Genet Resour Crop Evol (2012) 59:1465–1481
DOI 10.1007/s10722-011-9775-5

RESEARCH ARTICLE

Focused identification of germplasm strategy (FIGS) detects wheat stem rust resistance linked to environmental variables

Abdallah Bari · Kenneth Street · Michael Mackay · Dag Terje Filip Endresen · Eddy De Pauw · Ahmed Amri

Received: 8 May 2011 / Accepted: 17 November 2011 / Published online: 3 December 2011

Abstract Recent studies have shown that novel genetic variation for resistance to pests and diseases can be detected in plant genetic resources originating from locations with an environmental profile similar to the collection sites of a reference set of accessions with known resistance, based on the Focused Identification of Germplasm Strategy (FIGS) approach. FIGS combines both the development of *a priori* information based on the quantification of the trait-environment relationship and the use of this information to define a best bet subset of accessions with a higher probability of containing new variation for the sought after trait(s). The present study investigates the development strategy of the *a priori* information using different modeling techniques including learning-based techniques as a follow up to previous work where parametric approaches were used to quantify the stem rust resistance and climate variables relationship. The results show that the predictive power, derived from the accuracy parameters and cross-validation, varies depending on whether the models are based on linear or non-linear approaches. The prediction based on learning techniques are relatively higher indicating that the non-linear approaches, in particular support vector machine and neural networks, outperform both principal component logistic regression and generalized partial least squares. Overall there are indications that the trait distribution of resistance to stem rust is confined to certain environments or areas, whereas the susceptible types appear to be limited to other areas with some degree of overlapping of the two classes. The results also point to a number of issues to consider for improving the predictive performance of the models.

Keywords Focused identification of germplasm strategy · Geographic information systems · Learning-based modeling techniques · Receiver operating characteristics · Wheat stem rust

Abbreviations

AUC	Area under the ROC curve
GPLS	Generalized partial least squares
GIS	Geographic information systems
NN	Neural networks
PCA	Principal component analysis
PCLR	Principal component logistic regression
PLS	Partial least squares
RF	Random forest
ROC	Receiver operating characteristics
SVM	Support vector machine

A. Bari · K. Street (✉) · E. De Pauw · A. Amri
International Center for Agricultural Research in the Dry Areas (ICARDA), Aleppo, Syria
e-mail: k.street@cgiar.org

M. Mackay
Bioversity International, Rome, Italy

D. T. F. Endresen
Nordic Genetic Resources Center (NordGen), Alnarp, Sweden

Introduction

Stem rust caused by the fungi *Puccinia graminis* f. sp. *tritici*, has re-emerged as major threat to wheat (*Triticum aestivum* L. and *Triticum turgidum* ssp. *durum* L.) following the appearance of new virulent races. Ug99, a particularly virulent strain (or TTKSK using North American race notation) of stem rust, was found in Uganda in 1999 and has increased the vulnerability of the global wheat yields (CIMMYT 2005; Vurro et al. 2010; Fehser et al. 2010). This virulent strain is spreading and has already taken its toll on wheat production in Sub-Saharan East Africa, Yemen and Iran and is now threatening Central Asia and the Caucasus; an area that accounts for 37% of global wheat production (Vurro et al. 2010; Fehser et al. 2010). Because of its virulence and spread it has attracted both public and research community attention worldwide and global efforts to track and monitor its expansion are underway to counter its potential impact (Kolmer 2005; Vurro et al. 2010; Hodson and DePauw 2011). Plants usually react to virulent strains through the so-called R (resistance) genes group (Eckardt 2001). In the case of wheat, there are about 45–50 genes, known as Sr genes, which confer resistance to different races of stem rust (McIntosh et al. 2008, 2010; Vurro et al. 2010). Deployment of new sources of resistance to stem rust has been made a top priority by the Borlaug Global Rust Initiative that was established in 2005 (http://www.globalrust.org).

The appearance of a new virulence for a crop disease is a typical recurring scenario in agricultural production that can lead to severe yield losses (Qualset 1975; Leonard and Szabo 2005; Vurro et al. 2010; Fehser et al. 2010). Utilizing novel disease resistance genes found in *ex situ* germplasm collections, or genebanks, can help to avert these losses (Qualset 1975). Genetic resources, such as crop landraces and wild relatives, represent potential sources for pest and disease resistance critical for the stability and sustainability of global production (Dinoor 1975; Bonman et al. 2005). However such novel variation is often rare and may not be captured in a representative or fixed collections of germplasm such as core collections (Brown and Spillane 1999; Polignano et al. 2001; Gepts 2006; Dwivedi et al. 2007; Pessoa-Filho et al. 2010; Xu 2010). The need to rationalize the search for rare adaptive variation has led to the use of alternative approaches including the development of specific or thematic genetic resource collections (Gollin et al. 2000; Gepts 2006; Dwivedi et al. 2007; Pessoa-Filho et al. 2010; Xu 2010). Recent trait-based approaches to selecting germplasm from genebanks have shown that they are more likely to provide useful and novel genes (Street et al. 2008; Mackay and Street 2004; El-Bouhsini et al. 2009, 2010; Bhullar et al. 2009).

By using the eco-geographical data of a reference dataset of accessions with resistance to the sought after adaptive trait, such as resistance to either diseases or pests, the Focused Identification of Germplasm Strategy (FIGS) has successfully helped to identify a number of novel genes in germplasm from environmentally similar sites to those of the reference/template dataset (Mackay 1995; Mackay and Street 2004; El-Bouhsini et al. 2009, 2010; Bhullar et al. 2009). Relationships between adaptive traits and collection site environmental parameters have also been revealed by recent studies using multi-variate and multi-way models such as *N*-PLS (multi-linear Partial Least Squares) (Endresen 2010; Endresen et al. 2011). Modeling of stem rust resistance using geographical information system (GIS) approaches has also led to the detection of a relationship between geographical areas and incidence of resistance to stem rust (Bonman et al. 2007). Furthermore, some of the traits that have been found to carry strong climatic signals in wild species are being used to model the impact of climate changes (Barboni et al. 2004; Webb et al. 2010). However, although there have been trait-environment studies in the past, they were generally limited to a single or a small group of environmental variables (Pakeman et al. 2009).

FIGS is a trait-based and user-driven approach to select potentially useful germplasm for crop improvement. It searches for specific sought-after traits, using as surrogate the environment, based on the hypothesis that the germplasm is likely to reflect the selection pressures of the environment from which it was originally sampled (Mackay 1990, 1995; Mackay and Street 2004). The FIGS approach addresses the lack of available evaluation data as well as the temporal (the moment when the accession is evaluated) issue of evaluation as reported by Koo and Wright (2000). In a simulation of the economic impact of disease manifestation Koo and Wright (2000) found that it was faster to develop an improved variety by incorporating

novel resistant traits, provided the source of the resistance gene has already been identified. In terms of searching genetic resource collections for useful traits, Gollin et al. (2000) developed a theoretical model based on resistance to diseases and insects, including the Russian wheat aphid (RWA). The model highlighted that the search for a desirable trait is of equal importance to the process of transferring it into improved backgrounds. In their findings a focused search approach, which they called a specialized knowledge case, contributes positively to expected net benefits due to the increased probability of finding the desirable material and the associated cost savings. FIGS as a focused approach combines both the development of *a priori* information (dataset template or specialized knowledge as per Gollin et al. (2000)) based on the quantification of the trait-environment relationship and the use of this information to define a subset of accessions with a higher probability of containing the sought after genetic variation for adaptive traits (Mackay 1995; Mackay and Street 2004).

The distribution patterns of the adaptive trait might be, as in the case for taxonomic species distributions, the result of ecological and evolutionary factors, including, but not limited to, environmental factors, natural selection and local selection pressures that are hard to quantify, such as interactions with humans. However, according to Qualset (1975) and Dinoor (1975), disease resistance traits are more likely to be influenced by natural selection and thus have a restricted distribution. In this context, Hakes and Cronin (2011) assert that trait distribution patterns are not random and could be geographically and spatially structured. Compared to species distribution patterns, very few studies have been undertaken to identify key factors that influence specific trait distribution patterns (Chuine 2010). Furthermore, recent studies have shown that modeling the distribution for specific traits can improve the quality and predictive performance of plant species distribution models (Hanspach et al. 2010).

