Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA)
http://dx.doi.org/10.5424/sjar/2014122-4439

Spanish Journal of Agricultural Research 2014 12(2): 313-328
ISSN: 1695-971-X
eISSN: 2171-9292

RESEARCH ARTICLE

OPEN ACCESS

Predictive ability of machine learning methods for massive crop yield prediction

Alberto Gonzalez-Sanchez[1,2]*, Juan Frausto-Solis[1] and Waldo Ojeda-Bustamante[2]

[1] Tecnologico de Monterrey. Campus Cuernavaca. Autopista del Sol, km 104. Colonia Real del Puente. Xochitepec. Morelos, Mexico. [2] Instituto Mexicano de Tecnología del Agua. Paseo Cuauhnáhuac 8532. Col. Progreso. Jiutepec. Morelos, Mexico

Abstract

An important issue for agricultural planning purposes is the accurate yield estimation for the numerous crops involved in the planning. Machine learning (ML) is an essential approach for achieving practical and effective solutions for this problem. Many comparisons of ML methods for yield prediction have been made, seeking for the most accurate technique. Generally, the number of evaluated crops and techniques is too low and does not provide enough information for agricultural planning purposes. This paper compares the predictive accuracy of ML and linear regression techniques for crop yield prediction in ten crop datasets. Multiple linear regression, M5-Prime regression trees, perceptron multilayer neural networks, support vector regression and k-nearest neighbor methods were ranked. Four accuracy metrics were used to validate the models: the root mean square error (RMS), root relative square error (RRSE), normalized mean absolute error (MAE), and correlation factor (R). Real data of an irrigation zone of Mexico were used for building the models. Models were tested with samples of two consecutive years. The results show that M5-Prime and k-nearest neighbor techniques obtain the lowest average RMSE errors (5.14 and 4.91), the lowest RRSE errors (79.46% and 79.78%), the lowest average MAE errors (18.12% and 19.42%), and the highest average correlation factors (0.41 and 0.42). Since M5-Prime achieves the largest number of crop yield models with the lowest errors, it is a very suitable tool for massive crop yield prediction in agricultural planning.

Additional key words: regression trees; neural networks; support vector regression; k-nearest neighbor; multiple linear regression.

Introduction

Crop yield prediction (CYP) is a major problem in agriculture. Starting each growing season, agricultural planners require estimating the yield for all the involved crops (Frausto-Solis *et al*., 2009). Regrettably, CYP is difficult because it depends on many inter-related factors (Liu *et al*., 2001; Marinković *et al*., 2009). Moreover, yield is also affected by farmer decisions (such as applied irrigations, pest and fertilizers applications, crop rotation, and land preparation) and incontrollable factors (such as weather, subsidies and market). As stated by Ruß (2009), yield prediction traditionally has relied on farmers' long-term experience for specific fields, crops and climate conditions, which can be inaccurate. Simple estimators, such as the average of several previous yields or the last obtained yield, are also used. Nevertheless, crop yield varies spatially and temporally with a non-linear behavior (Liu *et al*., 2001; Drummond *et al.,* 2003; Schlenker & Roberts, 2006), introducing large deviations from one year to another. Thus,

* Corresponding author: alberto_gonzalez@tlaloc.imta.mx; albertogonzalez@itesm.mx
Received: 01-05-13. Accepted: 29-04-14.

Abbreviations used: ANN (artificial neural network); BAS (best attribute subset); CV (cross-validation); CYP (crop yield prediction); IWD (irrigation water depth); kNN (k-nearest neighbor); MAE (mean absolute error); ML (machine learning); MLP (multi-layer perceptron); MLR (multiple linear regression); PA (planning area); RF (rainfall); RMSE (root mean square error); RRSE (root relative square error); SDC (season duration cultivar); SDR (standard deviation reduction); SMO (sequential minimal optimization), SR (solar radiation); SVR (support vector regression).

more efficient methods have been developed, which can be classified as crop growth models and data-driven models. Crop models perform an abstraction of the dynamic mechanistic of the plant's physiological stages by fitting them into a mathematical model (Safa *et al*., 2004). Most of the mechanistic models are crop-specific: SOYGRO for legumes (Wilkerson *et al*., 1983); AFRCWHEAT2 (Porter, 1993) and Sirius (Jamieson *et al*., 1998a) for wheat; CERES-Maize (Jones & Kiniry, 1986) for corn. Some others, such as SUCROS (Spitters *et al*., 1988), SUCROS2 (Goudriaan & van Laar, 1994) and STICS (Brisson *et al*., 1998) are available for various crop types through parameter fitting. An evaluation performed by Jamieson *et al*. (1998b) for the wheat crop shows that some of these models has a reasonable accuracy, with deviations within 10%. Regrettably, as reported by Varcoe (1990) and Drummond *et al*. (2003), this kind of models is expensive in terms of time and money, being impractical for massive application and agricultural planning. On the other hand, data-driven models are built empirically, not requiring a deep knowledge about physical mechanisms that produced the data. Such techniques are inexpensive, relatively easy to apply, and do not need a predefined structure of the model. Consequently, data driven models have been widely applied in the last years using classical statistics (Dixon *et al*., 1994; Sudduth *et al*., 1996) and machine learning methods (Drummond *et al*., 2003; Roel & Plant, 2004; Irmak *et al*., 2006). Statistical models are represented by parametric structures tuned with sum-of-squares residuals, validated by hypotheses test and confidence intervals (Breiman, 2001).

The main regression application for CYP has been linear with rather weak results (Drummond *et al*., 2003); for instance, that implemented by Dixon *et al*. (1994) and Sudduth *et al*. (1996), obtained solutions ranging from bad to moderate results. Machine Learning (ML) techniques are based on non-parametric and semi-parametric structures, with validation relying on prediction accuracy (Breiman, 2001). Regression trees (Roel & Plant, 2004), artificial neural networks (Liu *et al*., 2001; Drummond *et al*., 2003; Irmak *et al*., 2006; Fortin *et al*., 2011) and support vector regression (Jaikla *et al*., 2008; Ruß, 2009) are common ML techniques applied for CYP purposes. Previous works suggest that data-driven models have better adaptability for cropping planning than crop growth models due to their friendly implementation and performance (Safa *et al*., 2004; Irmak *et al*., 2006).

Some comparisons among regression models for CYP have been made, looking for the most accurate technique. Drummond *et al*. (2003) and Fortin *et al*. (2011) have compared classical statistical models against Artificial Neural Networks (ANNs). Ruß (2009) compared ANNs, regression trees and support vector regression. Despite the high site-dependency, neural networks have been widely recognized as robust models, and they have obtained good results for CYP (Liu *et al*., 2001; Fortin *et al*., 2011). On the other hand, support vector regression models have proven to be more accurate than ANN and regression trees for some crop datasets (Ruß, 2009). Regrettably, previous comparisons consider only a very small number of crops. Thus, information for applying ML techniques as a planning tool is usually not enough. As a consequence, more extensive evaluations involving a greater number of crops are required. Besides, an important issue related to previous CYP techniques comparisons is the fact that the attribute subset is always the same for all the evaluated techniques. Different machine learning algorithms may perform better if they use a distinct set of attributes in the same training dataset (Kohavi, 1995). Therefore, a fairer comparison should include some metrics to measure their performance with the best attribute subset for each technique. Considering the above, this work presents a comparison among four popular machine learning techniques for CYP in a large irrigated area in Mexico with a typical cropping plan. The linear regression technique, which is frequently used to CYP is also assessed. The evaluation is made with ten crop datasets to provide a perspective of using ML techniques for real planning purposes. To ensure fairer conditions, the best attribute subset for each technique is determined. A complete algorithm enumerates all the attribute combinations, building a regression model of each subset. The models use the greater part of samples from the training dataset, except the most recent, which are used for measuring the performance. In this work, a popular set of ML and one statistical technique for CYP are ranked: multiple linear regression, M5-Prime regression trees, perceptron multilayer neural networks, support vector regression, and k-nearest neighbor. The models are evaluated with data held-out for testing. Results per technique are compared. The potential attributes considered for this work were planting area (ha), applied irrigation water depth (mm), cumulative rainfall (mm), cumulative global solar radiation (kWh m^{-2}), maximum, average and minimum

Figure 1. Irrigation module III-1, "Santa Rosa" localization.

temperatures (°C) and duration of the season-duration cultivar (short, medium, long). To build the models, historical data of ten crops were obtained from one irrigated zone in Sinaloa Mexico, developing the experimentation in realistic conditions. In the results section, we present the best CYP technique for massive crop datasets and the most influential attributes for each model.

