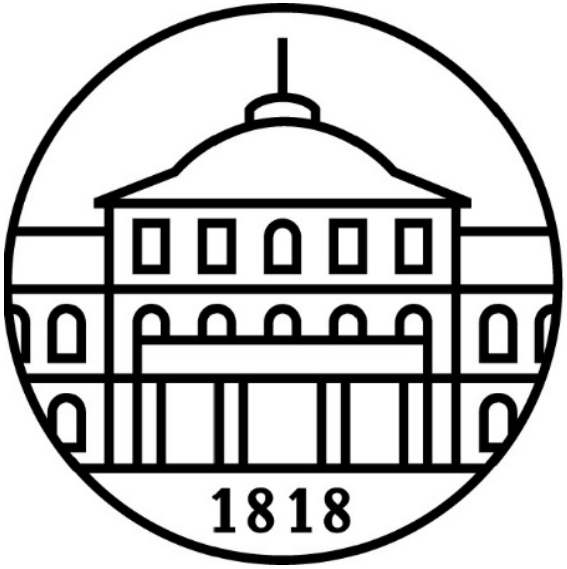
Modelling of data from multi-environment trials for predictions into new environments – From ANOVA-type models to factor-analytic models and regression on environmental covariates

Hans-Peter Piepho

Biostatistics Unit
University of Hohenheim
Germany



References

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

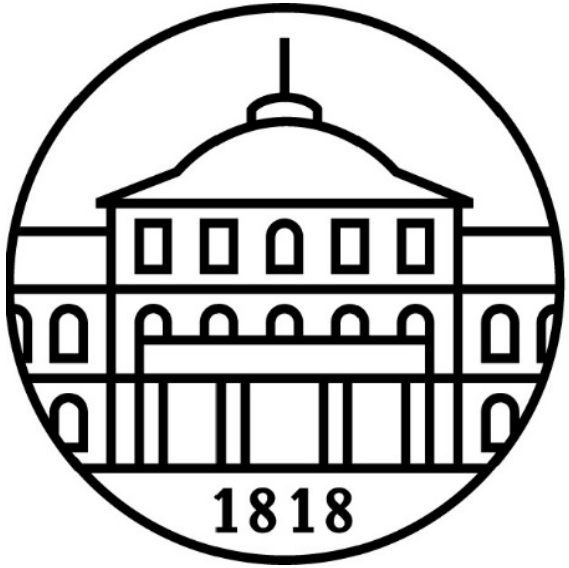
Piepho, H.P., Williams, E.R. (2024): Factor-analytic variance-covariance structures for prediction into a target population of environments. *Biometrical Journal* **66**, e202400008. https://doi.org/10.1002/bimj.202400008

