

# Machine learning methods for soil moisture prediction in vineyards using digital images

Chantal Saad Hajjar<sup>1,\*</sup>, Celine Hajjar<sup>2</sup>, Michel Esta<sup>3</sup> and Yolla Ghorra Chamoun<sup>1</sup>

<sup>1</sup> Ecole Supérieure d'Ingénieurs d'Agronomie Méditerranéenne - Université Saint-Joseph de Beyrouth, Beirut, Lebanon

<sup>2</sup> Ecole Supérieure d'Ingénieurs de Beyrouth - Université Saint-Joseph de Beyrouth, Beirut, Lebanon

<sup>3</sup> Institut de Gestion des Entreprises - Université Saint-Joseph de Beyrouth, Beirut, Lebanon

**Abstract.** In this paper, we propose to estimate the moisture of vineyard soils from digital photography using machine learning methods. Two nonlinear regression models are implemented: a multilayer perceptron (*MLP*) and a support vector regression (*SVR*). Pixels coded with *RGB* colour model extracted from soil digital images along with the associated known soil moisture levels are used to train both models in order to predict moisture content from newly acquired images. The study is conducted on samples of six soil types collected from Chateau Kefraya terroirs in Lebanon. Both methods succeeded in forecasting moisture giving high correlation values between the measured moisture and the predicted moisture when tested on unknown data. However, the method based on *SVR* outperformed the one based on *MLP* yielding Pearson correlation coefficient values ranging from 0.89 to 0.99. Moreover, it is a simple and noninvasive method that can be adopted easily to detect vineyards soil moisture.

## 1 Introduction

The vine growth and production depend on its water status directly related to the root system functionality and the availability of water and minerals in its soil. The soil fertility is directly related to the soil humidity [1-2]. Therefore, soil humidity determination is an important tool in terroir characterization, the latter being an essential procedure for viticulture development [3-4].

By definition, the soil moisture is the ratio of water mass in a sample to its total mass expressed as percentage. This is how the thermo-gravimetric method measures soil moisture [5], but it is a destructive method requiring the collection of the soil sample and its drying at 105°C during 24 hours. Agriculture uses different types of tools to detect soil moisture such as tensiometers, Frequency Domain Reflectometry (FDR) and Time-Domain Reflectometry (TDR). Tensiometers function well in humid to semi humid soils, but show uncertainty and even total dysfunction in dry soils such as vineyard soils at the end of the wine grapes growing season [6]. As for the TDR and FDR, they are accurate but expensive [5].

These constraints prompted several researchers to implement computer-based methods as a smart agriculture tool to predict soil moisture. In the field of machine learning, Altendorf *et al.* claimed that a neural network based model outperforms linear regression methods in forecasting soil moisture from soil temperature data [7]. In [8], a support vector machine built with meteorological data predicted soil moisture for four to seven days ahead. Meteorological data are also

used as input to a neural network that predicts the soil humidity in [9].

Based on the fact that soils get darker with increased moisture [10], several methods were proposed to predict soil moisture using image processing. A linear regression model is constructed in order to predict soil moisture from digital soil images where the predictors are the S and V values of the pixels coded with the HSV colour model [11]. Another linear regression model is proposed for each soil type to forecast soil moisture from soil images where the independent variables are chosen among the features of the RGB and HSV colour models as well as the digital number of panchromatic images [12]. A neural network based model is built to predict soil moisture of tropical soils where the network inputs are the R,G,B values of the pixels extracted from soil digital images [13].

The aim of the present study is to use machine learning methods in order to estimate vineyards soil moisture from soil digital photography. Soil samples having different moisture content were photographed. The colour information extracted from the photos and the measured moisture content were used to train two nonlinear regression methods. The first method is a neural network, more specifically a multilayer perceptron (*MLP*) of one hidden layer. The second method is a support vector machine used for nonlinear regression (*SVR*). Both methods succeeded in predicting the soil moisture yielding high values of the correlation coefficient R of the predicted moisture and the measured moisture when both models were tested on unknown data: The R coefficient ranged between 0.84 and 0.97 in the

\* Corresponding author: [chantal.hajjar@usj.edu.lb](mailto:chantal.hajjar@usj.edu.lb)

case of *MLP* and between 0.89 and 0.99 in the case of *SVR*.