Material and methods

Case study

There are two types of irrigation schemes in Mexico: irrigation units and irrigation districts. The former are about 20,000, covering an area of 2.9 million ha. These are managed directly by the agricultural producers. On the other hand, there are 83 irrigation districts covering an area of 3.5 million ha (Ojeda-Bustamante *et al*., 2007). These were formerly managed by the federal government and turned over to 474 water user associations (also called "irrigation modules"). Both, irrigation districts and units require modern technology and methods to improve their planning process. This technology should be inexpensive and effective enough to be applied to as many units or modules as possible. For this reason, this work was carried out in the Irrigation District 075 (ID-075), located in the north of the state of Sinaloa, México. Specifically, data from the irrigation module "Santa Rosa" was used. This is the largest module in the District, and is located near to the city of "Los Mochis", at 25° 55' 2.684" N and 109° 10' 25.297" W; with an average altitude of 15 m (Fig. 1). Two data were collected for this work: a) agricultural production data, and b) weather information data. The former included records at farm level regarding planting area, crop species and variety, day and quantity of irrigated water, starting and ending sowing dates and crop's yield. Such data were obtained from the Spriter-GIS system, a tool for irrigation management and statistical monitoring used in the module (Ojeda-Bustamante *et al*., 2007). The second dataset was composed of climatological variables such as rainfall, solar radiation, and temperatures. Weather data were collected from two meteorological stations located in the module. The two data-sources were merged into one single database, using the sowing date and the length of each growing-stage to integrate the climatological data. Each record in that database contains a crop with a cultivated area, agricultural production and the weather variables monitored during the crop season. Ten crop datasets were extracted from the merged database (Table 1).

Only crops and records from the fall-winter season in the years 1998-2006 were considered. The periods 1998-2004 and 1998-2005 were used for training, while 2005 and 2006 were used for testing (respectively). To simplify future references of these datasets, an ID is assigned in the first column. Table 1 describes the quantity of records and periods of time

Table 1. Testing and training samples distribution per crop dataset

Crop Dataset ID	Crop species[1]	Cultivar[2]	Total number of samples	Samples in training period (1999-2004)	Samples in testing period (2005-2006)	Testing percentage (%)
CD01	Pepper	Jalapeno	242	186	56	23.14
CD02	Common bean	Mayocoba	508	449	59	11.61
CD03	Chickpea	—	87	42	45	51.72
CD04	Chickpea	Blanco Sinaloa 92	586	435	151	25.77
CD05	Corn	Pioneer 30G54	2617	1685	932	35.61
CD06	Potato	Atlantic	1250	951	299	23.92
CD07	Potato	—	195	132	63	32.31
CD08	Tomato	—	156	108	48	30.77
CD09	Tomato	Saladette	461	413	48	10.41
CD10	Mexican husk tomatoes	—	115	90	25	21.74
Total			6217	4491	1726	

[1] Pepper (*Capsicum annuum*), common bean (*Phaseolus vulgaris*), chickpea (*Cicer arietinum*), corn (*Zea mays*), potato (*Solanum tuberosum*), tomato (*Solanum lycopersicum*), Mexican husk tomato (*Physalis ixocarpa*). [2] —: unspecified.

Table 2. Potential predictor attributes in crop datasets

Attribute code name	Attribute name	Attribute description
PA	Planting area (ha)	Amount of surface dedicated for sowing
IWD	Applied irrigation water depth (cm)	Amount of water directly applied over the sowed surface, accumulated during duration of the six crop growing stages
SR	Solar radiation (kWh m^{-2})	Average of accumulated daily radiation in the last three crop growing stages
RF	Rainfall (mm)	Amount of rainfall accumulated over the six crop growing stages (averaged and accumulated from the nearest meteorological stations)
MaxT	Maximum temperature (°C)	Average of daily maximum temperatures registered in the last three crop growing stages
AvgT	Average temperature (°C)	Average of daily mean temperatures registered in the last three crop growing stages
MinT	Minimum temperature (°C)	Average of daily minimal temperatures registered in the last three crop growth stages
SDC	Season-duration cultivar	Identifies the kind-duration cultivar of the crop (1 = short, 2 = medium, 3 = long). Duration time (in days) is different depending on each crop type

used for the training and testing stages. In order to maintain realistic conditions, the last year of available data was reserved for testing in each training-testing datasets.

Eight attributes were selected as potential predictor variables (Table 2), most of them have been considered important for CYP in previous works. For instance: IWD (Safa *et al*., 2004); SR (Dixon *et al*., 1994; Safa *et al.,* 2004); RF (Drummond *et al*., 2003; Safa *et al*., 2004; Irmak *et al*., 2006); temperature (Dixon *et al*., 1994; Drummond *et al*., 2003). In addition, we used two attributes in our work: Planning Area (PA), which should be used because the productivity of different crops depends on it; Season Duration Cultivar (SDC) is an important attribute in order to avoid mixing crops with different duration cultivar. For SDC attribute, we used an identifier denoting the kind of duration as it is described in Table 2. These variables are referred to as *potential* because this work uses a complete algorithm to find the best attribute subset for each regression technique. The weather attribute values (solar radiation, minimum, average and maximum temperatures) are estimated averaging the last three crop growing stages, the most influential in the crop

development. The target attribute is the yield expected at the end of the cropping season, measured in (t ha^{-1}).

Machine learning techniques

Machine Learning (ML) deals with problems where the relation between input and output variables is not known or hard to obtain. The "learning" term here denotes the automatic acquisition of structural descriptions from examples of what is being described (McQueen *et al*., 1995). Unlike traditional statistical methods, ML does not make assumptions about the correct structure of the data model, which describes the data. This characteristic is very useful to model complex non-linear behaviors, such as a function for crop yield prediction. ML techniques most successfully applied to CYP have been M5-Prime regression trees (Wang & Witten, 1997; Frausto-Solís *et al*., 2009; Marinković *et al*., 2009; Ruß & Kruse, 2010), artificial neural networks (Liu *et al*., 2001; Drummond *et al*., 2003; Safa *et al*., 2004; Fortin *et al*., 2011), support vector regression (Ruß, 2009) and k-nearest neighbor (Zhang *et al*., 2010). However, no comparisons covering all the aforementioned techniques have been made for a large amount of crops.

