Extending Finlay–Wilkinson regression with environmental covariates

Hans-peter Piepho, Justin Blancon

Received: 12 December 2022 | Revised: 31 May 2023 | Accepted: 3 July 2023

DOI: 10.1111/pbr.13130

ORIGINAL ARTICLE

Plant Breeding WILEY

Extending Finlay–Wilkinson regression with environmental covariates

Hans-Peter Piepho[1] | Justin Blancon[2]

[1]Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Stuttgart, Germany

[2]UMR GDEC, INRAE, Université Clermont Auvergne, Clermont-Ferrand, France

Correspondence
Hans-Peter Piepho, Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany.
Email: piepho@uni-hohenheim.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: PI 377/20-2; Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement, Grant/Award Number: I-SITE CAP 20-25

Abstract

Finlay–Wilkinson regression is a popular method for analysing genotype–environment interaction in series of plant breeding and variety trials. It involves a regression on the environmental mean, indexing the productivity of an environment, which is driven by a wide array of environmental factors. Increasingly, it is becoming feasible to characterize environments explicitly using observable environmental covariates. Hence, there is mounting interest to replace the environmental index with an explicit regression on such observable environmental covariates. This paper reviews the development of such methods. The focus is on parsimonious models that allow replacing the environmental index by regression on synthetic environmental covariates formed as linear combinations of a larger number of observable environmental covariates. Two new methods are proposed for obtaining such synthetic covariates, which may be integrated into genotype-specific regression models, that is, criss-cross regression and a factor-analytic approach. The main advantage of such explicit modelling is that predictions can be made also for new environments where trials have not been conducted. A published dataset is employed to illustrate the proposed methods.

KEYWORDS
factor-analytic model, factorial regression, partial least squares, reduced rank regression, singular value decomposition, synthetic covariate

1 | INTRODUCTION

The main challenge in the analysis of multienvironment trials (MET) in plant breeding and variety testing is modelling and exploiting genotype–environment interaction. One of the most popular methods for this purpose is a regression of genotype performances in the individual environments on the environmental mean. This method was originally proposed by Yates and Cochran (1938) and later popularized by the seminal paper by Finlay and Wilkinson (1963). We will henceforth refer to this approach as Finlay–Wilkinson (FW) regression. In this regression, the environmental mean serves as an index for the environmental conditions. These conditions are determined by a large array of environmental variables, and it therefore seems natural to replace the environmental mean with measurable environmental covariates. Early treatments of such regression models for MET, also known as factorial regression (FR) (Denis, 1988), are found in Abou-El-Fittouh et al. (1969) and Freeman and Perkins (1971). The main challenge with FR is that the number of environmental covariates may be large, which makes the FR model very complex. Conversely, the kernel approach (Jarquin et al., 2014) is based on a mixed model including an environmental covariance matrix that is computed from the environmental covariates, in the same way as a kinship matrix is computed from genotypic marker data. This model leads to the estimation of a single variance parameter for the genotype–environment interaction variance. Somewhere

between the FR approach at the complex end of the scale and the kernel approach at the simplistic end of the scale lies the idea to regress the environmental mean on covariates, instead of regressing the genotype–environment means on covariates separately for each genotype. This idea is so natural that it is hard to say who originally invented it. It is probably fair to say that the idea underlies and motivates the use of FW regression, even in cases where observable environmental covariates are not used in the analysis (Piepho, 2022). A rigorous treatment of the idea in an MET context was first put forward by Hardwick and Wood (1972), who showed how to estimate the model parameters by the method of least squares. A partial least squares (PLS) approach to fit the same kind of model was proposed by Aastveit and Martens (1986), and this may be of particular interest when the number of covariates exceeds that of the environments. Van Eeuwijk (1992, 1995) pointed out that these models are related to what is known as redundancy analysis in psychology and as reduced rank regression in other applied areas such as engineering (Davies & Tso, 1982). An early mathematical treatment of the approach is found in Rao (1964).

All of the references considered so far assume that the genotype–environment classification is complete and that the residuals from the regression are independent with constant variance so that the ordinary least squares method provides optimal estimates. It must be acknowledged, however, that MET data are often unbalanced. Moreover, the variance–covariance structure needed to fully represent the experimental design, as well as to meet the objectives of the analysis, may deviate from the simple structure assumed for ordinary least squares, calling instead for a linear mixed model with additional random effects. This may, in fact, be one reason why these methods have not yet found very widespread use, despite an urgent need for such methods in an era where environmental information is becoming readily available and breeders are keenly interested in leveraging such information to make better selections and predictions (Cooper & Messina, 2021; Costa-Neto et al., 2021; Diepenbrock et al., 2022; Resende et al., 2021; Xu, 2016). The purpose of this paper, therefore, is to review the classical work based on ordinary least squares and to explore ways in which the approach can be extended to deal with unbalanced data and the need to use a mixed model framework.

The rest of the paper is organized as follows. In Section 2, we briefly recapitulate the FW regression model and common methods for estimating it. Section 3 described the extension where the latent environmental score is regressed on environmental covariates and Section 4 considers the extension to more than one latent environmental score. In all three sections, the residual is assumed to be independent with constant variance so that ordinary least squares can be applied. Our exposition in these sections mainly focuses on the models and only sketches methods of estimation. In Section 5, we consider the extension to mixed models and focus on those estimation approaches in the preceding sections, which seem most suitable for this extension. For these select methods, we then provide more details of the estimation steps. An example is presented in Section 6 to illustrate and compare the methods. The paper ends with a discussion in Section 7.

2 | FINLAY–WILKINSON REGRESSION

The FW model assumes that the response of different genotypes in varying environments can be modelled using the linear predictor (Mandel, 1961, eq. 12)

$$\eta_{ij} = \alpha_i + \beta_i w_j \tag{1}$$

where η_{ij} is the expected performance of the i-th genotype ($i = 1, \ldots, n$) in the j-th environment ($j = 1, \ldots, m$), α_i and β_i are intercept and slope for the i-th genotype and w_j is a latent effect of the j-th environment. The slope β_i can be interpreted as a measure of sensitivity with small absolute values indicating stable responses over changing environments (Becker & Leon, 1988). If the latent effect w_j is mean-centred, then the intercept α_i assesses a genotype's mean performance. The observed data are assumed to be genotype–environment means y_{ij}, for which the model is $y_{ij} = \eta_{ij} + e_{ij}$, where e_{ij} are independently and identically distributed residuals. For balanced data, Finlay and Wilkinson (1963) estimated w_j by the arithmetic mean of all observed genotype mean yields, y_{ij}, that is, they used the estimator $\hat{w}_j = \bar{y}_{\bullet j}$. This, however, does not yield the least-squares fit of (1). The least squares fit can be obtained by a singular value decomposition (SVD) of the matrix $\{y_{ij} - \bar{y}_{i\bullet}\}$, extracting the first singular vector for environments (Hardwick & Wood, 1972; Williams, 1952; Yan & Kang, 2003). Again, this assumes balanced data.

