

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

Luis-Alexander Calvo-Valverde<sup>a,1</sup>, Mauricio Guzmán-Quesada<sup>b</sup>,  
José-Antonio Guzmán-Alvarez<sup>b</sup>, Pablo Alvarado-Moya<sup>c</sup>

<sup>a</sup>DOCINADE, Instituto Tecnológico de Costa Rica, Computer Research Center,  
Multidisciplinar program eScience, Cartago, Costa Rica

<sup>b</sup>Dirección de Investigaciones, Corporación Bananera Nacional S. A., Guápiles, Costa Rica

<sup>c</sup>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)

<sup>1</sup>Corresponding author. (506)70104420

14 control of the yellow Sigatoka in Cameroon, was later adapted by Ternesien  
 15 (1985) and Fouré (1988) for the black Sigatoka.

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

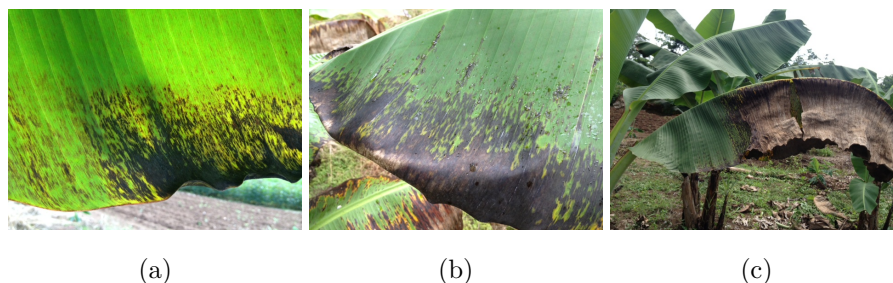


Figure 1: Examples of three disease stages of the black Sigatoka. (a) Initial stage. (b) Intermediate stage, and (c) Advanced stage.

18  
 19 The disease progression is then quantified according to Fouré’s scale of the  
 20 symptom stages (Fouré, 1988) by means of numeric coefficients that describe  
 21 the degree of incidence and the severity of the disease development. These  
 22 coefficients are then used to calculate two variables: gross sum and state of  
 23 evolution.

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

31 In Costa Rica the black Sigatoka is frequently treated with chemical fungi-  
 32 cides. Depending on the zone of production and the weather conditions, 45–  
 33 55 cycles/year of fungicide applications are required to keep this disease under  
 34 control and to produce the expected fruit quality for the international mar-

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

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

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

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

59 The outline of the paper is as follows: Section 2 presents related works and  
60 Section 3 summarizes the machine learning techniques selected for the analy-  
61 sis. In Section 4 we present the methodology used in this study and describe  
62 data used for its verification. The results and their discussion are presented in  
63 section 5. The Section ?? concludes this article and presents lines for future  
64 works.

## 65 2. Related works

66 Several efforts have been made to apply machine learning techniques in the  
67 automated discovery of relationships between environmental variables and quan-  
68 tified descriptors for variables of agricultural interest such as the progress of  
69 diseases. Huang et al. (2010) summarize in their survey the development of soft  
70 computing techniques in agricultural and biological engineering, especially in the  
71 soil and water context for crop management and decision support in precision  
72 agriculture, including fuzzy logic, artificial neural networks, genetic algorithms,  
73 Bayesian inference and decision trees. They do not present numeric results of  
74 each paper, only mention the main idea. Similarly, Kim et al. (2014) survey  
75 more recent prediction methods for crop pests using regression and machine  
76 learning approaches. Nor do they provide numerical results of each paper. In  
77 general, the machine learning methods applied to predict the evolution of plant  
78 diseases, can be classified in two main approaches: 1) those whose main inputs  
79 are images, and 2) Those whose main inputs are environmental and biological  
80 variables. Our study focuses in the second case. Romero Calderón (1995) re-  
81 lied on regression models using a stepwise procedure to predict incubation and  
82 disease latency periods for the black Sigatoka. He collected environmental data  
83 from two different farms in Costa Rica between December 1993 and August  
84 1995. The prediction models reached coefficients of determination  $R^2$  of 69%  
85 or 78% on the observed data for the incubation and disease latency periods,  
86 respectively; however, the cross validation on independent data sets failed. In  
87 contrast, our proposal presents a model that can be generalizable to other farms  
88 that have data.