The objective of this research was to detect whether there is a link between stem rust resistance and climate, the results of which will be used to (1) develop a subset of germplasm accessions with an increased probability of finding resistance to stem rust and (2) develop algorithms to use in subsequent applications of FIGS for 'trait mining' of large germplasm collections. Five modeling techniques were tested to quantify the hypothesized link. These included both parametric and non-parametric approaches as well as machine learning methods. The overall assumption is that the novel genetic variation will be confined to areas with similar environmental profiles to sites where stem rust resistance has been previously found.

Methods

The R language (R Development Core Team 2011) was used as a platform for the preparation and analysis of data. The data consisted of stem rust scores for bread and durum wheat accessions (trait data) and environmental or site data (climate data) describing where the accessions were originally sampled.

The trait data

The stem rust trait data used in this paper to develop the FIGS *a priori* information was taken from The United States Department of Agriculture (USDA) National Genetic Resources Program (NGRP) GRIN database. The data is an accumulation of results over six different years (during 1988–1994) in two different research stations in USA; the University of Minnesota Agricultural Research Station (44°59′17″ N, 93°10′48 W) and the Rosemount USDA Agricultural Research Station north of USA (44°43′01″ N, 93°05′56″ W). Dr Don V. McVey made all of the trait observations for both locations (Bonman et al. 2007; Endresen et al. 2011).

The accessions screened for stem rust originated from 2013 collection sites as one site was removed from the original data of 2,014 sites. Some of the sites lacking geographical coordinates were geo-referenced at ICARDA based on a description of the original collecting sites. The probability distribution for the disease scores shows that the number of susceptible accessions is more dominant than the number of resistant accessions. Cross-tabulation was carried out to assign a site's (i) trait attributes using the expected frequency (e_{ik}) of each score per site times the actual score count (y_{ik}) per site of each of the scores, k, (0–9) across all sites. The sites were then compared based on the frequency of either resistant accessions (0–4 scores) or susceptible accessions (5–9 scores).

$$Y_i = \begin{cases} 1, & \sum_{k=0}^{4} e_{ik} y_{ik} \geq \sum_{k=5}^{9} e_{ik} y_{ik} \\ 0, & otherwise \end{cases}$$

This was based on the results reported by Endresen et al. (2011) where they reclassified the trait states into 2 groups from 9 groups and used single sites as observations such that the score that is most common would be the score that is attributed to each site.

The climate data

The climate data was extracted from climatic maps generated from station data by co-splining (Hutchinson and Corbett 1995), a local interpolation method. The 'thin-plate smoothing spline' method of Hutchinson (1995), as implemented in the ANUSPLIN software (Hutchinson 2000), was used to convert the point climate data into climate surfaces. This is a smoothing interpolation technique in which the degree of smoothness of the fitted function is determined automatically from the data by minimizing a measure of the predictive error of the fitted surface, as given by the generalized cross-validation (GCV). The GCV is calculated by removing each data point and calculating the residual from the omitted data point of a surface fitted to all other data points using the same smoothing parameter value. The thin-plate smoothing spline method including GCV has been proven to be effective as it improves the accuracy of the interpolation similar to that of cokriging interpolation when the appropriate variogram is well selected (Wood 2000; Tait and Turner 2005; Wratt et al. 2006).

The generation of climatic maps was based on the use of terrain variables as auxiliary variables in the interpolation process, whereby these variables were first converted into digital elevation models (DEM) using GIS software. In contrast to the climatic target variables themselves, which are only known for a limited number of sample points, terrain variables have the advantage to be known for all locations in between. In addition, some climatic variables, such as temperature and precipitation, are highly correlated with elevation, which increases the precision of the interpolated values significantly (De Pauw et al. 2000).

The DEM used to generate the climate surfaces was GTOPO30, a global DEM with 30 arc-second (approximately 1 km) resolution (Gesch and Larson 1996). Parameter estimation was undertaken over a regular grid with the same dimensions and resolution as the user-provided DEM. The combination of point climatic data and terrain, in the form of a DEM, allows generating spatially or temporally linked derived variables, such as potential evapotranspiration (*pet*) and aridity (*ari*) index (Table 1). From the climate maps a total of 60 climatic variables were extracted for the georeferenced 2,013 locations, where accessions scored for stem rust, were originally sampled. These 60 variables represent monthly average minimum temperature (*tmin*), monthly average maximum temperature (*tmax*), monthly average precipitation (*prec*), monthly average evapo-transpiration (*pet*) and monthly average aridity index (*ari*) (Table 1).

Data preparation and data exploration

Prior to the analysis of the climate data each variable was examined individually. The climate variables appear to be mostly right- skewed and transformations were performed to tackle both the skewness and the different measurement scales to avoid the discrepancy between large and small values. Normality was not necessarily a prerequisite for some of the modeling techniques explored. Among the transformations used was the Box–Cox transformation, which is a power transformation included in Tukey's original family of transformations that use logarithmic transformations when the power value is equal to 0 (Tukey 1957; Osborne 2010). The Box–Cox transformation algorithm was applied individually to the aridity (*ari*), the precipitation (*prec*) and the evapo-transpiration (*pet*) monthly variables based on the equation:

$$f_\lambda(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(x) & \lambda = 0 \end{cases}$$

where λ power (of x) value is chosen to reduce non-normality by maximizing the $l(\lambda)$ function:

$$l(\lambda) = -\frac{n}{2}\log_e\left[\frac{1}{n}\sum\left(x_j^\lambda - \bar{x}^\lambda\right)^2\right] + (\lambda - 1)\sum_{j=1}^{n}\log_e(x_j)$$

$\bar{x}^\lambda$ is defined as the average of the newly transformed variables (Box and Cox 1964; Osborne 2010).

Table 1 The environment variables used in the study

Variable type	Variable name	Variable description	Unit	Number	Transformation (power value λ)[a]
Climatic	pet	Monthly potential evapo-transpiration	mm	12	[−0.40, 0.3]
	ari	Monthly moisture index (ari)		12	[0.12, 0.42]
	prec	Monthly precipitation cm	mm	12	[0.17, 0.56]
	tmin	Monthly minimum temperature	°C	12	–
	tmax	Monthly maximum temperature	°C	12	–
Geographic/topographic	lon	Longitude	°	1	–
	lat	Latitude	°	1	–
	alt	Elevation	m	1	–

[a] The transformation are based on the Box–Cox power transformation carried out for the skewed variables (*prec*, *pet* and *ari*)

Through this process the best option is selected from a range of transformations (Osborne 2010).

As a result of the Box–Cox transformation process the aridity variables were transformed with λ values ranging from 0.12 to 0.42, the precipitation variables with λ values ranging from 0.17 to 0.56 and the evapo-transpiration variables with λ values ranging from −0.40 to 0.3 (Table 1). Only the best value of λ, not tabulated here, for each of the 60 variables were used. Figure 1 demonstrates the effectiveness of this transformation. Mean annual temperatures (*tmax* and *tmin*) were not transformed since their distribution was approximately normal (Fig. 1).

All variables were standardized to a mean of zero and a standard deviation of 1. After the transformation the data was standardized and a comparison made between the transformed and non-transformed data. This data pre-processing was systematically and automatically carried out through the different models.

Uni-variate analysis (ANOVA) was carried out to detect the discriminating ability for each variable individually. In the ANOVA, the trait data (stem rust scores Y) was used as an independent variable while climate variables (X) were as dependent variables. Most of the variables taken individually showed significant discrimination between the different stem rust trait groups. Collinearly among the variables (X) was expected as indicated by its extremely high condition number value, which is equal to the square root of the largest eigenvalue divided by the smallest eigenvalue (Belsley 1991).

