Toward a Reliable Evaluation of Forecasting Systems for Plant Diseases: A Case Study Using Fusarium Head Blight of Wheat

S. Landschoot, KERMIT, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure links 653, BE-9000 Gent, Belgium, and Faculty of Applied Bioscience Engineering, University College Ghent, Valentin Vaerwyckweg 1, BE-9000 Gent, Belgium; **W. Waegeman**, KERMIT, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University; **K. Audenaert**, Faculty of Applied Bioscience Engineering, University College Ghent, and Department of Crop Protection, Laboratory of Phytopathology, Ghent University; **J. Vandepitte**, KERMIT, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University; **G. Haesaert**, Faculty of Applied Bioscience Engineering, University College Ghent, and Department of Crop Protection, Laboratory of Phytopathology, Ghent University; and **B. De Baets**, KERMIT, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University

Abstract

Landschoot, S., Waegeman, W., Audenaert, K., Vandepitte, J., Haesaert, G., and De Baets, B. 2012. Toward a reliable evaluation of forecasting systems for plant diseases: A case study using Fusarium head blight of wheat. *Plant Dis.* 96:889-896.

Despite great efforts to forecast plant diseases, many of the existing systems often fall short in providing farmers with accurate predictions. One of the main problems arises from the existence of year and location effects, so that more advanced procedures are required for evaluating existing systems in an unbiased manner. This paper illustrates the case of Fusarium head blight of winter wheat in Belgium. We present a new cross-validation strategy that enables the evaluation of the predictive performance of a forecasting system for years and locations that are different from the years and locations on which the forecast was developed. Four different cross-validation strategies and five regression

techniques are used. The results demonstrated that traditional evaluation strategies are too optimistic in their predictions, whereas the cross-year cross-location validation strategy yielded more realistic outcomes. Using this procedure, the mean squared error increased and the coefficient of determination decreased in predicting disease severity and deoxynivalenol content, suggesting that existing evaluation strategies may generate a substantial optimistic bias. The strongest discrepancies between the cross-validation strategies were observed for multiple linear regression models.

In the last decades, the development of accurate forecasting systems for the occurrence and intensity of crop diseases has been increasingly studied (3,12,14,16,24,43,45,54,55). Crop diseases can be considered as an important cause of crop yield loss, so modeling and predicting these diseases accurately can help in providing prior knowledge of the time and severity of the outbreak of diseases (44). Additionally, modeling aims to understand the main determinants of epidemic development in order to construct sustainable strategies for the management of plant diseases (55). Plant pathologists working in areas such as epidemiology, pathogen ecology, pathogen population biology, and disease management are challenged in uncovering patterns in multivariate data (48). Several statistical approaches exist to synthesize multivariate information and to model yield losses due to plant diseases (49).

Existing forecasting systems are principally based on the combined effect of host susceptibility, inoculum strength, and meteorological conditions on disease development (60). Many of these systems can be classified as descriptive or explanatory models, tracing the existence of relations between variables (44). Currently, a range of statistical techniques is used in plant pathology (20); multiple linear regression models, mechanistic models, and knowledge-based models are often applied (1,25,52–54). Ridge regression, a modification of multiple linear regression, is used in situa-

tions where predictor variables are inter-correlated (13,58). Conversely, if one aims to construct a system that delivers accurate predictions, substantial performance gains can be observed by utilizing more advanced data mining and machine learning models like artificial neural networks (10,12,28,38), ordinal regression models (24), and support vector machines (SVMs) (9,27). These models often provide fewer insights as a result of their black box nature, but they often exhibit a high coefficient of determination (R^2) and a low mean squared error (MSE) as an indication of predictive power (23).

Several strategies for evaluating plant disease prediction models can be distinguished in the literature. The simplest strategy consists of utilizing the same dataset for model construction and model evaluation. From a statistical perspective, it is well known that such an approach will always lead to over-fitting for complex nonlinear models that take many predictor variables into account (23). A straightforward way of avoiding over-fitting consists of considering separate and independent datasets for model construction and model evaluation, often denoted as the training set and the test set in predictive modeling, or the calibration and validation datasets in more applied research fields (1,17,24,29,41,47). This approach is much more reliable if the test set remains sufficiently large, but it has the obvious disadvantage that less data can be used to construct a model. Statistically speaking, this train/test splitting suffers from a high variance (6,23,59). The variance can be somewhat reduced by repeating this procedure several times (15,59), as considered in De Wolf and Francel (12) for predicting tan spot and Stagonospora blotch of hard red spring wheat.

To date, it remains very difficult to collect a sufficient number of observations in many plant disease prediction scenarios, so it would be highly undesirable to use only a part of the data during

Corresponding author: S. Landschoot, E-mail: sofie.landschoot@hogent.be

Accepted for publication 28 December 2011.

the model construction phase. Fortunately, a well-known alternative that utilizes the available data in an optimal way consists of cross-validation, a strategy in which multiple training and test sets are generated from the initial dataset (2,15,23,56,59). Unlike repeatedly splitting the data into training and test sets, cross-validation strategies guarantee that the different sets do not overlap, while almost the entire dataset can be utilized for constructing the model. Among other studies, random *K*-fold cross-validation has been applied for the prediction of the severity of Asian soybean rust epidemics (14), the severity of gray leaf spot of maize in Iowa (38), and the presence of the mycotoxin deoxynivalenol (DON) in mature winter wheat in the Netherlands (54).

In this article, the performance of five different regression techniques and four cross-validation strategies is illustrated on the case of *Fusarium* head blight (FHB) in winter wheat in Belgium. FHB of small grain cereals is an important disease worldwide, causing grain yield reductions, decreasing the quality of certain industrial production processes such as baking and brewing (e.g., reduced test weight, sticky doughs, and gushing of beer) and the quality of seed by reducing germination efficiency and causing root and crown rots. In addition, it leads to grain contamination with *Fusarium* mycotoxins such as DON, making the yield unsuitable for food and feed purposes (16). To ensure food and feed safety, the European Commission has set maximum limits for unprocessed cereals for human consumption (EC 1881/2006) and guideline limits for *Fusarium* mycotoxins in animal feed. Due to these maximum levels, EU farmers are obliged to take measures to minimize *Fusarium* mycotoxin contamination. Worldwide, several initiatives are on-going to develop quantitative models to predict DON content in mature wheat (44) and the relationship between DON content and FHB symptoms (40). The dynamic nature of the *Fusarium* population, inter- and intraseasonal variations (5,26,30,32,42,57,61), is a serious hindrance to controlling and predicting FHB during the growing season.

