Spanish Journal of Agricultural Research
15(2), e0205, 12 pages (2017)
eISSN: 2171-9292
https://doi.org/10.5424/sjar/2017152-9090
Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, O.A, M.P. (INIA)

RESEARCH ARTICLE

OPEN ACCESS

Machine learning applied to the prediction of citrus production

Irene Díaz[1], Silvia M. Mazza[2], Elías F. Combarro[1], Laura I. Giménez[2] and José E. Gaiad[2]

[1] *Universidad de Oviedo, C/ San Francisco 1, 33003 Oviedo, Asturias. Spain*

[2] *Universidad Nacional del Nordeste, 25 de Mayo 868, W3400BCH Corrientes. Argentina*

Abstract

An in-depth knowledge about variables affecting production is required in order to predict global production and take decisions in agriculture. Machine learning is a technique used in agricultural planning and precision agriculture. This work (i) studies the effectiveness of machine learning techniques for predicting orchards production; and (ii) variables affecting this production were also identified. Data from 964 orchards of lemon, mandarin, and orange in Corrientes, Argentina are analysed. Graphic and analytical descriptive statistics, correlation coefficients, principal component analysis and Biplot were performed. Production was predicted via M5-Prime, a model regression tree constructor which produces a classification based on piecewise linear functions. For all the species studied, the most informative variable was the trees' age; in mandarin and orange orchards, age was followed by between and within row distances; irrigation also affected mandarin production. Also, the performance of M5-Prime in the prediction of production is adequate, as shown when measured with correlation coefficients (~0.8) and relative mean absolute error (~0.1). These results show that M5-Prime is an appropriate method to classify citrus orchards according to production and, in addition, it allows for identifying the most informative variables affecting production by tree.

Additional keywords: lemon; mandarin; orange; M5-Prime; age; framework; irrigation.

Abbreviations used: KNN (K- Nearest-Neighbor); MAE (Relative Mean Absolute Error); ML (Machine Learning); PCA (Principal Component Analysis); PF (Precision Farming); R (Pearson Correlation Coefficient); RMSE (Root Mean Square Error); SDR (Standard Deviation Reduction); SVR (Support Vector Regression)

Authors' contributions: Conceived and designed the experiments: SMM; performed the experiments: SMM, LIG and JEG. Analyzed the data: ID, LIG and EFC. Contributed reagents/materials/analysis tools: ID, EFC and JEG. Wrote the paper: SMM, ID and EFC.

Citation: Díaz, I.; Mazza, S. M.; Combarro, E. F.; Giménez, L. I.; Gaiad, J. E. (2017). Machine learning applied to the prediction of citrus production. Spanish Journal of Agricultural Research, Volume 15, Issue 2, e0205. https://doi.org/10.5424/sjar/2017152-9090

Received:04 Dec 2015. **Accepted:** 01 Jun 2017

Funding: Northeastern University (project SGCYT PI 1713CA03); Spanish Ministry of Economy and Competitiveness/FEDER (project TEC2015-67387-C4-3-R)

Competing interests: The authors have declared that no competing interests exist.

Correspondence should be addressed to Silvia M.Mazza: smmazza@gmail.com

Introduction

Agriculture implies high levels of production risks and many variables must be considered to take decisions. In order to define management strategies and development programmes, an adequate knowledge about variables most directly affecting production is essential. Understanding the behaviour of variables is difficult due to the complexity of relationships and to the amount of factors involved. Citrus production becomes a special challenge due to the significant spatial and temporal variability present in orchards.

In citrus orchards, production is primarily defined by the amount and size of fruits. Production can be affected by both endogenous and exogenous factors. Endogenous factors are, for instance, genetic characteristics of species or varieties, and physiological issues. Among the exogenous factors, environmental and crop conditions, especially irrigation and fertilisation, are highlighted (Agustí, 2000, 2003). Production is also determined by trees' age, and their reaching a commercial production volume (> 50 kg/tree) at adult age (> 7 years after transplant) (Orduz-Rodríguez *et al*., 2007).

Citrus trees' development is possible between 10°C and 40°C and optimised between 24°C and 32°C. Fruit size and final set depend, among other factors, on the availability of carbohydrates for developing flowers. Thermal influence is very limited in the range of 22°C to 30°C. However, if leaf temperature rises above 32°C, the CO_2 assimilation rate decreases. Thermal influence on growth and competition between vegetative and

reproductive developments, emphasise problems from a limitation in the CO_2 fixation, such as the alternation of productivity between seasons, reducing fruits' size, and final fruits set (Agustí, 2003).

Maximum and average temperatures, reference evapotranspiration, wind speed, and relative humidity, are the meteorological variables with the greatest influence on fresh dough and equatorial diameter of fruits. In citrus orchards growing at temperate climates, autumn rains improve the fruits' final size and juice content, and reduce the concentration of sugars and free acids. Total annual rainfall between 900 mm and 1200 mm is enough to ensure fruit development. On the other hand, drought periods (even if short) tend to reduce the fruit size. When lower values or dry seasons occur, complementary irrigation is needed (Agustí, 2000, 2003). Irrigation absence mainly affects the fruit size, although the effect of this absence also depends on the phenological state (González-Altozano & Castel, 2003; García Petello & Castel, 2004; Gasque *et al*., 2010).

Many variables must be considered prior to making decisions about planting the framework. Tree vigour and growth habitat, as influenced by variety and rootstock, are important, and site quality in terms of climate, soil characteristics, and water availability must be considered. In general, higher density plantings that rapidly develop into a hedgerow appear to be advantageous, especially at the beginning of trees' production life. However, vigorous combinations with more spreading growth habits should be planted with wider spacing (Tucker *et al.*, 1994; Medina-Urrutia *et al.*, 2004).

Machine Learning (ML) is a branch of artificial intelligence that provides methods with the ability to learn from or to make predictions on data. These methods build a model from example inputs in order to make predictions or to take decisions (Mitchell, 1997). ML does not make any assumptions about the right structure of the data model, allowing the construction of complex non-linear models. There are many different paradigms in ML: lazy methods such as K-Nearest Neighbours (KNN) (Altman, 1992) methods, based on tree construction, as, for instance, C4.5 (Quinlan, 1993) or Neural or Bayesian networks (Mitchell, 1997). All of them have been successfully used in many different domains. For instance, neural networks have been successfully used to predict maximum dry density and unconfined compressive strength of cement-stabilised soil (Das *et al*., 2011), or to detect structural damage (Alavi *et al.*, 2016a,b).

In particular, some of these methods have been applied for comprehensive agricultural planning in precision farming (PF) (Arango *et al.*, 2015). PF techniques provide a complete knowledge about spatial variability and the different characteristics of a specific area, helping to define more efficient and rational crop management plans in relation to a more localised use of fertilisers and agrochemicals (Yu *et al*., 2010; Fernandez Quintanilla *et al.*, 2011).