For unbalanced data, Digby (1979) proposed obtaining the least squares fit by alternating least squares, also known as criss-cross regression (CCR) (Gabriel & Zamir, 1979), and Ng and Grunwald (1997; also see Ng & Williams, 2001) showed how to do this using nonlinear least squares. The model is not linear in the parameters, and some restrictions on the parameters is needed for the multiplicative term $\beta_i w_j$. A further method that can be interpreted as an expectation–maximization (EM) algorithm initially replaces the empty cells of the two-way classification with initial values, then applies SVD to the completed table, re-estimates the empty cells from the fitted model, and so forth until convergence (Gauch & Zobel, 1990). We here focus on the CCR and EM methods because they are easily generalized to mixed models and a regression on covariates.

3 | MODELLING THE ENVIRONMENTAL MEAN USING ENVIRONMENTAL COVARIATES

A downside of model (1) is that it cannot be used to predict the performance in unseen environments. If w_j can be replaced by an observable covariate, such predictions become possible. However, a single covariate rarely provides good predictions. Thus, a natural extension is to do a multiple regression on several covariates (Denis, 1988; Hardwick & Wood, 1972). Such an FR model quickly becomes very complex, because each genotype needs to have a separate regression coefficient for each environmental covariate.

For these reasons, it is desirable to consider more parsimonious alternatives. Specifically, one may consider regressing w_j on p observable covariates x_{jk} ($k = 1, ..., p$), that is (Guo et al., 2021; Li et al., 2018),

$$w_j = \theta_0 + \theta_1 x_{j1} + \theta_2 x_{j2} + ... + \theta_p x_{jp} \quad (2)$$

Importantly, this is just one multiple regression, instead of n multiple regressions for n genotypes. Inserting this into (1), we find

$$\eta_{ij} = \alpha_i + \beta_i(\theta_0 + \theta_1 x_{j1} + \theta_2 x_{j2} + ... + \theta_p x_{jp}) \quad (3)$$

It is seen that the regression model is not linear in the parameters either, and there is an overparameterization that needs to be resolved. Specifically, the intercept term α_i is fully confounded with the multiplicative term $\beta_i\theta_0$. In fact, because of this is confounding, we may drop the intercept term θ_0 without loss of generality, as this term will then be absorbed by the intercept α_i. The only caveat when dropping θ_0 is that the model no longer involves a regression on the environmental mean. Instead, we have a regression on a synthetic covariate formed as a linear combination of observable environmental covariates. As this view is the most useful one for several of the methods considered below, especially when extending the models to comprise more than one synthetic environmental covariate (Section 4), we here also state the corresponding reparameterized (and equivalent) model explicitly. Thus,

$$\eta_{ij} = \alpha_i + \beta_i z_j \quad (4)$$

where

$$z_j = \theta_1 x_{j1} + \theta_2 x_{j2} + ... + \theta_p x_{jp} \quad (5)$$

is a synthetic covariate. Note that when suitable observable covariates are chosen, z_j can be regarded as a stress index (Chapuis et al., 2012). Several methods are possible to fit this regression model, and they differ in the way they deal with the overparameterization. Here, we will consider several options, starting from simple but approximate methods assuming balanced data and ending with an approach that will yield the least squares fit, which also works for unbalanced data, and is readily extended to mixed models. Intermediate approaches give up optimality of the final estimate, with the benefit of a simplification of the pivotal step to find the values of the coefficients θ_k ($k = 1, ..., p$) for the synthetic covariate z_j. In this section, it will be assumed that all effects except the residual error term are fixed.

3.1 | Balanced data

(i) Consider the environmental averages based on (3):

$$\bar{\eta}_{\bullet j} = \bar{\alpha}_\bullet + \bar{\beta}_\bullet(\theta_0 + \theta_1 x_{j1} + \theta_2 x_{j2} + ... + \theta_p x_{jp}) \quad (6)$$

This model for environmental means suggests that a multiple regression of observed environmental means $\bar{y}_{\bullet j}$ on the covariates provides estimates of slopes $\tilde{\theta}_k = \bar{\beta}_\bullet \theta_k$ for covariates x_{jk} ($k = 1, ..., p$) and the intercept $\bar{\alpha}_\bullet + \bar{\beta}_\bullet \theta_0$. Without loss of generality, we may then use

$$z_j = \tilde{\theta}_1 x_{j1} + \tilde{\theta}_2 x_{j2} + ... + \tilde{\theta}_p x_{jp} \quad (7)$$

as our predictor for the environmental index in (4). This approach does not yield a least squares fit.

(ii) Instead of using environmental means, we can use CCR (Digby, 1979; Gabriel & Zamir, 1979; Hadasch et al., 2018) to iteratively estimate $\theta_1, ..., \theta_p$ based on (5) in the criss step, then pretend that these estimates are known constants, and estimate α_i and β_i in the cross step until convergence. This method provides a least squares fit.

(iii) We may fit the FR model

$$\eta_{ij} = \alpha_i + \gamma_{i1} x_{j1} + \gamma_{i2} x_{j2} + ... + \gamma_{ip} x_{jp} \quad (8)$$

and subsequently subject the matrix of fitted terms $\{\hat{\gamma}_{i1} x_{j1} + \hat{\gamma}_{i2} x_{j2} + ... + \hat{\gamma}_{ip} x_{jp}\}$ to an SVD. The first term of this decomposition provides the least squares fit for $\beta_i z_j$ in (1) with z_j as given in (5) (Davies & Tso, 1982; Hardwick & Wood, 1972; van Eeuwijk, 1992; Wood, 1976). The least squares fit for α_i in (1) is that obtained from the fit of (8). This approach is also known as reduced rank regression or redundancy analysis (RA).

(iv) We fit (4) using the first factor extracted by partial least squares (PLS; Aastveit & Martens, 1986; Vargas et al., 1998), regarding the genotypic responses in an environment as a single multivariate response. PLS is usually performed scaling both the response variables and the covariates to zero mean and unit variance. For MET data, it is preferable to preserve the original scale of the genotypic responses. When using a multivariate PLS routine that allows scaling to be suppressed, this analysis may be obtained by scaling the covariates but not the responses before submitting the data to the PLS routine with the scaling option switched off.

(v) We may fit (4) directly by nonlinear least squares (Ng & Grunwald, 1997; Ng & Williams, 2001). We do not use this method in the worked example in Section 6, because it is tedious to implement for more than one synthetic covariate.

3.2 | Unbalanced data

Among the five methods for balanced data reviewed in Section 3.1, method (i) cannot be used with unbalanced data. Methods (ii) and (v) work equally with unbalanced data. Methods (iii) and (iv) can be used with modification as described below.

(iii) Fit the FR model (8) and obtain the matrix of fitted terms $\{\hat{\gamma}_{i1} x_{j1} + \hat{\gamma}_{i2} x_{j2} + ... + \hat{\gamma}_{ip} x_{jp}\}$ of all cells, including the ones with no data. Then proceed with RA as in (iii) in Section 3.1.

(iv) We can use a method akin to the EM method for fitting the AMMI model (Gauch & Zobel, 1990). The method starts by filling the empty cells of the genotype–environment classification with some plausible values, for example, the genotype means. Then multivariate PLS is applied to the completed data and predictions are obtained to update the imputed values for the cells with missing data. This is repeated until predictions for the cells with missing data converge. For details, see Nelson et al. (1996). This EM method is implemented in the PLS procedure of SAS. It does not seem to be available in R but is easily programmed using any PLS package for complete data.

