FAKULTÄT AGRARWISSENSCHAFTEN

Aus dem Institut für Kulturpflanzenwissenschaften

Universität Hohenheim

Fachgebiet: Biostatistik

Prof. Dr. H.-P. Piepho

Model selection by cross-validation in multi-environment trials

Dissertation
zur Erlangung des Grades eines Doktors
der Agrarwissenschaften

vorgelegt

der Fakultät Agrarwissenschaften

von

Steffen Hadasch

aus Heilbronn

2017

Die vorliegende Arbeit wurde am 13.12.2017 von der Fakultät Agrarwissenschaften der Universität Hohenheim als "Dissertation zur Erlangung des Grades eines Doktors der Agrarwissenschaften" angenommen

Tag der mündlichen Prüfung: 13.06.2018

Leiter der Prüfung:	Prof. Dr. Thilo Streck
Berichterstatter, 1. Prüfer:	Prof. Dr. H.-P. Piepho
Mitberichterstatter, 2. Prüfer:	Dr. Johannes Forkman
3. Prüfer:	Dr. Tobias Würschum

Table of contents

in the field or to identify the most appropriate (prediction) model for an observed dataset. To evaluate these objectives, cross-validation (CV) can be used (Hastie and Tibshirani, 2001; James et al., 2013). In CV the data is divided into so-called training data and validation data. The model parameters are estimated from the training data and the validation data is predicted using the parameter estimates obtained by the training data. The goodness of the prediction is evaluated by a success criterion. The predictive ability of a model, defined as the Pearson correlation coefficient between the validation data and their predictions, or the test mean squared error (test-MSE) which is the mean of the squared differences between the validation data and their predictions (Hastie and Tibshirani, 2001; James et al., 2013) are often used as a success criterion.

CV can be used to evaluate different objectives the breeder may pursue with MET data by an appropriate choice of training and validation data. With models that use genetic marker information, either genotypes, environments or both may be sampled for validation in order to evaluate the predictive ability of the models in terms of predicting (i) unobserved genotypes, (ii) observed genotypes in an unobserved environment, and (iii) unobserved genotypes in an unobserved environment can be evaluated (Utz, 2000; Chapter 2).

In case of AMMI/GGE models, CV can be used to determine the number of multiplicative terms. With these models CV can be done by sampling one replicate of each genotype-environment combination randomly (Gauch and Zobel, 1988) such that the validation data come from different replications of the experiment. Another sampling strategy uses validation data from one complete replication of the experiment within each environment (Piepho, 1994). In this way, the validation data within an environment come from the same replication and furthermore this sampling strategy can be shown to add uncertainty only to the estimated environment effects while the sampling strategy by Gauch and Zobel (1988) adds uncertainty to the estimated interaction effects which may lead to an underestimation of the number of multiplicative terms (Piepho, 1994). The sampling procedure proposed by Piepho (1994) therefore mostly aims to improve the evaluation procedure by considering consequences associated with the model estimation; the model evaluation may further be improved by using a success criterion that is based on pairwise differences of the validation data (Piepho, 1998). Another approach to account for the experimental design in a CV is proposed in Chapter 3. In this approach, the data are adjusted for the design (replicate and block) effects before the application a CV scheme. This approach aims at improving the evaluation by an adjustment of the validation data such that the success criterion is largely unaffected by design effects