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 in many parts of Africa and Asia [5].

La bibliografia debe ser autor, año

This disease attacks the plant leaves producing a rapid deterioration of the leaf area, affects the growth and productivity of the plants due to the impairment of their photosinthetic ability, causes a reduction in the quality of the fruit, and promotes premature maturation of bunches, which is the major cause of product losses associated with the black Sigatoka. Figure 1 shows three stages of this disease.

maturation or ripening?

Phytopathological studies point out that precipitation, temperature, relative humidity and wind are the main climatic variables that affect its development

Email address: lualcava.sa@gmail.com (Luis-Alexander Calvo-Valverde)

[1] Corresponding author. (506)70104420

Preprint submitted to Journal of LaTeX Templates *April 2, 2016*

[5].

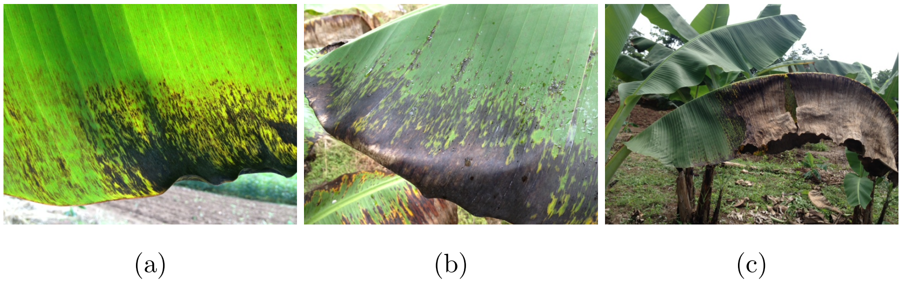
(a) (b) (c)

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

In Costa Rica the control of black Sigatoka makes use of 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 fruit quality expected in the international markets. This represent a cost per hectare per year in the range between US$1600 and US$2000; about 0.64–0.80 cents of the production costs for a 18.14 kg box. Overall, this represents 10%–12% of the total production cost.

The past and present rates of disease development can in principle be used to predict its future behavior and to determine whether a particular fungicide spray program will be able to effectively and economically control the disease Chuang and Jeger [2].

There are efforts to apply machine learning methods for decision-making in agriculture, including the control of crop diseases. For example, [Camargo et al.,2012] present an intelligent system for the assessment of crop disorders, [3] introduce a plant virus identification method based on neural networks with an evolutionary preprocessing stage, [4] summarize in their survey crop pests prediction methods using regression and machine learning approaches, while [7] present an intelligent agricultural forecasting system based on wireless sensor networks.

In this work, we compare five machine learning techniques (support vector regression (SVR), echo state networks (ESN), ridge regression, elastic-net regression and ordinary least squares linear regression) to predict the development rate of the black Sigatoka disease.

The main contribution of this work is a comparison between machine learning methods to forecast black Sigatoka development rate.

FALTA COMPLETAR esta parte

2. Materials and methods

2.1. Concepts

2.1.1. Black Sigatoka disease

Black Sigatoka, disease caused by the fungus Mycosphaerella fijiensis Morelet, is the main problem phytopathologic of banana and plantain crops in Central America [5].

This disease attacks the leaves of plants producing a rapid deterioration of the leaf area. It affects the growth and productivity of plants by decreasing photosynthetic capacity. Also causes a reduction in quality of the fruit [5].

The climate has a major effect on the behavior of the black Sigatoka. Precipitation, temperature, relative humidity and wind are the main climatic variables affecting the development of this disease [5].

2.1.2. Biological warning system

The early warning system for black Sigatoka is an adaptation of the yellow Sigatoka warning system developed by Ganry and Meyer and modified by Ganry and Laville to use for controlling yellow Sigatoka in Cameroon. Ternesien and Fouré later improved Ganry and Laville's system. The latter system is based on weekly observations of disease symtoms on young leaves of the plant, according to Fourés symptom (stages) descriptions. Arbitrary coefficientes, based on incidence and severity of disease development, are used to calculate two variables: gross sum and state of evolution. Gross sum is based on the stage present and an arbitrary coefficient, which increases with the advance of the symptoms and the juvenility of the leaf. The state of evolution is calculated using the gross sum and the foliar emission period. Although threshold levels were initially suggested as a guide to spray timing, the fluctuation of these two variables was found to better define appropiate times to spray [6].

2.1.3. Support Vector Regression (SVR)

From the perspective of Support Vector Regression (SVR) the regression function $y = f(s)$ for a given dataset $D = \{(s_i, y_i)\}_{i=1}^{n}$, is represented as a linear function of the form [8]:

$$f(s) = w^T s + b$$

where w and b are respectively the weight vector and the intercept of the model, and they are selected to find an optimal fit to the data available in D.

For nonlinear cases, one proceeds by mapping the input p-dimensional vectors via a nonlinear function $\phi : R^p \rightarrow F$, onto the feature space F. After nonlinear mapping, the regression function evolves to a pervasive form:

$$f(s) = w^T \phi(s) + b$$

SVR uses the ϵ-insensitive loss function:

$$l = |y - f(s)|_\epsilon = \begin{cases} 0 & |y - f(s)| \leq \epsilon \\ |y - f(s)| - \epsilon & \text{otherwise} \end{cases}$$

which ignores the error if the difference between the prediction value and the actual value is smaller than ϵ. The ϵ-insensitive loss function allows to find the coefficients w and b by solving a convex optimization problem, which balances the empirical error and the generalization ability. In SVR, the empirical error is measured by the loss function ϵ-insensitive and the generalization ability is measured by the Euclidean norm of w [9]. Then, the optimization problem to identify the regression model can be formulated by [8]:

$$
\begin{aligned}
\text{minimize} \quad & J(w,\xi_i,\xi_i^*) = \frac{1}{2}\Big|\Big|w\Big|\Big|^2 + C\sum_{i=1}^{n}(\xi_i,\xi_i^*) \\
& y_i - w^T\phi(s) - b \leq \epsilon + \xi_i \\
\text{subject to} \quad & w^T\phi(s) + b - y_i \leq \epsilon + \xi_i^* \quad i = 1,2,...,n \\
& \xi_i, \xi_i^* \geq 0
\end{aligned}
\tag{1}
$$

where C denotes the penalty parameter between empirical and generalization errors, and ξ_i, ξ_i^* are slack variables. Figure.2 shows this situation.

