Forecasting the black Sigatoka development rate: A comparison of machine learning techniques

Luis-Alexander Calvo-Valverde[a,1], Mauricio Guzmán-Quesada[b], José-Antonio Guzmán-Alvarez[b], Pablo Alvarado-Moya[c]

[a] *DOCINADE, Instituto Tecnológico de Costa Rica, Computer Research Center, Multidisciplinar program eScience, Cartago, Costa Rica*
[b] *Dirección de Investigaciones, Corporación Bananera Nacional S. A., Guápiles, Costa Rica*
[c] *DOCINADE, Instituto Tecnológico de Costa Rica, Cartago, Costa Rica*

Abstract

Pending.

Keywords: Machine learning, Black Sigatoka, Support vector regression, Banana disease prediction, Biological warning system

1. Introduction

The black Sigatoka disease caused by the fungus *Mycosphaerella fijiensis Morelet* is the major pathological problem of banana and plantain crops in Central America, Panama, Colombia and Ecuador, as well as in many parts of Africa and Asia (Marín Vargas and Romero Calderón, 1995).

This disease attacks the plant leaves producing a rapid deterioration of the leaf area. It affects the growth and productivity of the plants due to the impairment of the photosynthetic process. Furthermore, it causes a reduction in the quality of the fruit, and promotes premature ripening of bunches, which is the major cause of product losses associated with the black Sigatoka.

For these reasons, warning systems have been developed to detect the disease and monitor its progress. For instance, the early warning system developed by Ganry and Meyer (1983) and modified by Ganry and Laville (1972) for the

Email address: lcalvo@itcr.ac.cr (Luis-Alexander Calvo-Valverde)
[1] Corresponding author. (506)70104420

control of the yellow Sigatoka in Cameroon, was later adapted by Ternesien (1985) and Fouré (1988) for the black Sigatoka.

This biological warning system is based on weekly observations of the disease progression on young leaves of the plant. Figure 1 shows an example of three progressive stages of the black Sigatoka.

(a) (b) (c)

Figure 1: Examples of three disease stages of the black Sigatoka. (a) Initial stage. (b) Intermediate stage, and (c) Advanced stage.

The disease progression is then quantified according to Fouré's scale of the symptom stages (Fouré, 1988) by means of numeric coefficients that describe the degree of incidence and the severity of the disease development. These coefficients are then used to calculate two variables: gross sum and state of evolution.

The gross sum is based on the present disease progression stage and the numeric coefficients, which increase with the progression of the symptoms and the juvenility of the leaf. The state of evolution is calculated using the gross sum and the foliar emission period. Decades ago threshold levels on these variables were used as a guide to plan the spray schedules. Nowadays the fluctuation of these two variables seems to better suggest appropriate times to spray (Marín et al., 2003).

In Costa Rica the black Sigatoka is frequently treated with chemical fungicides. Depending on the zone of production and the weather conditions, 45–55 cycles/year of fungicide applications are required to keep this disease under control and to produce the expected fruit quality for the international mar-

kets. This represents a cost per hectare per year in the range from US$1600 to US$2000; about 0.64–0.80 cents of the production costs for a 18.14 kg box, which overall corresponds to 10%–12% of the total production costs.

The past and present rates of disease development can in principle be used to predict its future behavior and to determine if a particular fungicide spray program will be able to effectively treat the disease in an economically affordable way (Chuang and Jeger, 1987). Phytopathological studies point out that climate has a major effect on the development of the black Sigatoka, where the main variables affecting it are precipitation, temperature, relative humidity and wind (Marín Vargas and Romero Calderón, 1995). It can be expected that patterns in these variables correlate with the disease development and hence its automated discovery can support decision-making in the control of crop diseases.

In this work, we compare four machine learning techniques to predict the development rate of the black Sigatoka disease: support vector regression (SVR), echo state networks (ESN), elastic net regression and ordinary least squares linear regression, using input variables: maximal air temperature, minimal air temperature, mean air temperature, mean relative humidity, minimal relative humidity, maximal relative humidity, mean solar radiation, sum of precipitation, maximal wind speed and mean speed wind, to predict the evolution stage in the biological warning system.

The main findings were: 1) The highest R2 was 60% 2) The highest R2 were reached with linear models like support vector regression with linear kernel, 3) As little as three meteorological variables can be used because of the correlations detected among variables.

The outline of the paper is as follows: Section 2 presents related works and Section 3 summarizes the machine learning techniques selected for the analysis. In Section 4 we present the methodology used in this study and describe data used for its verification. The results and their discussion are presented in section 5. The Section 6 concludes this article and presents lines for future works.

2. Related works

Several efforts have been made to apply machine learning techniques in the automated discovery of relationships between environmental variables and quantified descriptors for variables of agricultural interest such as the progress of diseases. Huang et al. (2010) summarize in their survey the development of soft computing techniques in agricultural and biological engineering, especially in the soil and water context for crop management and decision support in precision agriculture, including fuzzy logic, artificial neural networks, genetic algorithms, Bayesian inference and decision trees. They do not present numeric results of each paper, only mention the main idea. Similarly, Kim et al. (2014) survey more recent prediction methods for crop pests using regression and machine learning approaches. Nor do they provide numerical results of each paper. In general, the machine learning methods applied to predict the evolution of plant diseases, can be classified in two main approaches: 1) those whose main inputs are images, and 2) Those whose main inputs are environmental and biological variables. Our study focuses in the second case. Romero Calderón (1995) relied on regression models using a stepwise procedure to predict incubation and disease latency periods for the black Sigatoka. He collected environmental data from two different farms in Costa Rica between December 1993 and August 1995. The prediction models reached coefficients of determination R^2 of 69% or 78% on the observed data for the incubation and disease latency periods, respectively; however, the cross validation on independent data sets failed. In contrast, our proposal presents a model that can be generalizable to other farms that have data.