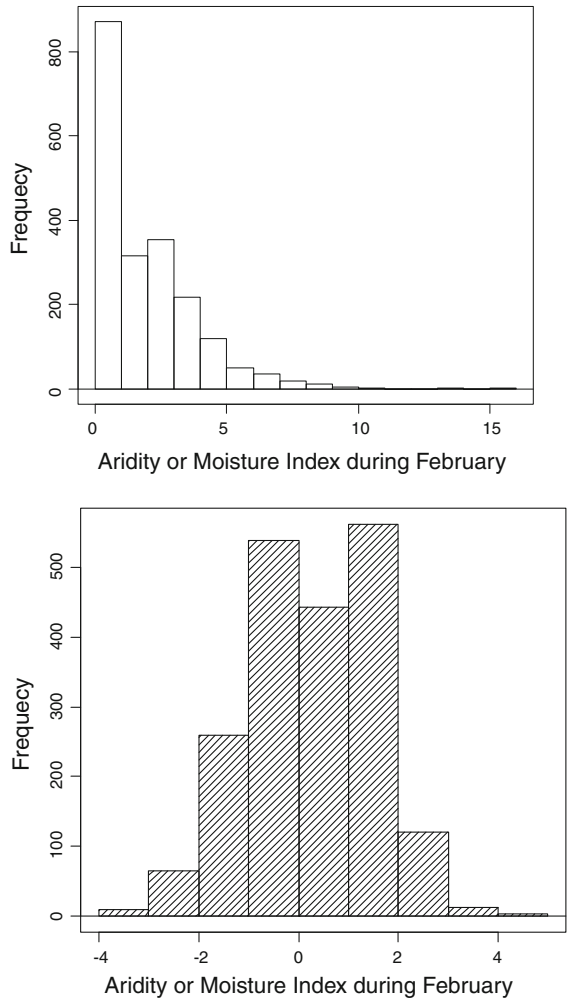

Fig. 1 Frequency histograms of climate (aridity/moisture index) variable before transformation (*above*) and after transformation with $\lambda = 0.13$ (*below*)

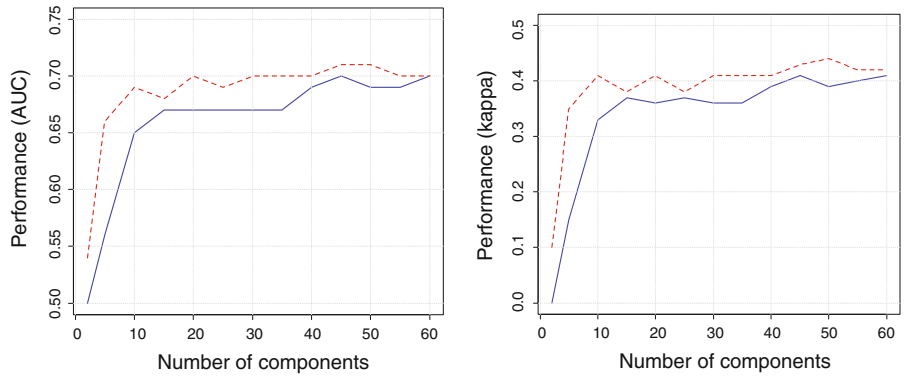

Fig. 2 Model performance for both logistic regressions (PCLR and GPLS) versus number of components using AUC (*left graph*) and Kappa (*right graph*). PCLR reaches AUC = 0.7 with less components (LVs = 20) while PCLR needs 45 PCs. With fewer components (less than 10) GPLS performs better that PCLR. Both models with LVs and PCs that explain only 95% of variance the performance is below accepted values. PCLR model is represented by the *continuous line* and GPLS as the *dashed line*

Modeling framework

The modeling framework refers to the context as well as the processes, ranging from training and tuning to testing and assessing the modeling techniques for their predictive power. The models are based on the paradigm that the value of the stem rust resistance state (Y) depends on the climatic variables (X), where $X = (x_1,\ldots, x_n)$. At first, the assumption is that Y is normally distributed with the mean being a linear function of X with a constant variance of σ_ϵ^2,

$$Y_i \sim N\left(\beta_0 + \sum_{i=1}^{n} \beta_i X_i, \sigma_\epsilon^2\right)$$

where β_i are coefficients. When the trait state is considered as resistant or susceptible, Y can take 2 possible values; either 0 (susceptible) or 1 (resistant) and thus the distribution of Y_i becomes

$$Y_i \sim \mathfrak{B}\left(1, \Phi\left(\beta_0 + \sum_{i=1}^{n} \beta_i X_i\right)\right)$$

The above equation describes a random Bernoulli function (Gollin et al. 2000), of which the standard normal output is the Probit model and the logistic distribution is the Logit model (Feelders 1999). Such relationships between the trait(s) and the environment can be modeled in a multiple regression framework, including the logistic regression (Webb et al. 2010) where the response variable Y is adjusted to a response vector *logit(p)* with $p = P(Y = 1)$. The *logit* stands for the logarithmic equation (Pohlmann and Leitner 2003):

$$logit(p) = ln(p/1 - p) = \beta_0 + \sum_{i=1}^{n} \beta_i X_i$$

which in turn leads to the mathematical expression of

$$p = \frac{\exp\left(\beta_0 + \sum_{i=1}^{n} \beta_i X_i\right)}{1 + \exp\left(\beta_0 + \sum_{i=1}^{n} \beta_i X_i\right)}$$

and this transformation assumes a linear relationship between the *logit* of the probability of $Y = 1$ and the climate variables. However, trait-based approach linear regression analysis may be more appropriate for exploratory data analysis (Webb et al. 2010). As a follow up to the early work by Endresen et al. (2011) this study was extended to a non-parametric framework where the phenomena are expected to be non-linear using the original response Y. The non-linear framework refers here to learning based techniques, which aim to overcome the problem of restrictive parametric paradigms on one hand and the prerequisite distribution assumptions on the other (Drake et al. 2006).

The accuracy of the models (or predictive power) was measured based on the ability of a model to accurately predict the number of times the fitted model classifies correctly the response for each of the two descriptor states (resistant or susceptible). The

modeling framework also includes a tuning process for optimal accuracy. The quality of the models was measured using cross-validation algorithms where the data is split into two subsets, the algorithms developed on the training test were used to predict for trait states in the test set.

Modeling techniques

The five modeling techniques that were assessed to quantify the hypothesized trait-environment relationship are described in the following sections (Table 2). In the two first techniques the response variable was adjusted to take into account that it is a binary variable and that the trait-environment relationship is described in terms of probability of $Y = 1$ instead of Y.

Multiple linear regression using principal component logistic regression (PCLR)

A PCLR analysis was performed on the transformed climatic variables and the stem rust trait prediction $\hat{Y}$ (value estimate of the probability that $Y_i = 1$) was generated from the PCA component scores (S_{pca}) instead of the original transformed climate variables (X). It is based on the PCLR equation $logit(p) = S_{pca}B + E$, where B consists of maximum likelihood estimates of the logistic regression coefficients (Aguilera et al. 2006). The approach aims to both reduce the number of predictor variables (multi-collinearity) and adjust the outcome or response variable. The prediction was initially carried out using the number of components (PCs) that account for 95% of explained variance. The optimization process of PCLR model was based on the accuracy measures which were examined by adding the components stepwise based on the PCs contribution to the overall explained variance, starting with those that explain a large amount of variance (Fig. 2). The components with high predictive value will lead to an increase of these accuracy measures while those that are largely "noise" will lead to their decrease.

Multiple linear regression using generalized partial least squares (GPLS)

PCLR as an extension of PCA approach eliminates only the collinearity but might not identify the optimum subset of candidate variables that can be used as predictors since their decomposition is carried out independently of the trait dataset. Thus a GPLS regression as an extension of PLS was used since PLS not only retains the original structure but also involves a decomposition into a product of factors and their loadings (regression coefficients), of both the environmental dataset and trait dataset simultaneously (Wold et al. 1984; Abdi 2010). GPLS as an extension to PLS it retains the rationale of PLS (Bastien et al. 2005). While PCA maximizes the variance of the scores PLS maximizes the covariance between the scores and the response variable. GPLS latent variables (LVs) would be thus more relevant for trait prediction as was demonstrated by Arif et al. (2007) for PLS to assess the correlation between morphology

Table 2 Models used in the study

Model	Tuning parameters	Library[a] (R language)	References
Principal component logistic regression (PCLR)	Number of principal components (PCs) (ncomp)	pls/stats (glm)	Mevik and Wehrens (2006)/R Core team (2011)
Generalized partial least squares (GPLS)	Number of latent variables (LVs)	gpls	Ding and Gentleman (2005)
Random forests (RF)	Number of trees (n.tree) number of predictors chosen at each node (mtry)	randomForest	Breiman (2001), Cutler et al. (2007), Prasad et al. (2006)
Neural networks (NN)	Number of hidden layers (number of perceptrons): size decay value (ε)	nnet	Venables and Ripley (2002)
Support vector machines (SVM)	gamma/sigma, cost (C)	svm (e1071)	Dimitriadou et al. (2010)
		ksvm (kernalab)	Karatzoglou et al. (2006)

[a] Caret library (Kuhn 2008), which stands for classification and regression training, was used across models

and climate variables. The optimization of the number of components selected in the prediction for this model followed the same procedure as that of the PCLR model.

