Agricultural and Forest Meteorology 205 (2015) 51–59

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet

A novel solution to the variable selection problem in Window Pane approaches of plant pathogen – Climate models: Development, evaluation and application of a climatological model for brown rust of wheat

David Gouache[a,c,*], Marie Sandrine Léon[a,b], Florent Duyme[c], Philippe Braun[d]

[a] ARVALIS – Institut du Végétal, Service Génétique Physiologie et Protection des Plantes, Rue de Noetzlin–Bât. 630, F-91405 Orsay Cedex, France
[b] Université Paris-Descartes, France
[c] ARVALIS – Institut du végétal, Station Expérimentale, F-91720 Boigneville, France
[d] ARVALIS – Institut du végétal, Domaine de la Bastide, Route de Generac, F-30900 Nimes, France

ARTICLE INFO

Article history:
Received 2 October 2014
Received in revised form 12 February 2015
Accepted 20 February 2015
Available online 27 February 2015

Keywords:
Puccinia triticina
Eheat
Window Pane
Model evaluation
Data mining
Elastic-net

ABSTRACT

A model for predicting brown rust severity in France was developed using the systematic screening of climatic variables of the Window Pane approach and data from 400 field trials spanning 30 years. The model was built using novel methods to manage the variable selection problem posed by the very large number of predictor variables generated by Window Pane, namely the elastic-net, and a systematic cross-validation to determine the most frequently retained variables. The model predicts the final severity of brown rust with an RMSEP (root mean square error of prediction) of 22.4%. The model's ability to predict treatment decisions was tested and exhibited a good performance as shown by an area under the receiver operator curve of 0.85. We also evaluated the suitability of our model for use in France by confronting the range of the climate variables in our dataset against the climatological range of these same variables in France. The final model also gives important insights into the key factors behind variations in brown rust disease pressure.

1. Introduction

The search for relationships between meteorological variables and epidemics in crop plants has been at the heart of plant pathology for many years, and has led to a great number of forecasting schemes (Coakley, 1988) based on hourly or daily weather records and equations linking these conditions to the different phases of pathogen life cycles. As reviewed by Coakley (1988), in the 1970s and 1980s, approaches linking climate variability to disease variability – between years and/or geographical areas – gained impetus. Using historical records of over ten years, Coakley's pioneering work on wheat stripe rust lead to the development not only of forecasting schemes for this disease (Coakley et al., 1982, 1984) and *Septoria* leaf blotch (Coakley et al., 1985), but also, and perhaps more importantly, on the advent of the "Window Pane" approach (Coakley et al., 1988). Briefly, the Window Pane approach consists of an algorithm that systematically calculates synthetic climate variables over overlapping time frames, ranging from a few weeks to a few months in length for every growing season. The procedure then calculates the correlation between each variable and observed yearly disease levels. This approach is used to identify critical periods in which the variations in specific climatological variables lead to variations in disease. These results serve to improve the understanding of the studied pathosystem, provide useful knowledge for managing the disease (Coakley, 1988; Te Beest et al., 2008), and finally lead to forecasting models, that in Coakley's work were built through multiple linear regression and stepwise variable selection among those selected from the Window Pane procedure.

Interest in Window Pane approaches has increased in recent years, with models for multiple wheat diseases being developed in the United Kingdom (Pietravalle et al., 2003; Te Beest et al., 2008, 2009). Pietravalle et al. (2003) addressed the risks of this approach, namely that by generating an extremely large number of climate variables, spurious correlations may occur. Moreover, variables from overlapping and adjacent time frames are often strongly correlated. In other words, the Window Pane algorithm

[*] Corresponding author at: ARVALIS – Institut du végétal, Service Génétique Physiologie et Protection des Plantes, Station expérimentale, France. Tel.: +33 686089432; fax: +33 16993039.
E-mail address: d.gouache@arvalisinstitutduvegetal.fr (D. Gouache).

http://dx.doi.org/10.1016/j.agrformet.2015.02.013
0168-1923/© 2015 Elsevier B.V. All rights reserved.

leads to a variable selection problem. Pietravalle et al. (2003) and Te Beest et al. (2008) approached this by extending the statistical criteria used in variable selection, still using a multiple regression framework however, and by analysing the relevance of the selected variables from a biological point of view. Luo (2008) approached this problem by introducing high-dimensional regression techniques, such as principal component regression and partial least squares regression. Another important feature of this work was to use cross-validation during the variable selection phase: the frequency of variable selection across all runs of cross validation was used to define the final model. This paper builds on these improvements of the Window Pane procedure by introducing a variable selection procedure created for high dimensional datasets, with the number of variables largely exceeding observations, and with groups of highly correlated predictors, such as micro-array data (Zou and Hastie, 2005). This technique was introduced to reduce the variable selection problem so as to be able to apply, in the final construction of the model, a classical multiple regression approach. Indeed, such models are simple to calculate and to understand, making them easily applicable and communicable to end-users.

