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 (**?**).

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 **?** and modified by **?** for the control of the yellow Sigatoka in Cameroon, was

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

later adapted by **?** and **?** 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 (**?**) 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 (**?**).

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 markets. 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 (**?**). 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 (**?**). 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 five machine learning techniques to predict the development rate of the black Sigatoka disease: support vector regression (SVR), echo state networks (ESN), ridge regression, elastic-net regression and ordinary least squares linear regression, using input variables such as: air temperature, relative humidity, solar radiation, precipitation and wind speed LISTO - completar las variables de entrada usa to predict the evolution stage in the biological warning system LISTO - completar las variables de salida - Modi

The main contribution of this work [2] is the selection of the best machine learning technique to forecast the black Sigatoka development rate according to the metric indicated below.

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] Definitivamente ese no es el aporte principal del artículo. Cuando esté listo el resto hay que volver aquí. Creo que va a ser algo como mostrar la relevancia de la etapa de preprocesamiento, o mostrar que a pesar de ser sistemas complejos, con variables caóticas, lo mejor en este caso son predictores lineales, o algo por el estilo - Pendiente

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. **?** 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, **?** survey more recent prediction methods for crop pests using regression and machine learning approaches. Nor do they provide numerical results of each paper.

Please review:

MODIFICADO - Falta indicar qué concluyen los dos surveys anteriores —— como indico, en realidad ellos mencionan qué métodos han usado en los paper, pero no dan R2 o rmse, ni detallan resultados en general

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.

? 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, **?** 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.

? 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 [3] 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 (**?**), machine learning assessments of soil drying for agricultural planning (**?**), and early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance (**?**).

Furthermore, there have been attempts to generate software tools. **?** presented an information system for the assessment of plant disorders (Isacrodi). They showed that human experts will attain a much accurate assessment than

[3] variabilidad de qué? - MODIFICADO: de los datos de cada uno

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, **?** 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. **?** 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.

Please review:

MODIFICADO - Importantísimo es explicar aquí brevemente por qué esos otros trabajos no se aplican al caso de la Sigatoka, o qué es lo que han hecho mal, que este trabajo sí hace bien

3. Compared regression techniques

In the prediction of the development rate of the black Sigatoka, we compare techniques such as least squares or ridge regression, commonly encountered in the agricultural literature with machine learning methods such as support vector regression, elastic 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[4] 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 $\mathbf{X}$ (**?**).

Please review:

Alex, no sé hasta donde hiciste en las pruebas ese "arreglo" de que los datos tengan todos una entrada igual a 1. De no ser así, el método no puede encontrar el offset, y sería como forzar que el modelo tenga que pasar por cero... a menos por supuesto que la constante se haya sacado de la ecuación al optimizar...

COMENTARIO: Modificado según lo conversado

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

3.2. Ridge regression

In contrast to the OLSR, for the ridge regression (RR) **?** proposed to add a term to penalize large weights into the error function

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

where the parameter $\alpha > 0$ controls how strong is the shrinking of the estimates towards zero. This shrinkage introduces some bias but helps to reduce the variance of the estimate. The solution of the optimization problem (2) in this case is given by $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$.

3.3. Elastic net regression

Instead of L_2 regularization prior ($\alpha \|\mathbf{w}\|_2^2$) included in the ridge regression, **?** 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 **?** 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 (**?**).

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.

Please review:

Algo que no entiendo es que las elastic net son en realidad una generalización del ridge, Lasso y OLSR. Así que en los experimentos con las selección adecuada de parámetros este método debería comportarse al menos igual o mejor que esos otros regresores! No sé hasta donde sea justificable usar este Y los otros, porque este ES los otros...

COMENTARIO: Revisando la documentación de la bibliteca (ver abajo), parece que por implementación no debería excluirse OLSR por separado. Estos son dos parámetros de las ElasticNet alpha : float. Constant that multiplies the penalty terms. Defaults to 1.0 See the notes for the exact mathematical meaning of this parameter. alpha = 0 is equivalent to an ordinary least square, solved by the LinearRegression object. For numerical reasons, using alpha = 0 with the Lasso object is not advised and you should prefer the LinearRegression object. l1ratio The ElasticNet mixing parameter, with 0 ¡ = l1ratio ¡ = 1. For l1ratio = 0 the penalty is an L2 penalty. For l1ratio = 1 it is an L1 penalty. For 0 ¡ l1ratio ¡ 1, the penalty is a combination of L1 and L2

Please review:

Alex, veo que usas mucho (**?**) para los métodos, pero ese es solo el paper de un tool y no los proponentes originales de los métodos. Usualmente uno hace referencia a algún artículo, libro o tutorial donde ojalá los que propusieron el método son los que lo explican. Ahí metí entonces en las referencias otros artículos por ese motivo.