More recently, Glezakos et al. (2010) used genetic algorithms (GA) and neural networks (NN) to identify the Tobacco Rattle Virus (TRV) and the Cucumber Green Mottle Mosaic Virus (CGMMV). The method was tested against some of the most commonly used classifiers in machine learning (Bayes classifiers, decision trees and k-nearest neighbors) via cross-validation and proved their applicability in these kind of problems. These authors do not prove their

methods in Sigatoka disease and they do classification. Instead we do regression.

Alves et al. (2011) used geoinformation techniques to develop predictive models in the study of risk areas to soybean rust, coffee leaf rust, and banana black Sigatoka, under consideration of Brazil's climatic characteristics and the distribution of soybean, coffee and banana crops. Temperature and rainfall data were acquired for the period from 1950 to 2000, and simulated data were generated for 2020, 2050 and 2080 using the SRES A2 climate change scenarios. Using principal components analysis, a single variable was generated as a linear combination of 57 input variables, in order to determine an index explaining 87%, 88% and 90% of the data variability of soybean, coffee and banana crops, respectively, in municipal districts across Brazil. The climatic model was used to generate the zoning of the three plant diseases, using temperature and leaf wetness as input. This methodology enabled the visualization of the changes in areas favorable for epidemics under possible future scenarios of climate change. How intermediate result, they characterized the monociclic process of the black Sigatoka using nonlinear regression. Although they do not present the detailed results, it no seem that they wanted to predict the progression of the black Sigatoka in one, two or more periods ahead, how we do.

Other applications of machine learning methods in precision agriculture include the use of support vector regression to predict carcass weight in beef cattle in advance to the slaughter (Alonso et al., 2013), machine learning assessments of soil drying for agricultural planning (Coopersmith et al., 2014), and early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance (Rumpf et al., 2010).

Furthermore, there have been attempts to generate software tools. Camargo et al. (2012) presented an information system for the assessment of plant disorders (Isacrodi). They showed that human experts will attain a much accurate assessment than the Isacrodi classifier, particularly when provided with samples from the affected crop. However, in those cases where such expertise is not available, the authors suggest that Isacrodi can still provide valuable support to farmers. Isacordi includes 15 crop disorders, but the black Sigatoka is none of

them. The prediction process is based on multi-class support vector machines.

Regarding the prediction of the black Sigatoka disease development with machine learning methods, Bendini et al. (2013) presented a study on the risk analysis of its occurrence based on polynomial models. A case study was developed in a commercial banana plantation located in Jacupiranga, Brazil. It was monitored weekly from February to December 2005. The data included the weekly monitoring of the disease's evolution stage, time series of meteorological data and remote sensing data. They obtained a model to estimate the evolution of the disease from satellite imagery. This model relates gray levels (NC) of the band 2 images of the Landsat-5 satellite, with the progress status or disease severity (EE). The authors claim to reach an R^2 of 90%.

There are also works related to the banana fruit. Soares et al. (2014) apply two techniques: artificial neural networks (ANNs) and multiple linear regression (MLR) in banana plant to predict the yield. Their results show that the neural network is more accurate in forecasting the weight of the bunch in comparison to the multiple linear regressors in terms of the mean prediction-error ($MPE = 1.40$), mean square deviation ($MSD = 2.29$) and coefficient of determination ($R^2 = 91\%$).

Although these studies have their contribution, none proposed the kind of preprocessing that we present, nor pose how to predict more than one period ahead without trying to predict climate.

3. Compared regression techniques

In the prediction of the development rate of the black Sigatoka, we compare techniques such as least squares regression and elastic-net regression, commonly encountered in the agricultural literature with machine learning methods such as support vector regression and echo state networks, where the parameter space of each technique is also taken into account.

3.1. Ordinary least squares regression

Given a data set

$$D = \{(\mathbf{x}_i, y_i) \mid i = 1 \dots n\} \tag{1}$$

composed of the d-dimensional[2] feature vectors $\mathbf{x}_i \in \mathbb{R}^d$ and the corresponding responses y_i. The ordinary least squares regression (OLSR) fits a linear model $\tilde{y}_i = f(\mathbf{x}_i) = \langle \mathbf{w}, \mathbf{x}_i \rangle$ such that the sum of squares of the residuals $(\tilde{y}_i - y_i)$ is minimized. Let $\mathbf{X}$ be the $n \times d$ feature matrix containing the i-th data sample $\mathbf{x}_i^T$ in its i-th row and $\mathbf{y}$ contain all the responses y_i corresponding to each row, then the least squares regression finds

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} E(\mathbf{w}) \tag{2}$$

with the error function

$$E(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2$$

The solution is found by means of the pseudoinverse $\hat{\mathbf{w}} = \mathbf{X}^+\mathbf{y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ or equivalently by the singular value decomposition of $\mathbf{X}$ (Press et al., 2007).