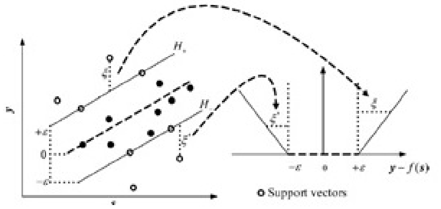

Figure 2: ϵ-insensitive loss function [8].

The solution of this optimization problem by the Lagrange method is given by:

$$f(s) = w^T\phi(s) + b = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(s,s_i) + b$$

where $\alpha_i - \alpha_i^*$ are the Lagrange multipliers of the optimization problem's dual form and $K(s_i, s_j)$ is the kernel function satisfying the Mercer condition, and holds:

$$K(s_i, s_j) = \big\langle \phi(s_i), \phi(s_j) \big\rangle$$

Operations in the kernel function $K(s, s_i)$ are performed in the input space rather than in the potentially high dimensional feature space of ϕ [10].

2.1.4. Ordinary least squares regression

This method fits a linear model with coefficients $w = (w1, .., wp)$ to minimize the residual sum of squares between the observed responses in the dataset, and the responses predicted by the linear approximation. Mathematically it solves a problem of the form [**?**]:

$$\min_{w}\left|\left|Xw - y\right|\right|_2^2$$

where X denotes the features matriz.

According Pedregosa et al. [11] the coefficient estimates for Ordinary Least Squares rely on the independence of the model terms. When terms are correlated and the columns of the design matrix X have an approximate linear dependence, the design matrix becomes close to singular and as a result, the least-squares estimate becomes highly sensitive to random errors in the observed response, producing a large variance. This situation of multicollinearity can arise, for example, when data are collected without an experimental design

2.1.5. Ridge regression

The ridge regression addresses some of the problems of ordinary least squares regression by imposing a penalty on the size of the coefficients. The ridge coefficients minimize a penalized residual sum of squares [11]:

$$\min_{w}\left|\left|Xw - y\right|\right|_2^2 + \alpha\left|\left|w\right|\right|_2^2$$

Here, $\alpha > 0$ is a complexity parameter that controls the amount of shrinkage: the larger the value of α, the greater the amount of shrinkage and thus the coefficients become more robust to collinearity.

2.1.6. Elastic-Net regression

Elastic-Net is a linear regression model trained with $L1$ and $L2$ prior as regularizer. This combination allows for learning a sparse model where few of the weights are non-zero like Lasso, while still maintaining the regularization properties of Ridge [11]. The convex combination of $L1$ and $L2$ is controled by using the $l1_ratio$ parameter.

Elastic-Net is useful when there are multiple features which are correlated with one another. Lasso is likely to pick one of these at random, while elastic-net is likely to pick both. A practical advantage of trading-off between Lasso and

Ridge is it allows Elastic-Net to inherit some of Ridge's stability under rotation. The objective function to minimize is [11]:

$$\min_{w} \frac{1}{2n_{samples}} \left|\left|Xw - y\right|\right|_2^2 + \alpha\rho \left|\left|w\right|\right|_1 + \frac{\alpha(1-\rho)}{2} \left|\left|w\right|\right|_2^2$$

2.1.7. Echo State Networks (ESN)

Recurrent Neural Networks (RNN) are useful for temporal patterns, but when they are trained with backpropagation methods, they are very slow. Echo State Network (ESN) is an alternative training method to solve that problem. ESN is based on the observation that if a random RNN possesses certain algebraic properties, training only a linear readout from it is often sufficient to achieve excellent performance in practical applications [12]. For a given training input signal $u(n) \in R^{N_u}$ a desired target output signal $y^{target}(n) \in R^{N_y}$ is known. Here $n = 1, ..., T$ is the discrete time and T is the number of data points in the training dataset. The task is to learn a model with output $y(n) \in R^{N_y}$, where $y(n)$ matches $y^target(n)$ as well as possible, minimizing an error measure $E(y, y^target)$, and, more importantly, generalizes well to unseen data. The untrained RNN part of an ESN is called a dynamical reservoir, and the resulting states x(n) are termed echoes of its input history [13]. Finally, these signals are sent to an output layer as shown in the Figure.3.

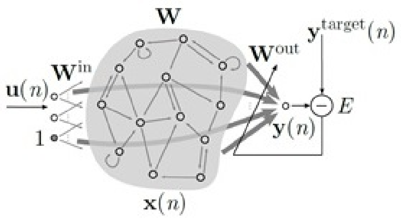

Figure 3: An echo state network [13].

The connections between the different elements of an Echo State Network have weights randomly generated. The weights of the internal connections of the reservoir (W) as well as the weights of the input layer (W_in), after being generated are set statically during all stages of implementation of the algorithm.

The weights between the reservoir and the output layer (W_out) are subject to changes of a supervised learning algorithm to correct the degree of error generated by the entire system [13].

2.1.8. Related works

Huang et al. [3] surveyed the development of soft computing techniques in agricultural and biological engineering, including fuzzy logic, artificial neural networks, genetic algorithms, bayesian inference and decision trees.

A related work, proposed by Romero [14] relies on regression models using a stepwise procedure to predict incubation and latency times of black Sigatoka. The author performed experiments on two farms located in Costa Rica (La Rita and Waldeck, the same as those used in this study but with different names). The study used data from: December 1993 to August 1995. Romero concluded that the model to predict the incubation period accounted a R^2 of 69% in his observed data but it was not a good predictor when it was validated against an independent dataset (cross validation). For latency, he developed two models that accounted a R^2 of 78% in the observed data, however, when validated against an independent dataset (cross validation), for La Rita obtained an adjusted R^2 of 82%, and for Weldeck, none of the models satisfactorily predicted the latent period; and then those predictions were not shown [14].