Among the huge number of issues related to PF, pest prediction is a task where different ML techniques have been successfully applied. In particular, Bayesian techniques have been adopted (Tripathy *et al.*, 2011; Pérez-Ariza *et al*., 2012). On the other hand, support vector machines are also extensively used (Wang & Ma, 2011).

When the variable to predict is continuous, ML methods more commonly used are CART (Breiman, 2001), M5 (Quinlan, 1992), M5-Prime (Wang & Witten, 1997), KNN (Altman, 1992), or support vector regression (SVR, see Basak *et al.*, 2007).

The model tree technique (see, for example, Frank *et al*., 1998, or Samadi *et al*., 2014) is based on combining decision trees with linear regression functions at the leaves. There are several techniques to predict numeric values instead of just a label. Standard regression imposes a linear relation on data; hence, it is not quite powerful. On the other hand, other paradigms not based on constructing a tree (such as Neural Networks, SVR or lazy classifiers) can be quite powerful, but their interpretability is low.

Regarding model tree techniques, the strategy to construct the tree is similar for all of them (El Gibreeb & Aksoy, 2015). The main differences among the methods are the splitting criteria, the pruning rules, and the mechanism to estimate the leaf value. CART uses variance as the splitting criteria, while M5 uses standard deviation reduction (SDR). In addition, the estimated value for a leaf is constant in CART. In contrast, M5 approximates the leaf values by linear regression models. In addition, it is able to improve predictions by introducing a smoothing procedure (Quinlan, 1992). Furthermore, trees generated with M5 are smaller than those generated with CART. Thus, M5 outperforms CART in accuracy and simplicity (Uysal & Altay, 1999). M5-Prime is an improvement over M5 that can deal with missing values and enumerated attributes (Wang & Witten, 1997) and has been used, for instance, to predict streamflow (Onyari & Ilunga, 2013), to model sediment yield (Goyal, 2014), to estimate the maximum scour depth at breakwaters (Pourzangbar *et al*., 2017) and to predict the compressive strength of high performance concrete (Behnood *et al*., 2017).

González-Sánchez *et al*. (2014) compared the predictive accuracy of ML and linear regression techniques for crop yield prediction in ten crop datasets. Multiple linear regression, M5-Prime model trees, Perceptron Multilayer Neural Networks, SVR,

and KNN methods, were ranked. M5-Prime and KNN techniques obtained the lowest errors and the highest average correlation factors. M5-Prime, which achieves the largest number of crop yield models with the lowest errors, was considered a very suitable tool for massive crop yield prediction in agricultural planning. In addition, it is more interpretable than KNN. Other approaches combine partial least square models and spectral imaging technology (Ye *et al*., 2007).

As production-predictive tasks require the learned model to predict a numeric value associated with a variable rather than the class the example belongs to, model regression trees are proposed. Hence, this work checks the effectiveness of ML techniques in order to determine the affecting variables and classify citrus orchards according to production. In particular, the predictive mechanism established in this work to characterise the variables involved, and to identify the most important factors affecting citrus production, is based on the M5-Prime method.

Material and methods

The studies have been conducted during seasons 2013 and 2014, with field information from 964 Citrus orchards in the province of Corrientes, Argentina, located at latitudes 57°W to 59°W, and longitudes 27°S to 31°S. Orchard tree canopies belong to several varieties of three species: lemon (*Citrus limon* Burman), mandarin (*Citrus reticulata* Blanco), and sweet orange (*Citrus sinensis* Osbeck), over diverse rootstocks.

Every orchard was characterised by the following variables: global position (latitude and longitude degrees, minutes and seconds); annual minimum and maximum average temperatures (ºC), annual total rainfall (mm) and annual total frost-free days defined from the corresponding isolines at orchards' location; environment, species, variety, age of trees, planting framework (between rows' distance, m; within rows' distance, m), presence or absence of irrigation (binary) and production by tree (kg/tree).

Lemon was present in 94 orchards (9.6%), placed at 28°S to 30°S and 57°W to 59°W, in Mesopotamic Park and savanna environments, with annual average temperatures between 18°C and 21°C, total annual rainfall between 1000 mm and 1200 mm, and 320 to 340 frost free days in the year. Two varieties of lemon were found in the studied orchards: 'Eureka' (71% of orchards) and 'Genova' (26% of orchards). In addition, 3% of orchard varieties could not be identified (Unknown). Only 26.5% of the orchards were under irrigation, with similar percentages in all varieties. The characteristics of these orchards are presented in Table 1.

Mandarin was present in 364 orchards (37.6%), placed at 28°S to 30°S and 57°W to 59°W, with annual average temperatures between 18°C and 21°C, total annual rainfall between 1000 mm and 1200 mm and 320 to 360 frost free days in the year, in mesopotamic park and savanna environments (however, 'Clemenules', 'Murcott', 'Criolla', 'Nova', 'Dancy' and 'Okitsu' varieties appeared in all locations; W Murcott is present only at 59°W, 29°S in mesopotamic park environment and the others only at 57°W, 30°S in savanna environment). Twelve varieties of mandarin were found in the studied orchards: 'Murcott' (24% of orchards), 'Ellendale' (20%), 'Okitsu' (15%), 'Nova' (12%), 'Dancy' (8%), 'Clemenules' (6%), 'Criolla' (5%), 'Encore' (3%), 'Ortanique' (2%), 'Malvacio' (1%), 'Montenegrina' (1%) and 'W Murcott' (1%). In 1% of orchards, the variety could not be identified (Unknown). Irrigation was present in 45% of orchards, with higher percentages in 'Montenegrina', 'W Murcott', 'Clemenules', 'Nova' and 'Murcott'. Table 1 presents a description of these orchards.

Orange was present in 509 orchards (52.8%), placed at 28°S to 30°S and 57°W to 59°W, in mesopotamic park and savanna environments, with annual average temperatures between 18°C and 22°C, total annual rainfall between 1000 mm and 1400 mm, and 320 to 360 frost free days in the year. Fourteen varieties of orange were found in the studied orchards: Valencia late (50% of orchards), Salustiana (8%), Valencia seedless (7%), Washington navel (7%), Delta seedless (5%), Valencia frost (4%), Criolla (3%), Lane late (2%), Navel late (2%), Navelina (2%), Robertson navel (1%), Newhall (0.2%), Hamlin (0.2%) and Westin (0.2%). In 7% of the orchards, the variety could not be identified (Unknown). Irrigation was present in 42.4% of the orchards, with higher percentages in Salustiana, Midknight, Navelina, Robertson Navel, and Newhall. Description of these orchards is presented in Table 1.