4 | MORE THAN ONE SYNTHETIC ENVIRONMENTAL COVARIATE

The model (4) may be extended as

$$\eta_{ij} = \alpha_i + \beta_{i1} z_{j1} + \ldots + \beta_{iq} z_{jq} \quad (9)$$

where $z_{j1},\ldots,z_{jq}$ are the q synthetic environmental covariates and $\beta_{i1},\ldots,\beta_{iq}$ are the corresponding genotype-specific slopes. This model may be estimated using the same methods as those in Section 3, excluding method (i). The only additional requirement is that estimability constraints, such as orthogonality, need to be imposed on estimates of both $z_{j1},\ldots,z_{jq}$ and $\beta_{i1},\ldots,\beta_{iq}$, when fitting explicit regressions on observable environmental covariates of the form

$$z_{jh} = \theta_{1h} x_{j1} + \theta_{2h} x_{j2} + \ldots + \theta_{ph} x_{jp} \ (h = 1,\ldots,q) \quad (10)$$

Note that we have dropped the intercept terms θ_{0h}, as these will be confounded with the genotype-specific intercepts. Where an SVD is used (methods iii and iv), constraints are automatically imposed as part of the decomposition, and we here extract the first q multiplicative terms. Where CCR or nonlinear least squares are used (methods ii and v), the constraints need to be actively imposed on the q multiplicative terms, e.g., by subjecting current estimates to an SVD on each iteration (Hadasch et al., 2018). Below, we propose an adaptation of CCR (ii in Section 3.1) to impose such constraints when multiple synthetic environmental covariates are estimated.

To initialize the algorithm, an SVD is applied to the residuals matrix from the additive model $\eta_{ij} = \alpha_i + u_j$ to obtain initial values for z_{jh} from the q first right vectors. In the last step of initialization, z_{jh} are fixed while α_i and β_{ih} are estimated. Iterations start with the criss step applied to shifted environment-genotype means $y_{ij}^* = y_{ij} - \tilde{\alpha}_i$, where $\tilde{\alpha}_i$ is the current estimate of α_i. An SVD is then applied to the matrix of fitted terms for $\beta_{i1} z_{j1} + \ldots + \beta_{iq} z_{jq}$ using (10), and right vectors of an SVD provide an estimate of the q values of z_{jh}. Similarly, for the cross step, z_{jh} are fixed while α_i and β_{ih} are re-estimated. An SVD is then applied to the matrix of fitted terms for $\beta_{i1} z_{j1} + \ldots + \beta_{iq} z_{jq}$ and left vectors of an SVD multiplied by singular values provide estimates of β_{jh}. Criss and cross steps are repeatedly iterated until convergence. This procedure adequately imposes orthogonality constraints that ensure estimability of the multiplicative terms.

5 | RANDOM-EFFECTS EXTENSIONS OF THE MODEL

If the regression model comprises random effects, it is referred to as a mixed model. There are various reasons why random effects may be needed with MET. Here, we will generally regard environments as a random factor, assuming that trial environments represent a random sample from a target population of environments (TPE). The assumption of random environments is at the heart of different concepts of phenotypic stability, among which FW regression and its extensions are one of the most prominent examples (Becker & Leon, 1988; Piepho, 1998). Furthermore, this assumption is needed to project the model to unseen environments in the TPE (Buntaran et al., 2021). This assumption will therefore be made throughout this section. A second reason for introducing random effects is to allow genetic markers to be used for modelling genotypic effects as in genomic prediction (Bernardo & Yu, 2007; Meuwissen et al., 2001). In this case, which is considered in Section 5.2, the genotype factor needs to be modelled as random as well.

5.1 | All parameters in η_{ij} are fixed

When modelling the variance over random environments, there are different aspects calling for random-effects modelling. For example, observed data may display heterogeneity of variance between genotypes in the deviations from the regression line. This genotype-specific variance has been proposed by Eberhart and Russell (1966) as an additional stability parameter to the regression coefficient β_{ik}. Furthermore, a covariance must usually be expected between genotypes in the same environment due to a residual main effect for the shared environment. This can be modelled, for example, by a random environmental main effect. Thus, our model for the mean response y_{ij} of the i-th genotype in the j-th environment may be written as

$$y_{ij} = \eta_{ij} + u_j + e_{ij} \quad (11)$$

where η_{ij} is as defined in (9), u_j is the random environmental main effect with variance σ_u^2, and e_{ij} is a random with stability variance $\sigma_{e(i)}^2$ for the i-th genotype (Eberhart & Russell, 1966; Shukla, 1972).

Of course this is just one possible mixed-model extension. The random deviations from the regression in (9) may be modelled in different ways depending on the data structure. For example, so far, we have assumed that the model is fitted to genotype–environment mean responses y_{ij}. Alternatively, replicate plot data may need to be modelled, requiring additional random design effects. With perennial crops or in long-term trials, it may be necessary to model serial correlation of observations on the same plot (Macholdt et al., 2023). All of these mixed-model extensions are straightforward with the general approaches suggested in this section. The distinction between

balanced and unbalanced data becomes a moot point here, because the variance–covariance structure used in the mixed model usually implies that ordinary least squares estimation and the use of simple arithmetic means for genotypes or environments in the estimation process are not usually a good option even when the data are balanced. Hence, we focus on the methods presented in Section 3.2. We do not present an adaptation of the PLS method (iv) to mixed models, because we are not aware of any proposed method or package that would provide this. Thus, we focus on methods (ii), (iii), and (v).

(ii) Following Nabugoomu et al. (1999), the CCR approach of Digby (1979) is easily extended in a mixed model framework. In the criss step, estimating θ_{hk} $(k=1,\ldots,p;h=1,\ldots,q)$, we fix the variance parameters at their current estimates to save computing time because this usually has fewer parameters to be estimated as fixed effects than the cross-step. These variance parameters are re-estimated in the cross step, estimating α_i and $\beta_{ih}(i=1,\ldots,n;h=1,\ldots,q)$, using residual maximum likelihood (REML). For an application of this method, see Macholdt et al. (2023).

(iii) We may fit FR model (8) under our assumed random-effects specification in (11) using REML. From the fitted model, we obtain the matrix of fitted terms $\{\widehat{\gamma}_{i1}x_{j1}+\widehat{\gamma}_{i2}x_{j2}+\ldots+\widehat{\gamma}_{ip}x_{jp}\}$ of all cells, including the ones with missing data. This matrix is subjected to an SVD to estimate the parameters in (9) and (10). Next, we compute z_{jk} according to (10) and refit (9) to estimate α_i and $\beta_{ih}(i=1,\ldots,n;h=1,\ldots,q)$. As a refinement, we can consider a weighted SVD using the approach of Hadasch et al. (2018), taking into account the variance–covariance matrix of predictions $\{\widehat{\gamma}_{i1}x_{j1}+\widehat{\gamma}_{i2}x_{j2}+\ldots+\widehat{\gamma}_{ip}x_{jp}\}$.