Multiple linear regression

Despite not being properly a ML technique, Multiple Linear Regression (MLR) has been applied frequently for CYP, reason why it is included in this comparison. MLR is a popular statistic technique which can be applied to predict the value of a dependent variable Y_i, using a set of independent or explanatory variables X_{ij} (Hair *et al*., 1987). As is indicated by Wasserman (2004), the MLR model is described by:

$$Y_i = \sum_{j=1}^{k} B_j X_{ij} + \in_i \qquad [1]$$

where k is the quantity of explanatory variables, B_j is the regression coefficient j, X_{ij} is the j value for the observation i and $\in_i$ the residual error. Assuming $X^T X$ is a ($k \times k$) non-singular matrix, an approximation for B ($\bar{\beta}$) can be obtained by Eq. [2]:

$$\bar{\beta} = \left(X^T X\right)^{-1} X^T Y \qquad [2]$$

And Eq. [1] can be written as

$$Y + X\bar{\beta} + \in \qquad [3]$$

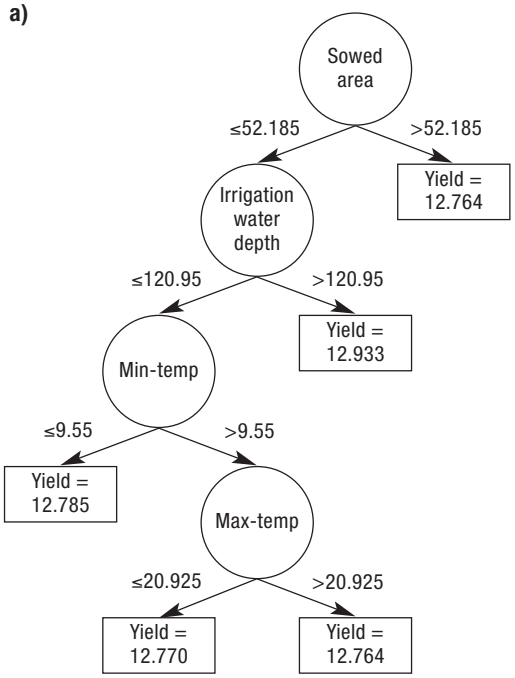
In expression [3] the individual contribution of attribute X_{ij} to the Y_i yield is given by the j-th element of vector β. The last expressions are important to describe the regression model.

Multiple linear regression and other classical linear methods have been compared to CYP problem (Sudduth *et al*., 1996; Drummond *et al*., 2003; Fortin *et al*., 2011). In contrast to previous works, this paper builds the MLR models using the best attribute subset, which improves the models' predictive accuracy.

Regression trees

MLR generates global models; there is a single predictive formula holding over the entire samples space. A regression tree uses an alternative approach, splitting recursively the samples' space in small regions until each region is small enough to be represented by a simple model (Quinlan, 1992). The first node in the tree is named the root node, which does not have incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called a test node and a node without outgoing edges is called a leaf node. Each test node splits the samples' space into two or more subspaces based on a set of conditions of the input attributes values. Conditions to splitting the samples are based on an impurity measure (such as the standard deviation or the Gini-Index). The leaf nodes assign a numerical value to the last partition of samples. Thus, new samples are evaluated by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path. A sample regression tree is shown in Fig. 2a. The most common algorithms to build regression trees are: CART (Breiman *et al*., 1984), M5 (Quinlan, 1992), and M5-Prime (Wang & Witten, 1997). Tree construction procedure (Fig. 2b) is similar for all these algorithms, introducing some differences in three main aspects: 1) the impurity measure on continuous attributes, 2) the prune rule and 3) the leaf value determination mechanism. Standard deviation reduction (SDR) is applied as impurity criterion in M5, instead of variance as used in CART. M5 has some particular characteristics for regression such as: a) it is able to handle linear models at leaf

a)



b)

```
M5-Prime(examples) {
  SD=sd(examples)
  for each k-valued enumarated attribute
    convert into k-1 synthetic binary
    attributes
  root=new_node
  root. Examples=examples
  split(root)
  prune(root)
  print_tree(root)
}
split(node) {
  if sizeof(node.examples)<4 or
     sd(node.examples)<0.05*SD
    node.type=LEAF
  else
    node.type=INTERIOR
    for each continuos and binary attribute
      for all possible split positions
        calculate the attribute's SDR
      node.attribute=attribute with max SDR
      split(node.left)
      split(node.right)
}
```

```
prune(node) {
  if node=INTERIOR
    prune(node.left_child)
    prune(nod.right_child)
    node.model=linear_regression(node)
  if subtree_error(node)>error(node) then
    node.type=LEAF
}
subtree_error(node) {
  l=node.left;
  r=node.right;
  if node=INTERIOR then
    return (sizeof(l.examples)*
      subtree_error(l)+sizeof(r.examples)*
        subtree error(r)/
        sizeof(node.examples))
  else
    return error(node)
}
```

Figure 2. a) A sample regression tree (Frausto-Solís *et al*., 2009); b) pseudo-code for the M5-Prime algorithm (Wang & Witten, 1997).

nodes instead of constant values; b) it can introduce a smoothing procedure to obtain better predictions (Quinlan, 1992). In addition, trees generated with M5 are smaller than those generated with CART. For these reasons, M5 overcomes CART algorithm in accuracy and simplicity (Uysal & Altay, 1999). M5-Prime is a "rational reconstruction" of M5. The former surpasses the performance of the latter by introducing a method for dealing with missing values and enumerated attributes (Wang & Witten, 1997). Despite these advantages, there are few M5-Prime applications for predicting crop yields. Thus, M5-Prime was selected over other algorithms in this work, being compared in accuracy with other regression techniques to CYP problem. The M5-Prime implementation in Weka (Witten *et al*., 1999) was used (Fig. 2b). The parameters to create regression trees without pruning and with a minimum of two samples in each leaf node were selected.

Artificial neural networks

Artificial neural networks (ANN) try to simulate the information processing capabilities of nervous systems (Rojas, 1996). As an analogy of the biological systems, an ANN is a group of simple processing units linked by weighted interconnections. Each processing unit has a certain number of inputs from the outer or from other processing units. Inputs are calibrated using the weights of its corresponding interconnections. Once calibrated, inputs are combined and transmitted to other processing units via the appropriate weighted interconnections. This process can be represented mathematically by a function that maps the set of inputs to a set of outputs. Commonly, the function obtained is non-additive and nonlinear. The iterative process performed to pound connections is called "training", which is guided for some error measure. There are several structures for build ANN, each one with a suitable group of training algorithms. One of the most popular structure-training algorithm combinations is the multilayer perceptron (MLP) and the back-propagation algorithm (Rumelhart *et al*., 1986). The back-propagation minimizes the error in weight space using the gradient descent method. This method requires an iterative computation of the gradient of the error function. In addition, the method requires an activation function, where the sigmoidal function $[1 / (1 + e^{-cx})]$ is the most commonly used. The MLP and back-propagation algorithms have been a popular choice to implement neural networks for crop

yield prediction (Liu *et al*., 2001; Drummond *et al*., 2003; Safa *et al*., 2004), for which this model is also used in the present work.

For MLP, the number of neurons in each layer defines the network topology; it strongly affects the network prediction accuracy. Regrettably, an automatic method to determine a suitable topology using the number of inputs and outputs is not available. Therefore, the quantity of neurons and hidden layers are commonly established by experimentation. In this sense, a topology based on previous works was selected (Sudduth *et al*., 1998; Irmak *et al*., 2006). A fully interconnected three-layered network with five neurons in one single hidden layer was used. Most recommended parameters were applied, such as the weight decay and numeric attribute normalization (Liu *et al*., 2001; Drummond *et al*., 2003). Training epochs, learning rate and the momentum were established by experimentation, being 500, 0.3, and 0.01 respectively. Quantity of neurons at the input layer depends on the number of attributes selected, while the output layer has only a neuron, the crop yield estimation.