This new approach was used to construct a climatological forecasting model for brown rust disease of wheat, caused by *Puccinia triticina*. Brown rust is the most important disease of winter wheat (*Triticum aestivum*) after *Septoria* leaf blotch (Jørgensen et al., 2008). It is also the most important disease of durum wheat (*Triticum durum*). In the absence of efficient protection, yield losses can amount to 50% in cropping regions worldwide (Huerta-Espino et al., 2011), with figures of 40% and 75% reported for these two crops in France by Caron (1993) and Zaka (2012). It is caused by a biotrophic foliar fungus. Its epidemics are polycyclic: they develop through successive infection cycles that progressively build up increasing loads of infectious spores within the field, if conditions are favorable to the dispersion and successful infection of the spores. Dispersal is essentially due to wind, although rain can also be involved (Sache, 2000). The infection process depends on high humidity and temperature, with an optimum temperature near 15 °C (de Vallavieille-Pope et al., 1995). In France, as in other wheat growing regions of the world, conditions generally become favorable in the spring, when the crop resumes its active growth. Previous to spring, however, conditions can be much less favorable. Indeed, Eversmeyer and Kramer (1998) showed that the importance of brown rust epidemics in the Central Great Plains of the United States is linked to more or less favorable conditions between the harvest of the previous crop and the following spring. Similar results were obtained from an initial Window Pane modeling approach on durum wheat in France (Thepot and Gouache, 2009).

We present the results of the evaluation and application of the model developed here. Model evaluation according to criteria such as mean square error of prediction (MSEP) obtained through cross-validation is a classical model evaluation technique (Wallach and Goffinet, 1987). However, for models that are intended for operational forecasting purposes, more may be required, such as suitability for the projected conditions of use and decisional capability. Our first consideration is that of model “suitability”. Indeed, statistical models are not meant for use outside the conditions represented in the data used to develop them. However, when in use, the model will be asked to predict disease for new weather sequences that may depart from those encountered previously. Coakley et al. (1988) had already pointed this problem out, and proposed that a minimum of 10 years of data was necessary to obtain a sufficiently wide range of weather conditions. The same reasoning applies to the geographical span of the model: Thepot and Gouache (2009), thus, showed that MSEP of a Window Pane brown rust model for the 4 durum wheat growing regions was lower when one model was fitted to all regions instead of one model being fitted to each region. To ensure that our new model would be sufficiently suitable, we verified the range of values for each selected weather variable against the observed range across the whole French territory during the past 25 years. Finally, a forecasting model is used to aid in making decisions, such as whether or not to apply fungicides. It is important that the decisional value of models be evaluated (Hughes et al., 1999). Hence, we will also present the decisional evaluation of our model with receiver operator characteristic (ROC) analysis.

2. Material and methods

2.1. General overview

We present the overall workflow of data management and analysis in Fig. 1. We will detail in the following paragraphs each of the 3 major phases of the work. Briefly, a first phase consisted in collecting diverse brown rust datasets from agronomical trials, curating them to answer to two different objectives, e.g., model building and model decisional evaluation, and calculating weather variables for each trial; the second phase consisted in building a brown rust prediction model through different successive variable selection steps; and the third phase consisted in evaluating the model according to three criteria, e.g., RMSEP (root mean square error of prediction), range of weather variables compared to climatological range, and decisional performance. All the statistical analyses presented below were fitted with R statistical software (R Core Team, 2012).

2.2. Phase 1: data

2.2.1. Disease data

We collected an extensive set of brown rust observations in over 400 field trials in France, spanning over 30 years (1980–2011 harvest years). These observations were collected from plots untreated by fungicides in trials aimed either at assessment of fungicide efficacy or disease resistance. Trials consisted in three replicate 20 m^2 plots (11 rows of 10 m long) managed according to standard local practices limiting all other biotic and abiotic stresses. Disease observations used are carried out on 20 stems per plot. On each stem and each leaf layer, disease severity is scored as the percentage of leaf area covered by brown rust lesions. The value for a given observation date and leaf layer is the mean across the replicates.

The dataset was then curated to answer to two objectives: model building and decisional evaluation. For decisional evaluation, the approach simply consisted in using the observations to answer the question: “at the date of observation, should the plot have received a fungicide treatment?” This question was answered using the decision threshold advised in France (Jørgensen et al., 2008; EuroWheat.org, accessed 24/07/2014), i.e., that a treatment should be made as soon as symptoms are observed on top three leaf layers after growth stage BBCH32 (2 nodes) (Lancashire et al., 1991). This led to a dataset of 903 true/false observations accross 400 site-years.

The dataset was also used to build a quantitative prediction model of brown rust severity. The objective was to obtain one quantitative value of the final severity of disease for as many site-years as possible. This led to retaining observations made on leaf layers one and/or two (one being the uppermost, i.e., flag leaf) at mid grain filling (BBCH 71, 75 and 85; Lancashire et al., 1991). Percentage values were transformed to obtain better normality (Ahrens et al., 1990):

$$\mathrm{Obs}_{\mathrm{transf}} = \arcsin\left(\sqrt{\frac{\mathrm{Obs}}{100}}\right), \quad (1)$$

where $\mathrm{Obs}_{\mathrm{transf}}$ is the transformed value and Obs the initial value.

We thus, obtained a dataset with observations on both bread and durum wheat, on a wide range of varieties, on 2 leaf layers