COMENTARIO: muchas gracias..

3.4. Support Vector Regression (SVR)

Please review:

Estoy seguro que los lectores van a solicitar reducir esto y dejar solo la referencia. Dejé solo lo que considero relevante, pero podés reducirlo más, si querés.

COMENTARIO: Ok, gracias

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 (**?**):

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

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 (**?**)

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

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* (**?**).

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 (**?**).

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.

Please review:

LISTO - Habría que insertar aquí los kernels utilizados en los experimentos

3.5. 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 (**?**).

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. [5]

[5] No puede ser! – Así está en el paper original, lo puse en el repositorio

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 (**?**).

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[6] : *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 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. However, for the current study weekly data were used. [7] 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.

[6] Both farms were also use in the study of **?**. Back then, *La Rita* was referred to as *Waldeck*.

[7] Hay que explicar por qué? → reducción de ruido, etc.

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

Please review:

Alex, Falta explicar cada variable y sus unidades y rangos usuales, lo que es relevante luego cuando se hable de la normalización.

Humidity ¿cuál se usa? ¿la relativa? ¿la absoluta?, ¿la específica? Las otras unidades las supuse, así que hay que revisar que estén bien!

No tengo idea qué es el "Sum precipitation" y en qué unidades estaría. Parece no ser algo estándar o medido directamente, ¿o sí? así que amerita una explicación aquí.

Dijiste que las variables se promedian para una semana. ¿Qué significan entonces las variables que dicen "mean"?

COMENTARIO: Modifiqué estas explicaciones para que quedara más claro el detalle de las variables. Me dirás qué te parece.

Please review:

The mean variables ...

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 [8] was used. The data collected did not exhibit outliers.

Due the fact that the variables measure meteorological or biological processes, they are discretized in order to reflect data trends [9]. The variation range of each variable is uniformly discretized. This discretization removes the effect of small variations in the data, either by inaccuracies of the instruments (meteorological variables) or by subjective bias introduced by the human collecting the data (biological warning system).

The coefficient of variation $C_v(x) = \sqrt{E((x - E(x))^2)}/E(x)$ of each variable x is used to determine the number n of discretization levels as $n = \lfloor 100\ C_v(x) \rfloor$, [10]

[8] falta una referencia

[9] Formalmente la discretización NO captura tendencias, sino la derivada de la función en el tiempo. Para que no se capture ruido uno simplemente usa un filtro pasabajas primero (los estadísticos le dicen diferente, pero no sé cómo). En otras palabras, si querés la tendencia, con usar la derivada de un filtro gaussiano sobre los datos ya se ve la tendencia muy bien con el signo de esa derivada. Con la varianza del filtro uno ajusta qué nivel de detalle le interesa, o lo qué es lo mismo, cuanto tiempo quiere usar para el pronóstico de tendencias (variable crece, está igual o decrece). El valor de salida de la derivada filtrada también se puede discretizar si solo interesa esa tendencia de forma difusa, pero quizá para la regresión es mejor mantenerla continua (si está filtrada, con un filtro o mejor aún con un banco de filtros, al estilo wavelets).

Eso es TOTALMENTE independiente de la cuantificación. Ahora, un cuantificador tiene en principio un efecto de filtro pasabajas, pero eso reduce el ruido y en realidad no dice nada de tendencias. En resumen, formalmente la cuantificación en sí no refleja nada de un comportamiento temporal, y "tendencia" es en principio un comportamiento temporal!

[10] Hay algo que no entiendo. ¿De dónde sale el 100? ¿Por qué 100? Así como está esto, el mismo rango de variación de temperatura produce diferentes n, con solo un cambio en el

where $\lfloor \cdot \rceil$ is the round operator.

Each feature was scaled[11] to fit in the range 0 and 1. The variable E_s to be predicted was not scaled.

Please review:

Aquí faltan la fórmula general de normalización usada. Lo que sigue lo estoy asumiendo, pero puede ser que no sea así

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 $b = -mx_{\min}$.