Using a sizeable dataset, including visual observations of FHB disease severity, DON content measurements, and more than 300 predictor variables, from 2002 onwards, we hypothesize that the commonly applied *K*-fold cross-validation strategy cannot be adopted for evaluating plant disease prediction models if year or location effects are present. These effects originate from the inclusion of agronomic variables such as previous crop and soil management information and weather variables, which vary spatially and temporally (44,53). In addition, we present a new cross-validation strategy that overcomes all the above-mentioned problems. This procedure allows obtaining an unbiased estimate of the model performance for future years and new locations. In this article, we will call this strategy cross-year cross-location validation. A formal definition can be found in Materials and Methods. This evaluation strategy, illustrated on FHB, can be generally adopted for assessing plant disease prediction models during the development phase, prior to building the final model that is applied in practice.

Materials and Methods

Description of the dataset. From growing season 2001–2002 until 2009–2010, different field trials of winter wheat were moni-

tored throughout Belgium (Table 1), under natural infection, to collect data on disease symptoms, DON content, and the composition of the *Fusarium* population. The surface of each elementary plot in the experimental field trials was about 15 m². At each location, at least 10 wheat varieties were grown, ranging from moderately resistant to highly susceptible for FHB, in a completely randomized block design with four replications. At all locations, wheat was produced under normal cropping conditions for Belgium.

The resistance of the different wheat varieties was determined through independent artificial inoculation tests with a mixture of spores from *Fusarium avenaceum*, *F. culmorum*, *F. graminearum*, *F. poae*, and *M. nivale* in separate field experiments. The varieties were classified into five risk classes for DON content and five risk classes for disease susceptibility. Varieties with a DON content 40% lower than the mean DON content are classified in resistance class 1, varieties with a DON content between 40 and 20% lower than the mean are classified in class 2, class 3 contains the varieties with a DON content between 20% lower and 20% higher than the mean. Class 4 varieties had a DON content between 20 and 60% higher than the mean, and the most susceptible varieties, in class 5, are those with a DON content more than 60% higher than the mean DON content.

Early July at growth stage (GS) 71 or GS 75 (62), all fields were evaluated for the presence of FHB symptoms. For each variety and replication, 100 ears chosen at random were evaluated for the presence of *Fusarium* symptoms by an ordinal scoring system. The infected ear areas were classified into severity classes 1, 2, 3, 4, or 5, meaning approximate infestation of 0%, 1 to 25%, 26 to 50%, 51 to 75%, or 76 to 100%, respectively. The disease index (DI) was then computed as:

$$DI = \frac{0n_1 + 1n_2 + 2n_3 + 3n_4 + 4n_5}{4n} \times 100\%$$

with *n* the number of evaluated ears and *n_i* the number of ears in class *i* (26). At harvest in August, the DON content was measured by enzyme-linked immunosorbent assay (ELISA). More details about the ELISA protocol can be found in Audenaert et al. (5). The number of ears in each disease class, the DI and DON content, together with the agronomic and weather variables are stored in a relational database. An in-depth descriptive statistical analysis of this dataset has been presented in Landschoot et al. (33). This analysis identified new variables that play an important role in FHB incidence and DON content. It demonstrated the effect of species interactions on DON content and weather conditions during the vegetative growth stage. Temperature from November until May was positively correlated with FHB and DON content, whereas relative humidity was negatively correlated with DON content in November and December (33).

The agronomic variables in the dataset included wheat variety resistance (resistance class 1 to 5), soil type, sowing date, harvest date, soil management technique, and previous crop. The previous crops were host plants (maize or wheat) or nonhost plants (onions, beans, beets, potatoes) for *Fusarium* spp. The weather variables included time series of weather observations for all years since

Table 1. Locations of the 78 experimental field trials from 2002 to 2010 in Belgium

2001-2002	2002-2003	2003-2004	2004-2005	2005-2006	2006-2007	2007-2008	2008-2009	2009-2010
Assent	Bottelare	Bottelare	Helkijn	Bottelare	Bottelare	Bottelare	Bottelare	Bottelare
Bottelare	Gistel	Gistel	Koksijde	Koksijde	Koksijde	Koksijde	Ciney	Ciney
Gingelom	Helkijn	Helkijn	Poperinge	Lierde	Lierde	Leefdaal	Havelange	Havelange
Helkijn	Poperinge	Poperinge	Tongeren	Poperinge	Poperinge	Lierde	Hollogne	Hollogne
Koksijde	Tongeren	Tongeren	Verrebroek	Tongeren	Tongeren	Poperinge	Koksijde	Koksijde
Nieuwenhove	Zuienkerke	Verrebroek	Zuienkerke	Tongeren	Verrebroek	Verrebroek	Lintor	Lintor
Poperinge	Zwevegum	Zuienkerke	Zwevegum	Verrebroek	Zuienkerke	Zuienkerke	Poperinge	Poperinge
Zuienkerke		Zwevegum		Zuienkerke	Zwevegum	Zwevegum	Tongeren	Tongeren
Zwevegum				Zwevegum			Verrebroek	Verrebroek
							Zuienkerke	Zuienkerke
							Zwevegum	Zwevegum

2001, including rainfall, temperature, leaf wetness duration, air pressure, wind speed, and relative humidity. Daily weather data were collected by automated weather monitoring equipment of 'Proefcentrum Aardappelteelt (PCA)' near (<5 km) the field experiments, and these data were supplemented with weather data from Wolfram Mathematica 7.0 (Wolfram Research, Champaign, IL). From these daily weather variables, summary statistics, such as monthly averages, minima, maxima, and percentiles (10th, 25th, 50th, 75th, 90th), were computed to extract the most relevant information (33). In addition, the number of freezing and rainy days per month were calculated. The tenth and ninetieth percentiles of weather variables can be used to estimate the impact of extreme weather conditions (22).