Random forest (RF)

RF is a type of recursive partitioning algorithm where the data is recursively split into groups of observations with similar response values, a procedure that does not require normality assumptions and deals well with a large number of variables (Strobl et al. 2009). It differs from standard tree classifier in that it "grows" many classification trees in the process. An object from an input vector is classified by all trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification of a given object having the most votes over all the trees in the forest. This approach has led to higher classification accuracy that can outperform other classifiers (Breiman 2001).

The data in the RF module is split intrinsically into a "training set" as a result of bootstrap sampling with replacement; the data that is not sampled to be part of the training set is referred to as the out-of-bag (OOB) set. The OOB set is used to test the predictive power of the RF module. A number, *mtrv*, is specified, which is less than the number of input variables (in this case the climatic parameters), such that at each node of the tree a *mtrv* number of variables are selected at random from the original variable set and the best split on these randomly selected variables is used to split the node. Each tree in the "forest" is grown to the largest extent possible without pruning until there are, *ntree,* number of trees. Varying the *mtrv* and *ntree* values is how the model is optimized. The optimization of the two *mtry* and *ntree* parameters is driven by monitoring the magnitude of the mean square prediction error (rate of classification error) observed in the OOB set; that is, the ability of each iteration to correctly classify a site as resistant or susceptible.

Neural network (NN)

In the NN model a neuron is described by its weight w_k and transfer function $f(x)$ that receives a set of numbers x_k as input, in this case climate variables, and generates a number, y, as output (Golden 1996; Bari et al. 2003), in this case the trait. Similar to a nervous system, NN consists of many processor (PE) units linked to communicate in a many-to-many connection structure where the computations are carried out in parallel by the units independently from each other. This parallelism and high connectivity are the characteristics of neural networks that help in overcoming the assumptions that are usually required in the case of linearity. In comparison to the human brain they were originally designed to identify patterns, even in the presence of noise (Warner and Misra 1996).

If the trait-climate relationship falls within a General Linear Model (GLM) context the NN weight w_k would correspond to the β_i coefficients. NN do not require an assumption of linearity between the trait variable (dependent variable) and climate variable (independent variables), it is the data that defines the functional form of this relationship (Warner and Misra 1996). After the climate data was transformed and scaled it was fed to the NN model [*nnet* model R library (Venables and Ripley 2002)]. Prior to the use of the NN model the tuning process was performed to define the number of neurons (NN size) and the decay parameter (ε) that measures the trade-offs between the weights (w_k) and the prediction error. The weights keep changing through the back-propagation iterative process until the reduction in error is optimized.

Support vector machines (SVM)

Support Vector Machines (SVM) is also a learning-based technique that maps input data to a high-dimensional space, and then optimally separates it into respective classes by isolating those inputs which fall close to the data boundaries (Cortes and Vapnik 1995; Principe et al. 2000). In this study SVM was used with a radial basis function (RBF) as the kernel function. RBF uses Gaussian transfer functions, the centres and widths of which are determined by unsupervised learning rules. RBF first carries out unsupervised clustering using a k-nearest neighbor algorithm, and then applies a supervised classification using the cluster number and width (radius, hence the name "radial"). Thus the sites are first split into k clusters and the size of each cluster is obtained from the structure of the input data. The centres of the clusters give the centres of the RBFs, while the distance between the clusters provides their widths (Silipo 1999; Bari et al. 2003).

SVM-RBF first assigns a "score" or a label to the combination of input variables, in this case the site climatic variables, during the unsupervised procedure and then the value of the label is matched with the actual Y value, in this case the trait value. The results of this comparison are fed back to the system and adjustments are made to the labels until the error between the predicted and the actual are minimized. Tuning the SVM involves adjusting the SVM parameters (γ) of the kernel function (RBF) and the cost (C) value chosen to be 0.1 and 1, respectively. SVM model was tuned by supplying parameter ranges for both parameters and the best values above were chosen by minimizing the error on the training data set.

Tuning the models and comparing outputs

Comparisons between models were made using the metric parameters derived from a confusion-matrix table and the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC) (Swets et al. 2000; Fawcett 2006). The confusion matrix parameters are derived from a 2 by 2 contingency table (Table 3). The comparison process involves two groups of algorithms for two different types of cross-validation, sensitivity and specificity.

Sensitivity, defined by a/ (a + c), and specificity, defined by d/(b + d), are indicators of the models ability to correctly classify observations as either susceptible or resistant. The higher the values of sensitivity and specificity the lower the error and thus the better the discriminating power of the model. The errors occur when resistance (R = 1) is classified as susceptible and *vice versa*. The former is a conditional probability notated by $P(R^* = 0|R = 1)$ while the latter is notated by $P(R^* = 1|R = 0)$. Thus sensitivity can be defined by $P(R^* = 1|R = 1)$ and specificity by $P(R^* = 0|R = 0)$.

The Kappa statistic was used to assess improvement over chance and measures the specific agreement in the confusion matrix table. A value of Kappa below 0.4 is an indication of poor agreement and a value of 0.4 and above is an indication of good agreement (Landis and Koch 1977). Thus a high value is an indication that the models performance is adequate for prediction purposes (Scott et al. 2002). A 90% confidence interval for the Kappa statistic was also used since it is asymptotically normally distributed. Since Kappa can be a threshold dependent "metric" parameter we also used the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC) plots to measure the models accuracy (Freeman and Moisen 2008). The AUC accuracy assesses improvement over randomness based on the ROC curve.

Table 3 Confusion matrix (2-by-2 contingency table)

		Observed	
		Resistant	Susceptible
Predicted	Resistant	a	b
	Susceptible	c	d

The ROC curve sensitivity, which is the conditional probability $P(R^* = 1|R = 1)$, is plotted against the conditional probability $P(R^* = 0|R = 1)$, which complements the sensibility $(1 - P(R^* = 0|R = 0))$. This is also known as the plot of true positive rate versus false positive rate, where the true positive rate is sensitivity and the false positive rate is 1- specificity. A ROC curve that rises nearly vertically at the origin towards the left corner of the graph has high true positive rate and a small false positive rate. Such a plot would have high AUC values indicating favorable model performance (Freeman and Moisen 2008). Thus in this study, the higher the AUC values the better the discrimination between collection sites yielding resistant or susceptible classes. An AUC value of 0.5 represents randomness while1represents peak model performance (Fawcett 2006).

Each model was tuned individually to define the most appropriate parameters to use for better prediction (Table 2). This involved the examination of the different errors (e.g. the root-mean-square error (RMSE) numerically or graphically using optimization and tuning algorithms. These algorithms invoke within cross-validation (10-fold cross-validation) using 10 random segments or folds, where 9 folds are used for learning purposes and the reminder onefold is used for validation purpose.

PCLR and GPLS optimization is based on the accuracy parameters. For RF, NN and SVM the parameters were tuned for optimal accuracy using separate tuning algorithms for each module. A range was provided for each parameter and the tuning algorithms selected the best value for each. The cross-validation was performed 10 times and the averages

for the performance indicators are reported in Tables 4 and 5.

The models were then compared based on the AUC and Kappa indicators for best predictive performance. The comparisons were based on the test data, which represents one third of the data (671 sites) while two thirds of the data (1,342 sites) was used to develop the five models. The modeling algorithms were left to run 10 times on the test set and the average with their confidence intervals were reported for AUC, Kappa and the overall correct classification for each model.

Results

The PCLR, GPLS, and RF models were all able to correctly classify sites that yield either resistant or susceptible genotypes with a 76% success rate, SVM and NNs improved this prediction by 1–2% (Table 5). The accuracy of the models is illustrated by the ROC plots shown in Fig. 3. A straight diagonal line is expected when a model is no better than random. While all models yielded plots that demonstrate a better than random performance the curves for the non-linear models (RF, SVM and NN) tended to be skewed more towards the left-hand side of the ROC plots indicating that they tend to classify the resistant trait relatively more correctly with less false positive error and thus will perform relatively better than the parametric models (Fawcett 2006).