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

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

96     Alves et al. (2011) used geoinformation techniques to develop predictive  
97 models in the study of risk areas to soybean rust, coffee leaf rust, and banana  
98 black Sigatoka, under consideration of Brazil’s climatic characteristics and the  
99 distribution of soybean, coffee and banana crops. Temperature and rainfall  
100 data were acquired for the period from 1950 to 2000, and simulated data were  
101 generated for 2020, 2050 and 2080 using the SRES A2 climate change scenarios.  
102 Using principal components analysis, a single variable was generated as a linear  
103 combination of 57 input variables, in order to determine an index explaining  
104 87%, 88% and 90% of the data variability of soybean, coffee and banana crops,  
105 respectively, in municipal districts across Brazil. The climatic model was used  
106 to generate the zoning of the three plant diseases, using temperature and leaf  
107 wetness as input. This methodology enabled the visualization of the changes in  
108 areas favorable for epidemics under possible future scenarios of climate change.  
109 How intermediate result, they characterized the monocyclic process of the black  
110 Sigatoka using nonlinear regression. Although they do not present the detailed  
111 results, it no seem that they wanted to predict the progression of the black  
112 Sigatoka in one, two or more periods ahead, how we do.

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

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

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

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

137     There are also works related to the banana fruit. Soares et al. (2014) apply  
138 two techniques: artificial neural networks (ANNs) and multiple linear regression  
139 (MLR) in banana plant to predict the yield. Their results show that the neural  
140 network is more accurate in forecasting the weight of the bunch in comparison  
141 to the multiple linear regressors in terms of the mean prediction-error ( $MPE =$   
142 1.40), mean square deviation ( $MSD = 2.29$ ) and coefficient of determination  
143 ( $R^2 = 91\%$ ).

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

### 147 **3. Compared regression techniques**

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

153 *3.1. Ordinary least squares regression*

Given a data set

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

composed of the  $d$ -dimensional<sup>2</sup> 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

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

with the error function

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

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

156 *3.2. Elastic net regression*

157 Instead of  $L_2$  regularization prior ( $\alpha \|\mathbf{w}\|_2^2$ ) included in the ridge regression  
 158 (RR), Tibshirani (1996) used an  $L_1$  term ( $\lambda \|\mathbf{w}\|_1$ ) for his lasso estimator, which  
 159 permits to select a subset of the available features by zeroing the weights of the  
 160 deselected features. If the dimension  $d$  of the data is larger than the number  $n$   
 161 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

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

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

---

<sup>2</sup>Without loss of generality assume that the first component of every vector  $\mathbf{x}_i$  is always 1.

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

### 168 3.3. Support Vector Regression (SVR)

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

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

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

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

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

173 Note that since OLSR, RR and ENR use a squared error function, data  
 174 outliers will have a strong influence on the resulting weights  $\mathbf{w}$ . On the SVR  
 175 formulation, however, the usage of the  $L_1$  norm and the slack variables consid-  
 176 erably 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)

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

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

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

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

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

189 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):

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

190 where  $\sigma$  is the parameter of gaussian model.

Sigmoid:

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

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

### 192 3.4. Echo State Networks (ESN)

193 Recurrent neural networks (RNN) are capable of learning temporal patterns  
 194 by feeding neuron outputs back into lower layers. Their training (usually by  
 195 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  $n = 1, \dots, T$  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<sup>3</sup> : *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 48th week of 2002 to the 17th week of 2015 (647 weeks); for 28 Miles the data was captured between the

---

<sup>3</sup>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]
$\bar{T}_a$	Mean air temperature	[°C]
$\bar{H}$	Mean relative humidity	[0 – 100]
$H_{min}$	Minimal relative humidity	[0 – 100]
$H_{max}$	Maximal relative humidity	[0 – 100]
$\bar{R}$	Mean solar radiation	[W/m <sup>2</sup> ]
$P$	Precipitation	[mm]
$W_{max}$	Maximal wind speed	[m/s]
$\bar{W}$	Mean speed wind	[m/s]
$E_s$	Biological warning system – Evolution Stage	> 0

Table 1: Variables available for the learning algorithms

37th week of 2003 and the 18th 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.

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

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

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

#### 246 4.3. Evaluation criteria

247 Although there are many types of indicators to assess the quality of the  
 248 prediction, here the coefficient of determination ( $R^2$ ) and the Root Mean Square  
 249 Error ( $RMSE$ ). This decision is supported by the widespread use of the former  
 250 indicator in the agriculture and the latter in machine learning (Soares et al.,  
 251 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)

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

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

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

#### 252 4.4. Programming environment

253 We use Python programming language in its interpreted 3.5.2 version, partic-  
 254 ularly with the libraries; Pandas (0.18.1 version) (McKinney, 2010) and Numpy  
 255 (1.11.1 version) (van der Walt et al., 2011).