Glezakos et al. [15] proposed to use Genetic Algorithms (GA) and Neural Networks (NN) to identify plant virus (Tobacco Rattle Virus (TRV) and the Cucumber Green Mottle Mosaic Virus (CGMMV)). This is achieved by the development of ana- lytical tools of evolutionary adaptive width, propelled by Genetic Algorithms (GAs) and Neural Networks (NNs). The method was tested against some of the most commonly used classifiers in machine learning (Bayes, Trees and k-NN) via cross-validation and proved its potential towards the identification.

In the agricultural context, Alves et al. [16] used geoinformation techniques to develop predictive models to study the areas of risk to soybean rust in soybean, coffee leaf rust in coffee, and black Sigatoka in banana, considering Brazil's climatic characterization and the distribution of soybean, coffee and banana crops. Temperature and rainfall data were obtained for the period from 1950 to 2000, and of simulations for 2020, 2050 and 2080 using the SRES A2 climate change scenarios. Using principal components analysis, a single variable was generated based on 57 variables, in order to determine an index explaining 87%, 88% and 90% of the 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. Areas of favorability for the diseases were plotted against the main coffee, soybean and banana growing areas in Brazil. This methodology enabled the visualization of the changes in areas favorable for epidemics under possible future scenarios of climate change.

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 [10], machine learning assessments of soil drying for agricultural planning [17], and early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance [18].

Furthermore, there have been attempts to generate software tools. Camargo et al. [Camargo et al.,2012] presented an information system for the assessment of plant disorders (Isacrodi). They proposed that experts will attain a much better accuracy than the Isacrodi classifier, particularly when provided with samples from the affected crop. However, those cases where such expertise is not available, they suggest that Isacrodi can provide valuable support to farmers. Isacordi includes 15 crop disorders, but the black Sigatoka no is one of them. The prediction process is based on multi-class support vector machines.

Regarding the prediction of the develpment of the black Sigatoka with machine learning methods, Bendini et al. [19] presented a study about the risk analysis of black Sigatoka occurrence based on polynomial models. A case study was developed in a commercial banana plantation located in Jacupiranga, Brazil. It was monitored weekly during the period from February to December 2005. 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%.

Also there are works related to banana fruit. Soares et al. [20] 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 proved to be more accurate in forecasting the weight of the bunch in comparison to the multiple linear regressions in terms of the mean prediction-error ($MPE = 1.40$), mean square deviation ($MSD = 2.29$) and coefficient of determination ($R^2 = 91\%$).

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.

2.1.9. Data

In this work we use data acquired in two research farms of Corbana in Costa Rica: 1) 28 Millas (previously called Waldeck and located at Matina) and La Rita (located at Pococí), both in the province of Limón, Costa Rica. The banana type is Musa AAA, subgroup Cavendish, cv. Grande Naine. The Table.1 shows the variables available.

The value to be predicted in all cases was E_s, that is the total measure of the biological warning system.

The data on the biological warning system are collected once a week. Although Corbana has meteorological stations that take data every five minutes, for these experiments, weekly averages generated by nearby stations to each of the farms were used.

The time intervals used for this study were: La Rita, week 48 of 2002 to week 17 of the 2015 (647 weeks) and for 28 Miles, week 37 of 2003 to week 18 of 2015 (605 weeks).

Table 1: Variables used in the study

Variable	**Meaning0'0'**
$T_{a_{max}}$	Max air temperature
$T_{a_{min}}$	Min air temperature
$\overline{T}_a$	Mean air temperature
$\overline{H}$	Mean Humidity
H_{min}	Min humidity
H_{max}	Max humidity
$\overline{R}$	Mean Solar radiation
P	Sum precipitation
W_{max}	Max speed wind
$\overline{W}$	Mean speed wind
L_2	Biological warning system – Leaf 2
L_3	Biological warning system – Leaf 3
L_4	Biological warning system – Leaf 4
E_s	Biological warning system – Evolution Stage

2.1.10. Data preprocessing

In 28 Miles farm, 1% of the data were missing, while in La Rita was 2.25%. To fill-in the missing values we use spline interpolation. The data collected did not exhibit outliers.

Due the fact that the variables measure meteorological or biological process, they are discretized in order to reflect trends in the data, i.e. the continuous values are not directly used. The coefficient of variation $C_v(x)$ of each variable x was used to determine the number n of discretization levels.

$$n = \lfloor 100\ C_v(x) \rfloor$$

where $\lfloor\ \rfloor$ is the round operator.

Each discretization range was uniformly partitioned. Besides enabling the capture of tendencies, the discretization removes the effect of small variations in the data collection, either by inaccuracies of the instruments (meteorological variables) or by subjective bias introduced by the human who collects the data (biological warning system).

Each feature was scaled to fit in a range between 0 and 1. The variable to be predicted was not scaled.

2.2. Evaluation criteria

Although there are many types of indicators to assess the quality of the prediction, we selected the determination coefficient (R^2) and the Root Mean Square Error ($RMSE$).

Given n records, let be y the actual value of the series, $\hat{y}$ the predicted value and $\acute{y}$ the mean of the observed data.

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

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

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

$$S_y^2 = S_R^2 + S_e^2$$

$$R^2 = \frac{S_R^2}{S_y^2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{n}}$$

This decision is supported by the widespread use in machine learning and agriculture areas [20], [21], [22] and [23].

2.3. Methodology

The selection of methods and their parametrisation was performed in two stages.

Phase one

In the phase one, we did ten-fold-cross-validation on a set of machine learning methods and different configurations:

- Patterns: n by m, where n from 1 to 8 and m from 1 to 2.
- Methods: support vector regression with the kernels functions: linear, gaussian and sigmoid; echo state networks; ordinary least squares linear regression, ridge regression and elastic-net regression.
- Variables included in the model:
  - All variables.
  - From the set $\{\overline{T}_a, \overline{H}, P, \overline{W}\}$ use the subsets with one, two or four elements. These variables are according to experts the ones having most impact on the disease development [5].