3.2. Elastic net regression

Instead of L_2 regularization prior ($\alpha \|\mathbf{w}\|_2^2$) included in the ridge regression (RR), Tibshirani (1996) used an L_1 term ($\lambda \|\mathbf{w}\|_1$) for his lasso estimator, which permits to select a subset of the available features by zeroing the weights of the deselected features. If the dimension d of the data is larger than the number n of data samples, lasso will select a maximum of d variables.

The elastic net regression (ENR) of Zou and Hastie (2005) combines both L_1 and L_2 priors of the ridge and lasso estimators such that the error function is now

$$E(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \alpha \|\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1$$

This combination of priors still allows to learn a sparse model with only a few weights being non-zero like in the case of lasso, but still maintaining the regularization properties of the ridge regression (Pedregosa et al., 2011).

[2] Without loss of generality assume that the first component of every vector $\mathbf{x}_i$ is always 1.

The elastic net is useful when multiple features are correlated: lasso will likely pick one of these at random, while the elastic net will still likely pick both.

3.3. Support Vector Regression (SVR)

From the perspective of Support Vector Regression (SVR) the regression function is usually formulated as

$$\tilde{y} = f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b \tag{3}$$

The weights are selected in a convex optimization problem (Smola and Schölkopf, 2004):

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ \text{subject to} \quad & \begin{cases} y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - b & \leq \epsilon + \xi_i \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i & \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq 0 \end{cases} \end{aligned}$$

where ϵ is the maximal allowed deviation of the targets $\tilde{y}_i$ from the responses y_i, the slack variables ξ_i and ξ_i^* allow to cope with otherwise unfeasible constraints for the optimization problem, and the constant $C > 0$ controls the trade-off between the flatness of f and the tolerance to deviations larger than ϵ.

Note that since OLSR, RR and ENR use a squared error function, data outliers will have a strong influence on the resulting weights $\mathbf{w}$. On the SVR formulation, however, the usage of the L_1 norm and the slack variables considerably restrict or completely block the influence of those outliers.

The SVR problem is reformulated by means of the dual optimization problem into (Smola and Schölkopf, 2004)

$$\mathbf{w} = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\mathbf{x}_i \tag{4}$$

where $\alpha_i, \alpha_i^* \in [0, C]$ are Lagrange multipliers subject to $\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0$. In this so-called *Support Vector expansion* the weights are expressed as a linear

combination of the data set patterns $\mathbf{x}_i$. Inserting (4) in (3) leads to

$$f(\mathbf{x}) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \langle \mathbf{x}_i, \mathbf{x} \rangle + b \tag{5}$$

The Lagrange multipliers α_i, α_i^* are both non-zero only for those data points where $|f(\mathbf{x}_i) - y_i| \geq \epsilon$. Hence, the expansion of $\mathbf{w}$ in terms of $\mathbf{x}_i$ is sparse. Those data points with non-vanishing coefficients are called *Support Vectors* (Wei et al., 2013).

Additionally, in (5) it is possible to employ the *kernel trick* and replace the terms $\langle \mathbf{x}_i, \mathbf{x} \rangle$ with the evaluation of any Mercer kernel $k(\mathbf{x}_i, \mathbf{x}) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle$, where $\phi(\mathbf{x})$ is a non-linear mapping of the input space onto a higher (even infinite) dimensional feature space. The kernel evaluation draws unnecessary the explicit evaluation of the non-linear mapping, and it allows to solve non-linear regressions in the input space by implicitly mapping the samples through the kernel into the higher dimensional space, where the linear regression occurs (Alonso et al., 2013).

Kernels used in this work were:

Linear kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = {\mathbf{x}_i}^T \mathbf{x}_j$$

Radial basis function (RBF):

$$K(\mathbf{x}_i, \mathbf{x}_j) = exp\left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$$

where σ is the parameter of gausian model.

Sigmoid:

$$K(\mathbf{x}_i, \mathbf{x}_j) = tanh\left[-c + \frac{\mathbf{x}_i \mathbf{x}_j}{\sigma^2}\right]$$

with $c \geq 0$ and σ^2 is the scaling vector.

3.4. Echo State Networks (ESN)

Recurrent neural networks (RNN) are capable of learning temporal patterns by feeding neuron outputs back into lower layers. Their training (usually by means of error backpropagation) is in general slow. Echo state networks (ESN)

are a particular type of recurrent neural network with a sparsely connected random hidden layer where only the weights of the output neurons are changed at training. The randomly selected weights at the input and middle layers (called *reservoir*) reproduce temporal patterns (*echoes*) that the output layer learns to select during the training (Lukoševičius and Jaeger, 2009).

For a given training input signal $u(n) \in \mathbb{R}^{N_u}$ a desired target output signal $y^{target}(n) \in \mathbb{R}^{N_y}$ is known. Here $n = 1, \ldots, T$ is the discrete time and T is the number of data points in the training dataset.

The training seeks to learn a model with output $y(n) \in \mathbb{R}^{N_y}$, where $y(n)$ matches $y^{target}(n)$ as close as possible, by means of the minimization of an error measure $E(y, y^{target})$ such that it also generalizes well to unseen data (Lukoševičius, 2012).