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

#### 260 4.5. Methodology

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

- 264 • Patrones de semanas: Se formaron patrones desde una semana de obser-  
 265 vación para predecir una semana adelante, hasta doce semanas de obser-  
 266 vación para predecir tres semanas adelante. Por tanto,  $n \times m$  combinations,  
 267 with  $n = 1 \dots 12$  and  $m = 1 \dots 3$ . Figure 2 shows the concept.

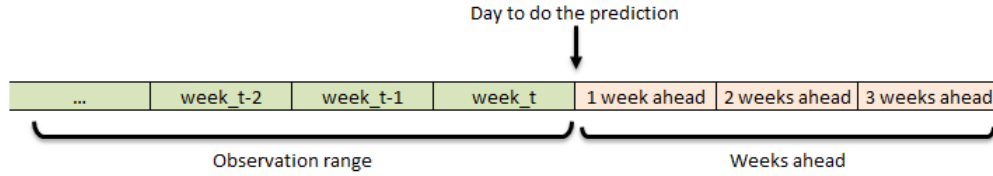


Figure 2: Week patterns explanation process

- 268 • Techniques: Se utilizaron las siguientes técnicas, revisando el espacio  
 269 paramétrico especificado a continuación:
  - 270 – SVR with linear kernel:  $C$  [0.001, 1, 10, 100, 1000],  $\epsilon$  [0.0, 0.4, 0.9].
  - 271 – SVR with gaussian kernel:  $C$  [0.001, 1, 10, 100, 1000],  $\epsilon$  [0.0, 0.4, 0.9],  
 272  $\gamma$  [0.0, 0.4, 0.9].
  - 273 – SVR with sigmoid kernel:  $C$  [0.001, 1, 10, 100, 1000],  $\epsilon$  [0.0, 0.4, 0.9],  
 274  $\gamma$  [0.0, 0.4, 0.9],  $\text{coef0}$  [0.0, 0.5, 5, 10].
  - 275 – Echo state networks:  $\text{LeakingRate}$  [0.02..0.9],  $\text{neurons}$  [1%..90%] de  
 276 la cardinalidad del training set,  $\text{InitLen}$  [0.1..0.8].
  - 277 – Ordinary least squares linear regression: No parameters.

- 278           – Elastic-net regression:  $\alpha$  [0..0.9],  $l1\_ratio$  [0..1.0].
- 279       • Variables included in the model:
- 280           – All variables.
- 281           – From the set  $\{\overline{T}_a, \overline{H}, P, \overline{W}\}$  use the subsets with one, two, three or
- 282           four elements. These variables have the largest impact on the disease
- 283           development (Marín Vargas and Romero Calderón, 1995).

## 284   5. Results and discussion

285       Realizados los experimentos, la Table.2 muestra los  $RMSE$  obtenidos, en la  
286       cual se puede observar que las ESN obtienen un resultado muy diferente a las  
287       demás técnicas. El  $RMSE$  promedio obtenido por las ESN 11085.47 es cercano  
288       a 20 veces el de las otras técnicas. Este resultado de las ESN se explica debido  
289       a que estas redes neuronales recurrentes parten de estados aleatorios para irse  
290       ajustando en el proceso de entrenamiento, y por tanto, los valores obtenidos  
291       son reflejo de que la cantidad de datos disponibles no logra que la red neuronal  
292       se ajuste. Adicionalmente, la Figure 3 muestra los box plots con respecto a los  
293        $RMSE$  obtenidos. Se grafica ESN por separado debido a la diferencia de escala.  
294       Con respecto a las otras técnicas, la regresión lineal, Elastic net, SVR with gaus-  
295       sian kernel y SVR with linear kernel presentan compartimientos muy similares  
296       y aunque el promedio del  $RMSE$  es diferente entre las fincas, La Rita cercano  
297       a 450 y 28 Millas cercano a 820, el comportamiento relativo de las técnicas es el  
298       mismo. Por su parte, SVR with sigmoid kernel presenta un promedio de  $RMSE$   
299       muy inferior a ESN, pero superior a las otras cuatro técnicas.