These indications are supported by the Kappa and AUC statistics. The PCLR and GPLS approaches have relatively low and similar Kappa average value (0.41). Further, the AUC values are nearly equal to what can be expected for a robust prediction model (0.7 and above). The SVM and NN approach had Kappa values of 0.44 and 0.45, respectively and AUC values of 0.71 and 0.72 with confidence intervals that indicate SVM and NN are significantly different than those obtained for the parametric approaches, in particular PCLR for either one of the accuracy parameters in Table 5.

The prediction density plots (Fig. 3) further support the inference we draw from the Kappa and AUC values. For the PCLR and GPLS plots there is higher degree of overlap between predictions for the two trait classes, while the non-parametric models yield plots that show a more pronounced degree of separation between the two classes.

The results of the tuning process for the PCLR and GPLS models, summarized in Table 4 and shown in Fig. 3, are noteworthy. The accuracy (AUC) and the rate of agreement (Kappa) values increase with the number of PCs and LVs until they reach a plateau after which the two models converge. However, GPLS reaches higher values with fewer components than the PCLR model. When the models are run with 5 PCs and 6 LVs, which represent 95% of total explained variance, GPLS performs better than PCLR, whilst the PCLR model with an AUC value close to 0.5 is similar to a random outcome. The minimum errors for both models were reached at 42 PCs and 22 LVs, respectively. Kappa values have a similar pattern to the AUC in relation to the number of PCs and LVs. Thus the GPLS model requires fewer components to reach a Kappa = 0.4 and AUC close to 0.7.

Discussion

This study allows us to confidently assert that resistance to stem rust in wheat landraces is not randomly distributed geographically but linked to agro-eco-climatic factors existing within collection site environments. This is supported by the findings of Bonman

Table 4 Model accuracy for PCLR and GPLS model using test data for PCs/LVs representing 95% of variance (PCs = 5 and LVs = 6) and both cases where PCA and PLS error reached the minimum error (PCs = 45 and LVs = 20) with high accuracy values

Model	AUC	AUC_L[a]	AUC_U[a]	K	K_L	K_U	O	O_L	O_U
PCLR(5)	0.56	0.55	0.57	0.15	0.13	0.17	0.69	0.68	0.70
PCLR(45)	0.70	0.68	0.71	0.41	0.39	0.43	0.76	0.75	0.77
GPLS(6)	0.67	0.66	0.67	0.35	0.34	0.37	0.74	0.73	0.75
GPLS(20)	0.70	0.69	0.71	0.41	0.39	0.44	0.76	0.75	0.77

K kappa, *O* overall correct classification

[a] (_L, _U) are 95% confidence limits (lower and upper limits)

Table 5 Model accuracy using test data for the 5 optimized and tuned models

Model[a]	AUC	AUC_L	AUC_U	K	K_L	K_U	O	O_L	O_U
PCLR	0.70	0.68	0.71	0.41	0.39	0.43	0.76	0.75	0.77
GPLS	0.70	0.69	0.71	0.41	0.39	0.44	0.76	0.75	0.77
RF	0.70	0.69	0.71	0.42	0.40	0.44	0.76	0.75	0.77
SVM	0.71	0.70	0.72	0.44	0.42	0.45	0.77	0.77	0.78
NN	0.72	0.71	0.73	0.45	0.43	0.47	0.77	0.76	0.78

[a] Tuning parameters values for PCLR: PCs = 45; GPLS: LVs = 20; RF: *mtry* = 20, *ntree* = 1,000; SVM: gamma = 0.1, cost = 1.0; and NN: size = 3, decay = 0.1

et al. (2007) who reported that the distribution of stem rust resistance in wheat landraces was linked to regions of origin. Thus, we can hypothesize that the emergence of resistance to other pathogens are also likely to be linked to long term environmental trends acting upon host pathogen systems. This is hardly surprising given that pathogens generally have optimal environmental conditions within which they will thrive, thus placing selection pressure on *in situ* host populations for the emergence of resistance. The assertion here is that the environment will strongly influence gene flow, natural selection and thus spatial/geographic differentiation for specific traits (Wu et al. 1975; Spieth 1979; Epperson 1990). In the case of powdery mildew, for example, Paillard et al. (2000) report that those populations of winter wheat with the highest level of resistance to powdery mildew originated from sites where powdery mildew pressure was high, due to environmental factors, while the reverse was true of those populations where the pressure was low. A practical application of this was demonstrated by Bhullar et al. (2009) where, after applying a FIGS approach to selecting wheat landraces for an eco-tilling exercise focused on the Pm3 region, found that forty percent of the collection sites chosen yielded genotypes resistant to the isolates used. The study went on to reveal 7 new resistance alleles for the Pm3 gene.

On the other hand, in a similar study aimed at investigating taxonomic and biogeographic predictivity for resistance to 32 pests and diseases of cultivated potato wild relatives, Spooner et al. (2009) report that resistance to only six pests and diseases could be reliably predicted by environmental variables. They concluded that the more efficient strategy to mining genetic resource collections is to carefully screen core collections. They mentioned, however, a number of factors that could impede their results such as the scale of climate grids from which the climate variables were extracted. A recent study by Endresen et al. (2011) indicated higher predictive performance when using finer resolutions for the climate data with grid sizes of 1, 4.5, 9.3, and 18.5 km. The grain or grid size has been an issue in ecological research where it is acknowledged that further studies are needed on the resolution (grid size)-dependency paradigm detailed in the MacArthur and Wilson (1967) bio-geographical "island theory", which originally led to the inclusion of the size of the area's paradigm in ecology (Malanson and Armstrongy 1990; Mann and Benwell 1995). Further it is suggested that trait-environment relationships may be influenced by the scale at which both independent and dependent variables are measured (Cushman and McGarigal 2004; Tautenhahn et al. 2008). In addition to the issue of scale, and as illustrated in Fig. 1, climatic variables also tend to be highly skewed and (in preliminary exploration of the modeling reported here) the modeling techniques performed better when the variables were transformed.

Notwithstanding the above, in preliminary work for this study it was clear that the predictability of the models decreased as the number of variables used decreased. This is supported by Stockwell (2007) who suggests that additional data layers are required to produce more efficient ecological models. Stockwell (2007) further asserts that some entities may not be modeled using restricted variables. In the study reported by Spooner et al. (2009), 38 variables were used including 12 each for rainfall, minimum and maximum temperatures. In this study 60 variables were used that include potential evapo-transpiration and aridity, both of which contain information about humidity, a factor which is widely reported to be of critical importance to the development of fungal pathogens such as stem rust.