4. Specification of data and methodology

Since the suitability of a machine learning technique to a particular problem is entirely depend on the nature of the data, we describe in this section, first, the data set employed in the study, followed by the methodology to compare the chosen techniques under consideration of their parameter space.

4.1. Data

The data used for the current study was acquired in two research farms of Corbana in Costa Rica[3] : *28 Millas* located in the region of Matina, and *La Rita* located in Pococí, both in the province of Limón, Costa Rica. Both farms produce banana fruit *Musa* sp. AAA group 'Grande Naine' (Cavendish subgroup).

The available input and output variables are summarized in Table 1.

The data was captured for La Rita between the 48 th week of 2002 to the 17 th week of 2015 (647 weeks); for 28 Miles the data was captured between the

[3]Both farms were also use in the study of Romero Calderón (1995). Back then, *La Rita* was referred to as *Waldeck*.

Symbol	Description	Units
$T_{a_{max}}$	Maximal air temperature	[°C]
$T_{a_{min}}$	Minimal air temperature	[°C]
$\overline{T}_a$	Mean air temperature	[°C]
$\overline{H}$	Mean relative humidity	$[0-100]$
H_{min}	Minimal relative humidity	$[0-100]$
H_{max}	Maximal relative humidity	$[0-100]$
$\overline{R}$	Mean solar radiation	$[W/m^2]$
P	Precipitation	$[mm]$
W_{max}	Maximal wind speed	$[m/s]$
$\overline{W}$	Mean speed wind	$[m/s]$
E_s	Biological warning system – Evolution Stage	> 0

Table 1: Variables available for the learning algorithms

37 th week of 2003 and the 18 th week of 2015 (605 weeks). The data on the biological warning system were collected once a week.

The meteorological stations of Corbana acquire data every five minutes. computed on the data collected by nearby stations in each farm. Experiments were carry out with daily periodicity in meteorological variables and the results proved do not improve the prediction. Besides, weekly data pretend to diminish noise due sensor accuracy, missing values and outliers no detected.

The value to be predicted in all cases is the evolution stage E_s, which is a measure of the level of disease progression.

4.2. Data preprocessing

Data taken on real farms during more than a decade is expected to contain outliers, noise and missing samples. These problems are caused by human errors or by technical defects on the instruments used. In the preprocessing step described in this section these problems need to be detected and fixed before moving them to the next processing stages.

In the farm 28 Miles 1% and in La Rita 2.25% of the data were missing. To fill-in the missing values spline interpolation was used ALGLIB® (2017). The data collected did not exhibit outliers.

Each variable $x \in [x_{\min}, x_{\max}]$ was normalized into the interval $[0, 1]$ with the linear map $x_n = mx + b$ with $m = 1/(x_{\max} - x_{\min})$ and $b = -mx_{\min}$.

The variable E_s to be predicted was not normalized. This normalization step was made because learning schemes, like regression methods, deals only with ratio scales because they calculate the distance between two instances based on the values of their attributes Witten et al. (2011).

4.3. Evaluation criteria

Although there are many types of indicators to assess the quality of the prediction, here the coefficient of determination (R^2) and the Root Mean Square Error ($RMSE$). This decision is supported by the widespread use of the former indicator in the agriculture and the latter in machine learning (Soares et al., 2013, 2014; Ibrahim and Wibowo, 2014; Demir and Bruzzone, 2014).

Given n records y_i, $i = 1 \dots n$ of the actual outcome of a process. The mean $\bar{y}$ of the observed data is given by

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Let $\hat{y}_i$ be the predicted value for y_i. Then, the mean square error (MSE) S_e^2 and the unexplained variance S_R^2 are estimated as University (2017)

$$S_e^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} \qquad S_R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{n}$$

The root mean square error is defined as $RMSE = \sqrt{S_e^2}$ and the coefficient of determination is

$$R^2 = \frac{S_R^2}{S_R^2 + S_e^2}$$

4.4. Programming environment

We use Python programming language in its interpreted 3.5.2 version, particularly with the libraries; Pandas (0.18.1 version) (McKinney, 2010) and Numpy (1.11.1 version) (van der Walt et al., 2011).

The implementation for SVR, elastic net and ordinary least squares regressions in scikit-learn (Pedregosa et al., 2011) were used. Adjustments to the ESN implementation code of Lukoševičius (2012) were necessary to allow its integration into our experimental framework.

4.5. Methodology

When the amount of data for training and testing is limited, is recommended to use cross-validation (Witten et al., 2011). We used ten-fold-cross-validation on the total set. El diseño experimental combinó los siguientes factores:

- Patrones de semanas: Se formaron patrones desde tres semanas continuas de observación para predecir una semana adelante, hasta doce semanas de observación para predecir tres semanas adelante. Por tanto, $n \times m$ combinations, with $n = 1 \dots 12$ and $m = 1 \dots 3$.
- Techniques: Se utilizaron las siguientes técnicas, revisando el espacio paramétrico especificado a continuación:
  - SVR with linear kernel: C $[0.001, 1, 10, 100, 1000]$, $epsilon$ $[0.0, 0.4, 0.9]$.
  - SVR with gaussian kernel: C $[0.001, 1, 10, 100, 1000]$, $epsilon$ $[0.0, 0.4, 0.9]$, $gamma$ $[0.0, 0.4, 0.9]$.
  - SVR with sigmoid kernel: C $[0.001, 1, 10, 100, 1000]$, $epsilon$ $[0.0, 0.4, 0.9]$, $gamma$ $[0.0, 0.4, 0.9]$, $coef0$ $[0.0, 0.5, 5, 10]$.
  - Echo state networks: $LeakingRate$ [0.02..0.9], $neurons$ [1%..90%] de la cardinalidad del training set, $InitLen$ [0.1..0.8].
  - Ordinary least squares linear regression: No parameters.
  - Elastic-net regression: $alpha$ [0..0.9], $l1\_ratio$ [0..1.0].
- Variables included in the model:
  - All variables.
  - From the set $\{\overline{T}_a, \overline{H}, P, \overline{W}\}$ use the subsets with one, two, three or four elements. These variables have the largest impact on the disease development (Marín Vargas and Romero Calderón, 1995).

5. Results and discussion

Realizados los experimentos, la Table.2 muestra los $RMSE$ obtenidos, en la cual se puede observar que las ESN obtienen un resultado muy diferente a las demás técnicas. El $RMSE$ promdio obtenido por las ESN 11085.47 es cercano a 20 veces el de las otras técnicas. Este resultado de las ESN se explica debido a que estas redes neuronales recurrentes parten de estados aleatorios para irse ajustando en el proceso de entrenamiento, y por tanto, los valores obtenidos son reflejo de que la cantidad de datos disponibles no logra que la red neuronal se ajuste. Adicionalmente, la Figure 2 muestra los box plots con respecto a los

Table 2: Promedio y desviación estándar de los RMSE obtenidos por Finca

Farm	**Technique**	$RMSE_mean$	$RMSE_std$
28 Millas	Elastic net	463.56	82.39
	SVR with linear kernel	465.92	85.70
	SVR with Gaussian kernel	466.63	89.53
	Linear regression	468.25	83.57
	SVR with sigmoid kernel	552.81	123.82
	ESN	11085.47	7965.66
La Rita	SVR with linear kernel	816.29	216.43
	Elastic net	817.98	211.14
	SVR with Gaussian kernel	820.92	231.21
	Linear regression	823.55	213.24
	SVR with sigmoid kernel	1070.58	331.89
	ESN	8329.72	5435.54

$RMSE$ obtenidos. Se grafica ESN por separado debido a la diferencia de escala. Con respecto a las otras técnicas, la regresión lineal, Elastic net, SVR with gaussian kernel y SVR with linear kernel presentan compartamientos muy similares y aunque el promedio del $RMSE$ es diferente entre las fincas, La Rica cercano a 450 y 28 Millas cercano a 820, el comportaniento relativo de las técnicas es el

mismo. Por su parte, SVR with sigmoid kernel presenta un promedio de $RMSE$ muy inferior a ESN, pero superior a las otras cuatro técnicas.

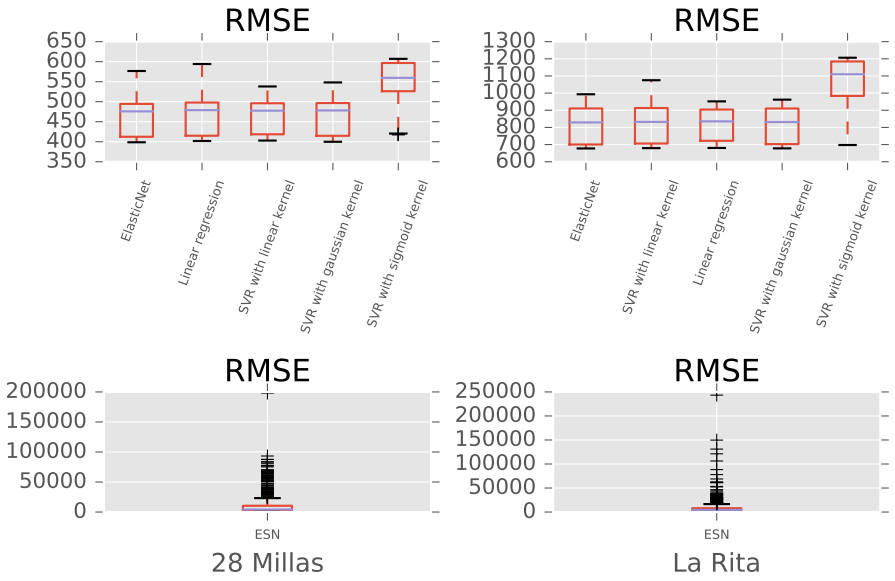

Figure 2: Box plots de los RMSE para cada una de las Fincas

Figure 3 shows the Pareto frontier for each farm with respect to R^2 and $RMSE$. La Rita obtains upper R^2 with respect to 28 Millas, but 28 Millas obtains better $RMSE$ than La Rita. This situation arise because $RMSE$ considers errors only with respect the prediction and in 28 Millas the average of Stage of Evolution is 4316.16, unlike, in La Rita the average is 5507.30. So, in La Rita we obtains higher errors in absolute values. R^2 is a relative metric between 0 thru 1 and it is less sensitive to absolute values.