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

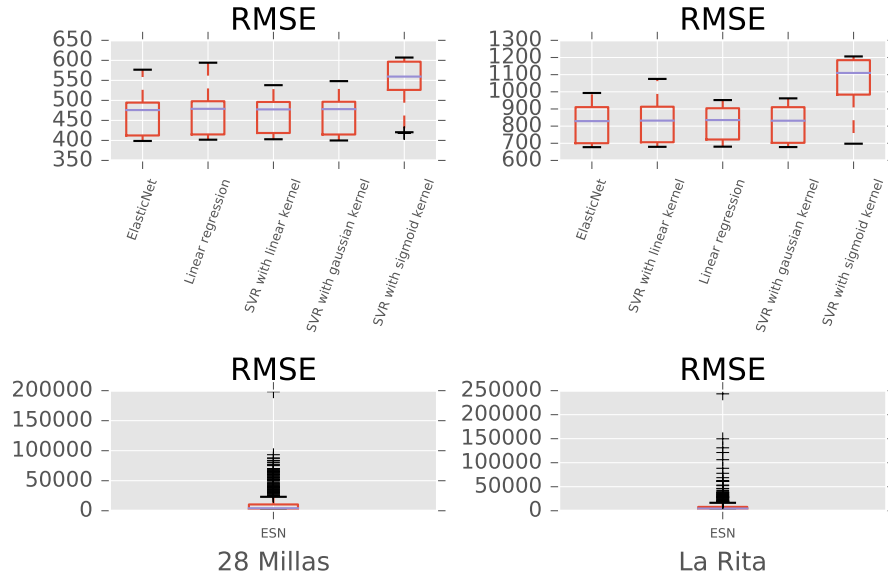


Figure 3: Box plots de los RMSE para cada una de las Fincas

Figure 4 shows the Pareto frontier for each farm with respect to  $R^2$  and  $RMSE$ . The red dots correspond to the observations belonging to Pareto frontier. 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.

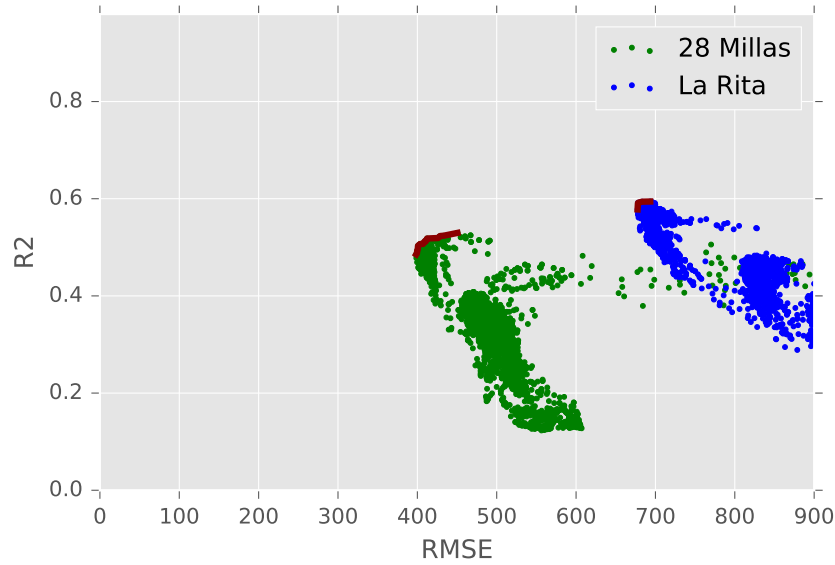


Figure 4: Pareto frontier for  $R^2$  and  $RMSE$

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.

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 5 compares the mean of the  $RMSE$  for each technique in



Table 3: Composition of the Pareto frontier - La Rita

Variable	Observation range - weeks	Weeks ahead	Technique	$RMSE$	$R^2$
$\overline{T}_a \overline{W}$	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%

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

Table 4: Composition of the Pareto frontier - 28 Millas

Variable	Observation range - weeks	Weeks ahead	Technique	$RMSE$	$R^2$
$\overline{T}_a \ \overline{H} \ \overline{W} \ P$	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%