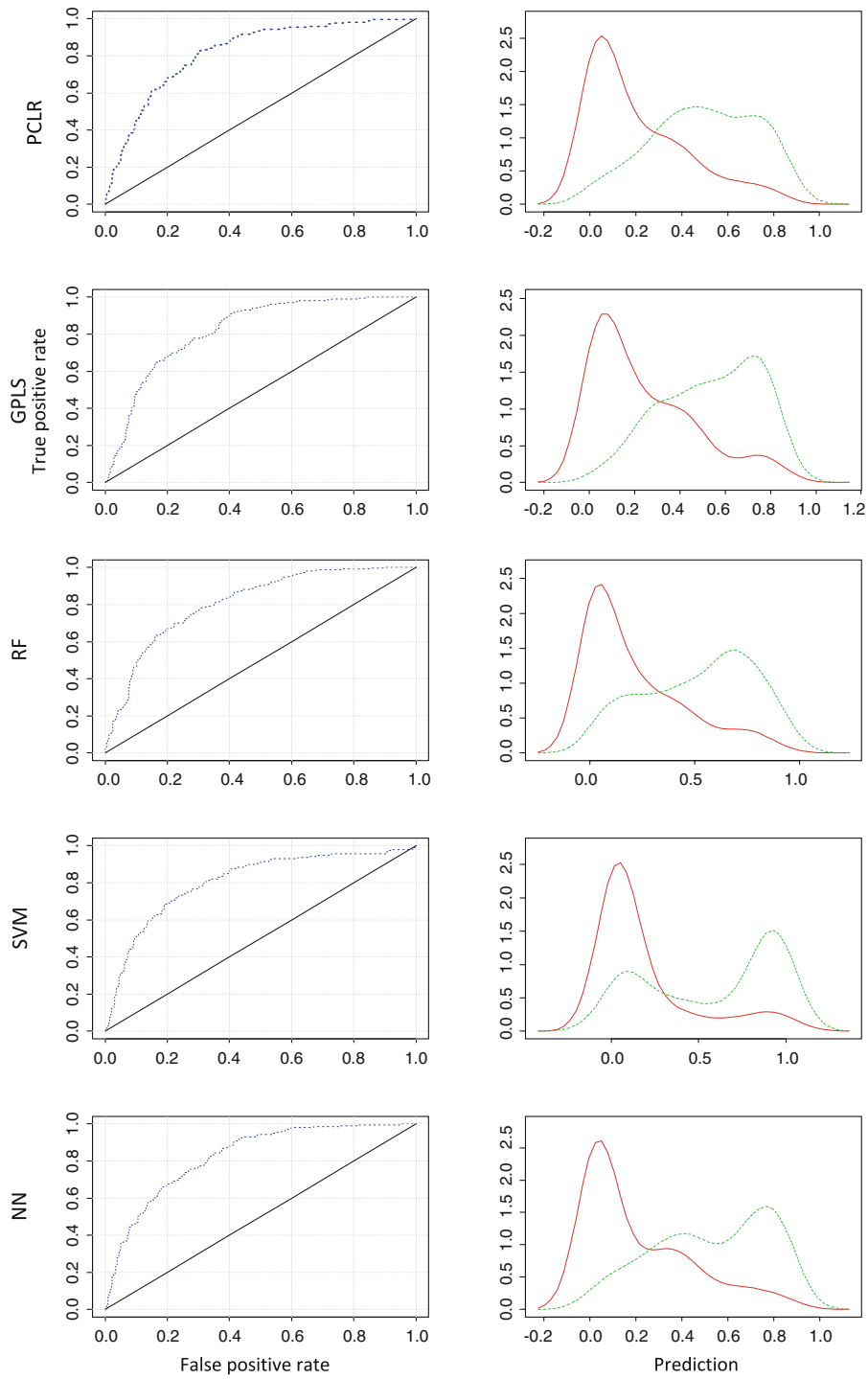

Fig. 3 ROC plots (*left*) and density plots of prediction for resistance and susceptible (*right*) for the 5 models using test set; *green curve* (*discontinuous line*) indicates the probability density distribution for resistance and *red curve* (*continuous line*) indicates susceptibility. Predictions fall out of range [0, 1] as a result of linearity/ interpolation in some of the models

In this context a discussion on the performance of the models used in this study is warranted. The results demonstrate that modeling techniques, such as those reported here, can provide a predictive framework to quantify trait-environment relationships and as such can be used to efficiently identify potentially valuable germplasm from genetic resource collections. The practical application of this has been clearly

demonstrated by El-Bouhssini et al. (2009, 2010) who have discovered multiple sources of resistances to the Syrian biotype of Russian wheat aphid and to Sunn pest in bread wheat for the first time using a FIGS approach after having unsuccessfully screened thousands of genebank accessions. Thus, the availability of environmental data provides an opportunity to improve the use of germplasm through quantifying such trait-environment relationships. In this context Webb et al. (2010) consider that the type and quality of data are so important that they are considered non-trivial limitations to such trait-based approaches.

For the PCLR and GPLS based regression models, the predictions derived by using the same number of PCs and LVs that account for 95% of the variance in the climate variables, are not much different from a random selection. This was unexpected and has implication in terms of setting up subsets of accessions for desirable traits using a PCLR approach to cluster accessions working within the generally accepted levels of variance explanation. Adding more PCs or LVs improved the prediction with both models, GPLS requiring fewer components than PCLR suggesting that some of the environmental variables that correlate with the trait response (resistant or susceptible) were not captured by PCLR. Common and expected is the pattern where PLS type of models needs fewer components than PCA based model as it has been found, for example, by Wu et al. (2008). Moreover, PLS logistic regression is more prone to lead to a coherent model (Bastien et al. 2005). The fact that NN and SVM perform better than PCA and PLS based models indicates, however, that a large amount of variance in the climate data is non-linear (Wu et al. 2008). The out-of-range predictions (presence of negative values to 0 and positive values greater than 1 indicate that the trait-environment relationship is more likely to be non-linear (Fig. 3).

The performance of the different models based on their discriminatory ability indicated by the AUC and Kappa values is also related to the decomposition of variables. When both trait and environmental variables are decomposed together using GPLS, the AUC and Kappa values are relatively higher indicating that the models retain relevant structure and information and thus allow a better discrimination. Similarly, when the prediction is applied directly to the data using machine learning techniques, the results of AUC and Kappa values are higher indicating their improved potential for discriminating between the two groups.

Although both AUC and Kappa values of testing datasets are, as expected, lower than those of the training data, the results show acceptable values. Confidence intervals of SVM and NN model in particular do not overlap with those of PLCR. Overall non-linear models perform better when compared to linear models even if all the data structure is retained for both types of models.

Machine-learning based techniques such as RF, NN and SVM in combination with fuzzy based approaches have the potential to yield better predictions (Kampichler et al. 2010). Such techniques are also more suited for data with a large number of variables. The only limitation, as in the case of SVM, may be due to the range of environmental variables (Drake et al. 2006). RF has the potential to yield better results, however it is computationally expensive as it requires large memory and significantly more run time. NNs are powerful predictive tools although difficult to interpret (Jeschke and Strayer 2008). SVM is more rapid and tends to distinguish optimally between groups and predicts entities while minimizing the loss of information (Guo et al. 2004; Karatzoglou et al. 2006). The advantage of learning-based techniques is that they need fewer assumptions and are more suitable when highly complex non-linear relationships are expected among input variables (Tirelli et al. 2009).

The results also suggest a number of other issues that may improve predictive performance. The data was composed of several taxa of wheat. The partitioning to sub-population or genetic background/lineages may lead to an increase in model accuracy, since compact or clumped distributions of population are easier to model than those of widespread and scattered distributions (Hernandez et al. 2006; Hanspach et al. 2010). Reclassification of trait states has been shown in previous work by Endresen et al. (2011) to improve predictions. Both the assessment of trait variation and the effect of probability distribution of the trait are to be further investigated. Trait distributions vary across any set of accessions and affect the optimal search process (Gollin et al. 2000). They may also vary depending on the type and degree of virulence of races or biotypes in the case of insects.

When optimally tuned all the models reported here displayed predictive powers superior to a random

selection and thus are adequate to explore genetic resource collections for novel sources of resistance to stem rust, and it is argued, to other pathogens. However, the indications are that the SVM and NN algorithms have a higher discriminating power and are more robust and thus will be used in further explorations of this kind. In fact, it is difficult to envisage further predictive gains being made through used of alternate models. Rather it is more likely that gains will come from an appropriate mixture of independent variables that are known to influence the selection pressure in question. For example, this study and that of Spooner et al. (2007) used maximum and minimum monthly temperatures. However, since pathogens are known to respond more to the diurnal temperature variation or average temperatures than to absolute max or min temperatures it is possible that greater resolution would be gained by expressing temperature in these terms.

Another line of research being explored by the authors is to consider climatic variables within the context of site specific growing seasons. For example, the minimum temperature in a given month may be important at one location but in another it may have no relevance because it falls outside of that site's growing season. It is proposed here that instead of using long term climatic averages expressed as monthly values they could instead be expressed as averages for stages in a crops development. Thus in the modeling process the noise created by differences in phenology between sites would be eliminated facilitating higher resolutions to detect environment—trait linkages. Further variables could be created by counting the number of days in a season that meet certain climatic conditions know to be favorable to the pathogen.

Conclusion

It is concluded here that the FIGS approach will improve the use of germplasm as *a priori* information becomes more available through improved modeling techniques or other approaches. Gollin et al. (2000) stated that such information (*a priori* information) is extraordinary valuable, so far it is a specialized knowledge, and the authors expect that technology will gradually provide further substitutes for such valuable information. This study demonstrated that modeling techniques such as those explored here provide a predictive framework to quantify the trait-environment relationship that will help in more effectively using genetic resources. The availability of environmental data is providing the opportunity to improve the use of germplasm through the quantification of such trait-environment relationships. As Webb et al. (2010) point out, the type and quality of data are so important that they must be considered non-trivial limitations of such trait-based approaches (Webb et al. 2010).