## 2 Materials and methods

The course of the study went through different phases. First, soil samples were collected from vineyards of six different terroirs. Then, soil samples were immersed in water for 48 hours and left to dry after drainage of excessive water. On a daily basis, the samples were photographed and weighed. When the masses reduction became negligible, the soil samples were dried completely in the oven and moisture contents were calculated according to the thermo-gravimetric method. RGB colour data extracted from photos along with the associated measured moisture were used to train two nonlinear regression models that can predict soil moisture content.

### 2.1 Soil sampling

The soil samples used for data collection were collected from Chateau Kefraya terroirs in Lebanon. This area was chosen because it is an important agricultural area cultivated mainly with wine grapes. Pure soil categories, from relatively wide units were selected randomly to conduct the study. Table 1 shows the soil types, the number of samples collected for each type and the colour of each type.

The samples collection was done in two steps: first, by removing the rocks from the surface, then by collecting around 3 kg of soils at a depth of 10 cm from each point. At this depth, the soil is arable, therefore its moisture can be investigated. Each sample was mixed and divided into two sub-samples: one was used to conduct the experiment and one was sent to the soil analysis laboratory of the Lebanese Agricultural Research Institute. The physical characteristics of the soils obtained from the laboratory analysis are simplified by calculating the average of all the replicates of each soil type as shown in Table 2.

To conduct the experiment, the soil samples were placed in plastic containers which dimensions are: 30 cm in length, 22 cm in width and 7 cm in depth. Each container is bottom perforated with 1 cm diameter holes to allow drainage of excessive water. The holes are covered with a fine mesh screen to prevent soil loss. Each container received equal amount of distilled water and then placed in a larger one, full of water, and soaked for 48 hours till it reached its full water capacity. Starting day one, each container was weighed and photographed daily in a dark room with a Canon EOS 1200D digital camera of 18 Mpx resolution. The camera was fixed on a tripod with its lens facing down parallel to the plan of the container, at a height of 0.5 m. A source of continuous light illuminates the soil sample at 45°. Four panels of white foam are placed around the container as reflectors to illuminate the sample with a continuous soft light. The custom white balance setting on the camera is used to calibrate the colours.

**Table 1.** Soil samples.

Code	Soil type	No. of replicates	Colour
Type-2	Calcic Cambisols	5	
Type-3	Eutric Luvisols	6	
Type-4	Eutric Leptosols	6	
Type-5	Skeletal Fluvisols	6	
Type-6	Chromic Luvisols	6	
Type-7	Vertic Cambisols	6	

**Table 2.** Soil types physical characteristics.

Soil	Sand	Silt	Clay	Organic Matter
Type-2	10%	23%	67%	1.05%
Type-3	23%	18%	59%	2.83%
Type-4	20%	23%	57%	2.47%
Type-5	27%	22%	51%	0.94%
Type-6	27%	16%	57%	1.86%
Type-7	10%	29%	61%	1.80%

Every sample mass was measured with a tare digital balance on a daily basis. The mass reduction became negligible on day 52. On day 53, the soil samples were completely dried in the oven at a temperature of 105-110°C. After complete dryness, the samples were weighed and photos were taken in the same methodology described above. The daily percentage of soil moisture content is calculated as such:

$$m = \frac{mass_{day} - mass_{dry}}{mass_{day}} \times 100 \quad (1)$$

Table 3 shows the maximum and minimum moisture content per type.

**Table 3.** Maximum and minimum moisture per type.

Soil type	Max moisture	Min moisture
Type-2	35.325%	0.072%
Type-3	42.841%	0.110%
Type-4	34.585%	0.040%
Type-5	30.414%	0.052%
Type-6	33.435%	0.083%
Type-7	36.849%	0.021%

## 2.2 Data acquisition

Each day of the experience yields 35 photos. In order to collect the data resulting from the photos, the following steps are performed:

- 1 Each photo is coded according to the *RGB* colour model. It is an additive colour model where the primary colours (red, green and blue) are added together to make 16,777,216 colours. Thus, each pixel is described by the red, green and blue features, each one ranging from 0 to 255.
- 2 A window of 500x500 pixels is extracted from the center of each photo.
- 3 Outlier pixels are removed from the cropped window. To do so, the sum of red, green, and blue values for each pixel is calculated. Then the first quartile (Q1) and third quartile (Q3) of all the sums are calculated. Pixels whose *RGB* sum is less than Q1 or greater than Q3 are removed [13].
- 4 The cropped window is divided into 9 sub-windows. For each sub-window, the mean of red, green and blue components of the pixels is computed.