Using this approach, more than 300 predictor variables were acquired. In an attempt to reduce the dimensionality and the complexity of the data, we retained only a subset of these variables for constructing predictive models. Hence, in a first step, a correlation analysis revealed the most important weather variables for our models. To this end, Pearson correlation coefficients were calculated for all combinations of the various weather variables in the different months. Per month, five weather variables with sufficiently low mutual correlations ($r < 0.7$) and a significant contribution in explaining the variance of DON content were selected for further analysis. These variables not only included the weather conditions around flowering time, but weather conditions from the whole year.

The selected variables were used to construct five types of models (linear regression, ridge regression, regression trees, boosting, and support vector regression) to predict DON content and visual disease symptoms by means of the DI. These models were validated with the four cross-validation strategies presented below.

Regression techniques. In this study, different predictive modeling techniques were applied. We focused on comparing five well-known regression techniques, as DON content and visual observations of FHB severity both lead to the prediction of a continuous response variable. These five models were multiple linear regression, ridge regression, regression trees, gradient boosting, and support vector machines.

Formally, the multiple linear regression model can be represented as follows:

$$Y = \beta_0 + \sum_{j=1}^p \beta_j X_j$$

where Y denotes the response variable, β_0 the intercept, β_j the j -th slope parameter, X_j the j -th predictor variable, and p the number of predictor variables.

The slope parameters β_j represent the independent contributions of each predictor variable to the prediction of the response variable, without taking potential interactions between predictor variables into account. In a simple linear regression setting, one aims to minimize the residual sum of squares (RSS) (7,23), which is calculated as:

$$RSS(\beta) = \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2$$

with N the sample size and p the number of predictor variables.

Ridge regression (23) is analogous to multiple linear regression, but in ridge regression a penalized residual sum of squares is minimized:

$$RSS(\beta, \lambda) = \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

where λ serves as a regularization parameter that requires fine-tuning for optimal predictive performance.

Regression trees (23) were analyzed as a third type of predictive model, which allows construction of nonlinear functions, unlike the two former regression techniques. Trees were inferred using a

process known as binary recursive partitioning. The algorithm recursively splits the data into two groups based on a splitting rule. Through the partitioning, one intends to increase the homogeneity of the two resulting subsets or nodes, based on the response variable. The partitioning stops when no splitting rule can improve the homogeneity of the nodes significantly.

Fourth, gradient boosting (19,23) was used, which is a relatively new and powerful data mining technique that has shown considerable success in predictive modeling (35,36). It involves training multiple models via an iterative procedure, in which the error function used to train a model in a given iteration depends on the performance of the model fitted in the preceding iteration. This can produce substantial improvements in performance compared to the use of a single model. In boosting, each base regression model is inferred using a weighted replicate of the training set, in which the weight of each observation depends on the error obtained in the previous iteration for that observation. In particular, observations that obtain a high error for one of the base regression models receive a greater weight when used to train the next regression model. Once all the base regression models have been trained, their predictions are combined by means of weighted averaging (7,23).

Finally, support vector machines (SVMs), a class of universal approximators based on statistical learning theory and quadratic programming, were analyzed. SVMs support both regression and classification tasks and can handle thousands of predictor variables of quite different natures. In a regression context, a quadratic program is solved to minimize a so-called ϵ -insensitive error function that is known to be a better alternative to the least-squares error function in terms of robustness to outliers. This setting is known as support vector regression (SVR). Moreover, SVMs can infer complex nonlinear functions by mapping the predictor variables to a high-dimensional space that can be represented by means of kernel functions. We used two common types of kernel functions: linear kernels, which lead to linear models, and Gaussian radial basis function (RBF) kernels, which represent nonlinear models by mapping the predictor variables to a high-dimensional space (11,51,52).

Evaluation strategies. Four different cross-validation strategies were empirically compared in the case study. First, the conventional random K -fold cross-validation strategy was considered, consisting of the following three steps (7,15,23,59):

- i. The observations were subdivided into K non-overlapping subsets, then for each of the K subsets, a predictive model was fit to the whole dataset without using the observations of the subset under consideration.
- ii. Observations in this subset were used to evaluate the predictive performance.
- iii. A global estimate of the predictive performance of the considered predictive modeling technique was obtained by averaging the results obtained for every individual subset.

Leave-one-out cross-validation refers to the specific situation where the number of folds K equals the sample size N , so that every observation gives rise to a separate test set.

A major shortcoming of random K -fold cross-validation is that it does not take potential year or location effects into account, because observations from the same year or the same location will end up in different folds. Consequently, this strategy does not evaluate the predictive performance for future years or new locations, and only the performance for past years and known locations can be reliably estimated, because observations from the same year or the same location are simultaneously used for model construction and model evaluation. As an alternative, the cross-year cross-location evaluation procedure is presented and can be summarized as follows:

- i. First, the available data were grouped per year and per location. In this way, $K \times L$ subsets could be distinguished for a dataset with observations from K years and L locations. It is important to include all locations and years. However, some of these subsets contained no data since not all locations are analyzed in all years.

- ii. Second, for each of the nonempty subsets, a predictive model was fit to the whole dataset without using the observations of the subsets from the same year or the same location.
- iii. Subsequently, the observations in the subset under consideration were used to evaluate the predictive performance.
- iv. Finally, a global estimate of the performance of the considered predictive modeling technique was obtained by weighted averaging of the results obtained for every individual subset, in which the weights were proportional to the number of observations in each subset (8,23).

In addition, we analyzed two simpler evaluation strategies, which allow assessing the predictive power for new years or locations, one at a time, but not for new years and new locations simultaneously. We refer to these strategies as cross-year validation and cross-location validation, respectively. They have been considered in Chakraborty et al. (10) to predict the severity of anthracnose disease of the tropical pasture legume *Stylosanthes scabra* and in Kaundal et al. (27) to develop weather-based prediction models for rice diseases. Cross-year and cross-location validation can be described as follows:

- i. First, the available data were grouped per year (location). In this way, K subsets were distinguished for a dataset with observations from K years (locations).
- ii. Second, for each subset with observations from a specific year (location), a predictive model was fit to the whole dataset without using the observations of the subset under consideration.
- iii. Subsequently, the observations in this subset were used to evaluate the predictive performance of the model.
- iv. Finally, a global estimate of the performance of the considered predictive modeling technique was obtained by weighted averaging of the results obtained for every individual subset, in which the weights were proportional to the number of observations in each subset (8,23).