Statistical analysis

Graphic and analytical descriptive statistical tools were used, and Pearson correlation coefficients (*R*) calculated, in order to define and characterise the relationships between all variables and production by tree. Principal component analysis (PCA) and Biplot graphics were performed to reduce dimension in a way that allows for examining data in a less dimensional space. PCA builds artificial axes (principal components) with maximum variability, enabling scatter plots of observations and/or variables

with optimum properties for the interpretation of the underlying variability and co-variability. In Biplots, observations and variables can be visualised in the same space, and possible associations between variables and observations can be identified (Di Rienzo *et al*., 2015). These analyses were performed with InfoStat 2015 (Di Rienzo *et al.*, 2015).

Learning approach

Based on endogenous (species, varieties, age of trees) and exogenous factors (global position, annual minimum and maximum average temperatures, total rainfall and total frost-free days, environment, planting framework and irrigation) (Agustí, 2003), citrus production was predicted via regression trees, which have been demonstrated as suitable methods to crop yield prediction.

Table 1. Characterisation of lemon, mandarin and sweet orange orchards: latitude degree (LATD), longitude degree (LONGD), annual average temperature (TAV), total rainfall (TR), frost-free days (FFD), trees' age (AGE) and trees' production (PROD), during seasons 2013 and 2014

Description	Minimum	Maximum
Lemon		
LATD (S)	28º	30º
LONGD (W)	57º	59º
TAV (ºC)	18º	21º
TR (mm)	1000	1200
FFD	320	340
AGE (years)	6	19
PROD (kg)	5.63	768.88
Mandarin		
LATD (S)	28º	30º
LONGD (W)	57º	59º
TAV (ºC)	18º	21º
TR (mm)	1000	1200
FFD	320	360
AGE (years)	1	62
PROD (kg)	1.11	2222.22
Sweet orange		
LATD (S)	27º	31º
LONGD (W)	57º	59º
TAV (ºC)	18º	22º
TR (mm)	1000	1400
FFD	320	360
AGE (years)	2	50
PROD (kg)	0.27	200

M5-Prime is a learner which constructs regression trees producing a classification, based on piece-wise linear functions (Wang & Witten, 1997). To do that, the space is partitioned into a set of regions. Further, the predicted value is fitted within each region using a linear model. The way this method works is the following: Assuming a training set with examples, each one defined by its value on a set of attributes (discrete or continuous) and a continuous target, the method constructs a model that relates the target values of the training examples to the values of the variables defining the example. This model can then be easily applied to predict the target variable: in the first phase, the decision tree (see, for example, the ones in Fig.s 2, 4, and 6) is used to classify the example into one of the groups; then, the linear equation associated with the particular group the example has been classified into, is used to predict the target variable (see Tables 2, 4, and 6 for examples of these equations).

M5-Prime selects the split that maximises the expected error reduction. Once the tree is constructed, a multivariate linear model is computed for the examples at each tree node with standard regression techniques and using only attributes that are referenced by tests or linear models somewhere in the sub-tree under this node. The main characteristics of this method are:

1. Regression tree construction:

a) Splitting criterion: Maximise SDR

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i)$$

T being the set of examples (orchards in this case) that reaches the node and T_1, T_2, … the subsets resulting from the node split according to the chosen attribute.

b) Stopping criterion: Standard deviation below a given threshold (small enough) in all nodes.

c) Pruning: Heuristic estimation of absolute error of linear regression models.

$$\frac{n + \upsilon}{n - \upsilon}$$

with n being the number of examples that reach the node and υ the number of parameters that represent the class value at that node. Pruning greedily removes terms from linear regression models to minimise the estimated error.

d) Smoothing is used to compensate discontinuities between the adjacent linear models at the leaves of the pruned tree. The smoothing process uses first the leaf model to compute the predicted value, and then it filters that value along the path back to the root, combining it with the value predicted by the linear

model for that node. The modified prediction p' is computed by

$$p' = \frac{nq + kr}{n + k}$$

with n being the number of examples at the smoothed node, k a constant, q and r are respectively the predictions passed to the studied node from below and the value predicted by the model at the studied node. Basically, this process propagates the effect of incorporating the ancestor models into the leaves.

2. The value at each leaf is estimated using a linear regression function.

3. At each node, it uses only a subset of the attributes occurring in the sub-tree.

The experiments were conducted using the RWeka Package, using the M5-Prime function with the standard configuration, *i.e.*, with pruning, smoothing, and with 4 being the minimum number of examples per node. Bootstrap resampling was used, that involves taking random samples from the dataset (with re-selection) to evaluate the model. In aggregate, the results reduce the effects of random selection. The experiments performed here were repeated 100 times. The models were trained with the original variables described at the beginning of the section. In addition, no feature reduction or extraction was applied since M5-Prime automatically selects the most relevant variables when building the decision trees.

The accuracy of this method was studied in terms of root mean square error (RMSE), correlation coefficient (R) and the relative mean absolute error (MAE). RMSE measures the difference between the real and the estimated value and MAE compares the average of the differences between the real and the estimated values to the average of the estimated values (Han & Kamber, 2006).

Results and discussion

Lemon

The R coefficients calculated indicate that production by tree is significantly associated in a positive way with trees' age (R=0.64; p<0.0001) and longitude (R=0.48; p<0.0001); and in a negative way with rainfall (R=-0.77; p<0.0001).

In Fig. 1, PCA Biplot associations between variables and observations can be identified. Angles between variable vectors and principal components indicate that the principal axis (containing 88.5% of variability) separates the different orchards by production by tree. On the right are more productive orchards, mostly belonging to Genova and Unknown varieties, with higher ages, latitudes, rainfall values and no irrigation. On the left are less productive orchards, primarily belonging to 'Eureka' variety, with higher longitudes, frost-free days, and within and between rows distance. Although orchards' locations present small variations, PCA results indicate variability in production associated to latitude and longitude.

Fig. 2 shows the regression tree obtained by M5-Prime algorithm and Table 2 presents the linear