Therefore, each photo gives nine three-dimensional data vectors. The soil moisture content measured at the day when the photo was taken is associated to the nine *RGB* data vectors. In order to limit the size of the training dataset, data from 24 regularly spaced days among the 53 were retained for each soil type, making a total of 1296 observations per soil type. This will prevent the prediction model from over fitting and will reduce the model training time.

Two nonlinear regression models based on machine learning are built to predict the soil moisture from the soil digital photos. The first one is an artificial neural network, more specifically a multilayer perceptron. The second one is a support vector machine that we use for regression. Supervised learning is used to train the models with the data collected from the experience where each data sample consists of a input vector ( $\mathbf{x}_i \in \mathcal{R}^3$ ) described by the red, green and blue components, and the corresponding measured moisture ( $tm_i$ ) also named the target moisture.

## 2.3 Multilayer perceptron method

A multilayer perceptron (*MLP*) [14] with one hidden layer of seven neurons is constructed in order to predict the soil moisture content from the soil digital images (Fig.1). The number of hidden neurons of the network is set to seven according to the Kolmogorov method which states that the number of hidden neurons in *MLPs* is equal to:  $2 \cdot (\text{Number of inputs}) + 1$  [15]. A higher number of hidden neurons might give better results but it could limit the generalization capabilities of the network. The input layer is fully connected to the hidden layer. The hidden layer is fully connected to the output neuron. The sigmoid function:

$$f(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

is the activation function of the hidden neurons, whereas the linear function is the activation function of the output neuron.

The network training is performed according to the Levenberg-Marquardt algorithm [16]. It is an algorithm based on the Newton method used in optimization to find the minimum of the error function. This algorithm is fast but requires more memory since it computes the Jacobian matrix, which is not a problem in this case since the network size is limited. The data vectors collected from the experience are divided randomly into three subsets.

The first one is the training subset (*T*) consisting of 70% of the data vectors. It is used to compute the gradient and update the network weights and biases. The second subset is the validation subset (*V*) consisting of 15% of the data vectors. It is used to stop the training when the validation error increases for a specific number of epochs, after being decreased during the training. The test subset (*Tt*), consisting of 15% of the data vectors, is not used during training. It is used to test the performance of the network and to compare different networks.

The supervised learning of the network consists of the following steps (Fig.2):

- 1 Initialize the network weights with random values.
- 2 Present the samples ( $\mathbf{x}_i, tm_i; i \in T$ ) of the training subset. The input data vector  $\mathbf{x}_i$  is presented to the network through the input layer and the value of the output neuron is the predicted moisture.
- 3 Compare the predicted moisture ( $pm_i$ ) to the target moisture ( $tm_i$ ) for all the training samples and calculate the overall error function:
- 4 Adjust the network weights and biases in order to minimize the error function:

$$E_t = \frac{1}{2} \sum_i (pm_i - tm_i)^2, i \in T \quad (3)$$

- 5 Compute the error function on the validation subset:

$$E_v = \frac{1}{2} \sum_i (pm_i - tm_i)^2, i \in V \quad (4).$$

Go to step 2 if  $E_v$  is decreasing. Stop the training if  $E_v$  increases for 6 consecutive epochs.

The steps from 2 to 5 form a training epoch.

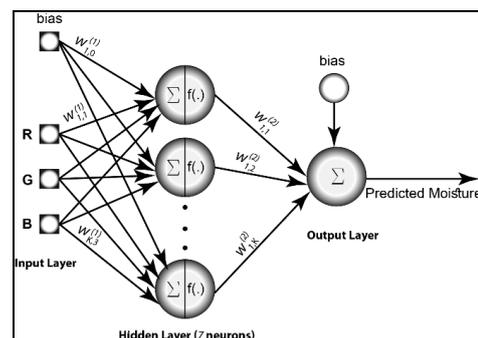


Fig. 1. MLP architecture.