Support vector regression

Support vector regression (SVR) technique is a classification method that arises from a nonlinear generalization of the Generalized Portrait algorithm developed by Vapnik & Lerner (1963). On its simplest form, the goal of the support vector technique is to obtain a linear function $f(x) = \langle w,x \rangle + b$ with $w \in R^N$ and $b \in R$ for a given training set $\{(x_1,y_1),...,(x_m,y_m)\}$. That function $f(x)$ should have at most one ε deviation from the current obtained targets y_i at the time that is as flat as possible (Smola & Schölkopf, 2004). Flatness can be obtained by a small value of w. Thus, the problem can be written as (Vapnik *et al*., 1997):

$$\text{minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\left(\xi_i + \xi_1^*\right)$$

$$\text{subject to} \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \quad [4]$$

where ξ_i and ξ_i^* are slack variables introduced to deal with infeasible constraints, and C is called the *regularization parameter*; C determine the quantity of deviations larger than ε that are accepted. In most of the cases, problem [1] can be easily solved in its dual formulation (Smola & Schölkopf, 2004):

$$\max_{\alpha,\alpha^*} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right)K\left(x_i - x_j\right) - \varepsilon\sum_{i=1}^{N}\left(\alpha_i - \alpha_i^*\right) + \sum_{i=1}^{N} y_i\left(\alpha_i - \alpha_i^*\right)$$

Subject to: [5]

$$\sum_{i=1}^{N}\left(\alpha_i - \alpha_i^*\right) = 0, \alpha_i, \alpha_i^* \in \left[0, C\right]$$

where $K(x_i,x_j)$ is known as the kernel function, which allows to project the original data into a higher-dimensional space to be linearly separable (Ruß, 2009). The most common kernel functions are the radial basis [6] and the polynomial [7]. To obtain good predictions, the parameters s and r of Eqs. [6] and [7] and the parameter C in Eqs. [4] and [5] should be tuned. Regrettably, there is no automatized method to find such optimal values. Thus, these parameters are commonly established by a trial and error procedure.

$$K(x, x_i) = e^{-\frac{[x - x_1]^2}{2\sigma^2}} \quad [6]$$

$$K(x, x_i) = \left(\langle x, x_i \rangle + 1\right)^{\rho} \quad [7]$$

This work uses the SVR implemented in Weka, which applies an improved version of the sequential minimal optimization (SMO) learning algorithm (Shevade *et al*., 2000). The selected kernel was polynomial (Eq. [7]), with a r value of 1. The C parameter was also assigned to 1 and was used for all the SVR models. Other values were also tested, such as those utilized in Ruß (2009), with non-favorable results.

k-Nearest neighbor

The *k*-Nearest neighbor (kNN) method classifies a new object with input vector y examining the k closest training dataset points to y, assigning the object to the class that has the majority of points among these k (Hand *et al*., 2001). In the case of regression, the response value is calculated as a weighted sum of the responses of all the k neighbors, where the weight is inversely proportional to the distance —generally normalized Euclidean— from the input record. The most basic form takes $k = 1$. However, this makes a prediction model rather unstable (high variance,

sensitive to data). Increasing k reduces the variance, but may increase the bias. Thus, the algorithm is sensitive to a proper selection of k.

The nearest neighbor method has several attractive properties. Beyond the choice of k and the distance metric, no optimization or training is required (Hand *et al.,* 2001). Also, the method is able to take full advantage of local information and form highly nonlinear, highly adaptive decision boundaries. Their disadvantages are the high computational cost in time and memory, since all the available data points (samples) should be scanned to find the nearest neighbors. The distance calculation becomes more difficult as the training dataset dimension increases. However, the method is popular due to its ease of implementation and the above-mentioned properties.

The kNN technique has been used to study the crops behavior, as it is shown in Zhang *et al*. (2010). Nevertheless, very few comparisons of kNN against other machine learning methods applied to CYP have been made. This work applies a kNN algorithm to predict the yield of ten crops, and the results are compared against MLR, ANN, SVR, and M5-Prime regression trees. A k value of 5 and the Euclidean distance were used as parameters for this technique.

Accuracy metrics

Four of the most common accuracy metrics of regression models were used: root mean square error (RMSE), root relative square error (RRSE), correlation coefficient (R) and the relative mean absolute error (MAE). Table 3 shows how these metrics are estimated (Han & Kamber, 2006). The RMSE measures the difference between the actual and estimates, exaggerating the presence of outliers (Han & Kamber, 2006). RMSE has been used to measure the performance of CYP models in previous works, such as Liu *et al*. (2001) or Drummond *et al*. (2003). In addition, this work applies the RRSE, which compares the model prediction against the mean. For this metric, a value below 100% indicates a better performance than the average. Thus, RRSE is easy to read by people unaccustomed to crop yield dimensions. Correlation coefficient (R) is also included, which measures the linear relationship between regression model predictions and the real values. MAE is the average of differences in estimations (in physical units). Because yield proportions are different among the crops, this metric is expressed as a percentage relative to the mean yield.

Table 3. Performance metrics (y = real value, $\hat{y}$ = yield estimation, i = observation, $\bar{y}$, $\bar{\hat{y}}$ = means)

Metric	Expression
RMSE	$\sqrt{\frac{\sum_{i=1}^{n}(y_i + \hat{y}_{\bar{y}})^2}{n}}$
RRSE (%)	$\sqrt{\frac{\sum_{i=1}^{n}(y_i + \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \cdot 100$
R	$\frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}$
MAE (%)	$\left(\frac{\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert}{(n)(\bar{y})}\right) \cdot 100$

RMSE: root mean square error. RRSE: root relative square error. R: correlation coefficient. MAE: mean absolute error

Evaluation method

Machine learning algorithms work with different heuristic or principles, being able to be influenced by different kinds of relationships on data (Kohavi, 1995). To ensure fairer conditions in evaluation, this work finds the best attribute subset (BAS) for each technique under analysis. The best subset is determined by using only the training dataset. To find the BAS, a combinatorial procedure enumerates and evaluates all the possible subsets $\{x_1, x_2, x_3 \ldots x_m\}$, where a x_k represents an unique combination of attributes obtained from the potential set of attributes $\{a_1, a_2, \ldots, a_n\}$ presented in Table 2. To evaluate each x_k subset, the training dataset conformed by the range of the years $[a,b]$ is divided into two datasets; the former with samples from the range $[a,b-1]$, and the second with samples from the year b. The first dataset is used to build a model with each x_k subset, while the other dataset is used to evaluate the models. The evaluation is performed according the holdout validation

```
// Procedure to evaluate the regression techniques with all the crop datasets
// listOfCropDS: A list of crop datasets identifiers
// potAttr: The set of potential attributes.
// algorithms: A list of algorithms (MLR, M5-Prime, ANN, kNN, SVR)
// firstTrainYear: First year in training dataset
// lastTrainYear: Last year in training dataset
// testYear: The year of the samples used for testing (commonly lastTrainYear + 1)

procedure evalAlgorithms(listOfCropDS, potAttr, algorithms, firstTrainYear, lastTrainYear, testYear) {
  for i=0 to 15
  begin
    for j=0 to 5
    begin
      // obtain samples for training
      trainSamples=GetSamples(listOfCropDS [i], firstTrainYear,lastTrainYear)
      // obtain the BAS for trainSamples using the algorithm in turn
      BAS=findBestAttrSubset(trainSamples, potAttr, algorithms[j], firstTrainYear, lastTrainYear)
      // obtain samples for testing
      testSamples=GetSamples(CropDS [i], testYear,testYear)
      // makes a model with the algorithm in turn, using trainSamples and
      // the best attribute subset
      model=makeRegressionModel(algorithms[i], trainSamples, BAS)
      // evaluates a regression model using testSamples
      errorMeasures=evalModel(model,testSamples)
      saveResults(CropDS [i],algorithms[j], errorMeasures)
    end
  end
}
```