Please review:

Explicar brevemente la razón de la normalización, ojalá con una referencia bibliográfica

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 (**????**).

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 ¿cómo se llama esto? ¿unexplained variance? S_R^2 are estimated as

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

promedio! Eso es muy extraño. Por otro lado, los rangos de las variables son tan diferentes, que el número n va a terminar siendo casi cualquier cosa. Yo hubiese esperado que la desviación estándar se use para discretizar las variables. Otro problema formal más serio, es que el coeficiente de variación solo se puede usar/tiene sentido si la escala de medición es de razón, y la temperatura, la precipitación y la humedad son usualmente mediciones de intervalo! Los L_i ni tan siquiera sé qué son porque creo que son algo subjetivos a la persona que los toma ¿no?, aunque creo que iban con el porcentaje de área afectado, lo que también sería medición de intervalo!

[11] ¿escaladas cómo? ¿sx? ¿$sx + b$? ¿$sf(x)$?

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

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

Please review:

Favor revisar. Vi una representación en términos de $R^2 = 1 - MSE/\sigma^2$ (con σ^2 la varianza), que no sé si está bien o no, pero lo mejor será poner la fuente de donde se tomó la anterior definición. Había en el S_R^2 un $\bar{y}_i$ pero el promedio no depende de i por lo que no tenía sentido. Supuse que era el promedio, pero no sé. El cuadrado en las S^2 ¿está bien?, o indirectamente estás diciendo que $RMSE = S_e$, que sería una mejor notación matemática (dejando RMSE solo como abreviatura textual)!

4.4. Programming environment

We use the Python programming language with the Integrated Development Environment (IDE) [12] Spyder (**?**), particularly with the libraries Pandas (**?**) and Numpy (**?**).

The implementation for SVR, ridge and ordinary least squares regressions in scikit-learn (**?**) were used. Adjustments to the ESN implementation code of **?** were necessary to allow its integration into our experimental framework.

All experiments were performed on a PC computer with an Intel® Core i7-4800MQ processor, 2.70 GHz, 16.0 GB RAM, under the operating system MS Windows 8 Pro.

4.5. Methodology

The evaluation of the techniques under consideration of their parameter space, was performed in two stages, described below.

[12] El IDE es para la reproducción de resultados irrelevante, y por lo tanto información circunstancial que es mejor evitar. Lo relevante es lo que tenga efectos directos en los resultados. Si yo prefiero usar un editor de texto corriente para programar, el restulado final es el mismo, si uso el mismo compilador/intérprete. Eso es la información relevante: ¿qué versión de Python concretamente? ¿Se compiló el Python o se usó interpretado? ¿Cuáles versiones de Pandas y Numpy en concreto se usaron?

4.5.1. Phase one

In the phase one, ten-fold-cross-validation on the total set of machine learning methods under a subset of configurations was evaluated:

- Patterns: $n \times m$, with $n = 1 \dots 8$ and $m = 1 \dots 2$.

  Please review:

  No se entiende que quiere decir eso!!

- Methods: support vector regression with the kernels linear, gaussian and sigmoid; echo state networks; ordinary least squares linear regression, ridge regression and elastic-net regression.

  Please review:

  Cada uno de esos métodos tiene a su vez un espacio paramétrico. Eso tiene que quedar claro aquí: cómo se barrió el espacio paramétrico, para saber el nivel de detalle de la prueba! Eso es importantísimo, porque si no no se puede decir nada respecto a los métodos, sino solo respecto la configuración concreta probada!

- 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 have the largest impact on the disease development (**?**).

4.5.2. Phase two

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

Please review:

Hmm, ¿entonces cuántas/cuáles semanas se usaron para la fase uno?

5. Results and discussion

In this section we present the main results for phase and their discussion.

Phase one

Figure 2 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.