All four cross-validation strategies were implemented using the R software for statistical computing, version 2.10.1 (21). Let us consider a simple example of data collected during 5 years at six locations to illustrate the differences among the four strategies (Fig. 1). The random cross-validation strategy subdivided the data

into K folds by ignoring year and location information. The cross-year validation strategy fits five models, for which four of the five subsets are used in the model construction phase. In contrast, the cross-location validation strategy fits six models, using five of the six subsets. Finally, the cross-year cross-location validation strategy fits $5 \times 6 = 30$ models, for which $4 \times 5 = 20$ subsets are used in model fitting and only one specific subset is used for evaluation. The remaining nine subsets were not considered in a particular iteration of the loop. In all four scenarios, more than one model was constructed for evaluating the modeling process. Subsequent to evaluation, one final model was constructed, which was then applied in practice. Hence, in predictive modeling, it is not the final model that is evaluated, but the modeling process. Cross-validation is a technique for assessing how the results of a statistical analysis will generalize to an independent dataset. Cross-validation can be used to optimize model parameters and to compare the performances of the different regression techniques.

Results

First, the results for predicting DON content were analyzed (Table 2). The random cross-validation strategy resulted in the lowest MSEs, ranging from 0.036 (SVR with RBF kernel) to 0.046 (regression tree and SVR with linear kernel) and the highest R^2 , ranging from 0.612 (SVR with RBF kernel) to 0.509 (regression tree). Subsequently, the more specific cross-validation strategies were compared. These three cross-validation strategies gave rise to substantially higher MSEs and lower R^2 values. For the cross-year cross-location validation strategy, the MSE ranged from 0.073 (ridge regression) to 27,883,943 (linear regression), and the R^2 ranged from 0.239 (regression tree) to <0.001 (SVR with RBF kernel). For the cross-year validation strategy, the MSE ranged from 0.040 (SVR with RBF kernel) to 84,910 (linear regression), and the R^2 ranged from 0.212 (ridge regression) and 0.026 (linear regression). For the cross-location validation strategy, the MSE ranged from 0.049 (SVR with RBF kernel) to 2,785 (linear regression), and the R^2 ranged from 0.277 (ridge regression) to 0.016 (linear regression). Indeed, the differences between the cross-validation strategies were most prominent for the multiple linear regression technique, where the MSEs were extremely

	Location 1	Location 2	Location 3	Location 4	Location 5	Location 6
Year 1	Train	Train	Train	Train	Test	Train
Year 2	Train	Train	Train	Train	Test	Train
Year 3	Train	Train	Train	Train	Test	Train
Year 4	Train	Train	Train	Train	Test	Train
Year 5	Train	Train	Train	Train	Test	Train

A

	Location 1	Location 2	Location 3	Location 4	Location 5	Location 6
Year 1	Train	Train	Train	Train	Train	Train
Year 2	Train	Train	Train	Train	Train	Train
Year 3	Train	Train	Train	Train	Train	Train
Year 4	Test	Test	Test	Test	Test	Test
Year 5	Train	Train	Train	Train	Train	Train

B

	Location 1	Location 2	Location 3	Location 4	Location 5	Location 6
Year 1	Train	Train	Train	Train	Omitted	Train
Year 2	Train	Train	Train	Train	Omitted	Train
Year 3	Train	Train	Train	Train	Omitted	Train
Year 4	Omitted	Omitted	Omitted	Omitted	Test	Omitted
Year 5	Train	Train	Train	Train	Omitted	Train

C

Fig. 1. A schematic representation of the **A**, cross-location validation strategy, **B**, cross-year validation strategy, and **C**, cross-year cross-location validation strategy. For one iteration of the respective cross-validation loops, the division into train and test set is shown on an artificial dataset collected in 5 years and at six locations. In cross-location validation, one location is used as test set and the other locations are used as training set. In cross-year validation, one year is used as test set and the other years are used as training set. In cross-year cross-location validation, all information from a particular year and location is omitted (dark gray) to build a model to make predictions for this year-location combination (indicated in gray), so only the white part of the dataset is considered to construct the model.

high (27,883,943) and the R^2 values were extremely low (<0.001) (Table 2). In the multiple linear regression technique, no regularization mechanism was used for tuning the complexity of the model. This problem was rectified in the ridge regression technique, in which the regularization parameter λ controls the effective complexity of the model.

Fig. 2. Influence of the penalty term in the ridge regression model on the prediction of deoxynivalenol (DON) content with cross-year cross-location validation.

The greater the regularization parameter, the smaller the range of predictions (Fig. 2). For example, all predictions lay between -0.23 mg/kg and 0.64 mg/kg when $\lambda = 1,000$ (without regularization in cross-year cross-location validation, predictions for DON content were situated between -941 mg/kg and $46,921$ mg/kg). When comparing support vector regression with a linear kernel and an RBF kernel, the RBF kernel performed better than a linear kernel. SVR with a Gaussian RBF kernel resulted in a MSE ranging from 0.036 to 0.100 , whereas SVR with a linear kernel resulted in a MSE ranging from 0.046 to 7.210 .

For the SVR model, validated with random cross-validation (Fig. 3A), DON predictions were situated between -0.06 mg/kg and 1.64 mg/kg; for the ridge regression technique validated with cross-year cross-location validation, all predictions ranged from -0.23 mg/kg to 0.64 mg/kg (Fig. 3B). The lower R^2 for the ridge regression technique resulted in points randomly distributed around the line $x = y$, while for random cross-validation, the points followed better the line $x = y$. In cross-year cross-location validation, the variance of the predictions for one year was smaller than for the random cross-validation, e.g., the predictions with random cross-validation for the year 2007 were situated between 0.08 and 0.87 mg/kg (Fig. 3A), while the predictions with cross-year cross-location validation varied between 0.44 and 0.64 mg/kg (Fig. 3B). This is because in cross-year cross-location validation, all DON contents for 1 year and location were predicted simultaneously with the same model, while in random cross-validation, predictions for a particular year and location were made with different models in the various cross-validation rounds.