Figure 3. Procedure to evaluate the algorithms.

technique (Han & Kamber, 2006) and the metrics described in material and methods section. The best attribute subset obtained from the training dataset is used for predicting yield on samples of testing dataset, which is composed with samples of $b + 1$ year. Fig. 3 provides a pseudo-code of the evaluation process. This process was applied twice, once for 1999-2004 and other for 1999-2005 training periods. Thus, each obtained BAS by technique was tested with 2005 and 2006 samples. The results were averaged and are shown in the next section.

As mentioned above, evaluation method in Fig. 3 uses the hold-out evaluation technique. This validation technique was selected by its implementation simplicity and low computational cost. These features fit well to the combinatorial procedure employed to test each attribute combination. Previous works related to CYP comparisons have been applied Cross-Validation (CV) technique (Drummond *et al*., 2003; Ruß, 2009; Ruß & Kruse, 2010). Nevertheless, some disadvantages of CV make it difficult to use in our evaluation scheme: a) the proper selection of the number of folds (k), to maintain the problem computationally tractable (Drummond *et al*., 2003), and b) the lack of consistency of a model selected by CV (Yang, 2008).

Results

As was mentioned before, the model comparison is made using four performance metrics. Table 4 shows the results for the RMSE and RRSE metrics for all evaluated techniques in all the crop datasets. Average RMSE and RRSE for each technique are shown at the bottom of Table 4. Average RMSE shows that kNN has the lowest mean error, followed closely by SVR and M5-Prime. Instead, there are a little differences with RRSE, placing M5-Prime first, followed closely by kNN and SVR. ANN has the highest mean error for both metrics. Table 4 shows in bold the best result per crop dataset. An individual counting per technique of such results indicates that M5-Prime achieves the largest quantity of models with the lowest RMSE and RRSE models [5] and [6], followed by kNN with four models in both metrics. ANN has one best model with

Table 4. Root mean square error (RMSE) and root relative square error (RRSE) metrics results for all the evaluated techniques. The first column shows the dataset identifier. The best result for each dataset and technique is shown in bold

Crop dataset ID	RMSE (%)					RRSE (%)				
	MLR	ANN	M5-Prime	kNN	SVR	MLR	ANN	M5-Prime	kNN	SVR
CD01	9.64	9.67	**8.54**	9.38	8.83	90.59	90.44	**80.62**	91.00	83.39
CD02	0.25	0.26	0.26	**0.23**	0.26	76.71	77.23	78.42	**70.48**	77.82
CD03	0.49	**0.47**	0.50	0.64	0.50	66.05	65.34	**63.79**	84.32	66.91
CD04	1.07	1.09	**0.97**	1.05	1.09	93.18	95.02	**83.08**	92.12	96.41
CD05	1.28	1.26	**1.25**	1.37	1.29	85.71	84.76	**83.73**	91.62	86.21
CD06	4.45	4.54	**3.97**	4.36	4.20	92.84	95.19	**82.06**	90.06	84.97
CD07	5.32	5.29	4.97	**4.40**	4.91	83.68	82.99	76.96	**64.27**	76.00
CD08	15.80	16.03	16.20	**12.87**	13.16	83.97	85.32	86.49	**68.91**	70.34
CD09	10.10	11.77	**10.03**	10.54	10.30	82.24	94.20	**80.75**	84.23	83.43
CD10	5.66	5.61	4.70	**4.26**	5.61	96.44	95.56	79.89	**71.97**	95.67
Average	5.41	5.60	5.14	**4.91**	5.02	84.54	86.18	**79.46**	79.78	81.97

MLR: multiple linear regression. ANN: artificial neural network. kNN: k-nearest neighbor. SVR: support vector regression.

Table 5. Correlation factor (R) and mean absolute error (MAE) metric results for all the evaluated techniques. The first column shows the dataset identifier. The best result for each dataset is shown in bold

Crop dataset ID	R					MAE (%)				
	MLR	ANN	M5-Prime	kNN	SVR	MLR	ANN	M5-Prime	kNN	SVR
CD01	0.27	0.43	**0.65**	0.55	0.54	25.30	27.04	**22.18**	23.95	22.37
CD02	0.29	0.35	0.23	**0.60**	0.29	9.07	9.13	9.37	**6.71**	8.94
CD03	0.65	0.63	**0.71**	0.35	0.63	16.76	**16.15**	16.24	22.84	17.27
CD04	0.11	0.15	**0.37**	0.33	0.14	24.99	25.02	**21.40**	22.91	25.54
CD05	0.15	0.15	**0.22**	0.15	0.10	8.24	8.00	**7.85**	9.21	8.40
CD06	0.25	0.31	**0.40**	0.00	0.30	12.57	12.90	**9.76**	10.43	10.41
CD07	0.03	–0.06	**0.40**	0.37	0.35	13.07	14.06	13.48	**9.97**	12.15
CD08	0.38	0.54	0.44	**0.72**	0.66	44.37	45.88	39.76	35.33	**34.80**
CD09	0.40	0.03	**0.45**	0.39	0.33	19.04	21.77	17.57	**16.61**	18.74
CD10	–0.05	–0.18	0.54	**0.74**	–0.04	46.54	44.97	39.35	**29.06**	46.33
Average	0.25	0.21	**0.42**	0.41	0.31	21.63	21.99	19.42	**18.12**	20.29

MLR: multiple linear regression. ANN: artificial neural network. kNN: k-nearest neighbor. SVR: support vector regression.

RMSE metric, while SVR and MLR do not have any model that improves overall anyone metric. However, the average performance of SVR is better than ANN, MLR in both metrics.

Table 5 shows the correlation factor (R) between real and estimated yield obtained per technique for all the modeled crop datasets. Average results of R (at the bottom of Table 6) shows that M5-Prime has the highest mean R value (0.42), followed by kNN (0.41), SVR (0.31), MLR (0.25), and ANN (0.21) respectively. Notice that M5-Prime is the only technique with all its R results positive. An individual counting of best results by crop dataset per each technique (highlighted in bold in Table 6) shows that M5-Prime obtains the largest quantity of best correlated models (seven models), followed by kNN with three models. Besides, ANN, SVR, and MLR never achieved the best place.