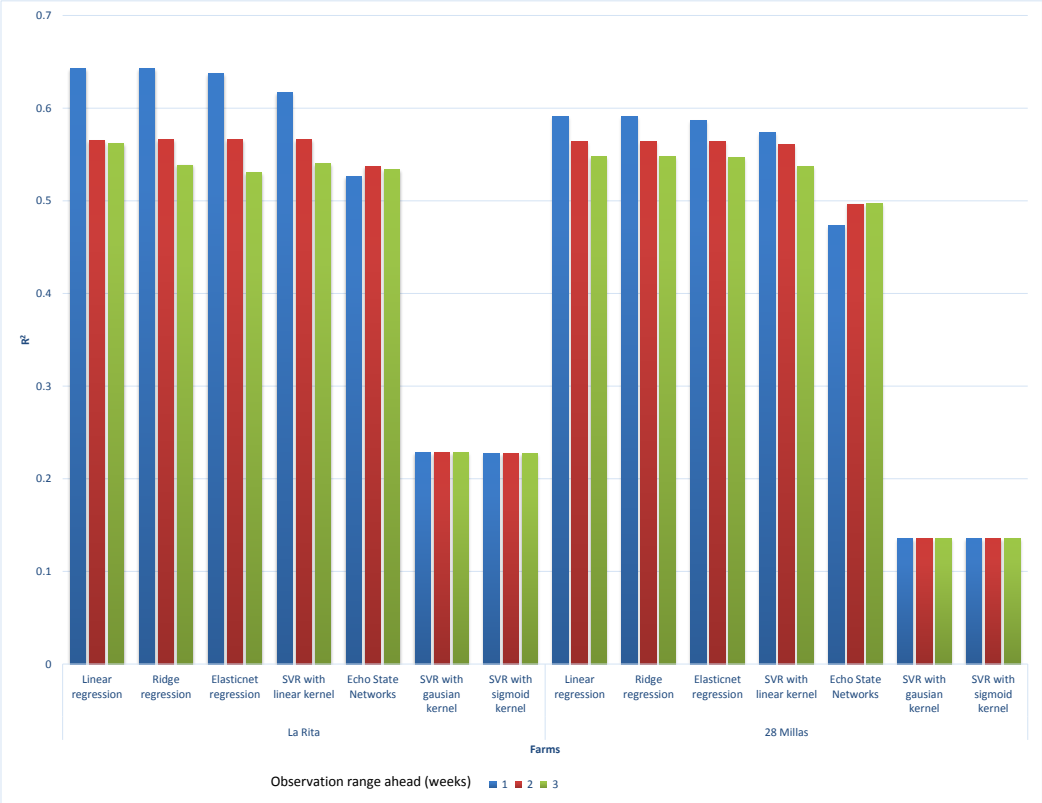

Figure 2: Phase one - Best R^2 for each algorithm

Figure 3 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

312 better R^2 for one week ahead than two weeks ahead and so on.