Next, the performance of the five modeling techniques for predicting the DI, validated with the four cross-validation strategies, was analyzed (Table 3). In general, similar trends as for DON pre-

Table 2. Effect of regression technique^a and cross-validation strategy^b on the mean squared error (MSE) and the coefficient of determination (R^2) for predicting deoxynivalenol (DON) content (mg/kg) in winter wheat^c

	Linear regression		Ridge regression		Regression tree		Boosting		SVR linear		SVR RBF	
	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
R	0.039	0.579	0.039	0.579	0.046	0.509	0.041	0.560	0.046	0.540	0.036	0.612
CYL	27883943	0.001	0.073	0.221	0.100	0.239	0.073	0.216	4.230	0.005	0.100	0.000
CY	84910	0.026	0.076	0.212	0.147	0.010	0.078	0.161	7.210	0.015	0.040	0.103
CL	2785	0.016	0.067	0.277	0.082	0.140	0.080	0.153	1.820	0.029	0.049	0.094

^a Multiple linear regression, ridge regression, regression tree, boosting, and support vector regression (SVR) with linear and Gaussian RBF kernels.

^b Random cross-validation (R), cross-year cross-location validation (CYL), cross-year validation (CY), and cross-location validation (CL).

^c The best results for the cross-year cross-location validation strategy are indicated in bold. The variance of DON content was 0.0972 .

Fig. 3. Impact of the cross-validation strategy used on predictions for deoxynivalenol (DON) content (mg/kg). **A**, Results obtained with random cross-validation (support vector regression [SVR] with radial basis function [RBF] kernel), **B**, results obtained with cross-year cross-location validation (ridge regression). The black line is the line $x = y$; points above this line represent underestimations and under this line represent overestimations.

diction with DI were observed. Random cross-validation resulted in the lowest MSE, ranging from 29 (regression tree and boosting method) to 39 (SVR with linear kernel), and the highest R^2 values, ranging from 0.797 (regression tree) to 0.740 (SVR with linear kernel). In the random cross-validation setting, the differences between the different regression techniques were very small. For the more advanced cross-validation strategies, clear differences could be observed. For the cross-year cross-location validation strategy, the MSE ranged from 55 (ridge regression) to 26,289 (linear regression), and the R^2 ranged from 0.621 (ridge regression) to <0.001 (linear regression and SVM with linear kernel). For the cross-year validation strategy, the MSE ranged from 135 (boosting method) to 433,207 (linear regression) and the R^2 ranged from 0.085 (boosting method) to 0.015 (multiple regression). The cross-location validation strategy resulted in lower MSEs and higher R^2 values than the cross-year and cross-year cross-location validation strategies. The MSE ranged from 61 (ridge regression) to 30,667 (linear regression) and the R^2 ranged from 0.582 (regression tree) to 0.003 (linear regression) (Table 3). For random cross-validation, the points follow more or less the line $x = y$, while for cross-year cross-location validation, the points are more randomly distributed, yet an increasing trend can be observed (Fig. 4).

Discussion

The potential for high yield losses through the presence of FHB have resulted in many studies to predict FHB symptoms and DON contamination. Most of these approaches have been site and year specific (37,38,44). The key finding of this study has been the demonstration of a new cross-year cross-location validation strat-

egy that can be used in disease prediction. This strategy enables the evaluation of the predictive performance of a forecasting system for years and locations different from the years and locations on which the forecast has been developed.

The predictive performance of random K -fold cross-validation was comparable to the performance of existing forecasting systems (25,38,46,54). However, it remains problematic to compare results of various forecasting systems developed for different geographical areas that are all characterized by specific properties. Here it was clearly demonstrated that the commonly used K -fold cross-validation strategy suffers from a substantial optimistic bias. The traditional multiple linear regression model provided the best illustration of potential pitfalls in evaluation. For this model, the largest discrepancy between random cross-validation and the other cross-validation strategies was observed. However, the same phenomenon occurred for the more advanced predictive models, both for DON content prediction and DI prediction. The optimistic random K -fold cross-validation strategy sketched an unreliable picture of the future performance of the model, while more specific cross-validation strategies such as cross-year cross-location validation were capable of achieving a more trustworthy estimate of the predictive performance for future years and locations.

Apart from the observed differences between the four cross-validation strategies, we also concluded that more advanced predictive models developed in areas like data mining and machine learning were able to outperform traditional multiple linear regression models substantially. For random cross-validation, the performance of multiple linear regression, ridge regression, regression trees, boosting, and SVR was comparable. With random cross-validation, we

Table 3. Effect of regression technique^a and cross-validation strategy^b on the mean squared error (MSE) and the coefficient of determination (R^2) for predicting the DI (%) in winter wheat^c

	Linear regression		Ridge regression		Regression tree		Boosting		SVR linear		SVR RBF	
	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
R	36	0.753	31	0.782	29	0.797	29	0.796	39	0.740	32	0.783
CYL	26289	0.000	55	0.621	64	0.267	61	0.576	1547	0.000	200	0.605
CY	433207	0.015	151	0.024	273	0.074	153	0.085	2104	0.025	200	0.066
CL	30667	0.003	61	0.578	62	0.582	66	0.558	259	0.167	85	0.510

^a Multiple linear regression, ridge regression, regression tree, boosting, and support vector regression (SVR) with linear and Gaussian RBF kernels.

^b Random cross-validation (R), cross-year cross-location validation (CYL), cross-year validation (CY), and cross-location validation (CL).

^c The best results for the cross-year cross-location validation strategy are indicated in bold. The variance of the DI was 154.

