomato yield caused by spatial temperature variations. Comput Electron Agr 65(2), 219-227.

BUTTROSE M.S., HALE C.R., KLIEWER W.M., 1971. Effect of Temperature on the Composition of 'Cabernet Sauvignon' Berries. Am J Enol Viticult 22, 71-75.

CASTEJÓN-LIMAS M., ORDIERES-MERÉ J.B., MARTÍNEZ-DE-PISÓN-ASCACÍBAR F.J., VERGARA-GONZÁLEZ E.P., 2004. Outlier detection and data cleaning in multivariate non-normal samples: the PAELLA algorithm. Data Min Knowl Disc 9(2), 171-187.

COOMBE, B.G., 1992. Research on development and ripening of the grape berry. Am J Enol Viticult 43, 101-110.

DUE G., MORRIS M., PATTISON S., COOMBE B.G., 1993. Modelling grapevine phenology against weather: considerations based on a large data set. Agr Forest Meteorol 65(1-2), 91-106.

EBADI A., MAY P., COOMBE B.G., 1996. Effect of short-term temperature and shading on fruit-set, seed and berry development in model vines of *V. vinifera,* cvs Chardonnay and Shiraz. Aust J Grape Wine Res 2(1), 2-9.

GIRONA J., MARSAL J., MATA M., DEL CAMPO J., BASILE B., 2009. Phenological sensitivity of berry growth and composition of tempranillo grapevines (*Vitis vinifera* L.) to water stress. Aust J Grape Wine Res 15(3), 268-277.

GORBAN A., KEGL B., WUNSCH D., ZINOVYEV A., 2007. Principal manifolds for data visualisation and dimension reduction, LNCSE 58. Springer, Berlin-Heidelberg-NY.

GREER D.H., WESTON C., 2010. Heat stress affects flowering, berry growth, sugar accumulation and photosynthesis of *Vitis vinifera* cv. Semillon grapevines grown in a controlled environment. Funct Plant Biol 37, 206-214.

HAYKIN S., 1999. Neural networks, a comprehensive foundation (2nd ed.). Prentice Hall, NJ, USA.

HUGLIN P., 1998. Biologie et ecologie de la vigne. Ed. Payot-Laussana, Paris. [In French].

JACKSON R.S., 2008. Wine science, 3rd Edition. Principles and applications. Elsevier Inc.

LI S.T., SHUE L.Y., 2004. Data mining to aid policy making in air pollution management. Expert Syst Appl 27, 331-340.

MACKAY D.J.C., 1998. Introduction to Gaussian processes. Dept. of Physics, Cambridge Univ, UK.

MARECA I., 1983. Origen, composición y evolución del vino. Ed. Alhambra S.A. [In Spanish].

MARTÍNEZ DE TODA F., SANCHA J.C., 1995. Variedades de vid cultivadas en Rioja a lo largo de la historia. Zubía monográfico 7, 9-13. [In Spanish].

OLLAT N., DIAKOU-VERDIN P., CARDE J.P., BARRIEU F., GAUDILLÈRE J.P., MOING A., 2002. Grape berry development: a review. J Int Sciences Vigne Vin 36(3), 109-131.

PASCUAL BELLIDO N., CABRERIZO CRISTÓBAL A., 1995. Distribución espacial del viñedo de rioja en relación con los condicionantes ambientales. Berceo 129, 75-95. [In Spanish].

PEYNAUD E., 1989. Enología práctica. Ed. Mundi-prensa. Madrid. [In Spanish].

PORTNOY S., KOENKER R., 1997. The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. Stat Sci 12(4), 279-300.

QUINLAN J.R., 1992. Learning with continuous classes. Proc Australian Joint Conf on Artificial Intelligence. World Scientific, Singapore. pp. 343-348.

RIBÉREAU-GAYON J., PEYNAUD E., SUDRAUD P., RIBÉREAU-GAYON P., 1982. Ciencias y técnicas del vino. Tomo 2: Características de los vinos. Maduración del racimo. Ed. Hemisferio Sur. [In Spanish].

RIBÉREAU-GAYON P., DUBOURDIEU D., DONÉCHE B., LONVAUD A., 2006. Handbook of Enology, Volume 1 - The microbiology of wine and vinifications, 2nd ed. Chapter 10: The grape and its maturation. John Wiley & Sons Ltd.

SHEVADE S.K., KEERTHI S.S., BHATTACHARYYA C., MURTHY K.R.K., 1999. Improvements to SMO algorithm for SVM regression. Technical Report CD-99-16. Control Division Dept of Mechanical and Production Engineering, Nat Univ Singapore.

SMOLA A.J., SCHOLKOPF B., 1998. A tutorial on support vector regression. NeuroCOLT2 Technical Report Series-NC2-TR-1998-030.

STOUT Q.F., 2008. Unimodal regression via prefix isotonic regression. Comput Stat Data Anal 53, 289–297.

VALDÉS-GÓMEZ H., CELETTE F., GARCÍA DE CORTÁZAR-ATAURI I., JARA-ROJAS F., GARY C., 2009. Modelling soil water content and grapevine growth and development with the STICS crop-soil model under two different water management strategies. J Int Sci Vigne Vin 43(1), 13-28.

WILKINSON G.N., ROGERS C.E., 1973. Symbolic descriptions of factorial models for analysis of variance. Appl Stat 22, 392-399.

WILLIAMS C.K.I., 1998. Prediction with Gaussian processes: from linear regression to linear prediction and beyond. In: Learning and inference in graphical models (Jordan M.I., ed). Kluwer Academic Press, pp. 599-621.

WITTEN I.H., FRANK E., 2005. Data mining: practical machine learning tools and techniques, 2nd ed. Morgan Kaufmann, San Francisco, CA, USA.