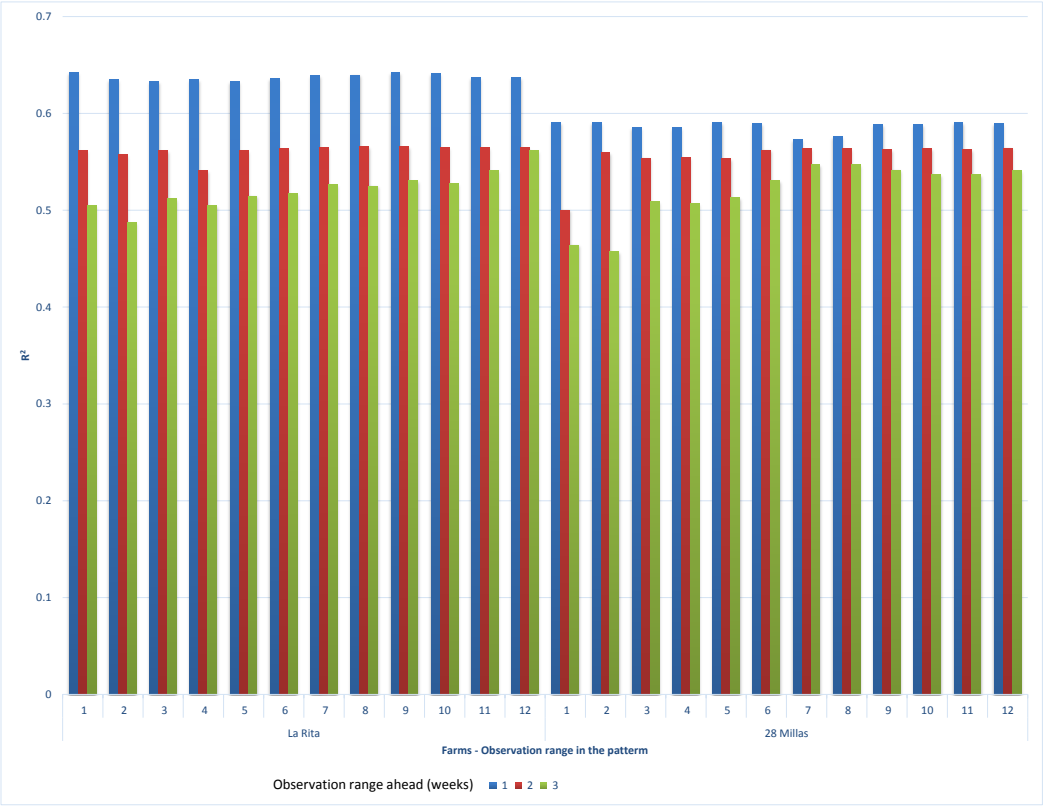

Figure 3: Phase one - Best R^2 for each observation range

313 Figure 4 shows the best R^2 for each variables combination. Results are group
314 by farm. The better results are obtained with $\overline{T}_a$ and the combination of $\overline{T}_a$
315 with $\overline{W}$, in both farms of similarly. You can note that the use of all variables
316 in the model or the inclusion of the four variables suggest for expert criteria do
317 not improve significantly the results, then the use of more sensors do not assure
318 a better result.

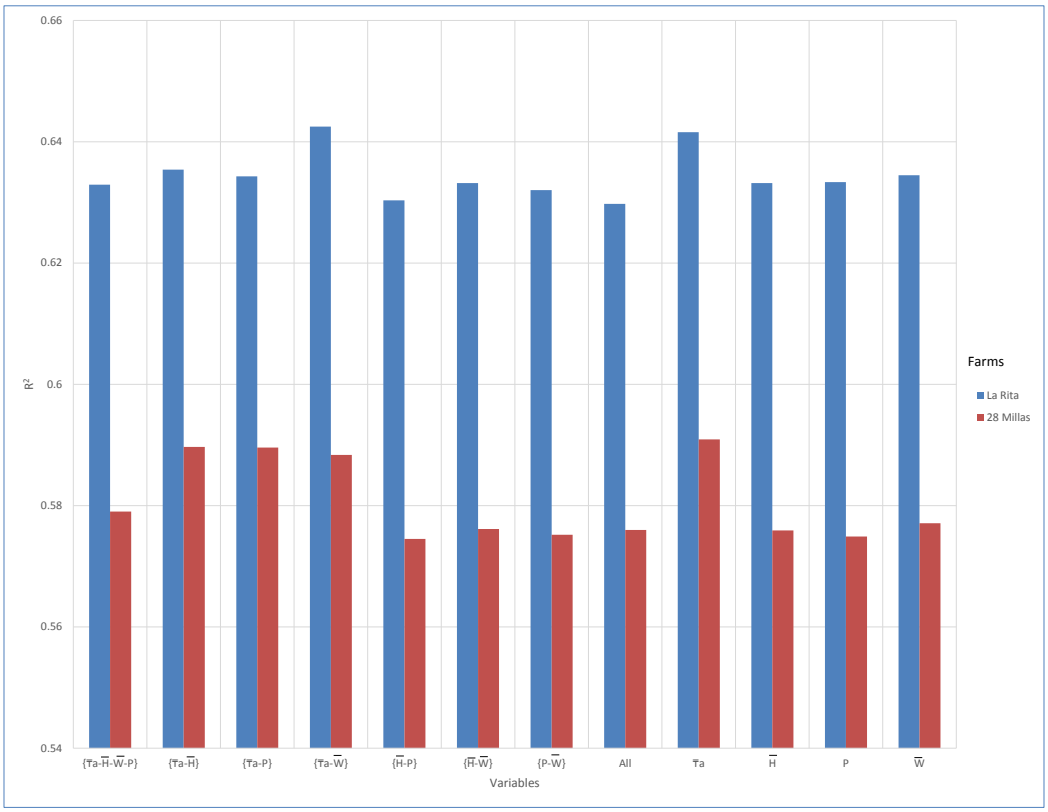

Figure 4: Phase one - Best R^2 for each variable combination

Figure 5 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.