(v) Fit (9) with (10) directly using full maximum likelihood (ML) (e.g., using NLMIXED in SAS; also see Piepho, 1999). Note that we cannot use REML because the model is nonlinear in the parameters for η_{ij} and hence the fixed effects cannot be removed by linear contrasts as required in REML. Full ML does not account for the degrees of freedom and hence leads to more biased variance parameter estimates than REML with linear mixed models. These problems are expected to carry over to the nonlinear mixed model (9) with (10). There is no REML equivalent because the model is intrinsically nonlinear. The important consequence is that in order to avoid the bias issues with full ML, we need to resort to approximate methods such as (ii) to (iv) that make use of REML.

Apart from the above modifications for mixed models, there is always the option to use methods (ii) to (v) in the same way as in Sections 3.2 and 4 but applying these to fitted means η_{ij} using a suitable mixed model for the variation of the observed data around these means. The main objective of applying those methods is to get coefficients θ_{kh} so we can compute the synthetic covariates z_{jh} in (10). Once these are available, we simply treat them as if they were known covariates. While this only constitutes an approximation and cannot be optimal, partly because the fact is ignored that z_{jh} involves coefficients estimated from the data, it does have the advantage of simplicity. So despite imperfections, this may be the most easily implemented approach in practice.

5.2 | Some or all genotypic parameters in η_{ij} are random

When the number of genotypes is large, it may be advantageous to fit both α_i and β_{ih} as random. This may be particularly worthwhile when marker information is available so kinship information can be used to perform genomic prediction of both α_i and β_{ih} by GBLUP (Resende et al., 2021). Furthermore, genotypes need to be modelled as random in case TPE is stratified into zones and we want to borrow strength across zones (Buntaran et al., 2021).

It is not as straightforward as it may seem to simply switch from fixed to random genotypes. This is because mixed model packages assume that all random effects have an expected value of zero. This is not a realistic assumption if we postulate the FW model (1) and its extensions as considered so far, such as model (4). To see this, assume the response indeed obeys (4); that is, we only have one synthetic environmental covariate, and we now take both the intercept α_i and the slope β_i to be random. We can assume that α_i has expected value μ_α and hence set $\alpha_i=\mu_\alpha+a_i$ with $E(a_i)=0$ and $\text{var}(a_i)=\sigma_a^2=\text{var}(\alpha_i)$. This can be fitted with a linear mixed model package because the intercept enters the model linearly. We may consider the same approach for the slope, setting $\beta_i=\mu_\beta+b_i$ with $E(b_i)=0$ and $\text{var}(b_i)=\sigma_b^2=\text{var}(\beta_i)$. The model (4) may then be rewritten as

$$\eta_{ij}=\mu_\alpha+a_i+\mu_\beta z_j+b_i z_j \tag{12}$$

The main challenge with this model is that the latent score, z_j, now appears both in the fixed-effects term $\mu_\beta z_j$ and the random effects term $b_i z_j$ (Piepho, 1999). The key point is that we cannot assume $\mu_\beta=0$ without loss of generality, because there is no fixed environmental main effect in (12) that would absorb this term. Also, $\mu_\beta z_j$ is a multiplicative regression term that may be worth fitting explicitly rather than absorbing it into an environmental main effect. Furthermore, to ensure invariance to shift transformations (translations) of the observable covariates, we need to allow for a covariance between intercept and slope, that is, $\text{cov}(a_i,b_i)=\sigma_{ab}$ (Piepho, 1999). Either genotypes are modelled as independent or they are modelled as correlated by kinship for GBLUP (Resende et al., 2021).

Two methods seem feasible for fitting this model and its extensions to more than one synthetic environmental covariate. We may apply a stage-wise approach by which we first model all parameters in (9) and (10) as fixed, using either of the methods (ii) to (iv). This provides estimates of z_j, which may then be held fixed and used in place of z_j in (12), using REML to fit μ_α, a_i, σ_a^2, μ_β, b_i, and σ_b^2. Extension to models with more than one latent environmental scale is straightforward. The other method is CCR, adapting the approach of Nabugoomu et al. (1999).

A further option is to model the genotype-specific intercepts a_i in (12) as fixed, while modelling the slopes b_i as random, thus obviating

the need to fit a covariance between intercept and slope (Piepho & Ogutu, 2002). This approach will be the focus of the remainder of this section. We further add a fixed environmental main effect ε_j that absorbs $\mu_\beta w_j$. Hence, the only random effect in η_{ij} is the slope b_i, and the model can be written as

$$\eta_{ij} = \alpha_i + \varepsilon_j + b_i z_j \tag{13}$$

with $E(b_i) = 0$ and $\text{var}(b_i) = \sigma_b^2 = \text{var}(\beta_i)$. This model is readily extended to comprise several synthetic environmental covariates:

$$\eta_{ij} = \alpha_i + \varepsilon_j + b_{i1} z_{j1} + \ldots + b_{iq} z_{jq} \tag{14}$$

It will now be shown how (14) can be cast as a random-coefficient FR model for the observed covariates, in which a factor-analytic (FA) variance–covariance structure is assumed for the random regression coefficients. Let $\boldsymbol{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jp})^T$, so that $z_{jh} = \boldsymbol{x}_j^T \boldsymbol{\theta}_h$ with $\boldsymbol{\theta}_h = (\theta_{1h}, \theta_{2h}, \ldots, \theta_{ph})^T$. Moreover, rewrite the random terms in (14) as

$$b_{i1} z_{j1} + \ldots + b_{iq} z_{jq} = \boldsymbol{z}_j^T \boldsymbol{b}_i \tag{15}$$

where $\boldsymbol{z}_j = (z_{j1}, z_{j2}, \ldots, z_{jq})^T$ and $\boldsymbol{b}_i = (b_{i1}, b_{i2}, \ldots, b_{iq})^T$. Now first consider a random coefficient regression of the form

$$\eta_{ij} = \alpha_i + \varepsilon_j + \boldsymbol{x}_j^T \boldsymbol{c}_i \tag{16}$$

where $\boldsymbol{c}_i = (c_{i1}, c_{i2}, \ldots, c_{ip})^T$ with $E(\boldsymbol{c}_i) = \boldsymbol{0}$ and $\text{var}(\boldsymbol{c}_i) = \boldsymbol{\Sigma}_c$ (Longford, 1993). Next assume that we approximate the variance–covariance matrix $\boldsymbol{\Sigma}_c$ by the model

$$\boldsymbol{\Sigma}_c = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T \tag{17}$$

where $\boldsymbol{\Lambda} = \{\lambda_{kh}\}$ is a $p \times q$ matrix of factor loadings (Buntaran et al., 2021; Tolhurst et al., 2022). This amounts to an FA structure of order q for the regression coefficients $\boldsymbol{c}_i$ without residual effects and associated specific variances. To ensure estimability, we impose the constraints $\lambda_{kh} = 0$ for $h > k$ (Jennrich & Schluchter, 1986). This structure is straightforward to fit using some mixed model packages. For example, in SAS, this is the FA0(q) structure (we use this acronym henceforth to denote the structure), and in ASReml-R, it is the rr() structure. With this approximation, the regression term can be rewritten as