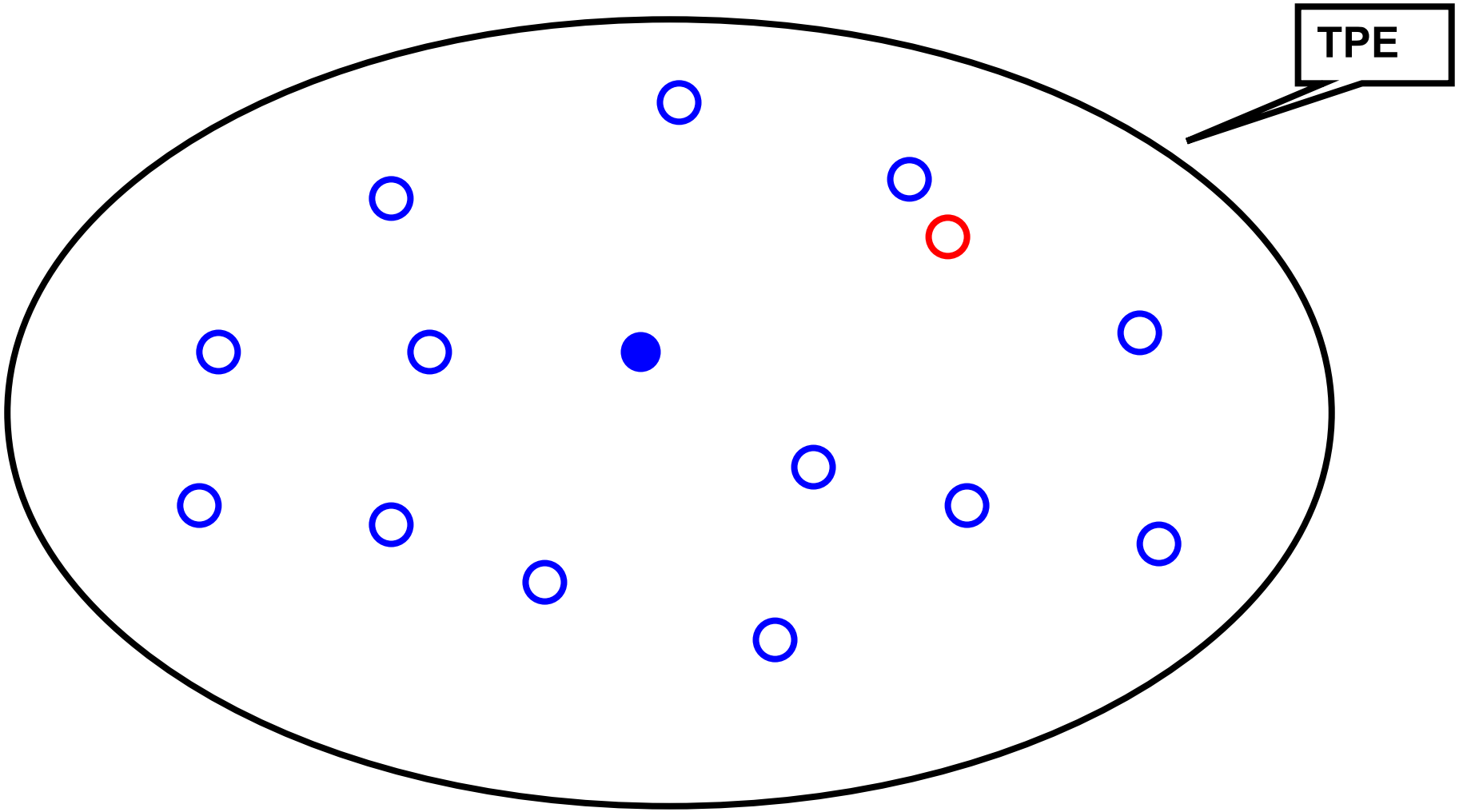
Hrachov, M., Piepho, H.P., Rahman, N.F., Malik, W. (2025): Regression approaches for modelling genotype-environment interaction and making predictions into a target population of environments. https://doi.org/10.48550/arXiv.2507.18125

ANOVA mixed models for MET

Hans-Peter Piepho

Biostatistics Unit
Universität Hohenheim
Germany





my farm

locations of trial network

"mean" of target population of environments (TPE) (Piepho & Williams 2024)

1. Introduction

Regarding an MET conducted at six locations over two years, Immer et al. (1934) asserted that

"insofar as these six stations constitute a random sample of conditions to be found over the entire state and that these 2 years are a random sample of weather conditions to be encountered in future years, general recommendations may be drawn up for the entire state with reasonable assurance that the variety or varieties recommended will prove to be consistently superior in most places of the state and in most years."

Yates and Cochran (1938):

"Any experimental programme which is instituted to assess the value of any particular treatment or practice or to determine the optimal amount of such treatment should therefore be so designed that it is capable of furnishing an accurate and unbiased estimate of the average response to this treatment in the various combinations of circumstances in which the treatment will subsequently be applied. The simplest and indeed the only certain way of ensuring that this condition shall be fulfilled is to choose fields on which the experiments are to be conducted by random selection from all fields which are to be covered by the subsequent recommendations."