Fig. 4. Impact of the cross-validation strategy used on predictions of the disease index (DI) (%). **A**, Results obtained with random cross-validation (regression tree), **B**, results obtained with cross-year cross-location validation (ridge regression). The black line is the line $x = y$; points above this line represent underestimations and points under this line represent overestimations.

achieved an MSE ranging from 0.036 to 0.046 and R^2 values ranging from 0.612 to 0.509 for DON content prediction model and an MSE ranging from 29 to 39 and R^2 values up to 0.797 for the DI prediction model. All these techniques yielded a similar performance for predictions within years and within locations. Conversely, when comparing our results for the other cross-validation strategies, we can see that multiple linear regression resulted in a very high MSE and a low R^2 , by comparison with the other techniques. This was due to over-fitting, because this method contains no complexity control mechanisms for the parameters, so the predictions for unseen data can become highly positive or negative (23). The ridge regression technique performed better since the penalty term controls for over-fitting, leading to lower and upper bounds for the possible values of the parameters, so that the predictions for new data lie in a narrower range. For DON content, the boosting technique resulted in the best predictions. Unlike the conclusions drawn in Kaundal et al. (27), support vector regression (with linear and RBF kernels) did not outperform the other techniques.

Cross-year validation resulted in a lower R^2 and a higher MSE than the cross-location strategy, so one can say that forecasting for new locations will be more accurate than forecasting for new years (cross-year validation). The cross-year cross-location validation strategy often performed slightly better than the cross-year validation strategy, except for the case of multiple linear regression. This was surprising, as models able to generalize for new years and new locations (cross-year cross-location validation) performed worse than models for a new year (cross-year validation). FHB severity and DON content highly depend on weather conditions (25,28,37,38,44,54,60), which are relatively similar within a small spatial resolution such as Belgium (30,528 km²), but highly temporally variable between years. Hence, predicting for new years and locations is of similar difficulty as predicting for new years. From a statistical point of view, the difference between both cross-validation strategies can be interpreted as insignificant. The location effect is likely to play a larger role if geographically larger areas are to be taken into account.

Furthermore, the results of our prediction models confirmed that in general it remains difficult to develop forecasting systems. Prediction models for visual disease symptoms were more accurate, which may be explained by the complex nature of DON production that follows infection when the appropriate *Fusarium* species and cultivars are present (50). Moreover, DON production is influenced by many biotic factors including host resistance, chemotype of isolates, aggressiveness of the prevalent fungal species, and interaction of the fungal species present (4,33). An in-depth discussion of different evaluation strategies can be found in Bishop (7) and Hastie et al. (23). In terms of estimating the predictive performance of a statistical model, one should look at the bias and the variance of the estimator used. Ideally, a performance estimator should be unbiased, meaning that it does not produce an optimistic or pessimistic estimate, and it should have a low variance, meaning that consistent results are to be expected when the experiments are repeated several times (7,23). In our case study, we observed that K -fold cross-validation yielded overly optimistic performance estimates for future years and new locations. The same optimistic bias can be expected if the same dataset is used for model construction and evaluation. Conversely, cross-year cross-location validation and splitting the data into a train and test set will both yield unbiased estimates.

Table 4. An overview of the statistical properties of different strategies in evaluating the predictive performance for future years and new locations

Strategy	Bias	Variance
One dataset	Very optimistic bias	High
Train and test set	Unbiased	Very high
Random K -fold cross-validation	Optimistic bias	Low
Cross-year cross-location validation	Unbiased	Low

The variance of the different evaluation strategies was not empirically analyzed, because it cannot be computed for real-world data. Simulation experiments on artificial data have to be conducted to get statistical insights into the variance. Cross-validation strategies may manifest a lower variance than train-test splitting procedures or strategies that employ the same dataset for training and testing. Hence, cross-year cross-location validation becomes the preferred strategy, because it has no bias and a low variance (7,23,59).

Many researchers who are confronted with plant disease forecasting will argue that the safest evaluation approach still consists of using a part of the available data for constructing the model and the remaining part for evaluating the model (1,17,24,29,40). These two separate datasets are commonly referred to as the training and test datasets (in statistics and predictive modeling [7,15,23,59]), or the calibration and the validation dataset (in more applied fields [18]). This splitting approach will be more useful if a lot of data can be collected, as it provides an unbiased estimate of the predictive performance, too, under the condition that training and test sets do not share observations from the same year or location. Nevertheless, the supply of data from many years and locations for training and testing will be limited in many applications, leading to a trade-off: as much data as possible should be used for estimating the parameters of the model accurately, while still a lot of data is needed for evaluation afterward. If the test set is small, it will lead to a performance estimate that is characterized by a high variance (6,15,23). Cross-validation strategies were introduced in the first place to circumvent this trade-off, and the most appropriate cross-validation strategy must be selected based on the amount of data available (8,23). If the amount of data is very small, optimistic leave-one-out cross-validation is the only way to perform a kind of validation (14,31). If sufficient data from different years and locations can be collected, cross-year or cross-location validation strategies should be overall preferred, because these strategies give unbiased predictions with a low variance (Table 4).

From these findings, we recommend the use of cross-year cross-location validation to assess plant disease prediction models in small-sample studies, as it uses the available data in an optimal way without biasing the conclusions. Its variance is smaller than train-test splitting, and it is unbiased for plant disease prediction since observations for a given year and location never participate in the model construction and the model evaluation process simultaneously.

Acknowledgments

This work was financially supported by the Flemish Institute for the Stimulation of Scientific-Technological Research in Industry project 70575 (IWT, Brussels, Belgium). Geert Haesaert, full professor at the Ghent University College, coordinates this study. Kris Audenaert is a postdoctoral fellow of the Ghent University Research Fund, and Willem Waegeman is supported as a postdoctoral fellow by the Research Foundation-Flanders. We thank the LCG Trial Network and the Centre wallon de Recherches agronomiques (CRA-W) for putting the wheat samples at our disposal.