$$\boldsymbol{x}_j^T \boldsymbol{c}_i = \boldsymbol{x}_j^T \boldsymbol{\Lambda} \boldsymbol{v}_i = \boldsymbol{z}_j^T \boldsymbol{v}_i \tag{18}$$

where $\boldsymbol{v}_i = (v_{i1}, v_{i2}, \ldots, v_{iq})^T$ with $E(\boldsymbol{v}_i) = \boldsymbol{0}$ and $\text{var}(\boldsymbol{v}_i) = \boldsymbol{I}_q$, and

$$\boldsymbol{z}_j = \boldsymbol{\Lambda}^T \boldsymbol{x}_j \tag{19}$$

Comparing coefficients between (15) and (18), it emerges that, apart from a difference in scaling, we can equate $\boldsymbol{v}_i$ with $\boldsymbol{b}_i$ and $\boldsymbol{\Lambda}$ with $\boldsymbol{\Theta} = \{\theta_{kh}\}$. The important practical consequence of this observation is that we can simply fit (16) with an FA0(q) structure for $\boldsymbol{c}_i$, extract the estimate of $\boldsymbol{\Lambda}$ and use this in (19) to compute the q synthetic covariates $\boldsymbol{z}_j$ for any environment j, including unobserved ones, so long as we have their covariate values $\boldsymbol{x}_j$. Furthermore, with this approach we can subsequently set $\text{var}(\boldsymbol{b}_i) = \boldsymbol{I}_q$, if genotypes are to be modelled as random. The full model (12) can then be refitted, taking intercepts a_i as random as well and allowing for a covariance among a_i and $\boldsymbol{b}_i$. Alternatively, we may model genotypes as fixed for the final analysis and regard the random-effects analysis as a convenient intermediate tool for obtaining $\boldsymbol{z}_j$. In either case, the fixed regression term $\boldsymbol{\mu}_\beta^T \boldsymbol{z}_j$ can now be estimated explicitly in the final step, rather than absorbing this into an environmental main effect as in (13). This approach for the final step is approximate, as it treats $\boldsymbol{z}_j$ as if these were observable covariates not involving parameters.

6 | EXAMPLE

To illustrate the random-effects modelling discussed in Section 5, we consider the lettuce data reported in van Eeuwijk (1992) as means per genotype and environment. All analyses were implemented in both SAS and R, and the full code for both packages is found in the Supporting Information. The response is nitrate concentration. The data is balanced and stems from a single replicated trial, in which eight genotypes were evaluated at 18 points in time, which are regarded as environments for our analyses. The data comprise eight observed environmental covariates, which are scaled here to zero mean and unit variance for all regression analyses. We start by fitting model (16), dropping the covariate terms, leaving the structure $\alpha_i + \varepsilon_j$. Our baseline model in Table 1 has independent (ID) residual effects e_{ij} with constant variance. Replacing this by an AR(1) model for serial correlation of observations on the same genotype across the 18 times (environments) leads to a substantial drop in the Akaike information criterion (AIC) (Wolfinger, 1996), indicating that serial correlation is important for this repeated measures data. Next, we add random effects $\boldsymbol{x}_j^T \boldsymbol{c}_i$ as in model (16) and fit an FA0(1) structure for the

TABLE 1 Model selection of covariance structure for lettuce data using fixed effects $\alpha_i + \varepsilon_j$.

Model	Random effects[a]	Structure for random effects	Residual error structure	−2 × residual log-likelihood (deviance)	AIC
M1	-	-	ID	52.2	54.2
M2	-	-	AR(1)	4.4	8.4
M3	$\boldsymbol{x}_j^T \boldsymbol{c}_i$	FA0(1)	AR(1)	−36.2	−16.2
M4	$\boldsymbol{x}_j^T \boldsymbol{c}_i$	FA0(2)	AR(1)	−52.0	−18.0

[a]The covariates were standardized to zero mean and unit variance.

covariance among random slopes c_i. This leads to a further marked drop in AIC, showing that the covariates are important. The variance parameter estimates of the FA0(1) model are given in Table 2 for illustration. The serial correlation of 0.40 is non-negligible. The estimated loadings are used to compute the synthetic variable $z_1 = \lambda_1 x_1 + \ldots + \lambda_8 x_8$ (see Equation 19). With this, we then fit model (4), adding a random main effect $u_j \sim N(0, \sigma_u^2)$ for environments and using the AR(1) model for the residual. Regression with this covariate has a significant interaction with genotype (Table 3), indicating there are differences in sensitivity. The sensitivities (slopes) as well as the intercepts for the eight genotypes are shown in Table 4. The variance explained is substantial, as a comparison of the models with and without covariate shows (Table 5). Adding a second latent factor using FA0(2) leads to a further improvement in fit (Table 1). Detailed results are omitted here for brevity.

For comparison, we also used CCR, PLS, and RA to obtain the synthetic environmental covariate z_1 (Tables 4 and 5). For further comparison, we also included classical FW regression results in Table 4, centring the environmental mean by subtracting the overall mean. The different methods result in very similar intercept estimates with a correlation equal to 1 between all methods. Even though there are differences of scale, the results in terms of sensitivities are also very similar, except for FW and RA which show some more notable differences (Table 4, Figure 1). The variance explained is also comparable between the different approaches (Table 5).

Table 6 gives an overview of the fits obtained by the different methods and models when using the full likelihood, thus permitting comparison of models with different fixed effects (Wolfinger, 1996). We followed Verbyla (2019) and plugged the REML estimates of the variance parameters into the full likelihood. The AIC values in Table 6 reveal that CCR provides the best fit among the methods using a single synthetic covariate (z_1) or two (z_1, z_2), closely followed by FA. These fits are also better than FR, which has a large number of parameters. Leave-one-environment-out cross-validation confirms that models with two synthetic covariates have an edge compared with models with just one synthetic covariate and outperform FR (Figure 2).

TABLE 2 REML estimate of variance parameters in Model M3 of Table 1 ($q = 1$) for lettuce data.

Parameter	Estimate	S.E.
λ_1	0.009574	0.02459
λ_2	−0.06102	0.06146
λ_3	0.2711	0.1355
λ_4	−0.2763	0.1173
λ_5	0.07636	0.07623
λ_6	0.01464	0.03586
λ_7	0.09994	0.06392
λ_8	0.07464	0.04965
ρ	0.3999	0.1128
σ^2	0.02897	0.005395

TABLE 3 Wald-type *F* tests for regression with synthetic variable *z* obtained from FA0(1) model. Lettuce data.

Effect	Numerator d.f.	Denominator d.f.[a]	*F*-value	*p*-value
Genotype	7	15.5	166.54	<0.0001
z	1	16.7	8.10	0.0113
Genotype × *z*	7	42.3	12.58	<0.0001

[a]Determined using the Kenward–Roger method.

TABLE 4 Intercepts and slopes of regression with a single synthetic environmental covariate $z_1 = \lambda_1 x_1 + \ldots + \lambda_8 x_8$ computed using different methods (FW, FA, PLS, CCR, RA). Lettuce data.