Phase two

In the second phase, the best configurations obtained in phase one are used to validate with the last 50 and 100 weeks.

This second phase intents to expose how these methods behave on a considerable climate in the years 2014 and 2015.

Poner una cita que fundamente este ultimo parrafo

2.4. Programming environment

We use the python programming language with the Integrated Development Environment (IDE) Spyder [24], particularly with the libraries pandas [25] and numpy [26]. For SVR, ridge and ordinary least squares regressions, we used sklearn [11] and for ESN the python-based code of Dr. *Lukoševičius* [13] on which the necessary were done for adjustments for the experiments of this work. The computer used a processor Intel(R) Core i7-4800MQ CPU 2.70GHz, 16.0 GB RAM, running Windows 8 Pro.

3. Results and discussion

In this section we present the main results for phase.

Phase one

Figure.4 shows the best R^2 for each algorithm in the experiment. Results are group by farm. Though La Rita obtains different results in magnitude than 28 Millas, the trend is similar. In both farms, the best results are for linear models, second position is occupied for Echo State Networks and SVR with gausian and sigmoid kernels are the worst results. In linear models, to predict one week ahead is better than two weeks ahead, and this is better than three weeks ahead.

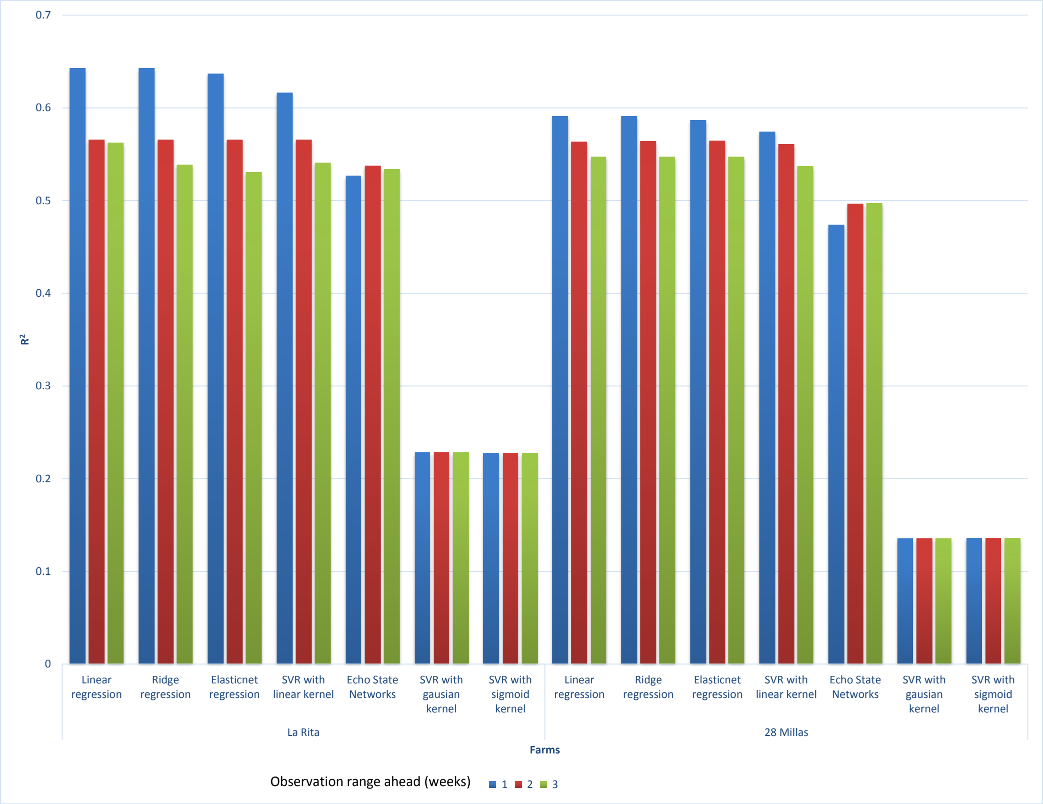

Figure 4: Phase one - Best R^2 for each algorithm

Figure.5 presents, for one, two and three weeks ahead, the best R^2. Results are group by farm. In general, to predict one week ahead is better than two weeks ahead and so on. The number of weeks consider in the observation range in the pattern is not the main discriminant factor, but it is clear that we get better R^2 for one week ahead than two weeks ahead and so on.