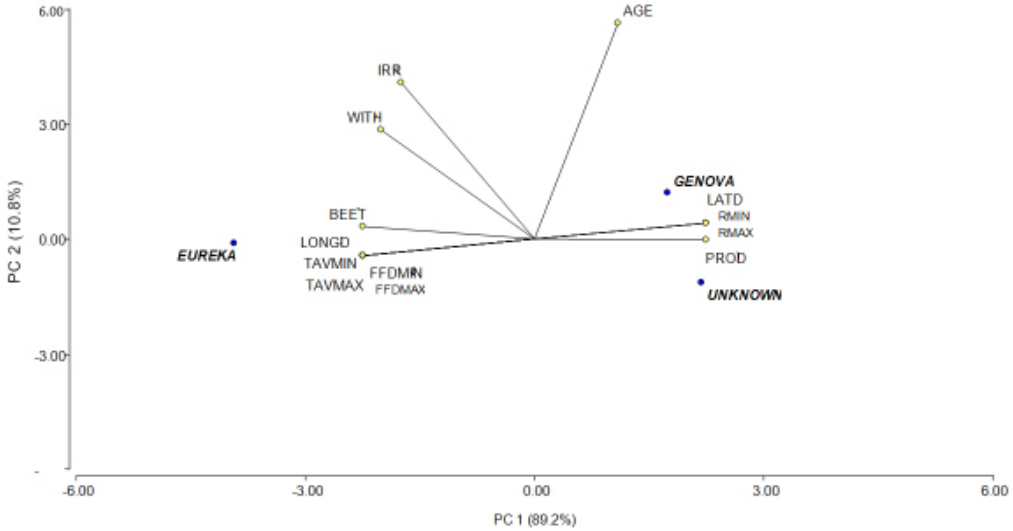

Figure 1. Principal Component Analysis Biplot of trees' production (PROD), trees' age (AGE), irrigation (IRR), latitude and longitude degrrres (LATD, LONGD), plantig framework (distance between rows, BEET and within row, WITH), annual minimum and maximum average temperatures (TAVMIN, TAVMAX), minimum and maximum total rainfall (RMIN, RMAX) and minimum and maximum frost-free days (FFDMIN, FFMAX) by variety (EUREKA, GENOVA, UKNOWN) in lemon orchards, during seasons 2013 and 2014.

Table 2. Linear regression equation associated to each leave of regression tree build by M5 algorithm in lemon orchards (IRR: irrigation, VAR: variety, WITH: within row distance, LATD: latitude degree, AGE: trees' age)

Variable	L1	L2	L3
IRR	0.0273	0.0273	0.2546
VAR=Eureka	-0.0163	-0.0153	
WITH	-0.0120	0.0008	0.0138
LATD	-0.0071	-0.0067	
AGE	0.0033	0.0032	0.0043
Constant	0.3130	0.2794	0.0187

regression equation associated to each leaf. According to that, the best variable to classify lemon production is the trees' age. Thus, orchards could be classified into three groups: L1, with trees' age of 9 years or below, the lowest production and high variability; L2, with trees' age between 9 and 21 years, an intermediate production and the highest variation; and L3, with trees' age over 21 years, the most homogeneous group with the maximum production by tree. Table 3 presents descriptive statistics of production by group.

Results obtained by all techniques related age with production by tree. However, in Biplot, other variables showed smaller angles with production, indicating stronger association. On the other hand, M5-Prime allows for grouping orchards according to production by tree and highlights age as the best classification variable.

In addition, M5-Prime defines groups primarily based on trees' age. Minimum and maximum temperatures, despite being below optimum values (Agustí, 2003), did not differ between orchards. According to Orduz-Rodríguez *et al.* (2007), orchards with trees' age below 9 years (L1), can be defined as pre-productive ones, with production of just over the minimum of 50 kg/

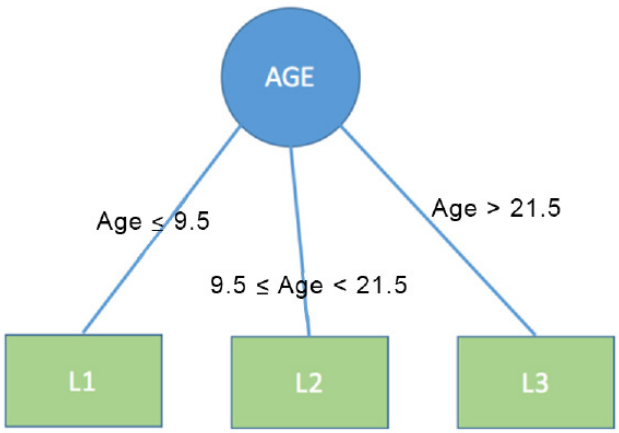

Figure 2. Regression tree of lemon orchard production classification by M5 algorithm (AGE: trees' age)

Table 3. Size (n) and descriptive statistics of production by tree (Av: average, CV: coefficient of variation, Min: minimum, Max: maximum, Med: median, Q1: first quartile, Q3: third quartile) by group in lemon orchards, during seasons 2013 and 2014

Statistics	L1	L2	L3
n	36	41	17
Av	65.44	123.61	418.49
CV	80.77	85.61	54.62
Min	10.00	5.63	30.77
Max	210.00	500.00	769.88
Med	49.86	100.00	405.56
Q1	28.57	58.82	210.00
Q3	80.00	182.19	591.55

tree. Orchards with trees between 9 and 21 years old (L2) (that are in a productive stage) almost doubled the production by tree.

Differences between L1 and L2 were mainly based on differences of weights associated to within rows (see Table 2), probably due to the effects of slightly strong planting in trees at the beginning of production life (trees' ages between 6 and 19 years), agreeing with Tucker *et al*. (1994) and Medina-Urrutia *et al.* (2004) (average distances L1: 6.33 m × 4.31 m; L2: 6.63 m × 4.05 m).

Orchards with tree age of over 21 years (L3), with the weakest planting framework, showed the largest production by tree. In this group, the main factor affecting production was irrigation. This can be deduced from the value of the corresponding coefficient in regression tree and from the fact that 76% of orchards in this group are irrigated. On the other hand, L1 and L2 orchards (< 25%) indicated that this practice is necessary and improves yield, according to Agustí (2000, 2003), González-Altozano & Castel (2003), García Petello & Castel (2004), and Gasque *et al*. (2010).

Differences in regression coefficients with L3 were mostly based on the inclusion of the 'Eureka' variety and latitude degree coefficients. In addition, the weight of irrigation, distance within rows and constant coefficients also influence these differences. Note that latitude and variety (specifically 'Eureka') are not relevant variables for trees with age over 21 years.

Mandarin

According to *R* coefficients, production by tree is significantly associated, in a positive way, with distance between rows (R=0.12; p<0.0274) and within rows (R=0.16; p<0.0023). However, coefficient values indicate a weak association.