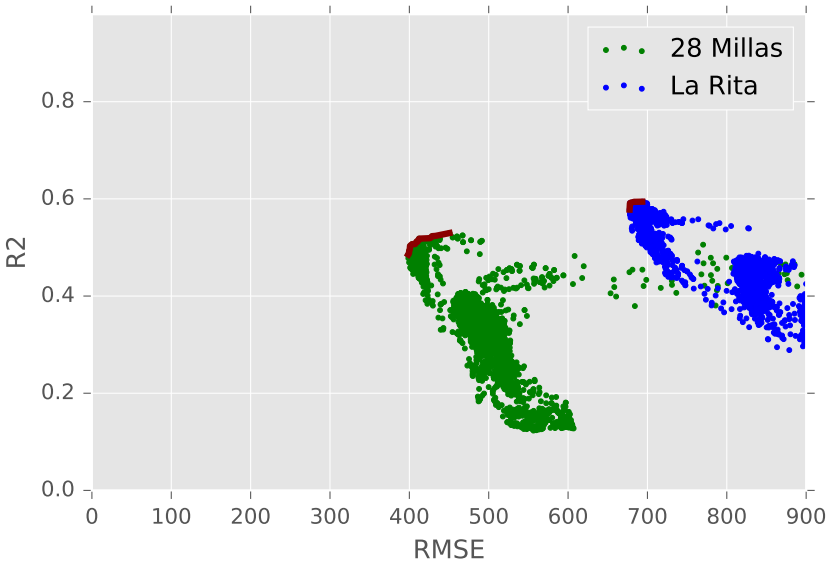

Figure 3: Pareto frontier for R^2 and $RMSE$

The Pareto frontier for the La Rita farm is composed by 5 elements. The Table.3 shows the composition about variables, observation ranges, techniques and weeks ahead.

Table 3: Composition of the Pareto frontier - La Rita

Variable	**Observation range**	**Weeks ahead**	**Technique**	R^2	$RMSE$
$\overline{T}_a$ $\overline{W}$	1	1	SVR with linear kernel	693.39	59.4%
	2	1	SVR with linear kernel	683.22	59.4%
$\overline{T}_a$	5	1	SVR with linear kernel	679.25	59.2%
$\overline{T}_a$ $\overline{H}$	5	1	Elastic net	677.49	57.8%
	5	1	SVR with linear kernel	677.37	59.1%

Similarly, the Pareto frontier for the 28 Millas farm is composed by 15 elements. The Table.4 shows the composition about variables and observation ranges.

Now, the Figure 4 compares the mean of the $RMSE$ for each technique in

Table 4: Composition of the Pareto frontier - 28 Millas

Variable	**Observation range**	**Weeks ahead**	**Technique**	R^2	$RMSE$
$\overline{T}_a$ $\overline{H}$ $\overline{W}$ P	9	1	Elastic net	398.65	48.5%
	8	1	Elastic net	399.57	48.8%
	7	1	Elastic net	399.65	49.1%
$\overline{T}_a$ $\overline{H}$ $\overline{W}$	9	1	SVR with linear kernel	399.86	49.3%
	7	1	SVR with linear kernel	400.87	50.3%
	6	1	SVR with linear kernel	403.40	50.7%
	3	1	SVR with gaussian kernel	407.25	90.7%
$\overline{T}_a$ $\overline{H}$	4	1	Elastic net	400.65	49.6%
	1	1	SVR with linear kernel	412.26	51.8%
	2	1	SVR with linear kernel	409.18	51.2%
$\overline{T}_a$ $\overline{H}$ P	2	1	SVR with linear kernel	410.71	51.2%
All	7	1	SVR with linear kernel	426.19	51.9%
	5	1	SVR with linear kernel	428.66	52.3%
	5	1	Elastic net	433.03	52.4%
	5	1	Linear regression	450.48	53.0%

the experiment. Se muestra el $RMSE$ para la predicción de 1, 2 ó 3 semanas adelante con respecto a la información de las variables climatológicas y del preaviso biológico observadas. Se aprecia que to predict one week ahead obtiene un $RMSE$ más bajo que two weeks ahead, and this is better than three weeks ahead, lo cual es lo esperado pues entre más semanas adelante se desee predecir, hay más incertidumbre.