Literature Cited

1. Aegerter, B. J., Nunez, J. J., and Davis, R. M. 2003. Environmental factors affecting rose downy mildew and development of a forecasting model for a nursery production system. *Plant Dis.* 87:732-738.
2. Airola, A., Pahikkala, T., Waegeman, W., De Baets, B., and Salakoski, T. 2011. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Comput. Stat. Data. An.* 55:1828-1844.
3. Andrade-Piedra, J. L., Forbes, G. A., Shtienberg, D., Grünwald, N. J., Chacón, M. G., Taipei, M. V., Hijmans, R. J., and Fry, W. E. 2005. Qualification of a plant disease simulation model: Performance of the LATEBLIGHT model across a broad range of environments. *Phytopathology* 95:1412-1422.
4. Audenaert, K., Callewaert, E., Höfte, M., De Saeger, S., and Haesaert, G. 2010. Hydrogen peroxide induced by the fungicide prothioconazole triggers deoxynivalenol (DON) production by *Fusarium graminearum*. *BMC Microbiol.* 10:112-125.
5. Audenaert, K., Van Broeck, R., De Witte, F., Heremans, B., Messens, K., Höfte, M., and Haesaert, G. 2009. Fusarium head blight (FHB) in Flanders: Population diversity, inter-species associations and DON contamination in commercial winter wheat varieties. *Eur. J. Plant Pathol.* 125:445-458.
6. Bengio, Y., and Grandvalet, Y. 2004. No unbiased estimator of K -fold cross-