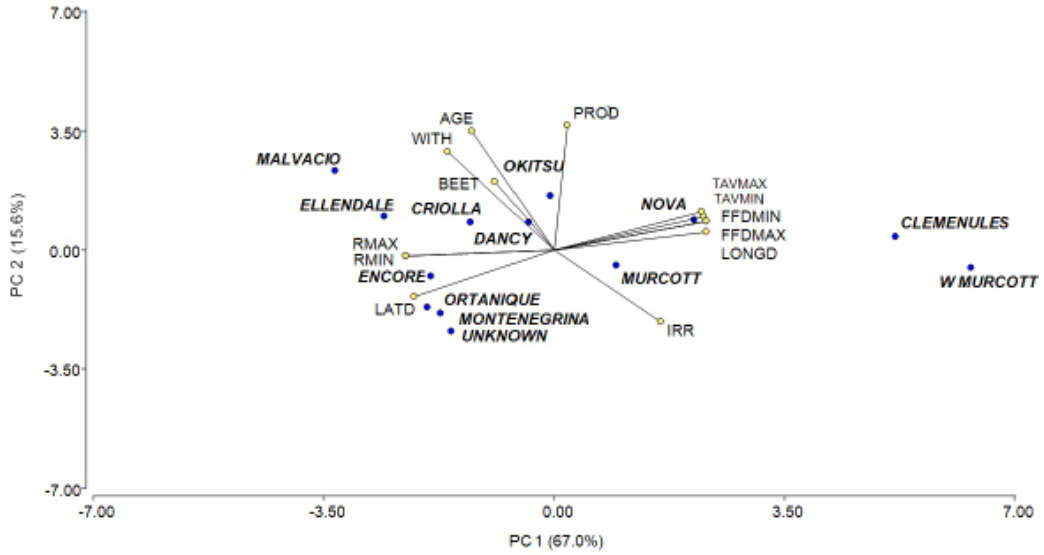

Figure 3. Principal Component Analysis Biplot of production tree (PROD), trees' age (AGE), irrigation (IRR), latitude and longitude degrees (LATD, LONGD), planting framework (between rows, BEET, within row, WITH), annual minimum and maximum average temperatures (TAVMIN, TAVMAX), minimum and maximum total rainfall (RMIN, RMAX) and minimum and maximum frost-free days (FFDMIN, FFDMAX) by variety in mandarin orchards, during seasons 2013 and 2014

Fig. 3 presents PCA Biplot, and associations between variables and observations could be identified. PCA and two coordinates Biplot showed that a principal axis (containing 64.9% of variability) separated the different orchards by location, climatic conditions and varieties, but it was not associated with production by tree. The second axis (conserving 15.6% of variability), separated orchards by productivity. On top were more productive orchards without irrigation, with older trees and higher planting frameworks.

Results obtained from M5-Prime indicate that tree age was the most informative variable to classify mandarin production, followed by irrigation, between and within rows distances as shown in Fig. 4 and Table 4. Although orchards' locations present small variations, PCA results indicate variability in production associated to latitude and longitude.

M5-Prime classified mandarin orchards into eight groups. For instance, M1, with tree age of 11.5 years or below, comprised the largest number of orchards, with one of the smallest productions by tree and high variation. The most relevant variable for the other seven groups associated to orchards older than 11.5 years, was irrigation. Descriptive statistics of production by group are presented in Table 5.

In mandarin orchards, not all techniques related age with production by tree. Age and production were non-significantly associated according to *R* coefficient

Table 4. Linear regression equation associated to each leave of regression tree build by M5 algorithm in mandarin orchards (LATD: latitude degree, LONGD: longitude degree, VAR: variety, AGE: trees' age, BEET: distance between rows, WITH: within row distance, IRR: irrigation)

	M1	M2	M3	M4	M5	M6	M7	M8
LATD	-0.0050	-0.0074	-0.0074	-0.0074	-0.0503	-0.0503	-0.0503	0.0372
LONGD	-0.0057	-0.0042	-0.0042	-0.0042	-0.0042	-0.0042	-0.0042	-0.0042
VAR=Ellendale	0.0480							
VAR=Okitsu					0.0797	0.0797	0.0797	0.0327
VAR=Murcott		-0.0388			-0.3261	-0.3314	-0.2733	
AGE								0.0229
BEET	0.0213	0.0328	0.0524	0.0510	-0.1912	-0.1912	0.2373	0.0024
WITH	0.0031	0.0103	0.0564	0.1006	0.1853	0.1557	0.1273	0.0222
IRR	0.0505	0.0201	0.0201	0.0201	0.0269	0.0269	0.0269	0.0269
Constant	0.4066	0.3280	-0.0595	-0.2581	2.7860	2.8520	3.1283	-1.2502

Table 5. Size (n) and descriptive statistics of production (Av: average, CV: coefficient of variation, Min: minimum, Max: maximum, Med: median, Q1: first quartile, Q3: third quartile) by group in mandarin orchards, during seasons 2013 and 2014

Statistics	M1	M2	M3	M4	M5	M6	M7	M8
n	149	103	13	14	6	5	6	63
Av	92.66	104.72	59.18	249.52	1272.38	313.02	122.08	135.25
CV	90.34	80.71	108.60	126.48	57.00	89.55	41.49	186.56
Min	1.11	6.67	7.00	37.50	280.00	105.00	32.79	3.32
Max	553.56	415.38	240.00	1251.35	2222.22	791.25	187.13	2000.00
Med	75.65	80.00	40.00	166.58	1133.33	210.00	131.47	87.33
Q1	40.43	40.00	20.00	52.91	865.38	135.33	110.17	47.96
Q3	120.00	157.50	80.00	240.00	2000.00	323.53	139.45	125.62

(R=0.08; p=0.1314). However, the fact that simple correlation coefficients associate a variable with another without considering other variables must be taken into account.

PCA and Biplot indicated a high association of production by tree with age, irrigation, within and between rows, and matching with the variables selected by M5-Prime.

Orchards were classified into eight groups by M5-Prime, seven of them associated to the orchards over 11.5-year-old M1 group. Orchards with trees' age under 11.5 years could be considered at the beginning of the commercial production (according to Orduz-Rodríguez *et al*., 2007, criterion). Finally, 75% of orchards were over 7 years.

For the orchards whose trees' ages were greater than 11.5 years, irrigation was an important characteristic. Groups M2, M3, and M4 present irrigation, but their productions were below higher, indicating that annual rainfall between 1000 and 1200 mm could be enough for citrus growth and production (Agustí, 2000, 2003). M2 group was associated to orchards with distance within rows of below 4.75 m. M3 and M4 with distances within rows of over 4.75m also differed on the distance between rows. Orchards in M4 presented the maximum production by tree followed by M2. From the viewpoint of the framework, these results were contrary to Tucker *et al.* (1994) and Medina-Urrutia *et al.* (2004).

Age was again an important variable for the groups with no irrigation, being 13.5 years the split point for age. For the groups associated to ages below 13.5 years (M5, M6 and M7), the distance between rows was relevant. The most productive orchards belonging to M5 and M6 groups presented ages of below 13.5 years and