	FW		FA		PLS		CCR		RA	
Genotype	Intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept	Slope
DM	3.0522	0.9742	3.0839	−2.9612	3.0875	0.2329	3.0858	1.9170	3.0873	2.2339
GT	2.4108	1.0172	2.4167	−1.0517	2.4188	0.1054	2.4163	0.7356	2.4187	1.7344
Ls	2.7496	1.2349	2.7131	−0.4989	2.7115	0.08166	2.7122	0.4153	2.7303	1.7400
Pa	3.4514	0.8500	3.4619	−2.6359	3.4601	0.1842	3.4632	1.6925	3.4651	1.8102
Pi	2.8582	0.8485	2.8677	−1.9980	2.8673	0.1589	2.8680	1.2769	2.8616	1.4612
RW	2.5074	0.9231	2.4858	−0.05262	2.4832	0.02304	2.4840	0.08799	2.4824	1.0372
Tr	1.7710	1.2111	1.7689	−1.9027	1.7700	0.1555	1.7697	1.2987	1.7886	2.3813
Wi	2.6687	0.9410	2.6687	−0.8141	2.6688	0.07303	2.6680	0.5761	2.6646	1.3852
Standard error[a]	0.1262	0.1156	0.1076	0.5911	0.1024	0.04263	0.1036	0.3485	0.1046	0.3531

[a]In each column for a parameter, the standard error is the same for all genotypes. Standard errors were adjusted using the Kenward–Roger method.

TABLE 5 Variance parameter estimates (REML) for intercept only model and for regression with a single synthetic environmental covariate $z_1 = \lambda_1 x_1 + \ldots + \lambda_8 x_8$ computed using different methods (FA, PLS, CCR, RA), as well as FR using the observed covariates x_1–x_8. Lettuce data.

		Without covariates		With covariate z_1								With covariates x_1–x_8	
				FA		PLS		CCR		RA		FR	
Parameter	Description	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
σ^2_u	Variance of environmental main effect	0.1959	0.07176	0.1466	0.05305	0.1277	0.04645	0.1311	0.04752	.04286	0.01660	0.07649	0.03738
ρ	Autocorrelation [AR(1) model]	0.6190	0.08591	0.4131	0.1061	0.3328	0.1070	0.4125	0.1072	0.5673	0.1157	0.5421	0.2480
σ^2	Residual error variance	0.06686	0.01529	0.02943	0.005464	0.03227	0.005426	0.02965	0.005544	0.05445	0.01369	0.03294	0.01568

FIGURE 1 Correlation between genotypic slopes (β_i) estimated with different models. The lower triangle shows the link between the slopes estimated with the different approaches, with the regression line in red, while the upper triangle shows the corresponding coefficient of correlation. [Color figure can be viewed at wileyonlinelibrary.com]

7 | CONCLUDING REMARKS

Among the different methods for obtaining synthetic environmental covariates to be used in model (4), the FA, CCR, and PLS approaches did best in the example. However, the FA approach, which models genotype-specific regression coefficients $\boldsymbol{c}_i$ for the observed environmental covariates $\boldsymbol{x}_j$ as random, is much easier to implement than the closest competitor, CCR, especially when use of more than one synthetic covariate is considered. Both of these methods for obtaining the synthetic covariates use a mixed model that is commensurate with the model finally fitted once the synthetic covariates are obtained. By contrast, the PLS and RA approaches do not fully take the final mixed model into account when estimating the synthetic covariates, as they essentially assume i.i.d. residual errors. The main advantage of the PLS approach over all other approaches considered here is that it can handle a larger number of observed covariates.

Using synthetic environmental covariates z_{jh} allows fitting more parsimonious regression models than when regressing directly on observed environmental covariates. Not only does this allow a more efficient analysis, but it also facilitates interpretation. When there are p observed environmental covariates x_{jk}, FR involves p regression coefficients for each genotype, which may be difficult to interpret when p is large. By contrast, if these p observed covariates are used to form a single synthetic covariate, a single regression coefficient is involved per genotype, and this can be interpreted as a sensitivity parameter, as with FW regression. It may be reiterated that in terms of numbers of parameters, our extended FW models with synthetic

TABLE 6 Deviance (full likelihood) and Akaike information criterion (AIC) for different models, plugging in REML estimates of variance parameters (Verbyla, 2019). Lettuce data.

Environmental covariates	Method to obtain synthetic covariate(s)	Deviance	Number of parameters	AIC
z_1	FA	−70.5	26	−18.5
z_1	PLS	−53.0	26	−1.0
z_1	CCR	−71.5	26	−19.5
z_1	RA	−37.0	26	15.0
z_1, z_2	FA2	−113.2	33	−47.2
z_1, z_2	PLS2	−66.9	33	−0.9
z_1, z_2	CCR2	−116.2	33	−50.2
z_1, z_2	RA2	−106.2	33	−40.2
x_1–x_8	-	−142.7	75	7.3
-	-	11.1	11	33.1

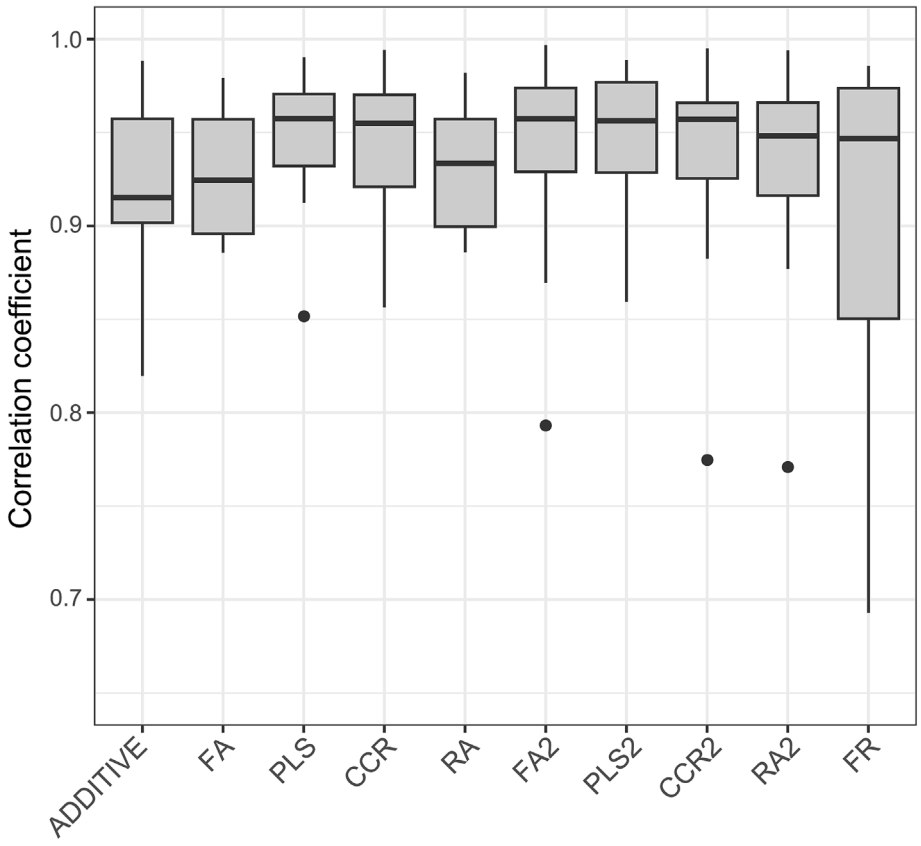

FIGURE 2 Predictive ability of different models for new environments, evaluated as the correlation coefficient between predicted and observed nitrate concentration in a leave-one-environment-out cross-validation scheme. RA2, FA2, CCR2, and PLS2 correspond respectively to RA, FA, CCR, and PLS models with two synthetic covariates. Horizontal lines in the box correspond to the medians, and circles indicate outliers. The box spans the interquartile range, and the whiskers correspond to 1.5 times the interquartile range.

covariates fall between FR (Denis, 1988) at the complex end of the scale and reaction-norm models (Jarquin et al., 2014), which essentially fit a single 'environmental kinship' matrix computed from all covariates, at the parsimonious end.