- validation exists. *J. Mach. Lear. Res.* 5:1089-1105.
7. Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer, New York, USA.
8. Browne, M. W. 2000. Cross-validation methods. *J. Math. Psychol.* 44:108-132.
9. Camargo, A., and Smith, J. S. 2009. Image pattern classification for the identification of disease causing agents in plants. *Comput. Electron. Agric.* 66:121-125.
10. Chakraborty, S., Ghosh, R., Ghosh, M., Fernandes, C. D., Charchar, M. J., and Kelemu, S. 2004. Weather-based prediction of anthracnose severity using artificial neural network models. *Plant Pathol.* 53:375-386.
11. Cristianini, N., and Shawe-Taylor, J. 2000. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK.
12. De Wolf, E. D., and Franc, L. J. 2000. Neural network classification of tan spot and stagonospora blotch infection periods in a wheat field environment. *Phytopathology* 90:108-113.
13. Debusho, L. K. 2008. Ridge regression technique to determine the environmental influences on tef (*Eragrostis tef*) grain yield. *Afr. J. Plant Soil.* 25:135-138.
14. Del Ponte, E. M., Godoy, C. V., Li, X., and Yang, X. B. 2006. Predicting severity of Asian soybean rust epidemics with empirical rainfall models. *Phytopathology* 96:797-803.
15. Dietterich, T. 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation* 10:1895-1924.
16. Eiblmeier, P., and von Gleissenthall, L. 2007. Risk evaluation of deoxynivalenol levels in Bavarian wheat from survey data. *J. Plant Dis. Prot.* 114:69-75.
17. Fidanza, M. A., Dernoede, P. H., and Grybauskas, A. P. 1996. Development and field validation for a Brown patch warning model for perennial ryegrass turf. *Phytopathology* 86:385-390.
18. Galvão, R. K. H., Araujo, M. C. U., José, G. E., Pontes, M. J. C., Silva, E. C., and Saldanha, T. C. B. 2005. A method for calibration and validation subset partitioning. *Talanta* 67(4):736-740.
19. Ganatra, A. P., and Kosta, Y. P. 2010. Comprehensive evolution and evaluation of boosting. *Int. J. Comp. Theo. Eng.* 2:931-936.
20. Garrett, K. A., Madden, L. V., Hughes, G., and Pfender, W. F. 2004. New applications of statistical tools in plant pathology. *Phytopathology* 94:999-1003.
21. Gentleman, R., and Ihaka, R. 1997. <http://www.r-project.org/>. Vienna University of Economics and Business Administration, Vienna, Austria.
22. Genton, M. G., and Furrer, R. 1998. Analysis of rainfall data by simple good sense: Is spatial statistics worth the trouble? *J. Geogr. Inform. Decision Anal.* 2:11-17.
23. Hastie, T., Tibshirani, R., and Friedman, J. 2008. *The Elements of Statistical Learning*. Springer Series in Statistics. 2nd ed.
24. Henderson, D., Williams, C. J., and Miller, J. S. 2007. Forecasting late blight in potato crops of southern Idaho using logistic regression analysis. *Plant Dis.* 91:951-956.
25. Hooker, D. C., Schaafsma, A. W., and Tamburic-Ilincic, L. 2002. Using weather variables pre- and post-heading to predict deoxynivalenol content in winter wheat. *Plant Dis.* 86:611-619.
26. Isebaert, S., De Saeger, S., Devreese, R., Verhoeven, R., Maene, P., Heremans, B., and Haesaert, G. 2009. Mycotoxin-producing *Fusarium* species occurring in winter wheat in Belgium (Flanders) during 2002-2005. *J. Phytopathol.* 157:108-116.
27. Kaundal, R., Kapoor, A. S., and Raghava, G. P. S. 2006. Machine learning techniques in disease forecasting: A case study on rice blast prediction. *BMC Bioinform.* 7:485.
28. Klem, K., Váňová, M., Hajšlová, J., Lancová, K., and Sehnalová, M. 2007. A neural network model for prediction of deoxynivalenol content in wheat grain based on weather data and preceding crop. *Plant Soil Environ.* 53:421-429.
29. Koch, S., Dunker, S., Kleinhenz, B., Röhrig, M., and von Tiedemann, A. 2007. A crop loss related forecasting model for *Sclerotinia* stem rot in winter oilseed rape. *Phytopathology* 97:1186-1194.
30. Köhl, J., de Haas, B., Kastelein, P., Burgers, S. L. G. E., and Waalwijk, C. 2007. Population dynamics of *Fusarium* spp. and *Microdochium nivale* in crops and crop residues of winter wheat. *Phytopathology* 97:971-978.
31. Krzanowski, W., and Hand, D. 1997. Assessing error rate estimators: The leave-one-out method reconsidered. *Aust. J. Stat.* 39(1):35-46.
32. Landschoot, S., Audenaert, K., Waegeman, W., Pycke, B., Bekaert, B., De Baets, B., and Haesaert, G. 2011. Connection between primary *Fusarium* inoculum and the final population on wheat ears in Flanders. *Crop Prot.* 30:1297-1305.
33. Landschoot, S., Waegeman, W., Audenaert, K., Vandepitte, J., Baetens, J. M., De Baets, B., and Haesaert, G. 2011. An empirical analysis of explanatory variables affecting *Fusarium* infection and deoxynivalenol content in wheat. *J. Plant Pathol.* In press.
34. Lian, Y., Zhang, F., Wang, J., Joshi, T., and Wang, Y. 2011. Prediction of drought-resistant genes in *Arabidopsis thaliana* using SVM-RFE. *Plos ONE.* 6(7):e21750. doi:10.1371/journal.pone.0021750.
35. Liu, K. Y., Lin, L., Zhou, X., and Wong, S. T. C. 2005. Boosting alternating decision trees modelling of disease trait information. *BMC Genet.* 6:S132.
36. Morra, J. H., Zhuowen, T., Apostolova, L. G., Green, A. E., Toga, A. W., and Thompson, P. M. 2010. Comparison of AdaBoost and Support Vector Machines for detecting Alzheimer's disease through automated hippocampal segmentation. *Med. Imaging* 29:30-43.
37. Moschini, R. C., and Fortugno, C. 1996. Predicting wheat head blight incidence using models based on meteorological factors in Pergamino, Argentina. *Eur. J. Plant Pathol.* 112:211-218.
38. Moschini, R. C., Piolib, R., Carmonac, M., and Sacchid, O. 2001. Empirical predictions of wheat head blight in the northern Argentinean pampas region. *Crop Sci.* 41:1541-1545.
39. Parker, B. J., Gunter, S., and Bedo, J. 2007. Stratification bias in low signal microarray studies. *BMC Bioinform.* 8:326.
40. Paul, P. A., Lipps, P. E., and Madden, L. V. 2006. Meta-analysis of regression coefficients for the relationship between *Fusarium* head blight and deoxynivalenol content of wheat. *Phytopathology* 96:951-961.
41. Paul, P. A., and Munkvold, G. P. 2005. Regression and artificial neural network modelling for the prediction of gray leaf spot of maize. *Phytopathology* 95:388-396.
42. Pereyra, S. A., Dill-Macky, R., and Sims, A. L. 2004. Survival and inoculum production of *Gibberella zeae* in wheat residue. *Plant Dis.* 88:724-730.
43. Pfender, W. F. 2003. Prediction of stem rust infection favorability, by means of degree-hour wetness duration, for perennial ryegrass seed crops. *Phytopathology* 93:467-477.
44. Prandini, A., Sigolo, S., Filippi, L., Battilani, P., and Piva, G. 2009. Review of predictive models for *Fusarium* head blight and related mycotoxin contamination in wheat. *Food Chem. Toxicol.* 47:927-931.
45. Rajni, J., Minz, S., and Ramasubramanian, V. 2009. Machine learning for forewarning crop diseases. *J. Ind. Soc. Agril. Statist.* 63:97-107.
46. Rossi, V., Giosuè, S., and Delogu, G. 2003. A model estimating risk for *Fusarium* mycotoxins in wheat kernels. *Asp. Appl. Biol.* 68:229-234.
47. Rykiel, E. J. 1996. Testing ecological models: The meaning of validation. *Ecol. Model.* 90:229-244.
48. Sango, S., and Yang, X. B. 2004. Overview of selected multivariate statistical methods and their use in phytopathological research. *Phytopathology* 94:1004-1006.
49. Savary, S., Teng, P. S., Willocquet, L., and Nutter, F. W. 2006. Quantification and modeling crop losses: A review of purposes. *Annu. Rev. Phytopathol.* 44:89-112.
50. Schaafsma, A. W., and Hooker, D. C. 2007. Climatic models to predict occurrence of *Fusarium* toxins in wheat and maize. *Int. J. Food Microbiol.* 119:116-125.
51. Schölkopf, B., and Smola, A. 2002. *Learning with Kernels, Support Vector Machines, Regularisation, Optimization and Beyond*. The MIT Press, Cambridge, MA.
52. Shawe-Taylor, J., and Cristianini, N. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK.
53. Smith, D. L., Hollowell, J. E., Isleib, T. G., and Shew, B. B. 2007. A site-specific, weather-based disease regression model for *Sclerotinia* blight of peanut. *Plant Dis.* 91:1436-1444.
54. Van Der Fels-Klerx, H., Burgers, S. L. G. E., Booi, C. J. H. 2010. Descriptive modelling to predict deoxynivalenol in winter wheat in the Netherlands. *Food Addit. Contam.* 27:636-643.
55. van Maanen, A., and Xu, X. M. 2003. Modelling plant disease epidemics. *Eur. J. Plant Pathol.* 109:669-682.
56. Varma, S., and Simon, R. 2006. Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* 23:7-91.
57. Waalwijk, C., Kastelein, P., de Vries, I., Kerényi, Z., van der Lee, T., Hesselink, T., Köhl, J., and Kema, G. 2003. Major changes in *Fusarium* spp. in wheat in the Netherlands. *Eur. J. Plant Pathol.* 109:743-754.
58. Wang, Q., Zahn, T., Cui, J., Wang, X., Zhou, H., Han, J., and Gislum, R. 2011. Path and ridge regression analysis of seed yield and seed yield components of Russian wildrye (*Psathyrostachys juncea* Nevski) under field conditions. *PLoS ONE.* 6(4):e18245. doi:10.1371/journal.pone.0018245.
59. Witten, I. H., Frank, E., and Hall, M. A. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd ed. Morgan Kaufmann, Waltham, MA.
60. Xu, X. M. 2003. Effects of environmental conditions on the development of *Fusarium* ear blight. *Eur. J. Plant Pathol.* 109:683-689.
61. Xu, X. M., Parry, D. W., Nicholson, P., Thomsett, M. A., Simpson, D., Edwards, S., Cooke, B. M., Doohan, F. M., Brennan, J. M., Moretti, A., Tocco, G., Mule, G., Hornok, L., Giczey, G., and Tatnell, J. 2005. Predominance and association of pathogenic fungi causing *Fusarium* ear blight in wheat in four European countries. *Eur. J. Plant Pathol.* 112:143-154.
62. Zadoks, J. C., Chang, T. T., and Konzak, C. F. 1974. A decimal code for growth stages of cereals. *Weed Resist.* 14:415-421.