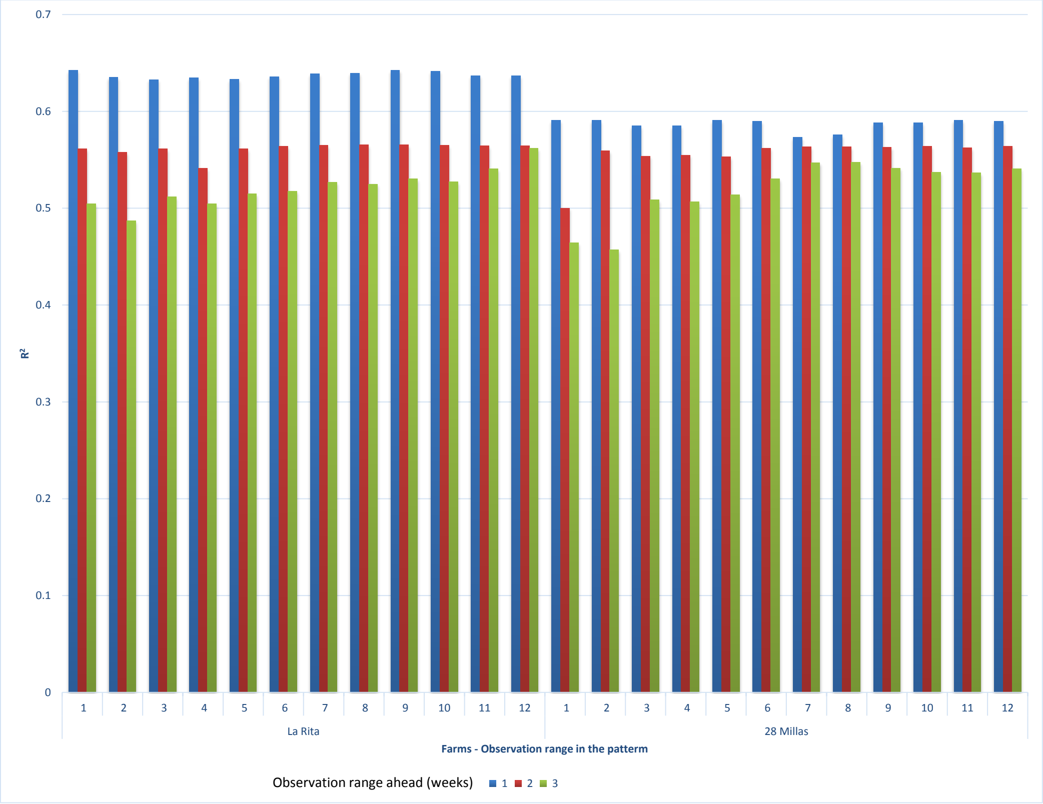

Figure 5: Phase one - Best R^2 for each observation range

Figure.6 shows the best R^2 for each variables combination. Results are group by farm. The better results are obtained with $\overline{T}_a$ and the combination of $\overline{T}_a$ with $\overline{W}$, in both farms of similarly. You can note that the use of all variables in the model or the inclusion of the four variables suggest for expert criteria do not improve significantly the results, then the use of more sensors do not assure a better result.

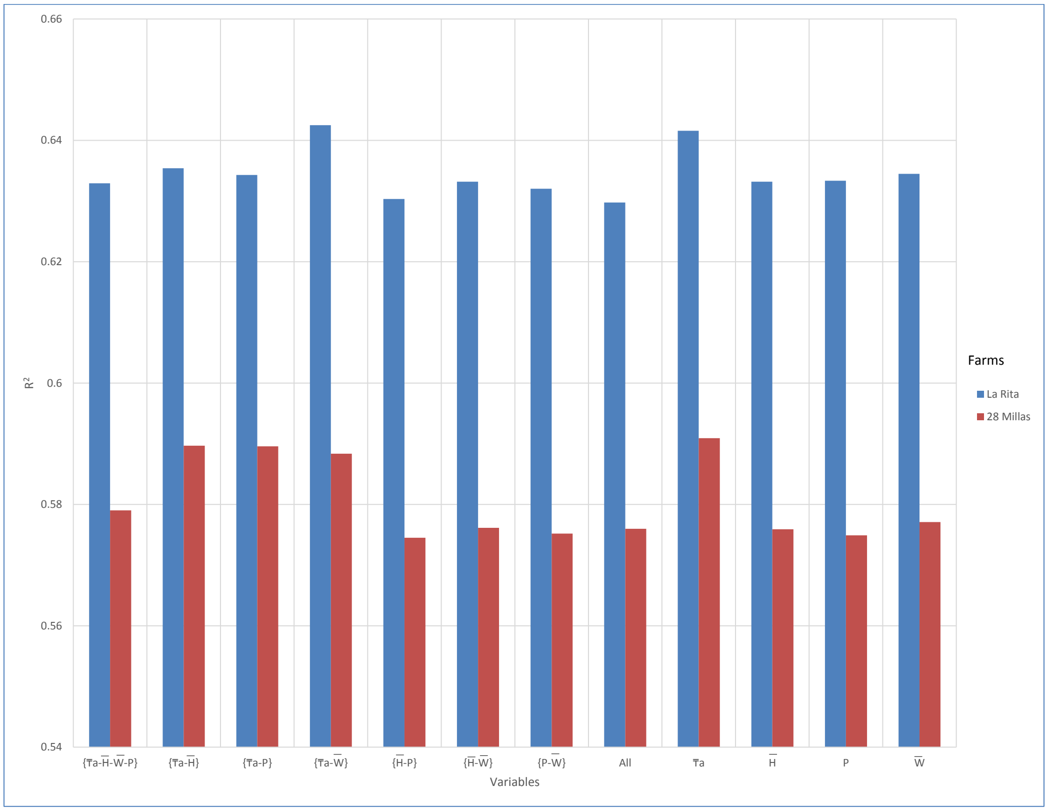

Figure 6: Phase one - Best R^2 for each variable combination

Figure.7 shows the Pareto frontier for each farm with respect to R^2 and $RMSE$. The 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. $R2$ is a relative metric between 0 thru 1 and it is less sensitive to absolute values.

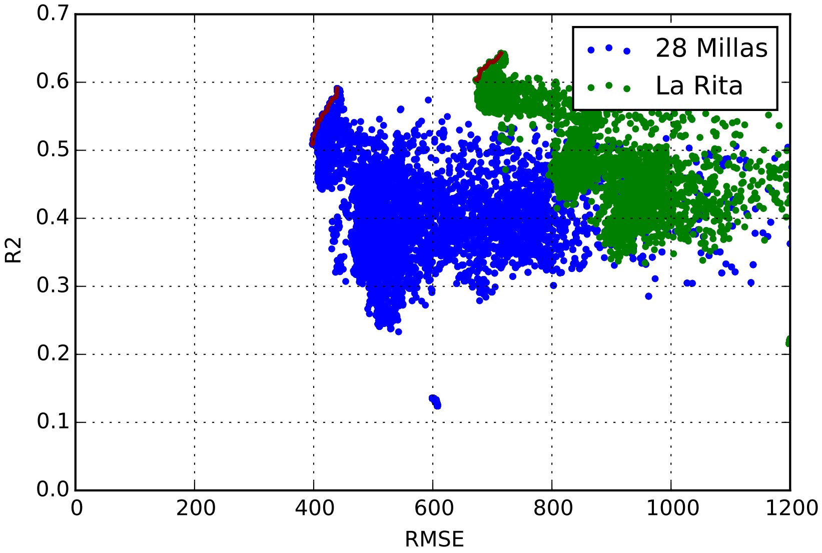

Figure 7: Phase one - Pareto frontier for R^2 and $RMSE$

The Pareto frontier for the La Rita farm is composed by 96 elements. The Table.2 shows the composition about variables and observation ranges.