The MAE (%) metric is also included at the left side of Table 5. The average value (at the bottom-left side of the table) indicates that kNN has the lowest mean error (18.12%), followed closely by M5-Prime (19.42%). All the other techniques have an average error above of 20.29%. ANN presents the highest average error (21.99%). A counting of the best results

Table 6. Attributes selected by evaluation method for each crop dataset and evaluated technique. A cell value of 1 indicates that attribute at such column header is selected for 2005 year. A value of 2 indicates that attribute is selected for 2006 year. A value of 3 indicates that attribute was selected for both years. A zero value indicates that such attribute was not included in any attribute set

Crop dataset ID	Technique[1]	Attribute[2]							
		PA	IWD	SR	RF	MaxT	AvgT	MinT	SDC
CD01	MLR	0	0	**3**	0	1	1	1	1
	ANN	0	**3**	2	2	0	0	0	1
	M5-Prime	1	**3**	0	2	1	2	**3**	1
	kNN	0	2	2	2	0	1	**3**	0
	SVR	2	3	2	1	0	2	2	1
CD02	MLR	0	0	1	**3**	1	0	1	2
	ANN	2	0	1	0	**3**	0	1	0
	M5-Prime	2	0	**3**	1	1	0	1	2
	kNN	2	**3**	2	2	2	**3**	**3**	**3**
	SVR	2	0	**3**	1	**3**	2	1	0
CD03	MLR	**3**	**3**	**3**	0	**3**	**3**	2	0
	ANN	**3**	**3**	**3**	0	0	0	2	0
	M5-Prime	2	**3**	**3**	0	2	0	0	0
	kNN	2	0	2	**3**	1	**3**	1	**3**
	SVR	**3**	**3**	1	0	1	2	2	0
CD04	MLR	0	0	1	2	**3**	**3**	2	0
	ANN	**3**	**3**	**3**	2	1	0	0	0
	M5-Prime	**3**	0	1	0	0	0	0	1
	kNN	1	2	2	2	2	2	2	**3**
	SVR	0	2	**3**	2	1	1	2	0
CD05	MLR	**3**	0	1	2	1	1	**3**	0
	ANN	1	0	1	**3**	1	1	2	1
	M5-Prime	0	0	1	0	1	2	1	0
	kNN	0	0	0	0	1	1	0	2
	SVR	**3**	**3**	1	1	1	1	**3**	2
CD06	MLR	0	0	1	0	0	1	1	**3**
	ANN	0	2	1	0	0	0	0	**3**
	M5-Prime	2	2	1	1	1	2	1	**3**
	kNN	0	0	0	1	0	0	1	**3**
	SVR	1	2	2	**3**	**3**	2	2	**3**
CD07	MLR	0	0	1	2	1	1	0	2
	ANN	0	1	1	**3**	2	0	0	**3**
	M5-Prime	2	1	1	2	2	**3**	1	0
	kNN	1	1	0	2	0	0	2	**3**
	SVR	**3**	1	1	2	1	**3**	2	0
CD08	MLR	2	0	0	0	0	**3**	0	0
	ANN	**3**	0	0	0	2	0	2	1
	M5-Prime	**3**	2	1	1	2	0	2	0
	kNN	2	2	1	**3**	1	1	0	**3**
	SVR	**3**	0	0	0	1	0	1	1
CD09	MLR	0	0	2	1	2	**3**	1	0
	ANN	1	0	2	**3**	2	1	**3**	1
	M5-Prime	1	0	1	1	**3**	1	1	0
	kNN	1	2	0	0	2	1	1	0
	SVR	0	0	2	1	1	0	1	1

Table 6 (cont.). Attributes selected by evaluation method for each crop dataset and evaluated technique. A cell value of 1 indicates that attribute at such column header is selected for 2005 year. A value of 2 indicates that attribute is selected for 2006 year. A value of 3 indicates that attribute was selected for both years. A zero value indicates that such attribute was not included in any attribute set

Crop dataset ID	Technique[1]	Attribute[2]							
		PA	IWD	SR	RF	MaxT	AvgT	MinT	SDC
CD10	MLR	0	0	0	2	2	**3**	2	0
	ANN	1	0	1	2	0	1	2	2
	M5-Prime	**3**	0	0	0	0	1	2	0
	kNN	2	2	0	2	0	2	**3**	0
	SVR	1	0	0	0	2	2	0	0

[1] Evaluated techniques were multiple linear regression (MLR), artificial neural network (ANN), M5-Prime, k-nearest neighbor (kNN) and support vector regression (SVR). [2] The attributes considered were planting area (PA), irrigation water depth (IWD), solar radiation (SR), rainfall (RF), maximum temperature (MaxT), average temperature (AvgT), minimum temperature (MinT) and season-duration cultivar (SDC).

by crop dataset for each technique indicates that kNN and M5-Prime have the largest quantity of models with the lowest errors (four models each one). Both ANN and SVR obtain only one model with the lowest MAE. Individually, MLR never overcomes the other techniques using the MAE metric.

The RMSE, RRSE, R and MAE metrics are useful to measure individual aspects of models' accuracy. For this reason, it is unfair to use only one of them to indicate which technique is most suitable for massive crop yield prediction. For example, consider the average RMSE in the last row in Table 4, which indicates that kNN is the best algorithm. However, mean RRSE in this same row indicates that M5-Prime performs better. RMSE is measured in physical dimensions and is more susceptible to be biased with high values when it is use to compare errors in crop datasets with different yield ranges. In contrast, RRSE is measured against the mean, expressed as percentage deviation. Thus, this last metric would be more suitable to select the most accurate algorithm. In this work, we apply a filter for obtaining the algorithm which achieves the higher quantity of models with the lowest RMSE, RRSE, and MAE errors, and also the highest R values (shown in bold in Tables 4 and 5). In this individual count, M5-Prime obtains four best models, while kNN only two.

An interesting aspect of this research is to compare the best attribute set (BAS) found for each technique. As is explained in the evaluation method section, two BAS were obtained by technique, one per each evaluated year. These attribute sets are shown in Table 6. The best attribute sets are shown separately by crop dataset, because attributes influencing each crop yield are different. A first issue that can be pointed out is the lack of consistency among the BAS selected by technique. Table 6 reveals that majority of numbers are 1 or 2, indicating that BAS usually changes from one year to another in the majority of crop datasets. Not even the techniques that present the best predictions and the closest results show significant coincidences (M5-Prime and kNN). This is easier to see in Table 7, where attributes appearing in BAS of 2005 and 2006 for all crop datasets are accumulated. Only AvgT and SDC attributes were repeatedly selected in more than five crop datasets for MLR and kNN techniques. There are no evidence to show preference of a technique to always include a particular attribute in the BAS. Thus, consistency in BAS is more related to the crop datasets than the specified applied algorithm. Consider for example results from CD03 and CD10 in Table 6, where the former presents a tendency to include PA, IWD and SR attributes almost independently of technique, while the second do not repeat an attribute in more than one algorithm. CD06 for instance, includes SDC attribute for all the evaluated techniques.

Discussion

Several evaluations of ML methods applied to CYP have been made in the literature, each one with different researching purposes. Some works measure the ML performance using a particular attribute set (Liu *et al*., 2001; Safa *et al*., 2004; Marinković *et al*.,

Table 7. Number of times that attributes are selected by evaluation method in 2005 and 2006 for all the crop datasets

Technique[1]	Attribute[2]							
	PA	**IWD**	**SR**	**RF**	**MaxT**	**AvgT**	**MinT**	**SDC**
MLR	2	1	2	1	2	5	1	1
ANN	3	3	2	3	1	0	1	2
M5-Prime	3	2	2	0	1	1	1	1
kNN	0	1	0	2	0	2	3	6
SVR	4	2	2	1	2	1	1	1
Total	10	9	8	7	4	9	7	11

[1] Evaluated techniques were multiple linear regression (MLR), artificial neural network (ANN), M5-Prime, k-nearest neighbor (kNN) and support vector regression (SVR). [2] The attributes considered were planting area (PA), irrigation water depth (IWD), solar radiation (SR), rainfall (RF), maximum temperature (MaxT), average temperature (AvgT), minimum temperature (MinT) and season-duration cultivar (SDC).