References

Abdi H (2010) Partial least squares regression and projection on latent structure regression (PLS regression). Wiley Interdiscip Rev Comput Stat 2(1):97–106. doi:10.1002/wics.51

Aguilera AM, Escabias M, Valderrama MJ (2006) Using principal components for estimating logistic regression with high-dimensional multicollinear data. Comput Stat Data Anal 50:1905–1924

Arif S, Adams DC, Wicknick JA (2007) Bioclimatic modeling, morphology, and behavior reveal alternative mechanisms regulating the distributions of two parapatric salamander species. Evol Ecol Res 9:843–854

Barboni D, Harrison SP, Bartlein PJ, Jalut G, New M, Prentice IC, Sanchez-Goñi M-F, Spessa A, Davis B, Stevenson AC (2004) Relationships between plant traits and climate in the Mediterranean region: a pollen data analysis. J Veg Sci 15:635–646

Bari A, Martin A, Boulouha B, Barranco D, Gonzalez-Andujar JL, Trujillo I, Ayad G (2003) Image feature extraction combined with a neural networks approach for the identification of olive cultivars. In: Proceeding of the 3rd IASTED international conference on visualization, imaging and image processing, pp 613–620. ACTA Press

Bastien P, Vinzi VE, Tenenhaus M (2005) PLS generalized linear regression. Comput Stat Data Anal 48(1):17–46

Belsley DA (1991) A guide to using the collinearity diagnostics. Comput Sci Econ Manag 4:33–50

Bhullar NK, Zhang Z, Wicker T, Keller B (2009) Wheat gene bank accessions as a source of new alleles of the powdery mildew resistance gene *Pm3*: a large scale allele mining project. BMC Plant Biol 10:88. doi:10.1186/1471-2229-10-88

Bonman JM, Bockelman HE, Jackson LF, Steffenson BJ (2005) Disease and insect resistance in cultivated barley accessions from the USDA national small grains collection. Crop Sci 45:1271–1280

Bonman JM, Bockelman HE, Jin Y, Hijmans RJ, Gironella A (2007) Geographic distribution of stem rust resistance in wheat landraces. Crop Sci 47:1955–1963

Box GEP, Cox DR (1964) An analysis of transformations. J R Stat Soc Series B (Methodol) 26(2):211–252

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Brown AHD, Spillane C (1999) Implementing core collections principles, procedures, progress, problems and promise. In:

Johnson RC, Hodgkin T (eds) Core collections for today and tomorrow. International Plant Genetic Resources Institute, Rome, pp 1–9

Chuine I (2010) Why does phenology drive species distribution? Phil Trans R Soc B 365:3149–3160

CIMMYT (2005) Sounding the alarm on global stem rust. An Assessment of race ug99 in Kenya and Ethiopia and the potential for impact in neighbouring regions and beyond. Resource Document. Accessed 17 Feb 2011. http://www.globalrust.org/db/attachments/about/2/1/Sounding%20the%20Alarm%20on%20Global%20Stem%20Rust.pdf

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297. doi:10.1007/BF00994018

Cushman SA, McGarigal K (2004) Patterns in the species-environment relationship depend on both scale and choice of response variable. Oikos 105:117–124

Cutler DR, Edwards TC Jr., Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random Forests for classification in ecology. Ecology 88:2783–2792

De Pauw E, Goebel W, Adam H (2000) Agrometeorological aspects of agriculture and forestry in the arid zones. Agric Forest Meteorol 103:43–58

Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A (2010) R library (e1071). The R foundation for statistical computing. ISBN: 3-900051-07-0

Ding BY, Gentleman R (2005) Classification using generalized partial least squares. J Comput Graphical Stat 14(2):280–298

Dinoor A (1975) Evaluation of sources of resistance. In: Frankel OH, Hawkes JD (eds) Crop genetic resources for today and tomorrow. Cambridge University Press, Cambridge, pp 201–210

Drake JM, Randin C, Guisan A (2006) Modelling ecological niches with support vector machines. J Appl Ecol 43:424–432

Dwivedi SL, Crouch JH, Mackill DJ, Xu Y, Blair MW, Ragot M, Upadhyaya HD, Ortiz R (2007) The molecularization of public sector crop breeding: progress, problems, and prospects. Adv Agron 95:163–318

Eckardt NA (2001) Functional evolutionary genetics and plant adaptation linking phenotype and genotype. Plant Cell 13(6):1249–1254

El-Bouhssini M, Street K, Joubi A, Ibrahim Z, Rihawi F (2009) Sources of wheat resistance to Sunn pest, *Eurygaster integriceps* Puton, in Syria. Genet Resour Crop Evol 56(8):1065–1069

El-Bouhssini M, Street K, Amri A, Mackay M, Ogbonnaya FC, Omran A, Abdalla O, Baum M, Dabbous A, Rihawi F (2010) Sources of resistance in bread wheat to Russian wheat aphid (*Diuraphis noxia*) in Syria identified using the focused identification of germplasm strategy (FIGS). Plant Breed 130:96–97

Endresen DTF (2010) Predictive association between trait data and ecogeographic data for Nordic barley landraces. Crop Sci 50(6):2418–2430. doi:10.2135/cropsci2010.03.0174

Endresen DTF, Street K, Mackay M, Bari A, De Pauw E (2011) Predictive association between biotic stress traits and ecogeographic data for wheat and barley landraces. Crop Sci 51:2036–2055

Epperson BK (1990) Spatial autocorrelation of genotypes under directional selection. Genetics 124(3):757–771

Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27:861–874. doi:10.1016/j.patrec.2005.10.010

Feelders AJ (1999) Statistical concepts. In: Berthold M, Hand DJ (eds) Intelligent data analysis: an Introduction. Springer, Berlin, pp 15–66

Fehser S, Beike U, Stoveken J, Pretorius ZA, Van der Westhuizen A, Moersbacher B (2010) Histological and initial molecular analysis of Ug99, the new Sr31-breaking race of the wheat stem rust fungus threatening global wheat production. J Plant Pathology 92(3):709–720

Freeman EA, Moisen GG (2008) A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. Ecol Model 217:48–58

Gepts P (2006) Plant genetic resources conservation and utilization: the accomplishments and future of a societal insurance policy. Crop Sci 46:2278–2292

Gesch DB, Larson KS (1996) Techniques for development of global 1-kilometer digital elevation models. On-line document: http://edcdaac.usgs.gov/gtopo30/README.html

Golden RM (1996) Mathematical methods for neural network analysis and design. Massachusetts Institute of Technology, Cambridge, MA

Gollin D, Smale M, Skovmand B (2000) Searching an *ex situ* collection of wheat genetic resources. Am J Agric Econ 82(4):812–827

Guo Q, Kelly M, Graham CH (2004) Support vector machines for predicting distribution of Sudden oak death in California. Ecol Model 182(1):75–90

Hakes AS, Cronin JT (2011) Environmental heterogeneity and spatiotemporal variability in plant defense traits. Oikos 120:452–462. doi:10.1111/j.1600-0706.2010.18679.x

Hanspach J, Kühn I, Pompe S, Klotz S (2010) Predictive performance of plant species distribution models depends on species traits. Perspect Plant Ecol Evol Syst 12(3):219–225. doi:10.1016/j.ppees.2010.04.002

Hernandez PA, Graham CH, Master LL, Albert DL (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29:773–785

Hodson D, DePauw E (2011) Use of GIS applications to combat the threat of emerging virulent wheat stem rust races. In: Sharon A (ed) GIS applications in agriculture, vol 3. Clay CRC Press, Boca Raton, pp 129–157

Hutchinson MF (1995) Interpolating mean rainfall using thin plate smoothing splines. Int J Geogr Inf Syst 9:385–403

Hutchinson MF (2000) ANUSPLIN version 4.1. User Guide. Center for resource and environmental studies. Australian National University, Canberra

Hutchinson MF, Corbett JD (1995) Spatial interpolation of climatic data using thin plate smoothing splines. Co-ordination and harmonisation of databases and software for Agroclimatic applications, FAO Agrometeorology Series 13. FAO, Rome, pp 211–224