Table 2: Composition of the Pareto frontier - La Rita - Phase one

Variable	**Observation range**	**Quantity**	**Max R^2**	**Min $RMSE$**
Pair $\overline{T}_a$ $\overline{W}$	1 to 1	36	64.25%	714.51
	2 to 1	6	62.97%	695.10
All	1 to 1	18	62.98%	701.95
	2 to 1	12	61.76%	679.92
	3 to 1	6	60.60%	676.42
	5 to 1	2	60.37%	672.39
$\overline{T}_a$	1 to 1	12	63.60%	708.77
	2 to 1	4	62.23%	689.55

Similarly, the Pareto frontier for the 28 Millas farm is composed by 75 elements. The Table.3 shows the composition about variables and observation ranges.

Table 3: Composition of the Pareto frontier - 28 Millas - Phase one

Variable	**Observation range**	**Quantity**	**Max R^2**	**Min $RMSE$**
Pair $\overline{T}_a$ $\overline{W}$	1 to 1	8	57.80%	438.09
All	9 to 1	2	50.93%	397.93
	10 to 1	2	50.97%	398.81
	8 to 1	6	51.62%	398.93
	7 to 1	2	52.25%	400.28
	6 to 1	2	53.16%	404.14
	4 to 1	2	54.32%	407.54
$\overline{T}_a$	1 to 1	8	59.09%	439.44
Pair $\overline{T}_a$ $\overline{H}$	1 to 1	8	57.51%	428.61
	2 to 1	20	56.91%	414.37
	3 to 1	3	54.41%	411.55
	4 to 1	3	53.34%	406.65
Pair $\overline{T}_a$ P	3 to 1	9	56.23%	422.76

We can conclude that the best configuration in both farms is to consider the climate and the evolution stage of the current week to predict the evolution stage of the next week.

Phase two

In the second phase, the best configurations obtained in phase one were used to validate with the last 50 and 100 weeks.

Figure.8 shows the best R^2 for each algorithm in the experiment. Results are group by farm. Even if linear models continue with good R^2, Echo State Networks improve their scores because in 50 and 100 last weeks validation, we are in presence of climate change, then the behaviour is less lineal.

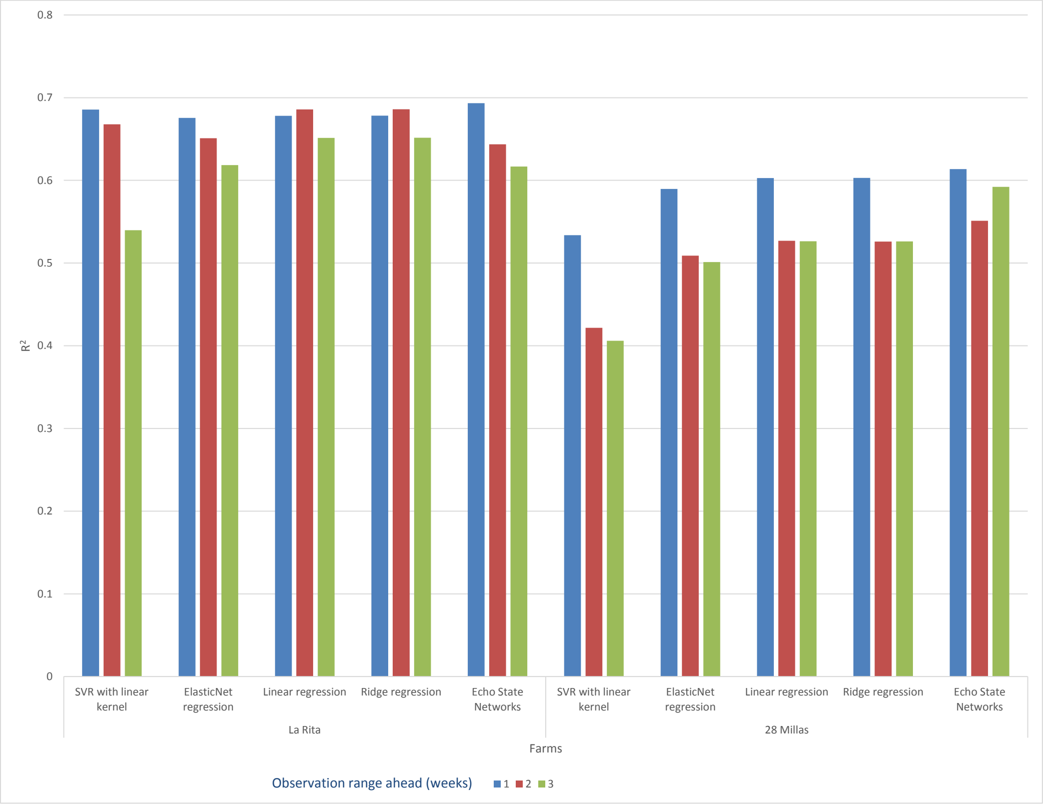

Figure 8: Phase two - Best R^2 for each algorithm

Figure.9 presents, for one, two and three weeks ahead, the best R^2. Results are group by farm. The results confirms that, in general, to predict one week ahead is better than two weeks ahead, two than three and so on, this for both farms.