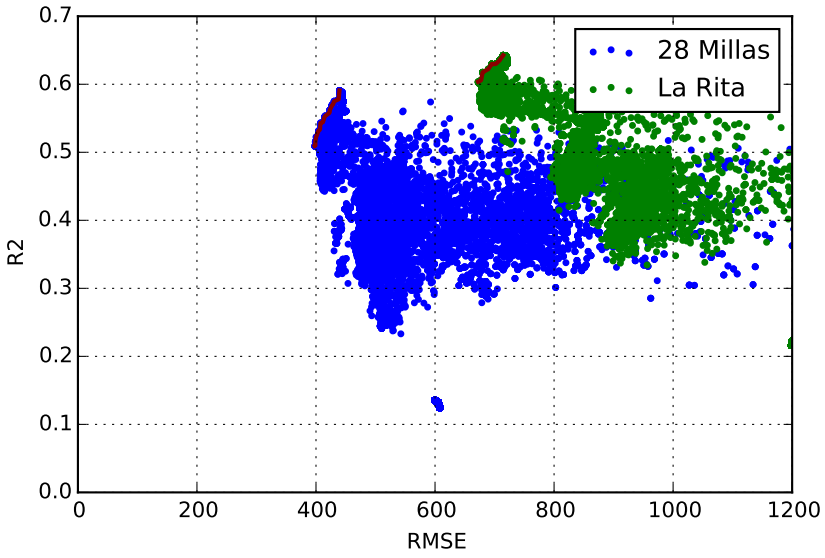

Figure 5: Phase one - Pareto frontier for R^2 and $RMSE$

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 6 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.

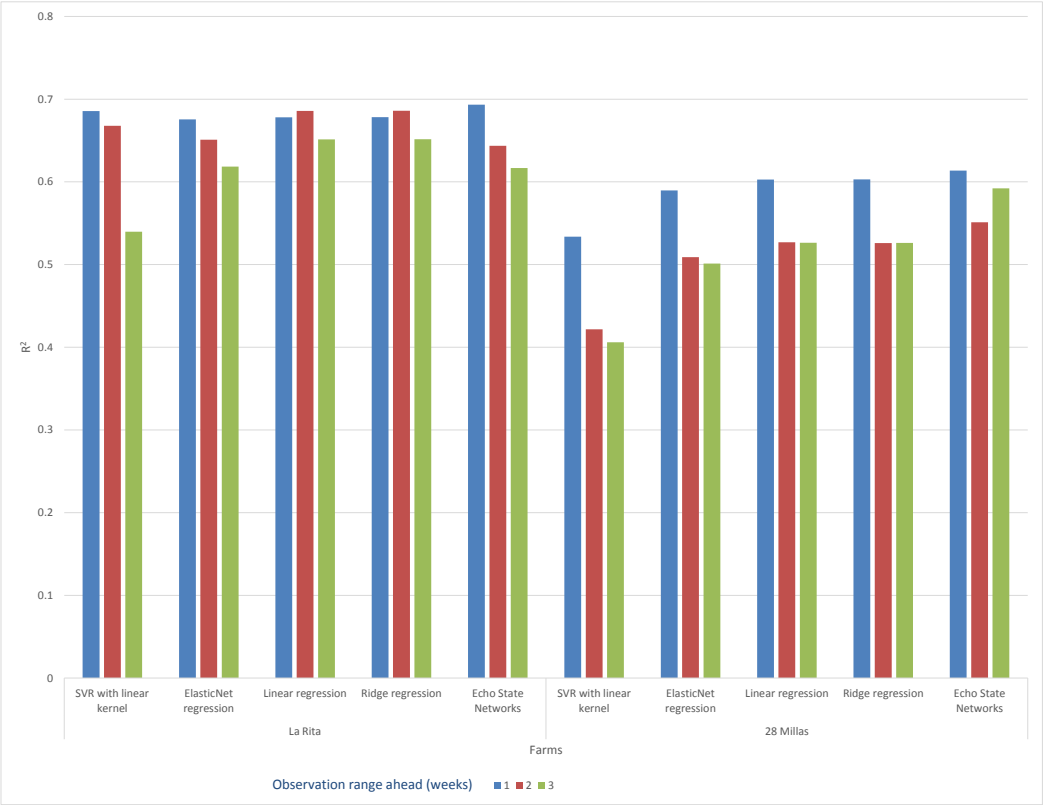

Figure 6: Phase two - Best R^2 for each algorithm

Figure 7 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.