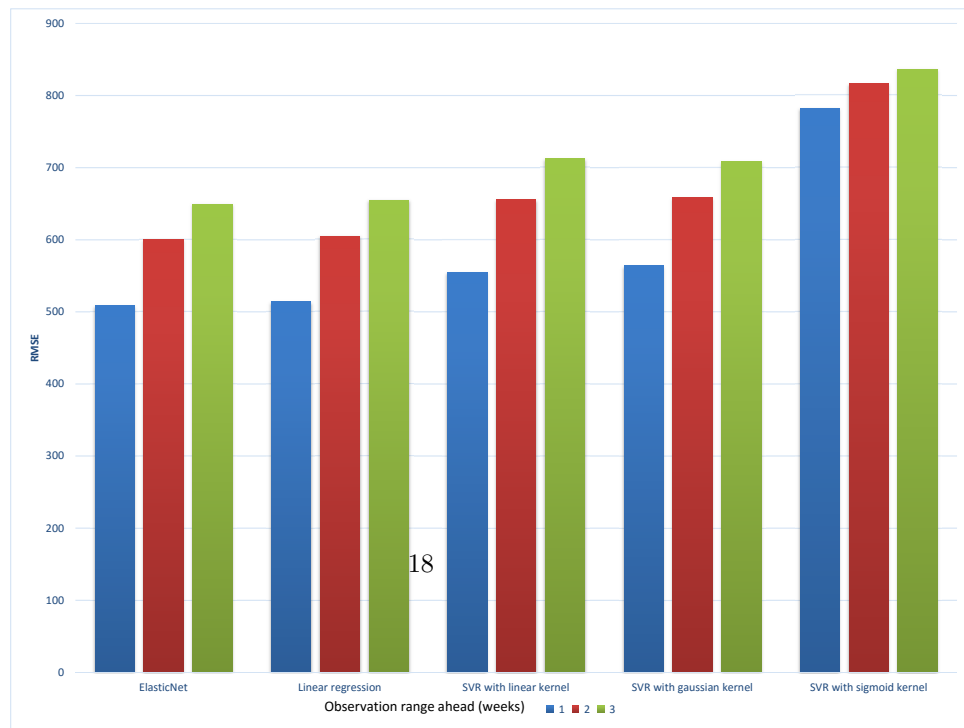


Figure 6 shows the the mean  $RMSE$  for each variables combination. The better results are obtained with  $\bar{T}_a$ ,  $\bar{W}$  and  $P$ . Incluir más variables en el modelo no asegura mejorar la métrica, lo cual es importante pues este resultado sugiere que the use of more sensors do not assure a better result of prediction.

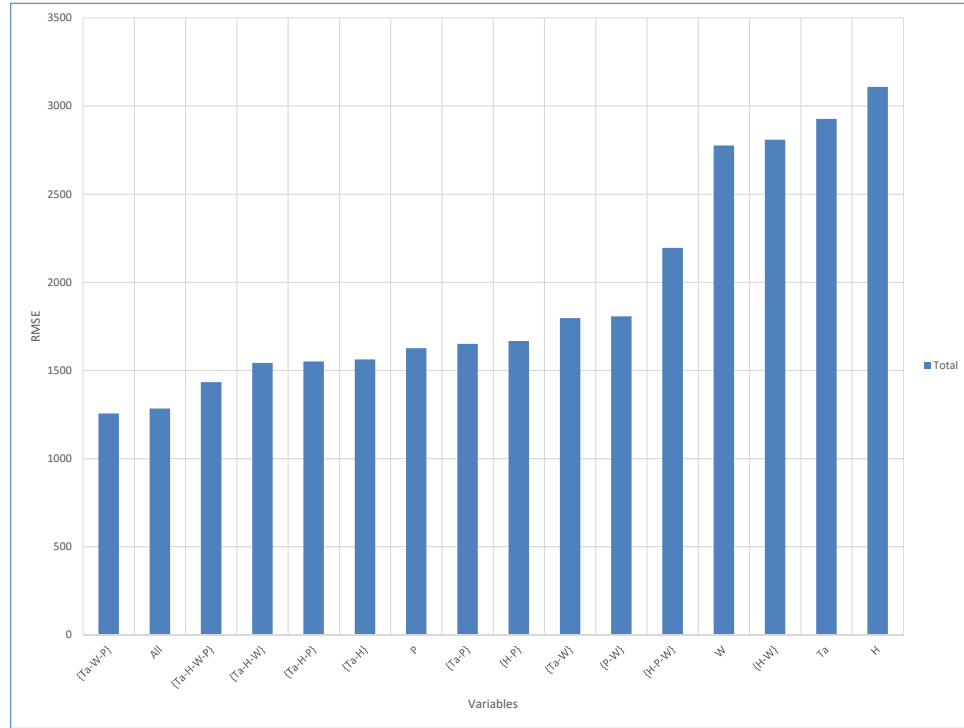


Figure 6:  $RMSE$  for each variable combination

## 6. Conclusions and Future work

Este estudio presentó la comparación de varias técnicas de machine learning en la predicción de la tasa de desarrollo de la enfermedad llamada Sigatoka negra. The following conclusions can be drawn:

1. Si bien los valores absolutos del  $RMSE$  y el  $R^2$  difieren entre las dos fincas en estudio, se mostró que al seleccionar la técnica, la cantidad de

semanas de observación, la cantidad de semanas hacia adelante a predecir y las variables a incluir en el modelo, coinciden los resultados. Lo cual es muestra de que estas características están asociadas al fenómeno más que a una finca en particular.