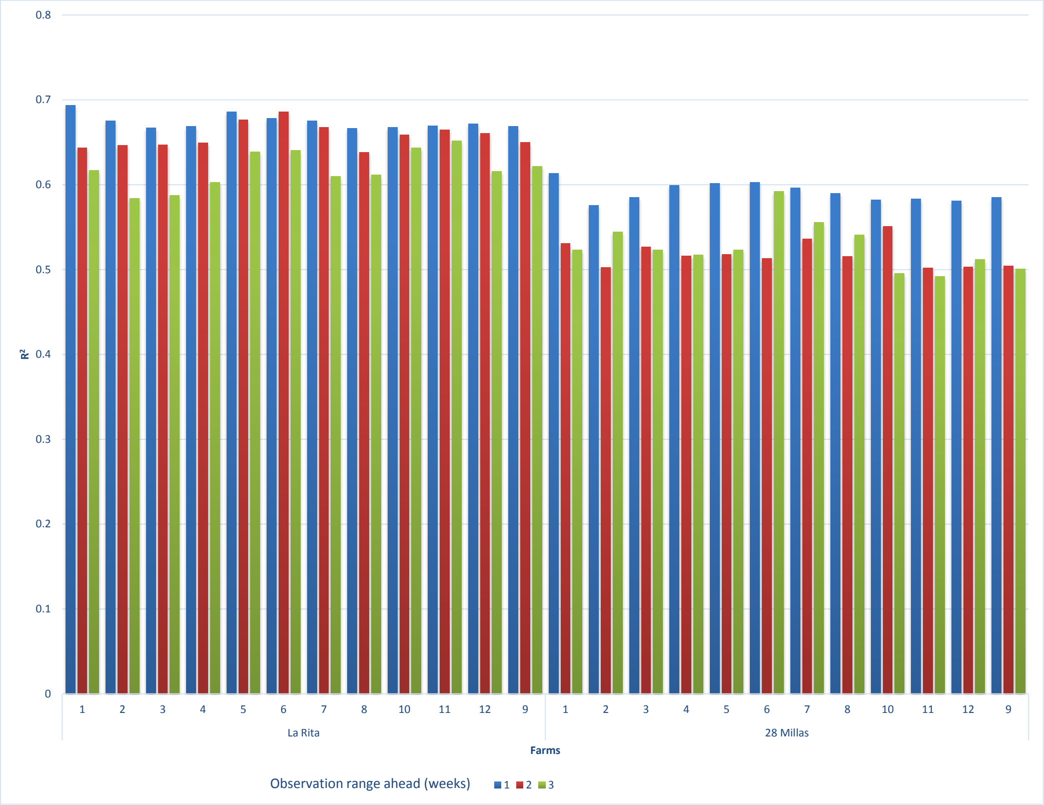

Figure 9: Phase two - Best R^2 for each observation range

Figure.10 shows the best R^2 for each variables combination. Results are group by farm. This results confirm that $\overline{T}_a$ and the combination of $\overline{T}_a$ with $\overline{W}$, in both farms are the best variables combinations.

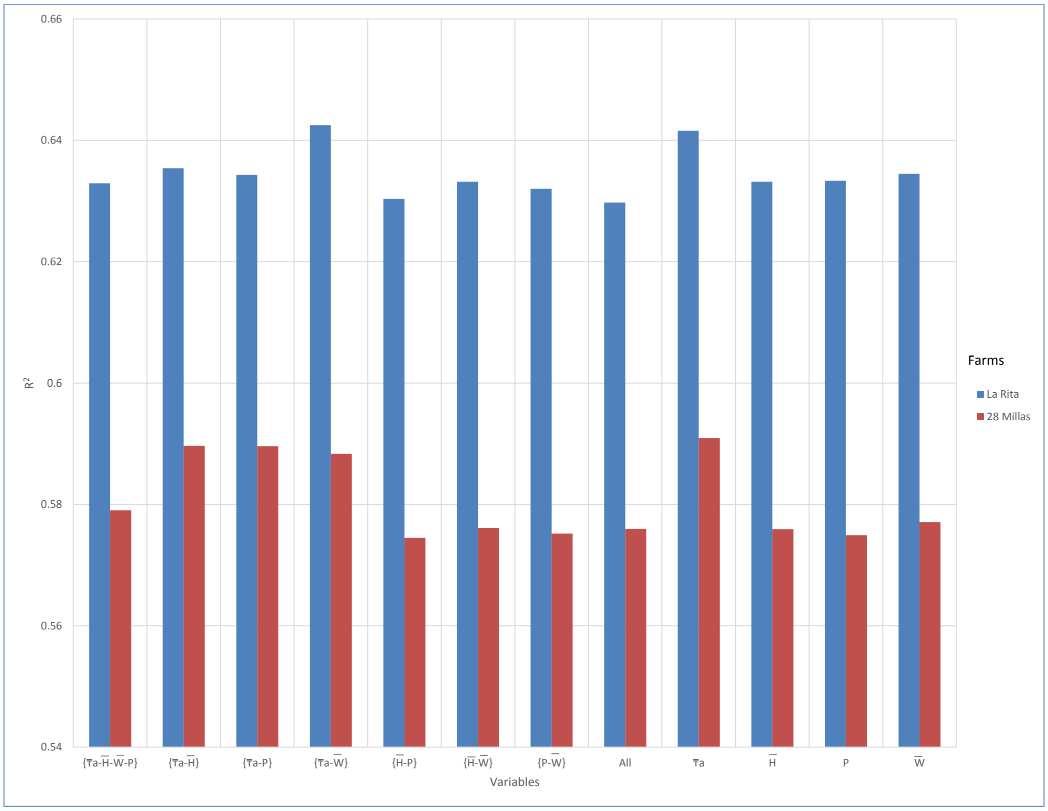

Figure 10: Phase two - Best R^2 for each variable combination

Figure.11 shows the Pareto frontier for each farm with respect to R^2 and $RMSE$. You can note that the behaviour of R^2 and $RMSE$ is similar to the phase one.