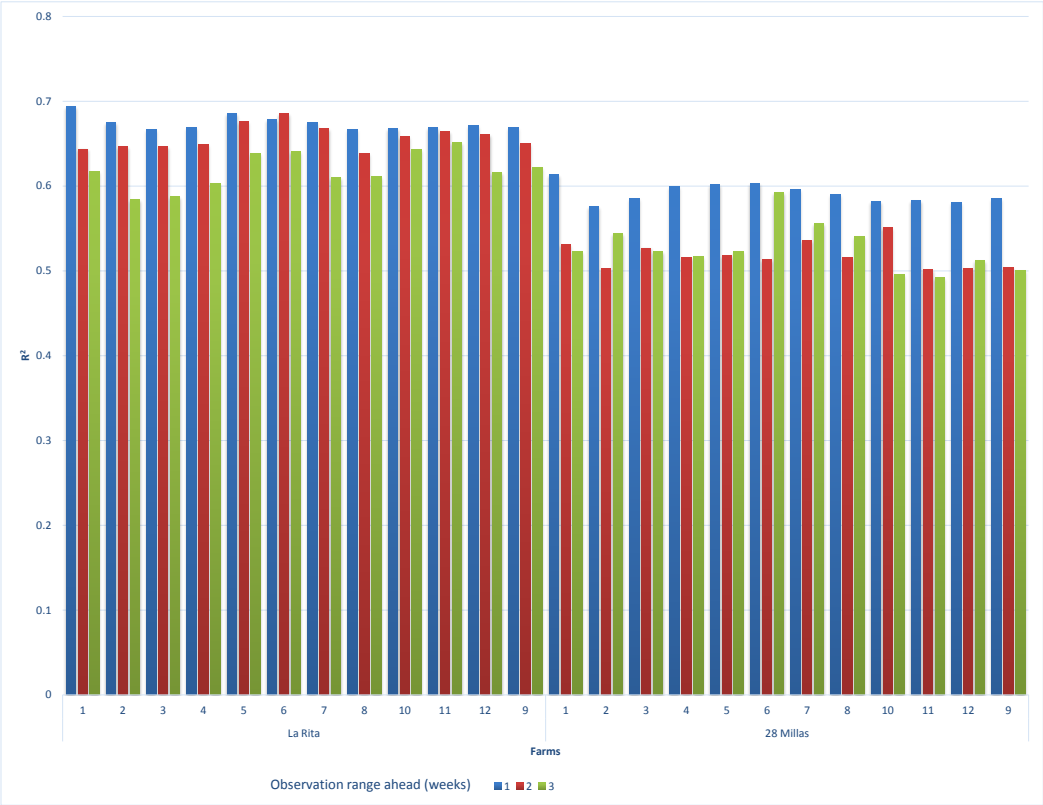

Figure 7: Phase two - Best R^2 for each observation range

Figure 8 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.