Our FA model in Section 5.2 has similarities with the model proposed by Tolhurst et al. (2022). Those authors initially model environments as fixed (see their model (1)), whereas we model environments as random throughout (except when extracting synthetic covariates using FA). When extending their model to allow for environmental covariates, the authors do consider a random environmental effect for deviations from the fixed-effects regression on covariates $\boldsymbol{x}_j$ to model the environmental main effect. The regression on observed covariates $\boldsymbol{x}_j$ would absorb our term $\mu_\beta z_j$. However, their model does not employ the more parsimonious regression on the synthetic covariates in the fixed part of the model, which is a major difference from our model. Also, synthetic variables $\boldsymbol{z}_j$ do not feature explicitly in the random part of their model, but they do so implicitly. Conversely, one aspect of our approach is that not only synthetic variables $\boldsymbol{z}_j$, but also the genotypic sentivities β_i and the genotypic mean performance α_i emerge directly from the model, which eases the interpretation of genotype–

environment interactions and the variety evaluation. Another difference is that Tolhurst et al. (2022) model genotypes as random throughout. Specifically, they fit random intercepts a_i (our notation) throughout, allowing for a covariance with slopes $\boldsymbol{c}_i$. Hence, the intercept is included in their FA structure for the random regression on the observed covariates (see their eq. (17)). From this, one could also extract coefficients for $\boldsymbol{z}_j$, but because of the presence of the random intercept, these would be different from the method used in the present paper. By contrast, in our example, we only fit random effects for slopes $\boldsymbol{c}_i$, while modelling intercepts α_i as fixed, thus obviating the need to fit the covariances among a_i and $\boldsymbol{c}_i$, which can be numerically challenging (Buntaran et al., 2021). The primary purpose of fitting our random-effects regression on observed environmental covariates is to estimate coefficients for the synthetic environmental covariates $\boldsymbol{z}_j$ in Equation (19). This is just an intermediate step before fitting the final model, which may have fixed genotypic slopes for the regression on $\boldsymbol{z}_j$.

AUTHOR CONRIBUTIONS

Hans-Peter Piepho conceptualized and wrote the paper and implemented all analyses in SAS. Justin Blancon implemented all methods in R, helped editing the manuscript. Hans-Peter Piepho developed the methods. For the CCR method, Justin Blancon developed the important extension to more than one synthetic covariate. Justin Blancon generated the figures.

ACKNOWLEDGEMENTS

Hans-Peter Piepho thanks the German Research Foundation (DFG) for financial support (grant PI 377/20-2). Justin Blancon thanks I-SITE CAP 20-25 and INRAE for financial support. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

DATA AVAILABILITY STATEMENT

The data used in this paper are available in literature and are included with the full computer code that allows reproducing all results.

ORCID

Hans-Peter Piepho https://orcid.org/0000-0001-7813-2992

REFERENCES

Aastveit, A. H., & Martens, H. (1986). ANOVA interactions interpreted by partial least squares regression. *Biometrics*, *42*, 829–844. https://doi.org/10.2307/2530697

Abou-El-Fittouh, H. A., Rawlings, J. O., & Miller, P. A. (1969). Genotype by environment interactions in cotton—Their nature and related environmental variables. *Crop Science*, *9*, 377–381. https://doi.org/10.2135/cropsci1969.0011183X000900030042x

Becker, H. C., & Leon, J. (1988). Stability analysis in plant breeding. *Plant Breeding*, *101*, 1–23. https://doi.org/10.1111/j.1439-0523.1988.tb00261.x

Bernardo, R., & Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, *47*, 1082–1090. https://doi.org/10.2135/cropsci2006.11.0690

Buntaran, H., Forkman, J., & Piepho, H. P. (2021). Projecting results of zoned multi-environment trials to new sites using environmental covariates with random coefficient models. *Theoretical and Applied Genetics*, *134*, 1513–1530. https://doi.org/10.1007/s00122-021-03786-2

Chapuis, R., Delluc, C., Debeuf, R., Tardieu, F., & Welcker, C. (2012). Resiliences to water deficit in a phenotypic platform and in the field: How related are they in maize? *European Journal of Agronomy*, *42*, 59–67. https://doi.org/10.1016/j.eja.2011.12.006

Cooper, M., & Messina, C. D. (2021). Can we harness "enviromics" to accelerate crop improvement by integrating breeding and agronomy? *Frontiers in Plant Science*, *12*, 735143. https://doi.org/10.3389/fpls.2021.735143

Costa-Neto, G., Galli, G., Fanelli Carvalho, H., Crossa, J., & Fritsche-Neto, R. (2021). EnvRtype: A software to interplay enviromics and quantitative genomics in agriculture. *G3 Genes|Genomes|Genetics*, *11*(4), jkab040. https://doi.org/10.1093/g3journal/jkab040

Davies, P. T., & Tso, M. K. S. (1982). Procedures for reduced-rank regression. *Applied Statistics*, *31*, 244–255. https://doi.org/10.2307/2347998

Denis, J. B. (1988). Two way analysis using covariates. *Statistics*, *19*, 123–132. https://doi.org/10.1080/02331888808802080

Diepenbrock, C. H., Tang, T., Jines, M., Technow, F., Lira, S., Podlich, D., Cooper, M., & Messina, C. D. (2022). Can we harness digital technologies and physiology to hasten genetic gain in US maize breeding? *Plant Physiology*, *188*, kiab527.

Digby, P. G. N. (1979). Modified joint regression analysis for incomplete variety × environment data. *Journal of Agricultural Science*, *93*, 81–86. https://doi.org/10.1017/S0021859600086159

Eberhart, S. A., & Russell, W. A. (1966). Stability parameters for comparing varieties. *Crop Science*, *6*, 36–40. https://doi.org/10.2135/cropsci1966.0011183X000600010011x

Finlay, K. W., & Wilkinson, G. N. (1963). The analysis of adaptation in a plant-breeding programme. *Australian Journal of Agricultural Research*, *14*, 742–754. https://doi.org/10.1071/AR9630742

Freeman, G. H., & Perkins, J. M. (1971). Environmental and genotype-environmental components of variability. VIII. Relations between genotypes grown in different environments and measures of these environments. *Heredity*, *27*, 15–23. https://doi.org/10.1038/hdy.1971.67

Gabriel, K. R., & Zamir, S. (1979). Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, *21*, 489–498. https://doi.org/10.1080/00401706.1979.10489819

Gauch, H. G. Jr., & Zobel, R. W. (1990). Imputing missing yield trial data. *Theoretical and Applied Genetics*, *79*, 753–761. https://doi.org/10.1007/BF00224240

Guo, X., Dutta, S., & Nettleton, D. (2021). A hierarchical spatial Finlay-Wilkinson model for analysis of multi-environment field trials. Talk presented at AgStat conference 2021, Gainesville, Florida.