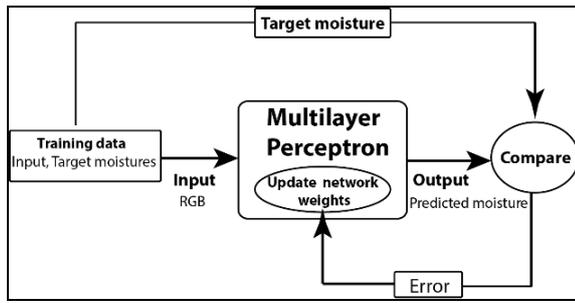


Fig. 2. MLP training.

### 2.4 Support vector regression method

Support Vector Machines (*SVM*) is a popular machine learning tool for classification and regression [17]. The *SVM* formulates a quadratic optimization problem that ensures a global minimum, which makes them outperform traditional learning algorithms. In this study, we use the  $\epsilon$ -intensive *SVM* or ( $\epsilon$ -*SVR*) regression in order to predict soil moisture from *RGB* predictors. The goal is to find a function  $f(x)$  that has at most  $\epsilon$  deviation from the actually obtained targets moisture  $tm_i$  for all the training data, and that is as flat as possible (Fig.3).

A nonlinear regression is achieved by using a Gaussian kernel function that map data into a higher dimensional space.

The data collected from the experience is split into two subsets: the training subset consisting of 85% of the data used to construct the model, whereas remaining 15% are used as test subset to assess the model. The observations that constitute the test subset are the same for the test subset used in the case of *MLP*.

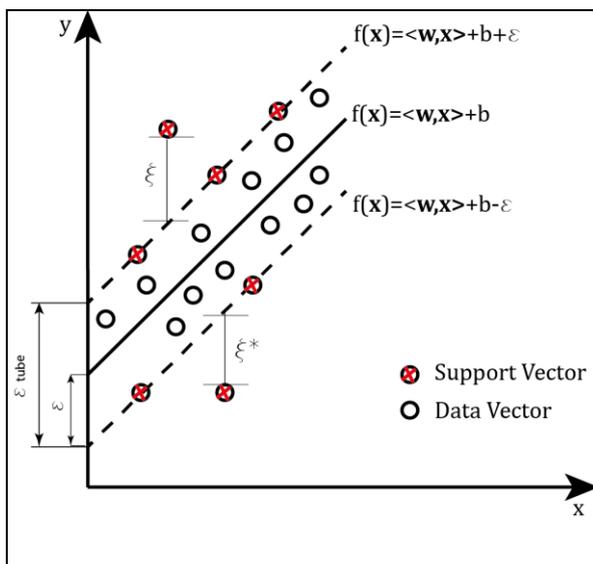


Fig. 3.  $\epsilon$ -intensive *SVR*.

## 3 Results

In the following, the results of using the *MLP* and the *SVR* models are exposed. Comparison between both

models and between other methods proposed in similar works is also discussed. The Pearson correlation coefficient  $R$  of the predicted moisture and the target moisture is used to evaluate the performance of a model. Values close to one of this coefficient indicate a good prediction model. The mean squared error (*MSE*) between the predicted moisture and the measured moisture is also used to evaluate the network performance and to compare different models. Lower values of *MSE* indicate better prediction model.

### 3.1 Multilayer perceptron results

The Matlab Deep Learning toolbox is used to construct and train the *MLP* model [18]. In order to overcome the problem of local minima, the *MLP* is trained 50 times, and the replicate that yields the smallest error  $E_i$  (Eq.(3)) is chosen.

Table 4 shows  $R$  and *MSE* values of the test subset, as well as the number of training epochs when the *MLP* is trained for the six soil types individually as well as for all soil types. The best results are obtained for type-4 ( $R=0.972$ ,  $MSE=4.0729e-4$ ). Less promising results are obtained for type-6 and when training is done with data from all types. Fig.4 and Fig.5 show the regression plots of the target moisture and the predicted moisture on the test subset when the *MLP* is trained with data collected from type-4 alone and from all types respectively.