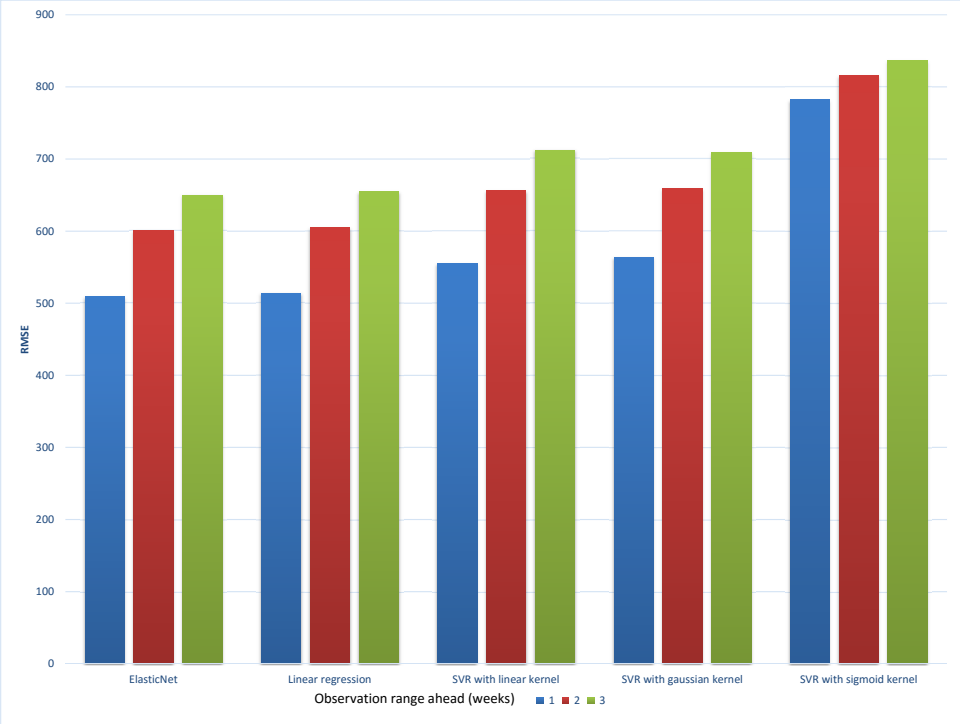

Figure 4: $RMSE$ for each algorithm

Figure 5 shows the the mean $RMSE$ for each variables combination. The better results are obtained with $\overline{T}_a$, $\overline{W}$ and P. Incluir más variables en el modelo no asegura mejorar la métrica, lo cual es importante pues este resultado sugiere que the use of more sensors do not assure a better result of prediction.

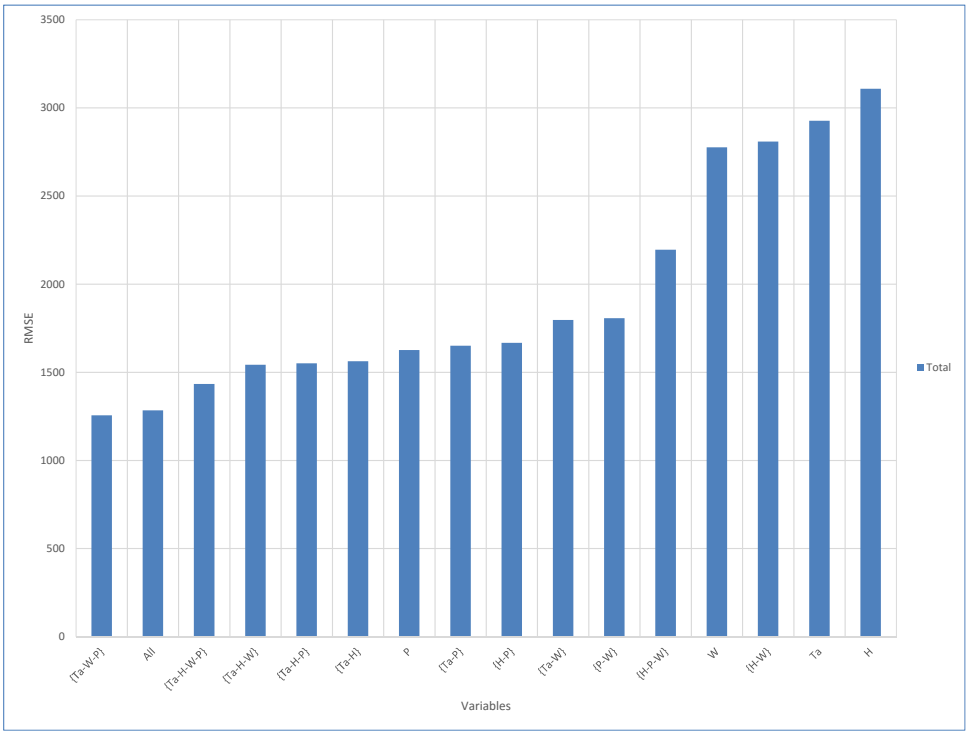

Figure 5: $RMSE$ for each variable combination

6. Conclusions

Este estudio presentó la comparación de varias técnicas de machine learning en la predicción de the la tasa de desarrollo de la Sigatoka negra. Al respecto se presentan las conclusionez al respecto:

-
-
-
-

A comparison of machine learning techniques The main findings were: 1) The highest R2 was 60% 2) The highest R2 were reached with linear models like support vector regression with linear kernel, 3) As little as three meteorological variables can be used because of the correlations detected among variables.

6.1. Future work

Currently this research is focused on the extension of these findings to other farms and the extension of the model to be applied in an early warning system, which does not require the regression task.

Acknowledgements

The authors would like to thank Corbana S. A. for providing the data for this research.

References

ALGLIB®, 2017. Spline interpolation. Online; accessed 28-April-2017. URL `http://www.alglib.net/interpolation/spline3.php`

Alonso, J., Rodríguez Castañón, Á., Bahamonde, A., 2013. Support vector regression to predict carcass weight in beef cattle in advance of the slaughter. Computers and Electronics in Agriculture, 116–120.

Alves, M., de Carvalho, L., Pozza, E., Sanches, L., Maia, J., 2011. Ecological zoning of soybean rust, coffee rust and banana black Sigatoka based on Brazilian climate changes. Procedia Environmental Sciences 6, 35–49.