2009); some others compare a ML technique against classical statistical methods (Drummond *et al*., 2003) or other ML methods (Ruß, 2009; Ruß & Kruse, 2010). However, in these works, research is commonly limited to one or two crops, and their results are hard to extrapolate to other crops or fields (Liu *et al*., 2001). The crops are commonly selected by its economic importance, or by some particular scientific interest. Nevertheless, the agricultural planning process requires a yield estimation of several crops, and not only most economically productive. In this sense, ten crops were selected for this work using the data availability as the main criterion. Thus, a crop was selected when enough data samples appeared in the range of years under analysis.

Commonly, comparisons of regression methods apply the same attribute set for all the evaluated techniques. This could bias the results in favor of some techniques, because ML methods respond differently depending on the utilized attribute set (Kohavi, 1995). The work presented in this paper applies an exhaustive method to find the best attribute subset for each technique, starting from a potential set of attributes (surface, irrigated water, solar radiation and minimal, maximal and average temperatures). This approach is allowed only for datasets with a low quantity of attributes. However, some CYP datasets have relatively few attributes and an exhaustive approach can be applied for model comparison purposes (Ruß & Kruse, 2010). In addition, a bigger number of real cases can be addressed each time a new computer generation arises.

Evaluating regression techniques requires performance metrics. The R and the RMSE are the most commonly used metrics (Drummond *et al*., 2003). As an average, RMSE is dominated by its large individual terms. For this reason, in this work, the RRSE was also considered. Unlike metrics that are measured in physical units, RRSE provides a reference point, being easy to understand by people not related to agricultural forecasting metrics. This reasoning was also applied to MAE metric, which is expressed in percentage respect to the mean.

The evaluated techniques were ranked, from the best to the worst, according to RMSE, RRSE, R, and MAE results, in the following order: M5-Prime, kNN, SVR, ANN and MLR. An evaluation described by Ruß (2009) ranked M5-Prime in the last position, below multi-layer perceptron and the SVR techniques. These discrepancies can be due to: a) the attribute set of Ruß (2009) containing only soil features (fertilizers, vegetation and electric conductivity) and b) the regression trees parameter values related to the samples of leaf nodes and pruning procedure are different. The SVR parameters used in Ru (2009) were tested using our data with very bad results. Regrettably, Ruß (2009) did not report the number and species of crops, and a fair comparison is difficult.

ANN have reported a better performance than classical statistical methods (Drummond *et al*., 2003) and regression trees (Ruß, 2009; Ruß & Kruse, 2010). However, ANN obtained a poor performance in our work; their average values for RMSE and MAE were the highest, while for R were the lowest. The

performance of neural networks depends on several factors, such as the network structure, training parameters and samples' quality. Nevertheless, an important difference between our work and the previous ones is that we did not include any attributes to identify the crop field. This is because: a) the fields are not always cultivated or their crops are not always the same and b) the agricultural planning is made at the global level in our study case. Our results are according to a previous conclusion that site-dependency of ANNs makes difficult to obtain good performance when field identification is missing (Liu *et al*., 2001). In addition, it should be considered that the same structure for all the developed ANN models was used. To develop a different ANN structure for each crop is impractical due the number of crops and the training time required. This practice does not follow the ANNs' experimental nature, which requires several trials to obtain a good predictor model.

As mentioned before, kNN technique provides comparable average results than M5-Prime. In our rank, only the individual counting of the best CYP models by technique places kNN below M5-Prime. Thus, this technique deserves further research to explore their potential in agricultural planning.

The average measures of MLR models are between ANN and SVR mean results. However, individually none of the MLR models' results surpass any other of the remaining techniques. This situation occurs with all the error metrics implemented. Despite the unfairness of the comparison (after all, only one linear technique versus four non-linear techniques are compared), the general assumption is that ML techniques are more suitable to extract complex relationships from data than MLR.

As it can be seen in Table 4 and Table 5, M5-Prime achieves the largest number of models with the lowest RMSE and RRSE and the highest R. In practice, each crop has a different economic profit, and a modeling technique that maximizes the number of accurate models is important. Therefore, M5-Prime is selected as a very suitable modeling tool for massive CYP. A second choice is kNN, for their good obtained results.

However, in a strict sense, in the group of evaluation techniques, none of them exceeds the entire set. Although M5-Prime and kNN achieve mostly the models with the lowest errors, SVR and ANN techniques obtain some good models as well. Thus, for maximizing the CYP efficiency, a combination of ML models can be the best strategy for agricultural planning. Further research is to develop an integrated framework for selecting the most appropriate regression technique for each crop dataset.

As final conclusions, we have that the best attribute set obtained from each training dataset varies a lot from one year to another, once new data are incorporated. Thus, is difficult to establish a constant set of attributes that guarantee good results all the time for all techniques, although this issue is variable for crop dataset. In our experiment, only two datasets (CD03, CD06) show some tendency to include repeatedly the same attributes for all techniques. Related with this, is the fact that some crop datasets may be very difficult to model for any regression technique. The individual count applied to obtain the models with all the lowest errors shows that some crop datasets do not meet any model that accomplishes these criteria. This is the case for the CD03, CD07, CD08 and CD09 datasets. The most likely reason for that is the natural complexity of the crop-cultivar yield behavior, which is not reflected by the attributes used. Proper information of the agricultural area can help to test this assumption, proposing a different set of attributes with these crop-cultivars.

Also, it should be considered that ensuring a fair comparison for evaluating all the ML techniques is difficult. The evaluation method utilized considers the fact that a different set of attributes changes the performance of ML models. Nevertheless, the same configuration for each ML technique is used to generate all its models. Most of the parameter values were taken from the literature. Evidently, some techniques are more experimental than others, and such treatment can be inequitable for some methods. Further research will focus on trying an extensive number of configurations and parameter values for all the ML techniques. This "massive calibration of models" should be reproducible for practical agricultural planning cases.

Finally, it is necessary to point out that this work deals only with comparing the predictive accuracy of the above-mentioned techniques. Machine learning techniques are complex, and several factors are related with their performance measuring. Some examples of these factors are the model structure, knowledge representation, implementation cost, missing data handling and training time. Further research will be dedicated to compare these characteristics of ML algorithms and their compatibility with agricultural planning.

Acknowledgements

The authors acknowledge the support from the Mexican Institute of Water Technology (IMTA) and particularly the Department of Irrigation and Drainage, which provided the datasets used in this research.

References

Breiman L, 2001. Statistical modeling: the two cultures (with discussion). Statist Sci 16: 199-231.

Breiman L, Friedman JH, Olshen RA, Stone CJ, 1984. Classification and regression trees. Wadsworth, Belmont, CA, USA.

Brisson N, Marry B, Ripoche D, Jeuffory MH, Ruget F, Nicoullaud B, Gate P, Devienne BF, Antonioletti R, Durr C *et al.,* 1998. STICS: a generic model for the simulation of crops and their water and nitrogen balance. 1. Theory and parameterization applied to wheat and corn. Agronomie 18: 311-346.

Dixon BL, Hollinger SE, Garcia P, Tirupattur V, 1994. Estimating corn yield response models to predict impacts of climate change. J Agr Resour Econ 19(1): 58-68.

Drummond ST, Sudduth KA, Joshi A, Birrel SJ, Kitchen NR, 2003. Statistical and neural methods for site-specific yield prediction. T ASABE 46 (1): 5-14.