Jeschke JM, Strayer DL (2008) Usefulness of bioclimatic models for studying climate change and invasive species. Ann N Y Acad Sci 1134:1–24

Kampichler C, Wieland R, Calmé S, Weissenberger H, Arriaga-Weiss S (2010) Classification in conservation biology: a comparison of five machine-learning methods. Ecol Inform 5(6):441–450

Karatzoglou A, Meyer D, Hornik K (2006) Support vector machines in R. J Stat Softw 15(9)
Kolmer JA (2005) Tracking wheat rust on a continental scale. Curr Opin Plant Biol 8(4):441–449
Koo B, Wright BD (2000) The optimal timing of evaluation of genebank accessions and the effects of biotechnology. Am J Agric Econ 82(4):797–811
Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28(5)
Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174
Leonard KJ, Szabo LJ (2005) Stem rust of small grains and grasses caused by *Puccinia graminis*. Mol Plant Pathol 6:99–111
MacArthur RH, Wilson EO (1967) The theory of island biogeography. Princeton University Press, Princeton
Mackay MC (1990) Strategic planning for effective evaluation of plant germplasm. In: Srivastava JP, Damania AB (eds) Wheat genetic resources: meeting diverse needs. Wiley, Chichester, pp 21–25
Mackay MC (1995) One core collection or many? In: Hodgkin T, Brown AHD, Van Hintum TJL, Morales AAV (eds) Core collections of plant genetic resources. Wiley, Chichester, pp 199–210
Mackay MC, Street K (2004) Focused identification of germplasm strategy—FIGS. In: Black CK, Panozzo JF, Rebetzke GJ (eds) Proceedings of the 54th Australian cereal chemistry conference and the 11th wheat breeders' assembly, pp 138–141. Royal Australian Chemical Institute, Melbourne
Malanson GP, Armstrongy MP (1990) Improving environmental simulation models to assess climate change impacts. University of Iowa, Department of Geography discussion paper no. 43, p 35
Mann S, Benwell GL (1995) Geographic information systems in environmental management, AURISA/ 7th colloquium of the Spatial Information Research Centre, pp 295–310, Palmerston North
McIntosh RA, Yamazaki Y, Dubcovsky J, Rogers J, Morris C, Somers DJ, Appels R, Devos KM (2008) Catalogue of gene symbols for wheat. In: Appels R, Eastwood R, Lagudah E, Langridge P, Mackay M, McIntyre L, Sharp P (eds) Proceedings of the 11th international wheat genetics symposium, Brisbane
McIntosh R, Dubcovsky J, Rogers W, Morris C, Appels R, Xia X (2010) Catalogue of gene symbols for wheat: 2010 supplement. http://www.shigen.nig.ac.jp/wheat/komugi/genes/macgene/supplement2010.pdf
Mevik BH, Wehrens R (2006) The pls package: principal component and partial least squares regression. J Stat Softw 18(2):1–24
Osborne JW (2010) Improving your data transformations: applying Box–Cox transformations as a best practice. Pract Assess Res Eval 15(12):1–9
Paillard S, Goldringer I, Enjalbert J, Trottet M, David J, de Vallavieille-Pope C, Brabant P (2000) Evolution of resistance against powdery mildew in winter wheat populations conducted under dynamic management. II. Adult plant resistance. Theoretical Appl Genet 101:457–462
Pakeman R, Leps J, Kleyer M, Lavorel S, Garnier E, VISTA consortium (2009) Relative climatic, edaphic and management controls of plant functional trait signatures. J Veg Sci 20:148–159
Pessoa-Filho M, Rangel PHN, Ferreira ME (2010) Extracting samples of high diversity from thematic collections of large gene banks using a genetic-distance based approach. BMC Plant Biol 10:127
Pohlmann JT, Leitner DW (2003) A comparison of ordinary least squares and logistic regression. Ohio J Sci 103(5): 118–125
Polignano GB, Uggenti P, Scippa G (2001) Diversity analysis and core collection formation in Bari faba bean germplasm. FOA/Bioversity PGR Newsl 125:33–38
Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9:181–199
Principe JC, Euliano NR, Lefebvre WC (2000) Neural and adaptive systems: fundamentals through simulations. Wiley, New York
Qualset CO (1975) Sampling germplasm in a center of diversity: an example of disease resistance in Ethiopian Barley. In: Frankel H, Hawkes JD (eds) Crop genetic resources today and tomorrow. Cambridge University Press, Cambridge, pp 81–96
R Development Core Team (2011) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna. ISBN: 3-900051-07-0
Scott JM, Heglund PJ, Morrison ML (2002) Predicting species occurrences: issues of accuracy and scale. Island Press, Covelo California
Silipo R (1999) Neural networks. In: Berthold M, Hand DJ (eds) Intelligent data analysis: an Introduction. Springer, Berlin, pp 217–268
Spieth PT (1979) Environmental heterogeneity: a problem of contradictory selection pressures, gene flow, and local polymorphism. Am Nat 113(2):247–260
Spooner DM, Jansky SH, Simon R (2009) Tests of taxonomic and biogeographic predictivity: resistance to disease and insect pests in wild relatives of cultivated potato. Crop Sci 49:1367–1376
Stockwell D (2007) Niche modeling: predictions from statistical distributions. Chapman and Hall, CRC. ISBN: 9781584884941
Street K, Mackay M, Zuev E, Kaur N, El Bouhssini M, Konopka J, Mitrofanova O (2008) Diving into the genepool: a rational system to access specific traits from large germplasm collections. In: Appels R, Eastwood R, Lagudah E, Langridge P, Mackay M (eds) Proceedings of the 11th international wheat genetics symposium, pp 28–31, Brisbane
Strobl C, Malley J, Tutz G (2009) An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods 14(4):323–348
Swets JA, Dawes RM, Monahan J (2000) Better decisions through science. Sci Am 283:82–87
Tait AB, Turner RW (2005) Generating multi-year gridded daily rainfall over. NZ J Appl Meteorol 44:1315–1323
Tautenhahn S, Heilmeier H, Götzenberger L, Klotz S, Wirth C, Kühn I (2008) On the biogeography of seed mass in Germany distribution patterns and environmental correlates. Ecography 31:457–468

Tirelli T, Pozzi L, Pessani D (2009) Use of different approaches to model presence/absence of *Salmo marmoratus* in Piedmont (Northwestern Italy). Ecol Inform 4:234–242

Tukey JW (1957) On the comparative anatomy of transformations. Ann Math Stat 28(3):602–632

Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York

Vurro M, Bonciani B, Vannacci G (2010) Emerging infectious diseases of crop plants in developing countries: impact on agriculture and socio-economic consequences. Food Sec 2:113–132

Warner B, Misra M (1996) Understanding neural networks as statistical tools. Am Stat 50(4):284–293

Webb CT, Hoeting JA, Ames GM, Pyne MI, LeRoy Poff N (2010) A structured and dynamic framework to advance traits-based theory and prediction in ecology. Ecol Lett 13:267–283

Wold S, Ruhe A, Wold H, Dunn WJ (1984) The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J Sci Stat Comp 5:735–743

Wood SN (2000) Modelling and smoothing parameter estimation with multiple quadratic penalties. J R Stat Soc (B) 62(2):413–428

Wratt DS, Tait A, Griffiths G, Espie P, Jessen M, Keys J, Ladd M, Lew D, Lowther W, Mitchell N, Morton J, Reid J, Reid S, Richardson A, Sansom J, Shankar U (2006) Climate for crops: integrating climate data with information about soils and crop requirements to reduce risks in agricultural decision-making. Meteorol Appl 13:305–315

Wu L, Bradshaw AD, Thurman DA (1975) The potential for evolution of heavy metal tolerance in plants. III. The rapid evolution of copper tolerance in *Agrostis stolonifera*. Heredity 34(2):165–187

Wu Y, Johnson GL, Gomez SM (2008) Data-driven modeling of cellular stimulation, signaling and output response in RAW 264.7 cells. J Mol Signaling 3:11. doi:10.1186/1750-2187-3-11

Xu Y (2010) Plant genetic resources: Management, evaluation and enhancement. In: Molecular plant breeding. CAB International, Wallingford, UK, pp 151–194