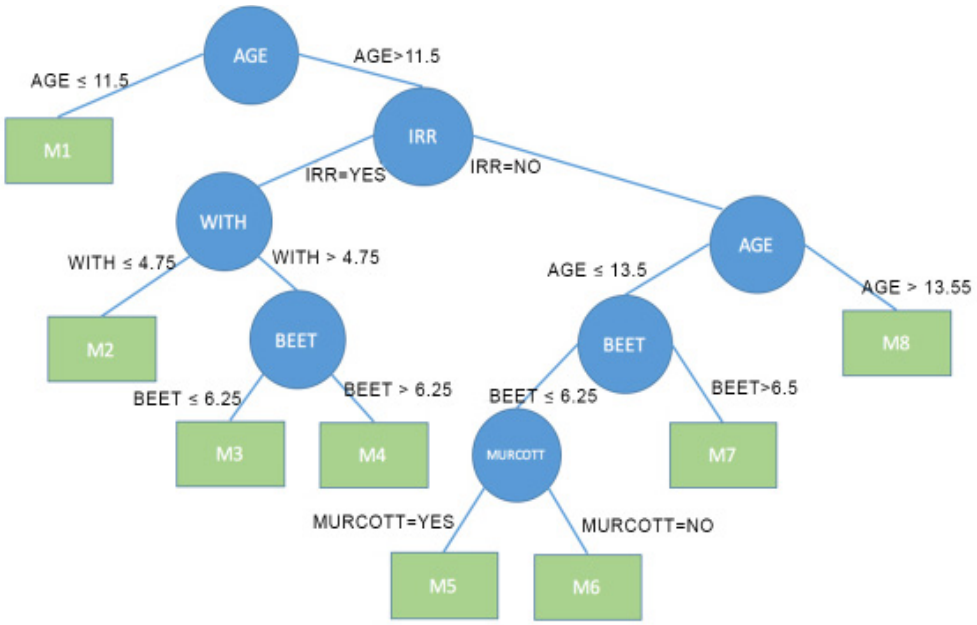

Figure 4. Regression tree of mandarin orchards classification by M5 algorithm (AGE: trees' age, IRR: irrigation, WITH: distance within rows, BEET: distance between rows, MURCOTT: variety 'Murcott' or not).

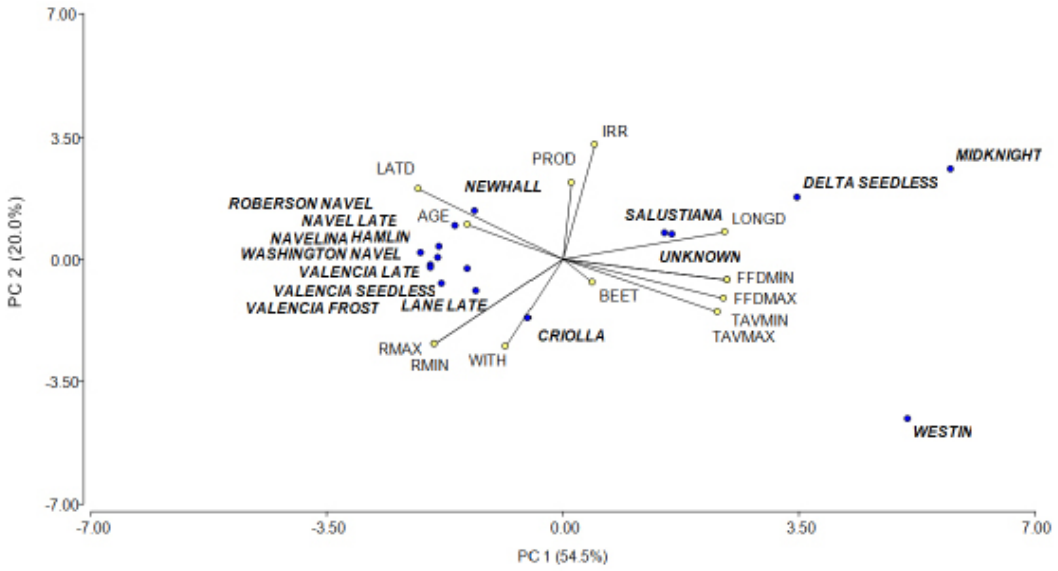

Figure 5. Principal Component Analysis Biplot of production tree (PROD), trees' age (AGE), irrigation (IRR), latitude and longitude degrees (LATD, LONGD), planting framework (between rows, BEET, within row, WITH), annual minimum and maximum average temperatures (TAVMIN, TAVMAX), minimum and maximum total rainfall (RMIN, RMAX) and minimum and maximum frost-free days (FFDMIN, FFDMAX) by variety in orange orchards, during seasons 2013 and 2014.

distance between rows of below 6.5 m. These results strongly agree with the results of Tucker *et al.* (1994) and Medina-Urrutia *et al.* (2004) of higher density advantages at the beginning of the trees' production life. Note that the variety was also relevant for those orchards with trees' age between 11.5 and 13.5 years and with distances between rows of below 6.5 m.

Sweet orange

Production by tree was significantly associated, in a positive way, with age (R=0.14; p=0.0013), and in a negative way with minimum and maximum rainfall (R=-0.10; p=0.02; R=-0.10; p=0.02, respectively); however, coefficients values indicate weak association.

Orange PCA Biplot is shown in Fig. 5, where the associations between variables and observations can be identified. PCA and two coordinates Biplot show that a principal axis (containing 54.1% of variability) separates the different orchards by location, climatic conditions and varieties, but is not associated with production by tree. The second axis (conserving 20.4% of variability) separates orchards by productivity. On

Table 6. Linear regression equation associated to each leave of regression tree build by M5 algorithm in sweet orange orchards (AGE: trees' age, BEET: distance between rows, IRR: irrigation, LATD: latitude degree, VAR: variety, WITH: within row distance)

Variable	O1	O2	O3	O4	O5
AGE	-0.2446	0.0054	0.0023		
BEET	-1.5401	-0.0905	-0.0776	-0.0025	-0.0020
IRR	0.5003	-0.0475	-0.0522	0.0465	0.0351
LATD	-0.0216	-0.0216	-0.0216	0.0398	-0.0036
VAR=Delta seedless				0.0831	
VAR=Robertson navel				-0.0071	-0.0095
VAR= Valencia late	0.6882	0.0601	0.0601	-0.0005	-0.0016
VAR=Valencia seedless				-0.0038	-0.0052
VAR=Washington navel		0.0491	0.0161	-0.0033	-0.0045
WITH	0.0802	-0.0326	-0.0202		
Constant	10.9084	1.3352	1.1707	-1.0694	0.2476

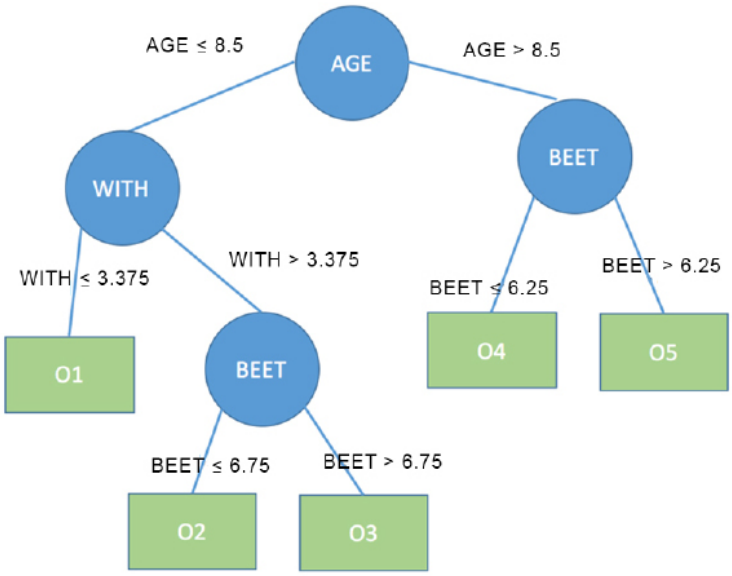

Figure 6. Regression tree of orange orchards classification by M5 algorithm (AGE: trees' age, WITH: distance within rows, BEET: distance between rows)

top are more productive orchards with irrigation, older trees, and higher between-rows distance.