Fortin JG, Anctil F, Parent L, Bolinder MA, 2011. Site-specific early season potato yield forecast by neural network in Eastern Canada. Precis Agr 12(6): 905-923.

Frausto-Solís J, Gonzalez-Sanchez A, Larre M., 2009. A new method for optimal cropping pattern. Proc. 8th Mex Int Conf on Artificial Intelligence, pp: 566-577.

Goudriaan J, van Laar H, 1994. Modelling potential crop growth processes. Kluwer Acad. Publ., Dordrecht, the Netherlands.

Hair JF Jr, Anderson RE, Tatham RL, 1987. Multivariate data analysis, 2nd edition. MacMillan Publ. Co.

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

Hand D, Mannila H, Smyth P, 2001. Principles of data mining. MIT Press.

Irmak A, Jones JW, Batchelor WD, Irmak S, Boote KJ, Paz JO, 2006. Artificial neural network model as a data analysis tool in precision farming. T ASABE 49(6): 2027-2037.

Jaikla R, Auephanwiriyakul S, Jintrawet A, 2008. Rice yield prediction using a support vector regression method. ECTI-CON 5th Int Conf, Vol. 2, pp: 29-32.

Jamieson PD, Semenov MA, Brooking IR, Francis GS, 1998a. Sirius: a mechanistic model of wheat response to environmental variation. Eur J Agron 8: 161-179.

Jamieson PD, Porter JR, Goudriaan J, Ritchie JT, van Keulen H, Stol W, 1998b. A comparison of the models AFRCWHEAT2, CERES-Wheat, Sirius, SUCROS2 and SWHEAT with measurements from wheat grown under drought. Field Crops Res 55: 23-44.

Jones CA, Kiniry JR, 1986. CERES-Maize: A simulation model of maize growth and development. Texas A&M Univ. Press, College Station, Texas, USA. 194 pp.

Kohavi R, 1995. Wrappers for performance enhancement and oblivious decision graphs. Doctoral dissertation, Stanford Univ., Comp. Sci. Dept.

Liu J, Goering CE, Tian L, 2001. Neural network for setting target corn yields. T ASAE 44(3): 705-713.

Marinković B, Crnobarac J, Brdar S, Antić B, Jaćimović G, Crnojević V, 2009. Data mining approach for predictive modeling of agricultural yield data. Proc. First Int Workshop on Sensing Technologies in Agriculture, Forestry and Environment (BioSense09), Novi Sad, Serbia, October, pp: 1-5.

McQueen RJ, Garner SR, Nevill-Manning CG, Witten IH, 1995. Applying machine learning to agricultural data. Comput Electron Agr 12(4): 275-293.

Ojeda-Bustamante W, González-Camacho JM, Sifuentes-Ibarra E, Isidro E, Rendón-Pimentel L, 2007. Using spatial information systems to improve water management in Mexico. Agr Water Manage 89: 81-88.

Porter JR, 1993. AFRCWHEAT2: a model of the growth and development of wheat incorporating responses to water and nitrogen. Eur J Agron 2: 69-82.

Quinlan JR, 1992. Learning with continuous classes. Proc. AI'92, 5th Aust. Joint Conf. on Artificial Intelligence (Adams & Sterling, eds.), World Scientific, Singapore, pp: 343-348.

Roel A, Plant RE, 2004. Factors underlying yield variability in two California rice fields. Agron J 96: 1481-1494.

Rojas R, 1996. Neural networks - A systematic introduction. Springer-Verlag, Berlin, NY.

Rumelhart DE, Hinton GE, Williams RJ, 1986. Learning internal representations by error propagation. In: Parallel distributed processing: explorations in the microstructure of cognition, (Rumelhart DE, McClelland JA, eds), vol. 1, chapter 8. The MIT Press, Cambridge, MA (USA). pp: 418-362.

Ruß G, 2009. Data mining of agricultural yield data: a comparison of regression models. Proc. 9th Indust. Conf. on Advances in Data Mining-Applications and Theoretical Aspects, July 20-22, Leipzig, Germany.

Ruß G, Kruse R, 2010. Feature selection for wheat yield prediction. In: Research and development in intelligent systems XXVI (Bramer M *et al.,* eds.), Springer-Verlag, London.

Safa B, Khalili A, Teshnehlab M, Liaghat A, 2004. Artificial neural networks application to predict wheat yield using climatic data. Proc. 20th Int. Conf. on IIPS, Jan. 10-15, Iranian Meteorological Organization, pp: 1-39.

Schlenker W, Roberts MJ, 2006. Estimating the impact of climate change on crop yields: The importance of non-linear temperature effects. Discussion Papers 0607-01, Columbia University, Dept. Economics.

Shevade SK, Keerthi SS, Bhattacharyya C, Murthy KRK, 2000. Improvements to the SMO algorithm for SVM regression. IEEE Transactions on Neural Networks 11(5): 1188-1193.

Smola A, Schölkopf B, 2004. A tutorial on support vector regression. Stat Comput 14(3): 199-222.

Spitters CJT, van Keulen H, van Kraailingen DWG, 1988. A simple but universal crop growth simulation model, SUCROS87. In: Simulation and systems management in crop protection (Rabbinge R, Van Laar H & Ward S, eds). PUDOC, Wageningen. pp: 87-98.

Sudduth KA, Drummond ST, Birrell SJ, Kitchen NR, 1996. Analysis of spatial factors influencing crop yield. Proc. 3rd Int. Conf. on Precision Agriculture (Robert PC, Rust RH, & Larson WE, eds.) ASA-CSSA-SSSA, Madison, WI, USA, pp: 129-140.

Sudduth K, Fraisse C, Drummond S, Kitchen N, 1998. Integrating spatial data collection, modeling and analysis for precision agriculture. First Int. Conf. on Geospatial Information in Agriculture and Forestry, vol. 2, pp: 166-173.

Uysal I, Altay HG, 1999. An overview of regression techniques for knowledge discovery. Knowl Eng Rev 14: 319-340.

Vapnik V, Lerner A, 1963. Pattern recognition using generalized portrait method. Automat Remote Contr 24: 774-780.

Vapnik V, Golowich S, Smola A, 1997. Support vector method for function approximation, regression estimation, and signal processing. In: Advances in neural information processing systems (Mozer M, Jordan M, & Petsche T, eds), MIT Press, Cambridge, MA, USA, pp: 281-287.

Varcoe VJ, 1990. A note on the computer simulation of crop growth in agricultural land evaluation. Soil Use Manage 6(3): 157-160.

Wang Y, Witten I, 1997. Inducing model trees for continuous classes. Proc. 9th Eur. Conf. Machine Learning (van Someren M & Widmer G, eds), pp: 128-137.

Wasserman L, 2004. All of statistics. A concise course in statistical inference. Springer.

Wilkerson GG, Jones JW, Boote KJ, Ingram KT, Mishoe JW, 1983. Modeling soybean growth for crop management. T ASAE 26: 63-73.

Witten IH, Frank E, Trigg L, Hall M, Holmes G, Cunningham SJ, 1999. Weka: Practical machine learning tools and techniques with Java implementations. ICONIP/ANZIIS/ANNES'99 Int. Workshop (Kasabov H & Ko K, eds), Dunedin.

Yang Y, 2008. Consistency of cross validation for comparing regression procedures. Ann Stat 35 (6): 2450-2473.

Zhang L, Zhang J, Kyei-Boahen S, Zhang M, 2010 Simulation and prediction of soybean growth and development under field conditions. Am-Euras J Agr Environ Sci 7(4): 374-385.