### 3.2 Support vector regression results

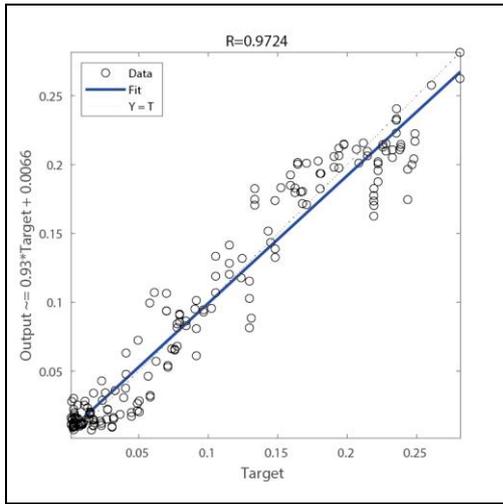
A support vector regression model is built using the Matlab Statistics and Machine Learning Toolbox [19] in order to predict the soil moisture from the soil digital images. The values of the model parameters are the default values set by the toolbox.

Table 5 shows the values of  $R$ , *MSE* as well as the percentage of support vectors, when a *SVR* model is trained for the 6 individual soil types as well as for all types combined.

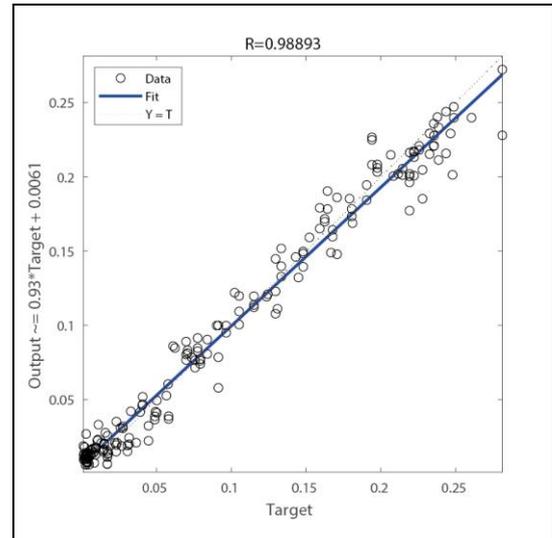
Similarly, as in the *MLP* case, the highest value of  $R$  (0.989) and lowest value of *MSE* ( $1.7818e-04$ ) are obtained for type-4. Less promising results are obtained for type-6 and when training is done with data from all types. Fig.6 and Fig.7 show the regression plots of the target moisture and the predicted moisture on the test subset when the *SVR* model is constructed with data collected from type-4 alone and from all types respectively.

Table 4. MLP prediction results.

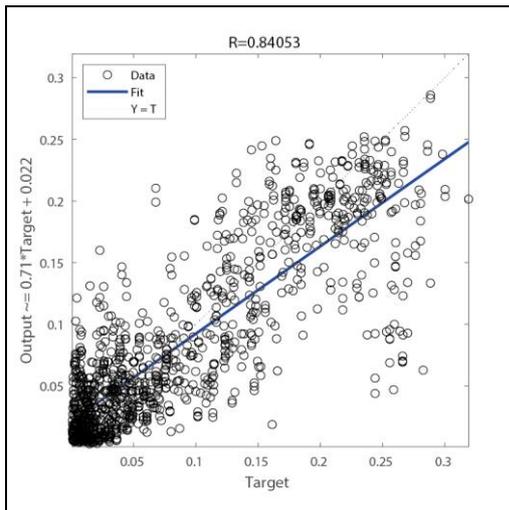
Soil type	$R$	<i>MSE</i>	No. of Epochs
Type-2	0.965	4.5939e-4	105
Type-3	0.945	8.9160e-4	72
Type-4	0.972	4.0729e-4	56
Type-5	0.882	6.4402e-4	176
Type-6	0.841	0.0023	54
Type-7	0.877	0.0018	37
All-types	0.840	0.0021	76