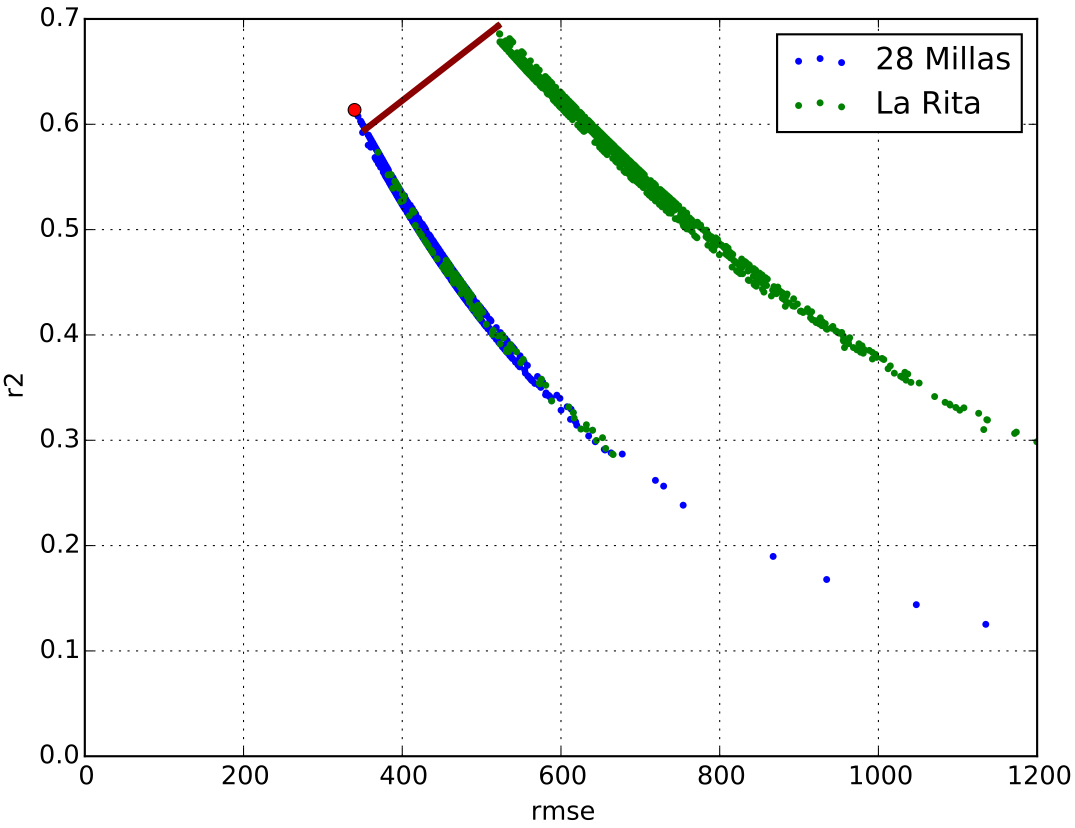

Figure 11: Phase two - Pareto frontier for R^2 and $RMSE$

In phase two, the Pareto frontier for the La Rita farm is composed by 2 elements. The Table.4 shows the composition about variables and observation ranges.

Table 4: Composition of the Pareto frontier - La Rita - Phase two

Variable	**Observation range**	**Quantity**	**Max R^2**	**Min $RMSE$**
All	1 to 1	2	69.33%	353.33

In 28 Millas Farm, the Pareto frontier is composed by 1 element. The Table.5

details the result.

Table 5: Pareto frontier - 28 Millas - Phase two

Variable	**Observation range**	**Quantity**	**Max R^2**	**Min $RMSE$**
Pair $\overline{T}_a$ P	1 to 1	1	61.36%	339.89

Again, similar to phase one, we can conclude that the best configuration in both farms is to consider the climate and the evolution stage of the current week to predict the evolution stage of the next week, one week to predict one week ahead with combinations of variables listed above.

4. Acknowledgements

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

5. References

[Camargo et al.,2012] 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. doi:10.1016/j.compag.2012.02.017.

[2] 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.

[3] 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. doi:10.1016/j.compag.

[4] 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. doi:10.1016/j.ieri.2014.03.009.

[5] Marin Vargas, D., Romero Calderón, R. 1995. El combate de la Sigatoka Negra. Boletín Departamento de Investigaciones, Corbana Costa Rica.

[6] Marin, D., Romero, R., Guzman, M, Sutton, T. 2003. Black Sigatoka: An increasing threat to banana cultivation. Plant Disease, 87(3), 208-222.

[7] Zhao, L., He, L., Harry, W., Jin, X. 2013. Intelligent Agricultural Forecasting System Based on Wireless Sensor. Journal of Networks(8), 1817–1824. doi:10.4304/jnw.8.8.1817-1824.

[8] 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. doi:10.1016/j.ress.2013.05.021.

[9] Cristiani, Nello and Shave-Taylor, John. An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press. Ninth printing, United Kingdom, 2005.

[10] 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.

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

[12] Lukosevicius, M. and Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. Computer Science Review(3), 127–149. doi:10.1016/j.cosrev.2009.03.005.

[13] Lukosevicius, M. (2012). A Practical Guide to Applying Echo State Networks. Neural Networks: Tricks of the Trade. 1-20

[14] Romero, R. (1995). Dynamics of fungicide resistant populations of Mycosphaerella fijiensis and Epidemiology of black Sigatoka of banana. Costa Rica: North Carolina State University.

[15] 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.

[16] 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.

[17] 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. http://doi.org/10.1016/j.compag.2014.04.004

[18] 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. http://doi.org/10.1016/j.compag.2010.06.009

[19] Bendini, H., Moraes, W., da Silva, S., 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, 035-043.

[20] 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 journal, 167, 84-90.

[21] 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.

[22] Ibrahim, N. and 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.

[23] Demir, B. and Bruzzone, L. (2014). A multiple criteria active learning method for support vector regression. Pattern Recognition, 2558–2567. doi:10.1016/j.patcog.2014.02.001

[24] Continuum Analitycs. (2015). Anaconda. Retrieved from https://www.continuum.io/

[25] McKinney W. (2010). Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 51-56

[26] van der Walt S., Colbert C. and Varoquaux G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation, Computing in Science & Engineering, 13, 22-30 (2011), DOI:10.1109/MCSE.2011.37