According to Fig. 6 and Table 6, M5-Prime model indicates that age was again the most relevant variable for predicting sweet orange production. Between and within-rows distances were also relevant variables for this species.

M5-Prime classifies orchards into five groups. O1, O2, and O3 are groups with tree age below 8.5 years, with the lowest values of production, and the other groups are over this age. Descriptive statistics of production by groups are presented in Table 7.

Not all techniques related age with production by tree. The *R* coefficients between production and age and rainfall were significant, but low values indicate a weak association. PCA and Biplot indicate a high association of production by tree with irrigation and longitude. Production was related in a negative way with rain, within and between rows. Nevertheless, age appears as a variable, weakly related with production. Only within and between rows' distance matched with the variables selected by M5-Prime.

Table 7. Size (n) and descriptive statistics of production (Av: average, CV: coefficient of variation, Min: minimum, Max: maximum, Med: median, Q1: first quartile, Q3: third quartile) by group in sweet orange orchards, during seasons 2013 and 2014

Statistics	O1	O2	O3	O4	O5
n	34	55	11	252	157
Av	103.38	37.19	99.44	127.00	139.43
CV	178.98	184.06	95.24	135.06	64.81
Min	5.00	0.27	16.00	2.62	8.00
Max	1000.00	123.3	300.00	2000.00	441.18
Med	55.96	30.00	56.9	92.93	136.36
Q1	32.16	7.00	34.00	56.19	70.00
Q3	95.59	56.00	201.12	153.33	194.24

Groups O1, O2, and O3, with tree age of 8.5 years or below (that can be considered by Orduz-Rodríguez *et al.* (2007)´s criterion as initiating the commercial production), exhibit the lowest productions. Group O1, associated to orchards with distance within rows of below 3.375m, presented the higher production in this set indicating the advantages of stronger density planting in younger orchards, agreeing with Tucker *et al*. (1994) and Medina Urrutia *et al*. (2004). Groups O2 and O3 differed on the distance between rows (for O2 was less or equal than 6.75 m and over this value for O3). These results disagree with the criterion that strongly planting framework is associated with more productive orchards, perhaps due to the higher influence of distance within rows over distance between rows (Tucker *et al*., 1994; Medina-Urrutia *et al*., 2004). Finally, the production for oldest trees (>8.5 years), considered commercial production orchards, depended on the distance between rows, being the split point of 6.25 m.

Table 8 shows the performance metrics associated to M5-Prime. The highest *R* was reached when M5-Prime predicts Sweet Orange production (0.828), followed by lemon prediction (0.813). Finally, the worst *R* is obtained for predicting mandarin production. All these values were good enough, compared to the values obtained in González-Sánchez *et al*. (2014), for yield prediction. Regarding RMSE, the highest error was obtained when the production was predicted for sweet orange (0.297), whereas the lemon prediction was the most accurate (0.072). MAE lowest values were

Table 8. Performance or M5-Prime for each species, measured by correlation coefficient (R), root mean square error (RMSE) and relative mean absolute error (MAE)

	Lemon	Mandarin	Sweet orange
R	0.813	0.744	0.828
RMSE	0.072	0.165	0.297
MAE	0.107	0.081	0.102

obtained when production was predicted for mandarin (0.081), obtaining values around 0.10 for lemon and sweet orange production. These values were similar to those obtained in González-Sánchez *et al*. (2014). Both RMSE and MAE were computed as the average, obtained over the 100 repetitions of the bootstrapping process.

Thus, M5-Prime is demonstrated as appropriate to classify citrus orchards and allows for defining more informative, i.e., more relevant, variables affecting tree production. For all the studied species, the most informative variable is tree age; in mandarin and orange orchards, age is followed by between and within rows distances; irrigation also affects mandarin production.

Conclusions

In this work, the factors affecting sweet orange, lemon, and mandarin production were studied using different techniques. In particular, statistical methods such as correlation coefficient, principal component analysis, and Biplot were employed, to identify such factors. In addition, in order to provide a more complete and interpretable point of view, a machine learning technique (known as M5-Prime) was applied.

M5-Prime is demonstrated appropriate to classify citrus orchards and allows for defining more informative, *E*., more relevant, variables affecting tree production. For all the studied species, the most informative variable is tree age; in mandarin and orange orchards, age is followed by between and within rows distances; irrigation also affects mandarin production.

In all species studied, in younger orchards, higher productions are associated with stronger planting densities, mainly distance within rows.

Future studies would involve a more thorough investigation in the possibility of using ML techniques for the prediction of citrus yield, and comparing the effectiveness and efficiency of several different paradigms and learning methods, such as regression trees, SVR, neural networks… as well as combinations of them with techniques such as bagging, boosting or random forests.

New, complementary variables will also be incorporated, such as those obtained from hyperspectral satellite imagery, which have been already used successfully in Precision Farming problems (Arango *et al.*, 2015). Finally, the possibility of extracting qualitative information from the data (for instance, with methods such as the self-organising maps proposed in Kohonen (1982) will be explored in this case.

References

Agustí M, 2000. Crecimiento y maduración del fruto. In: Fundamentos de Fisiología Vegetal. McGraw Hill, Madrid. 669 pp.

Agustí M, 2003. Citricultura. Ed. Mundi-Prensa, Madrid. 456 pp.

Alavi AH, Hasni H, Lajnef N, Chatti K, Faridazar F, 2016a. An intelligent structural damage detection approach based on self-powered wireless sensor data. Aut Construc 62: 24-44. https://doi.org/10.1016/j.autcon.2015.10.001

Alavi AH, Hasni H, Lajnef N, Chatti K, Faridazar F, 2016b. Damage detection using self-powered wireless sensor data: An evolutionary approach. Measurement 82: 254-283. https://doi.org/10.1016/j.measurement.2015.12.020

Altman NS, 1992. An introduction to kernel and nearest-neighbor nonparametric regression. The Amer Statist 46 (3): 175-185. https://doi.org/10.1080/00031305.1992.10475879

Arango RB, Díaz I, Campos AM, Combarro EF, Canas EF, 2015. On the influence of temporal resolution on automatic delimitation using clustering algorithms. Appl Math Inf Sci 9 (2L): 339-347.

Basak D, Pal S, Patranabis DC, 2007. Support vector regression. Neural information processing. Letters and Reviews 11 (10): 203-224.