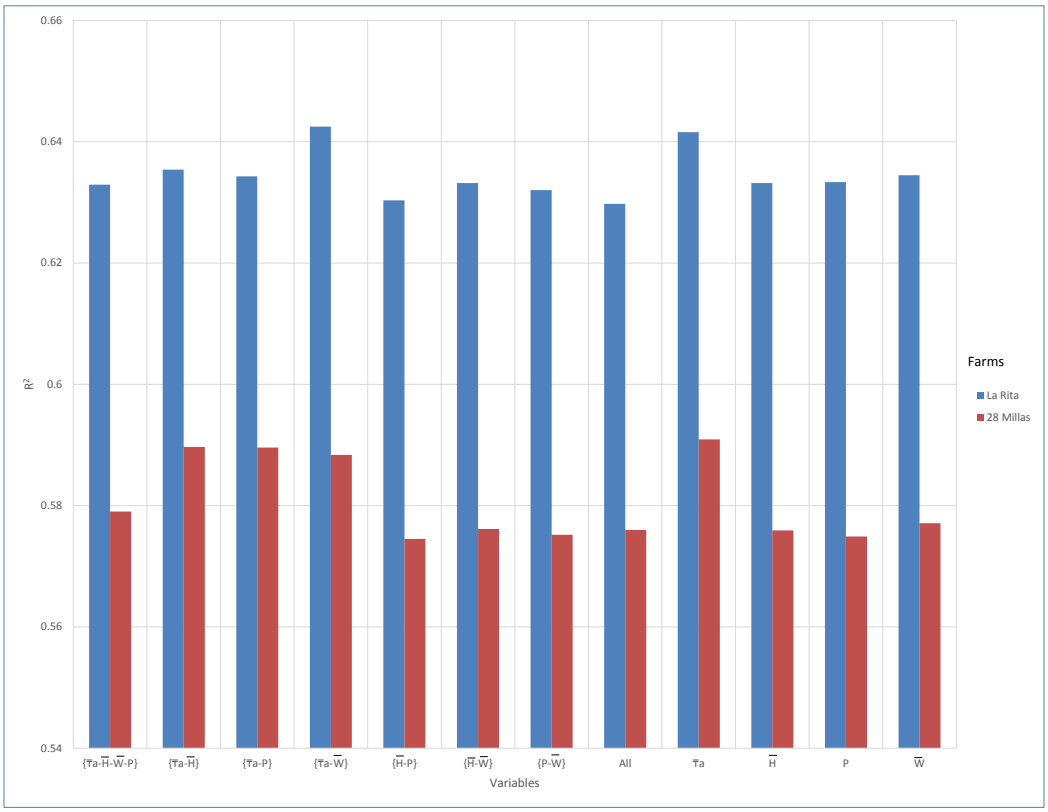

Figure 8: Phase two - Best R^2 for each variable combination

Figure 9 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.

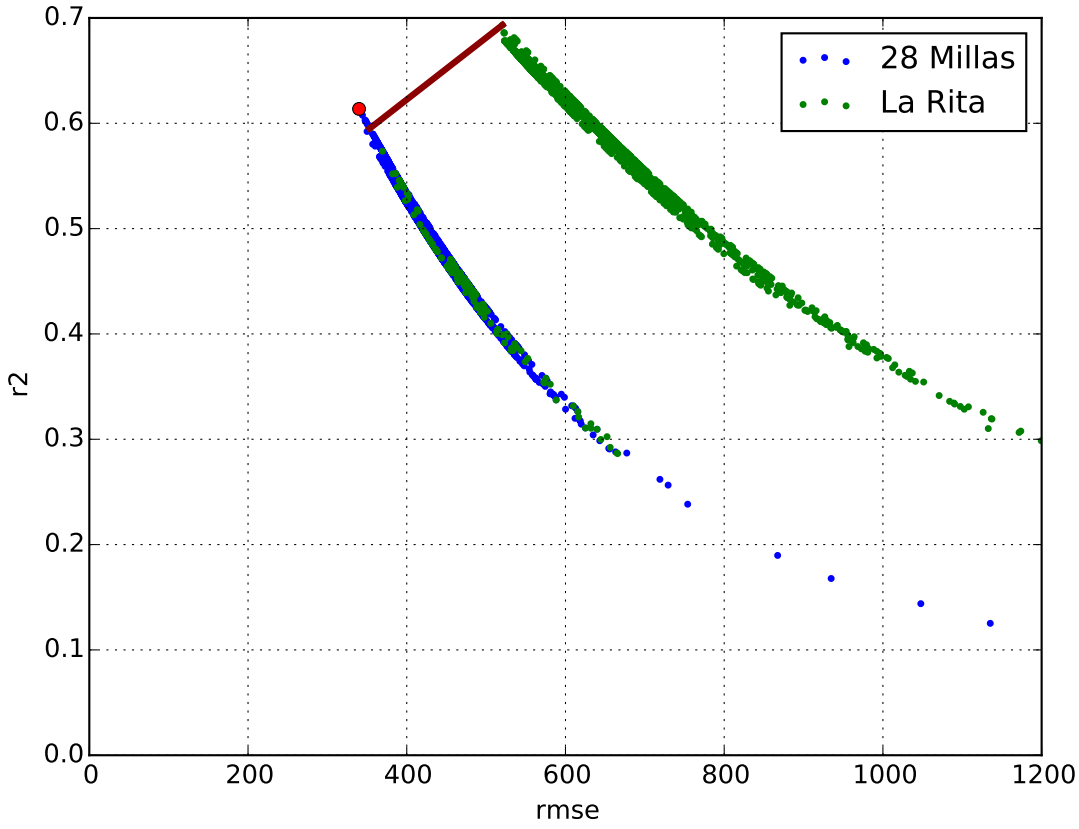

Figure 9: Phase two - Pareto frontier for R^2 and $RMSE$

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.

6. Conclusions

6.1. Future work

Acknowledgements

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

References