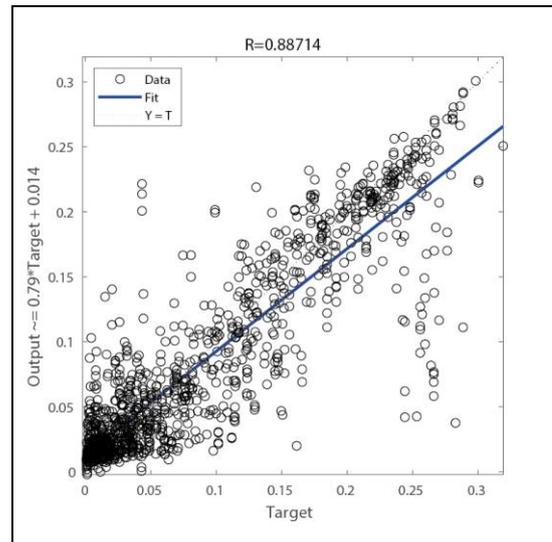
**Fig. 4.** MLP type-4 regression plot.



**Fig. 6.** SVR Type-4 regression plot.



**Fig. 5.** MLP all types regression plot.



**Fig. 7.** SVR All types regression plot.

Table 5 shows the values of  $R$ ,  $MSE$  as well as the percentage of support vectors, when a SVR model is trained for the 6 individual soil types as well as for all types combined.

Similarly, as in the MLP case, the highest value of  $R$  (0.989) and lowest value of  $MSE$  ( $1.7818e-04$ ) are obtained for type-4. Less promising results are obtained for type-6 and when training is done with data from all types. Fig.6 and Fig.7 show the regression plots of the target moisture and the predicted moisture on the test subset when the SVR model is constructed with data collected from type-4 alone and from all types respectively.

**Table 5.** SVR prediction results.

Soil type	$R$	$MSE$	Percent.of SVs
Type-2	0.975	$3.3496e-04$	39.56%
Type-3	0.976	$4.2378e-04$	49.00%
Type-4	0.989	$1.7818e-04$	37.02%
Type-5	0.914	$4.8239e-04$	61.71%
Type-6	0.8921	0.0017	57.17%
Type-7	0.9428	$9.0781e-04$	52.72%
All-types	0.8871	0.0015	63.04%

## 4 Results interpretation

According to Tables 4 and 5, and Fig. 4, 5, 6 and 7, the results obtained with the SVR method are better than those obtained with the MLP method. Moreover, SVR ensure global minimum and is deterministic whereas MLPs are at risk to be stuck in local minima and their outputs depend on the initial connection weights.

By comparing the results obtained for the different soil types (Tables 4 and 5), we notice that for type-4 (Eutric Leptosols) and for type-2 (Calcic Cambisols), the obtained  $R$  values are the highest and the obtained  $MSE$  values are the lowest. It is probably due to their colour which is lighter than the colour of other types (Table 1), an evidence assessed by [11].

When comparing our work to other similar ones, we find that the proposed methods perform better than the ones proposed in [13] and in [12] in term of correlation coefficient. According to [13], the correlation coefficients ranged between 0.703 and 0.909. But, it is important to mention that they used a camera resolution of 7.1 Mpx

which is lower than the one used in this study. Besides, the experiments were conducted on tropical soils having different soil compositions. Santos *et al.* built a different linear regression model for each soil type and the resulting correlation coefficients varied between 0.8538 and 0.9506 [12]. Persson obtained better results with correlation coefficients ranging from 0.965 to 0.995 [11], yet, the investigated soils contained higher percentages of sand (above 40%) that increase their reflectance which is not the case with the present study (Table 2). It is obvious that better prediction results are obtained when the regression models are trained with data of individual soil types. Since soil colour may be also affected by the soil physical, biological and chemical properties [20-21], it would be wise to include some soil properties in addition to colour data to train the prediction model especially if soils of different types are involved.

## 5 Conclusion

In this paper, we implemented a multilayer perceptron and a support vector regression to predict vineyards soil moisture from digital images. The experiments were conducted on six soil types collected from Chateau Kefraya terroirs in Lebanon. The training data consisted of *RGB* pixel values extracted from the soil images and of the associated measured soil moisture by means of the thermo-gravimetric method. The *SVR* method predicted the soil moisture better than the *MLP* one and better than other regression methods found in earlier studies. As prospects to this work, the models might be tested in real time by collecting digital photos on the site and by comparing the predicted moisture to the real measured moisture.