Bendini, H., Moraes, W., da S. Silva, Tezuka, E., Cruvinel, P., 2013. Análise de risco da ocorrência de sigatoka-negra baseada em modelos polinomiais: um estudo de caso. Tropical Plant Pathology 38 (1), 35–43.

Camargo, A., Molina, J., Cadena-Torres, J., Jiménez, N., Kim, J., 2012. Intelligent systems for the assessment of crop disorders. Computers and Electronics in Agriculture 85, 1–7.

Chuang, T., Jeger, M., 1987. Predicting the rate of development of black Sigatoka (*Mycosphaerella fijiensis var. difformis*) disease in southern Taiwan. Phytopathology 77, 1542–1547.

Coopersmith, E. J., Minsker, B. S., Wenzel, C. E., Gilmore, B. J., 2014. Machine learning assessments of soil drying for agricultural planning. Computers and Electronics in Agriculture 104, 93–104.

Demir, B., Bruzzone, L., 2014. A multiple criteria active learning method for support vector regression. Pattern Recognition, 2558–2567.

Fouré, E., 1988. Stratégies de lutte contre la cercosporioses noire des bananiers et plan-tains provoquée par *Mycosphaerella fijiensis Morelet*. Lávertissement biologique au Cameroun. Evaluation des possibilités d'amélioration. Fruits 43 (5), 269–274.

Ganry, J., Laville, E., 1972. La lutte controlée contre le Cercospora aux antilles. Bases climatiques de lávertissement. Fruits 27, 665–676.

Ganry, J., Meyer, J., 1983. Les cerco-sporioses du bananier et leurs traitements. Evolution des méthods de traitement. Fruits 38, 3–20.

Glezakos, T., Moschopoulou, G., Tsiligiridis, T., Kintzios, S., Yialouris, C., 2010. Plant virus identification based on neural networks with evolutionary preprocessing. Computers and Electronics in Agriculture 70, 263–275.

Huang, Y., Lan, Y., Thomson, S., Fang, A., Hoffmann, W., Lacey, R., 2010. Development of soft computing and applications in agricultural and biological engineering. Computers and Electronics in Agriculture 71 (2), 107–127.

Ibrahim, N., Wibowo, A., 2014. Time series support vector regression with missing data treatment based variables selection for water level prediction of Galas River in Kelantan Malaysia. International Journal of Applied Research in Engineering and Science 3, 25–36.

Kim, Y., Yoo, S., Gu, Y., Lim, J., Han, D., Baik, S., 2014. Crop pests prediction method using regression and machine learning technology: Survey. IERI Procedia 6, 52–56.

Lukoševičius, M., 2012. A practical guide to applying echo state networks. Neural Networks: Tricks of the Trade, 1–20.

Lukoševičius, M., Jaeger, H., 2009. Reservoir computing approaches to recurrent neural network training. Computer Science Review 3, 127–149.

Marín, D., Romero, R., Guzmán, M., Sutton, T., 2003. Black Sigatoka: An increasing threat to banana cultivation. Plant Disease 87 (3), 208–222.

Marín Vargas, D., Romero Calderón, R., 1995. El combate de la Sigatoka Negra. In: Boletín Departamento de Investigaciones. Corbana, Costa Rica, pp. 1–23.

McKinney, W., 2010. Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference. pp. 51–56.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12, 2825–2830.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., 2007. Numerical Recipes. The Art of Scientific Computing, tercera edición Edition. Cambridge University Press.

Romero Calderón, R., 1995. Dynamics of fungicide resistant populations of *Mycosphaerella fijiensis* and epidemiology of black Sigatoka of banana. Ph.D. thesis, Department of Plant Pathology, North Carolina State University.

Rumpf, T., Mahlein, A.-K., Steiner, U., Oerke, E.-C., Dehne, H.-W., Plümer, L., 2010. Early detection and classification of plant diseases with support vector

machines based on hyperspectral reflectance. Computers and Electronics in Agriculture 74 (1), 91–99.

Smola, A. J., Schölkopf, B., 2004. A tutorial on support vector regression. Statistics and Computing 14, 199–222.

Soares, J., Pasqual, M., Lacerda, W., 2013. Utilization of artificial neural networks in the prediction of the bunches' weight in banana plants. Scientia Horticulturae 155, 24–29.

Soares, J., Pasqual, M., Lacerda, W., Silva, S., Donato, S., 2014. Comparison of techniques used in the prediction of yield in banana plants. Scientia Horticulturae 167, 84–90.

Ternesien, E., 1985. La cercorporioses des bananiers et plantains. Methodes de lutte-Avertissements. Perspectives au Cameroun. Memoire de fin d'etudes.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B 58 (1), 267–288.

University, T. P. S., 2017. The coefficient of determination r2. Online; accessed 28-April-2017.
URL `https://onlinecourses.science.psu.edu/stat501/node/29/`

van der Walt, S., Colbert, C., Varoquaux, G., 2011. The NumPy Array: A structure for efficient numerical computation. Computing in Science & Engineering 13, 22–30.

Wei, Z., Tao, T., ZhuoShu, D., Zio, E., 2013. A dynamic particle filter-support vector regression method for reliability prediction. Reliability Engineering & System Safety, 109–116.

Witten, I. H., Eibe, F., Hall., M. A., 2011. Data Mining, 3rd Edition. CMorgan Kaufmann Publishers, United States.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B 67, 301–320.