2. Los menores *RMWE* were reached with linear models.
3. As little as three meteorological variables can be used because of the correlations detected among variables.
4. Las ESN obtuvieron *RMWE* muy altos, lo cual es producto de la cardinalidad del conjunto de datos, que no es suficiente para que la red neuronal se ajuste durante el entrenamiento.
5. Valores de  $R^2$  cercanos al 60% se alcanzaron en las pruebas.

Future work of this research is focused on:

1. The extension of these findings to other farms.
2. The extension of the model to be applied in an early warning system, which does not require the regression task.
3. The use of image to increase the feature sets.

## 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.

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

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

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

365 Chuang, T., Jeger, M., 1987. Predicting the rate of development of black Siga-  
366 toka (*Mycosphaerella fijiensis* var. *difformis*) disease in southern Taiwan. *Phy-*  
367 *topathology* 77, 1542–1547.

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

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

373 Fouré, E., 1988. Stratégies de lutte contre la cercosporioses noire des bananiers  
374 et plan-tains provoquée par *Mycosphaerella fijiensis* Morelet. *L’avis*  
375 *biologique au Cameroun*. Evaluation des possibilités d’amélioration. *Fruits*  
376 43 (5), 269–274.

377 Ganry, J., Laville, E., 1972. La lutte contrôlée contre le *Cercospora* aux antilles.  
378 Bases climatiques de l’avis. *Fruits* 27, 665–676.

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

381 Glezakos, T., Moschopoulou, G., Tsiligridis, T., Kintzios, S., Yialouris, C.,  
382 2010. Plant virus identification based on neural networks with evolutionary  
383 preprocessing. *Computers and Electronics in Agriculture* 70, 263–275.

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

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

391 Kim, Y., Yoo, S., Gu, Y., Lim, J., Han, D., Baik, S., 2014. Crop pests predic-  
392 tion method using regression and machine learning technology: Survey. *IERI*  
393 *Procedia* 6, 52–56.

394 Lukoševičius, M., 2012. A practical guide to applying echo state networks. *Neu-*  
395 *ral Networks: Tricks of the Trade*, 1–20.

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

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

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

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

404 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,  
405 Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos,  
406 A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-  
407 learn: Machine learning in python. *Journal of Machine Learning Research* 12,  
408 2825–2830.

409 Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., 2007. Nu-  
410 merical Recipes. *The Art of Scientific Computing*, tercera edición Edition.  
411 Cambridge University Press.

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

415 Rumpf, T., Mahlein, A.-K., Steiner, U., Oerke, E.-C., Dehne, H.-W., Plümer, L.,  
416 2010. Early detection and classification of plant diseases with support vector  
417 machines based on hyperspectral reflectance. Computers and Electronics in  
418 Agriculture 74 (1), 91–99.

419 Smola, A. J., Schölkopf, B., 2004. A tutorial on support vector regression. Statis-  
420 tics and Computing 14, 199–222.

421 Soares, J., Pasqual, M., Lacerda, W., 2013. Utilization of artificial neural net-  
422 works in the prediction of the bunches’ weight in banana plants. Scientia  
423 Horticulturae 155, 24–29.

424 Soares, J., Pasqual, M., Lacerda, W., Silva, S., Donato, S., 2014. Compari-  
425 son of techniques used in the prediction of yield in banana plants. Scientia  
426 Horticulturae 167, 84–90.

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

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

431 University, T. P. S., 2017. The coefficient of determination  $r^2$ . Online; accessed  
432 28-April-2017.  
433 URL <https://onlinecourses.science.psu.edu/stat501/node/29/>

434 van der Walt, S., Colbert, C., Varoquaux, G., 2011. The NumPy Array: A struc-  
435 ture for efficient numerical computation. Computing in Science & Engineering  
436 13, 22–30.

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

- 440 Witten, I. H., Eibe, F., Hall, M. A., 2011. Data Mining, 3rd Edition. CMorgan  
441 Kaufmann Publishers, United States.
- 442 Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic  
443 net. Journal of the Royal Statistical Society: Series B 67, 301–320.