The *SVR* based model succeeded in predicting the soil moisture from digital images, especially if individual soil types are investigated. It constitutes a simple smart tool for soil moisture prediction in vineyards which simplifies and automates viticulture terroirs characterization.

The authors would like to thank the Lebanese National Council for Scientific Research (CNRS) for funding the whole research, and the management of Chateau Kefraya for the technical support they provided.

## References

1. S. Priori, S. Pellegrini, R. Perria, S. Puccioni, P. Storchi, G. Valboa and E. A. C. Costantini, *Scale effect of terroir under three contrasting vintages in the Chianti Classico area (Tuscany, Italy)*, *Geoderma*, **334**, 99-112 (2019).
2. E. A. C. Costantini, R. Lorenzetti and G. Malorgio, A multivariate approach for the study of environmental drivers of wine economic structure, *Land use policy*, **57**, 53-63 (2016).
3. P. Clingeffer, *Terroir: The Application of an Old Concept in Modern Viticulture*, 277-288 (2014).
4. C. V. Leeuwen, 9 - Terroir: the effect of the physical environment on vine growth, grape ripening and wine sensory attributes, in *Managing Wine Quality*, A. G. Reynolds, Ed., Woodhead Publishing, 273-315 (2010).
5. S. U. Susha Lekshmi, D. N. Singh and M. S. Baghini, A critical review of soil moisture measurement, *Measurement*, **54**, 92-105 (2014).
6. P. Dobriyal, A. Qureshi, R. Badola and S. A. Hussain, A review of the methods available for estimating soil moisture and its implications for water resource management, *J. Hydrol* **458-459**, 110-117 (2012).
7. C. T. Altendorf, R. Elliott, E. W. Stevens and M. L. Stone, Development and validation of a neural network model for soil water content prediction with comparison to regression techniques, *T ASABE*, **42**, 691-700 (1999).
8. M. K. Gill, T. Asefa, M. W. Kemblowski and M. McKee, Soil moisture prediction using support vector machines, *J AM WATER RESOUR AS*, **42**, 1033-1046 (2006).
9. A. Elshorbagy and K. Parasuraman, On the relevance of using artificial neural networks for estimating soil moisture content, *J. Hydrol.*, **362**, 1-18 (2008).
10. M. Persson, Estimating Surface Soil Moisture from Soil Colour Using Image Analysis, *VADOSE ZONE J. - VADOSE ZONE J*, **4**, 1119-1122, 11 (2005).
11. J. F. C. d. Santos, H. R. F. Silva, F. A. C. Pinto and I. R. d. Assis, Use of digital images to estimate soil moisture, *Rev. bras. eng. agríc. ambient.*, **20**, 1051-1056, 12 (2016).
12. L. Pasolli, C. Notarnicola and L. Bruzzone, Estimating Soil Moisture With the Support Vector Regression Technique, *IEEE GEOSCI REMOTE S*, **8**, 1080-1084, 11 (2011).
13. S. S. Haykin, *Neural networks and learning machines*, Third ed., Pearson Education (2009).
14. I. N. Silva, D. Hernane Spatti, R. Andrade Flauzino, L. H. B. Liboni and S. F. Reis Alves, *Multilayer Perceptron Networks, in Artificial Neural Networks: A Practical Course*, Cham, Springer International Publishing, 55-115 (2017).
15. M. T. Hagan and M. B. Menhaj, Training feedforward networks with the Marquardt algorithm, *IEEE T NEUR NET LEAR*, **5**, 989-993, 11 (1994).
16. V. N. Vapnik, *Methods of Function Estimation, in The Nature of Statistical Learning Theory*, New York: Springer New York, 181-224 (2000).
17. *Deep Learning Toolbox*, MATLAB, R2018b.
18. *Statistics and Machine Learning Toolbox*, MATLAB, R2018b.
19. M. Aitkenhead, C. Cameron, G. Gaskin, B. Choisy, M. Coull and H. Black, *Digital RGB photography*

and visible-range spectroscopy for soil composition analysis, *Geoderma*, **313**, 265-275 (2018).

20. Y. N. Vodyanitskii and A. T. Savichev, The influence of organic matter on soil colour using the regression equations of optical parameters in the system CIE-L\*a\*b\*, *Ann. Agrar. Sci.*, **15**, 380-385 (2017).