Hadasch, S., Forkman, J., Malik, W. A., & Piepho, H. P. (2018). Weighted estimation of AMMI and GGE models. *Journal of Agricultural, Biological, and Environmental Statistics*, *23*, 255–275. https://doi.org/10.1007/s13253-018-0323-z

Hardwick, R. C., & Wood, J. T. (1972). Regression methods for studying genotype-environment interaction. *Heredity*, *28*, 209–222. https://doi.org/10.1038/hdy.1972.26

Jarquin, D., Crossa, J., Lacaze, X., du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Peréz, P., Calus, M., Burgueno, J., & de los Campos, G. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, *127*, 595–607. https://doi.org/10.1007/s00122-013-2243-1

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, *42*, 805–820. https://doi.org/10.2307/2530695

Li, X., Guo, T., Mu, Q., Li, X., & Yu, J. (2018). Genomic and environmental determinants and their interplay underlying phenotypic plasticity. *Proceedings of the National Academy of Science*, *115*, 6679–6684. https://doi.org/10.1073/pnas.1718326115

Longford, N. T. (1993). *Random coefficient models*. Oxford University Press.

Macholdt, J., Hadasch, S., Macdonald, A. J., Perryman, S., Piepho, H. P., Scott, T., Styczen, M., & Storkey, J. (2023). Climatic drivers of long-term trends in yield variability of grassland depending on lime x fertilizer treatments. *Agronomy for Sustainable Development*, *43*, 37. https://doi.org/10.1007/s13593-023-00885-w

Mandel, J. (1961). Non-additivity in two-way analysis of variance. *Journal of the American Statistical Association*, *56*, 878–888. https://doi.org/10.1080/01621459.1961.10482132

Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*, 1819–1829. https://doi.org/10.1093/genetics/157.4.1819

Nabugoomu, F., Kempton, R. A., & Talbot, M. (1999). Analysis of series of trials where varieties differ in sensitivity to locations. *Journal of Agricultural, Biological, and Environmental Statistics*, *4*, 310–325. https://doi.org/10.2307/1400388

Nelson, P. R. C., Taylor, P. A., & MacGregor, J. F. (1996). Missing data methods in PCA and PLS: Score calculation with incomplete observations. *Chemometrics and Intelligent Systems*, *35*, 45–65. https://doi.org/10.1016/S0169-7439(96)00007-X

Ng, M. P., & Grunwald, G. K. (1997). Nonlinear regression analysis of the joint-regression model. *Biometrics*, *43*, 1366–1372. https://doi.org/10.2307/2533503

Ng, M. P., & Williams, E. R. (2001). Joint-regression analysis for incomplete two-way tables. *Australian & New Zealand Journal of Statistics*, *43*, 201–206. https://doi.org/10.1111/1467-842X.00165

Piepho, H. P. (1998). Methods for comparing the yield stability of cropping systems—A review. *Journal of Agronomy and Crop Science*, *180*, 193–213. https://doi.org/10.1111/j.1439-037X.1998.tb00526.x

Piepho, H. P. (1999). Fitting a regression model for genotype-by-environment data by methods for nonlinear mixed models. *Biometrics*, *55*, 1120–1128. https://doi.org/10.1111/j.0006-341X.1999.01120.x

Piepho, H. P. (2022). Prediction of and for new environments: What's your model? *Molecular Plant*, *15*, 581–582. https://doi.org/10.1016/j.molp.2022.01.018

Piepho, H. P., & Ogutu, J. O. (2002). A simple mixed model for trend analysis in wildlife populations. *Journal of Agricultural, Biological, and Environmental Statistics*, *7*, 350–360. https://doi.org/10.1198/108571102366

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhyā: The Indian Journal of Statistics, Series A*, *26*, 329–358.

Resende, R. T., Piepho, H. P., Rosa, G. J. M., Silva-Junior, O. B., e Silva, F. F., de Resende, M. D. V., & Grattapaglia, D. (2021). Enviromics: Applications and perspectives on envirotypic assisted breeding. *Theoretical and Applied Genetics*, *134*, 95–112. https://doi.org/10.1007/s00122-020-03684-z

Shukla, G. K. (1972). Some statistical aspects of partitioning genotype-environmental components of variability. *Heredity*, *29*, 237–245. https://doi.org/10.1038/hdy.1972.87

Tolhurst, D. J., Gaynor, R. C., Gardunia, B., Hickey, J. M., & Gorjanc, G. (2022). Genomic selection using random regressions on known and latent environmental covariates. *Theoretical and Applied Genetics*, *135*, 3393–3415. https://doi.org/10.1007/s00122-022-04186-w

van Eeuwijk, F. A. (1992). Interpreting genotype-environment interaction using redundancy analysis. *Theoretical and Applied Genetics*, *85*, 92–100. https://doi.org/10.1007/BF00223849

van Eeuwijk, F. A. (1995). Linear and bilinear models for the analysis of multi-environment trials: I. An inventory of models. *Euphytica*, *84*, 1–7. https://doi.org/10.1007/BF01677551

Vargas, M., Crossa, J., Sayre, K., Reynolds, M., Ramírez, M. E., & Talbot, M. (1998). Interpreting genotype × environment interaction using partial least squares regression. *Crop Science*, *38*, 679–689. https://doi.org/10.2135/cropsci1998.0011183X003800030010x

Verbyla, A. P. (2019). A note on model selection using information criteria for general linear models estimated using REML. *Australian & New Zealand Journal of Statistics*, *61*, 39–50. https://doi.org/10.1111/anzs.12254

Williams, W. J. (1952). The interpretation of interactions in factorial experiments. *Biometrika*, *39*, 65–81. https://doi.org/10.1093/biomet/39.1-2.65

Wolfinger, R. D. (1996). Heterogeneous variance: Covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics*, *1*, 205–230. https://doi.org/10.2307/1400366

Wood, J. T. (1976). The use of environmental variables in the interpretation of genotype-environment interaction. *Heredity*, *37*, 1–7. https://doi.org/10.1038/hdy.1976.61

Xu, Y. (2016). Envirotyping for deciphering environmental impacts on crop plants. *Theoretical and Applied Genetics*, *129*, 653–673. https://doi.org/10.1007/s00122-016-2691-5

Yan, W., & Kang, M. S. (2003). *GGE biplot analysis*. CRC Press.

Yates, F., & Cochran, W. G. (1938). The analysis of groups of experiments. *Journal of Agricultural Science*, *28*, 556–580. https://doi.org/10.1017/S0021859600050978

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Piepho, H.-P., & Blancon, J. (2023). Extending Finlay–Wilkinson regression with environmental covariates. *Plant Breeding*, *142*(5), 621–631. https://doi.org/10.1111/pbr.13130