Behnood A, Behnood V, Gharehveran MM, Alyamac KE, 2017. Prediction of the compressive strength of normal and high-performance concretes using M5P model tree algorithm. Constr Build Mater 142: 199-207. https://doi.org/10.1016/j.conbuildmat.2017.03.061

Breiman L, 2001. Statistical modeling: The two cultures (with discussion). Statist Sci 16 (3): 199-231. https://doi.org/10.1214/ss/1009213726

Das SK, Samui P, Sabat AK, 2011. Application of Artificial Intelligence to maximum dry density and unconfined compressive strength of cement stabilized soil. Geotech Geol Eng 29 (3): 329-342. https://doi.org/10.1007/s10706-010-9379-4

Di Rienzo JA, Casanoves F, Balzarini MG, González L, Tablada M, Robledo CW, 2015. InfoStat versión 2015. Grupo InfoStat, FCA, Universidad Nacional de Córdoba, Argentina. http://www.infostat.com.ar.

El Gibreen H, Aksoy MS, 2015. Classifying continuous classes with reinforcement learning rules. In: Intelligent Information and database systems; Nguyen NT, Trawinski B, Kosala R (eds.), pp: 116-127. Springer Int.

Fernández-Quintanilla C, Dorado J, San Martín C, Conesa-Muñoz J, Ribeiro A, 2011. A five-step approach for planning a robotic site-specific weed management program for winter wheat. Proc. Robotics and Associated High-Technologies and Equipment for Agriculture; Gonzalez de Santos P & Rabatel G (eds.), Montpellier (France), pp. 3-12.

Frank E, Wang Y, Inglis S, Homles G, Witten I, 1998. Using model trees for classification. Mach Learn 32 (1): 63-76. https://doi.org/10.1023/A:1007421302149

García-Petello J, Castel JR, 2004. The response of Valencia orange trees to irrigation in Uruguay. Span J Agric Res 2 (3): 429-443. https://doi.org/10.5424/sjar/2004023-98

Gasque M, Granero B, Turegano JV, González-Altozano P, 2010. Regulated deficit irrigation effects on yield, fruit quality and vegetative growth of ´Navelina´ citrus trees. Span J Agric Res 8 (S2): S40-S51. https://doi.org/10.5424/sjar/201008S2-1347

González-Altozano P, Castel JR, 2003. Riego deficitario controlado en ´Clementina de Nules´. Efectos sobre la producción y la calidad de la fruta. Span J Agric Res 1 (2): 81-92. https://doi.org/10.5424/sjar/2003012-24

González-Sánchez A, Frausto-Solís J, Ojeda-Bustamante W, 2014. Predictive ability of machine learning methods for massive crop yield prediction. Span J Agric Res 12 (2): 313-328. https://doi.org/10.5424/sjar/2014122-4439

Goyal, MK, 2014. Modeling of Sediment yield prediction using M5 model tree algorithm and wavelet regression. Water Resour Manage 28: 1991-2003. https://doi.org/10.1007/s11269-014-0590-6

Han J, Kamber M, 2006. Data mining: concepts and techniques, 2nd ed. Morgan Kaufmann Publ.

Kohonen T, 1982. Self-organized formation of topologically correct feature maps. Biol Cybern 43: 59-69. https://doi.org/10.1007/BF00337288

Medina-Urrutia VM, Becerra-Rodríguez S, Ordaz-Ordaz E, 2004. Crecimiento y rendimiento del limón mexicano en altas densidades de plantación en el trópico. vista Chapingo Serie Horticultura 10 (1): 43-49.

Mitchell T, 1997. Machine learning. McGraw Hill.

Onyari EK, Ilunga FM, 2013. Application of MLP neural network and M5P model tree in predicting stream flow: A case study of Luvuvhu Catchment, South Africa. Int J Innov Manage Technol 4 (1): 11-15.

Orduz-Rodríguez JO, Chacón-Díaz A, Linares-Briceño VM, 2007. Evaluación del potencial de rendimiento de tres especies y un híbrido de cítricos en la región del Arari del Departamento del Meta (Colombia) durante doce años, 1991-2003. Orinoquia 11 (2): 41-48. http://www.redalyc.org/pdf/896/89611204.pdf.

Pérez-Ariza C, Nicholson A, Flores M, 2012. Prediction of coffee rust disease using Bayesian networks. Proc. 6th Eur Workshop on Probabilistic Graphical Models, Granada (Spain), pp: 259-266.

Pourzangbar A, Brocchini M, Saber A, Mahjoobi J, Mirzaaghasi M, Barzegar M, 2017. Prediction of scour depth at breakwaters due to non-breaking waves using machine learning approaches. Appl Ocean Res 63: 120-128. https://doi.org/10.1016/j.apor.2017.01.012

Quinlan JR, 1992. Learning with continuous classes. Proc Aust Joint Conf on Artificial Intelligence, Hobart (Tasmania), Nov 16-18, pp: 343-348.

Quinlan JR, 1993. C4.5: Programs for machine learning. Morgan Kaufmann Publ.

Samadi M, Jabbari E, Azamathulla HM, 2014. Assessment of M5′ model tree and classification and regression trees for prediction of scour depth below free overfall spillways. Neural Comput Appl 24 (2): 357-366. https://doi.org/10.1007/s00521-012-1230-9

Tripathy AK, Adinarayna J, Sudharsan D, Merchant SN, Desai UB, Vijayalaksmi K, Raji-Reddy D, Screenivas G, Ninomiya S, Hirafuji M, Kiura T, Tanaka K, 2011. Data mining and wireless sensor network for agriculture pest/disease predictions. Proc World Cong on Information and Communication Technologies, Mumbai (India), pp: 1229-1234. https://doi.org/10.1109/wict.2011.6141424

Tucker DPH, Wheaton TA, Muraro RP, 1994. Citrus tree spacing. University of Florida. Fla Coop Ext Serv.

Uysal I, Altay HG, 1999. An overview of regression techniques for knowledge discovery. Knowl Eng Rev 14: 319-340. https://doi.org/10.1017/S026988899900404X

Wang Y, Witten IH, 1997. Induction of model trees for predicting continuous classes. 9th Eur Conf on Machine Learning, Prague (Czech Republic).

Wang H, Ma Z, 2011. Prediction of wheat stripe rust based on support vector machine. Proc 7th Int Conf on Natural Computation, Shanghai (China), pp: 259-266. https://doi.org/10.1109/icnc.2011.6022095

Ye X, Sakai K, Manago M, Asada S, Sasao A, 2007. Prediction of citrus yield from airborne hyperspectral imagery. Precis Agric 8 (3): 111-125. https://doi.org/10.1007/s11119-007-9032-2

Yu H, Liu D, Chen G, Wan B, Wang S, Yang B, 2010. A neural network ensemble method for precision fertilization modelling. Math Comput Model 51 (11): 1375-1382. https://doi.org/10.1016/j.mcm.2009.10.028