Yates and Cochran (1938) go on to acknowledge that

"it is usually impossible to secure a set of sites selected entirely at random. An attempt can be made to see that the sites actually used are a "representative" selection, but averages of the responses from such a collection of sites cannot be accepted with the same certainty as would the averages from a random collection."

- MET are routinely analysed using linear mixed models.

- A key step in the use of such models is to decide for each factor whether it is to be regarded as fixed or random.

- Once the status of each factor as either fixed or random has been established, one may use the convention that if an effect involves at least one random factor, it should be modelled as random (Nelder 1977).

- Hence, if environment is considered as a random factor, it follows that variety $\times$ environment interactions are to be modelled as a random effect.

Immer et al. (1934):

"Unless the mean square for varieties exceeds significantly the mean square for interaction of varieties × stations no general recommendations can be made for the entire state. In like manner, unless the variety mean square exceeds significantly the mean square for varieties × years, we have no assurance that the varieties recommended will consistently prove superior in subsequent years."

Fisher (1935, Section 65):

"In fact, if our concern is to ascertain not merely the best variety on the aggregate of the (...) fields actually used, but to ascertain which is the best over the whole area deemed suitable for this type of crop, within the region from which the sites of the experiment have been selected, the comparison between varieties V and interaction of varieties and places VP will be the more appropriate. For if the (…) sites have been chosen at random from this area, a significant difference in this comparison would indicate, at the level of significance used, varietal differences applicable to the whole area."

Yates and Cochran (1938) assert that

"in the ideal case where the chosen places are a strictly random selection from all possible places (…) it would appear to be legitimate (…) to compare the mean square for varieties with that for varieties × places …"

2. ANOVA models

$$\eta_{ij} = \alpha_i + E_j + (\alpha E)_{ij} \quad (1)$$

α = genotype (fixed)

E = environment (random)

$$\eta_{ijk} = \alpha_i + L_j + Y_k + (LY)_{jk} + (\alpha L)_{ij} + (\alpha Y)_{ik} + (\alpha LY)_{ijk} \quad (2)$$

L = location (random)

Y = year (random)

(Greek letters: fixed factors; Latin letters: random factors)

Difference of two varieties *i* and *h* in a given location *j* and year *k*:

$$\eta_{ijk} - \eta_{hjk} = \alpha_i - \alpha_h + (\alpha L)_{ij} - (\alpha L)_{hj} + (\alpha Y)_{ik} - (\alpha Y)_{hk} + (\alpha LY)_{ijk} - (\alpha LY)_{hjk} \quad (3)$$

Expected value of a variety difference in the TPE:

$$E(\eta_{ijk} - \eta_{hjk}) = \alpha_i - \alpha_h \quad (4)$$

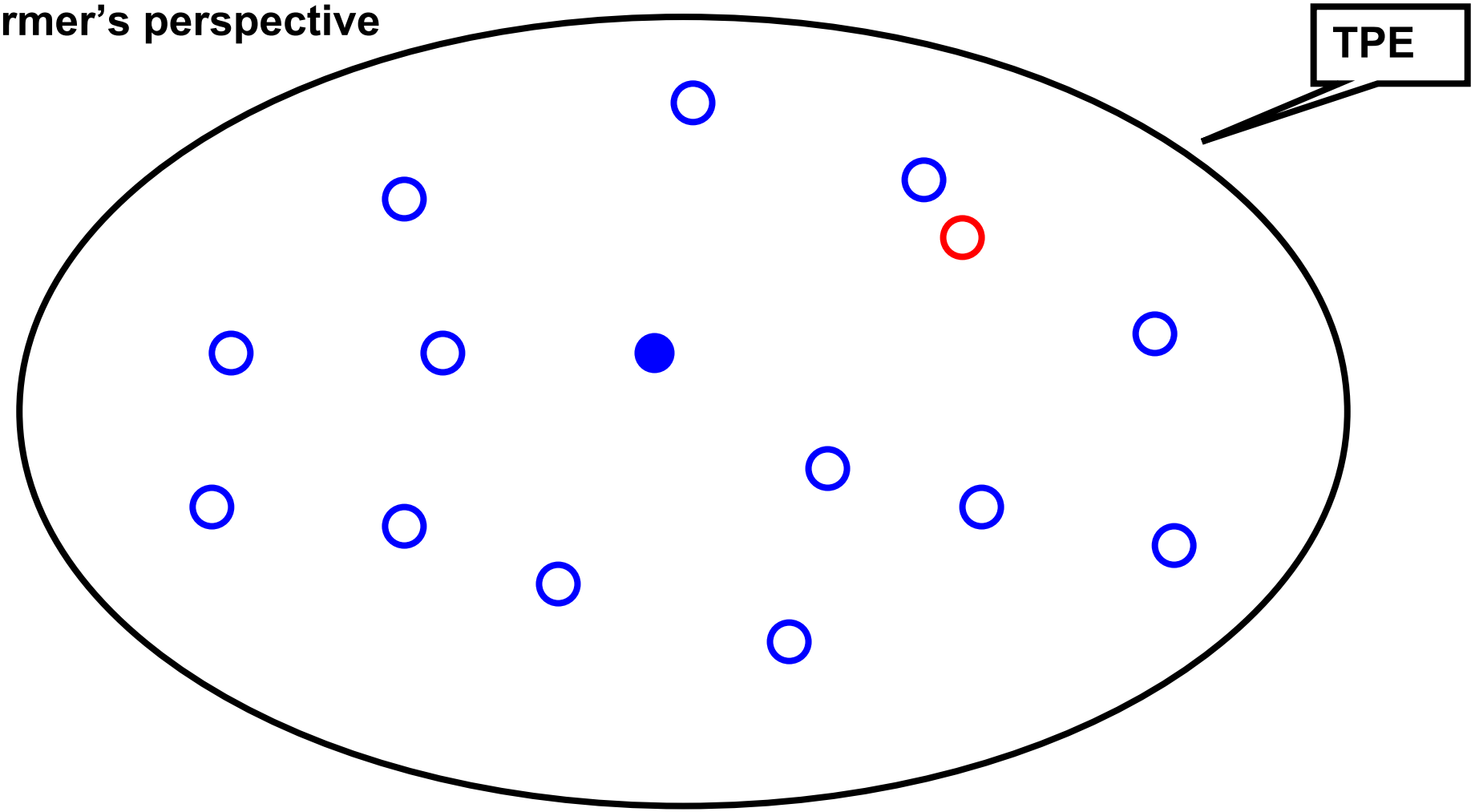
If all random effects are independently and identically distributed (i.i.d.), we find

$P(\eta_{ijk} - \eta_{hjk} > 0) > 0.5$ in a randomly selected environment if $\alpha_i - \alpha_h > 0$

(Eskridge & Mumm 1992).

$\Rightarrow$ rationale for recommending the variety with the largest mean across environments for the whole TPE $\Rightarrow$ **fine from breeder's perspective**

Farmer's perspective



- my farm
- locations of trial network
- "mean" of target population of environments (TPE)

In the best-case scenario, the farmer's location is identical to one of the trial locations:

$$E\left(\eta_{ijk} - \eta_{hjk} \middle| j\right) = \alpha_i - \alpha_h + (\alpha L)_{ij} - (\alpha L)_{hj} \qquad (5)$$

$$P\left(\eta_{ijk} - \eta_{hjk} > 0 \middle| j\right) > 0.5 \quad \text{if} \quad \alpha_i - \alpha_h + (\alpha L)_{ij} - (\alpha L)_{hj} > 0$$

$\Rightarrow$ a farmer only needs to identify the variety with the largest long-term mean at his or her location.

But: farmer's location $\neq$ trial locations (even the "closest" one)

The key question then is how to best estimate this mean

Option 1: Long-term mean at *j*-th location

$$\mathrm{var}\left[\hat{\alpha}_i - \hat{\alpha}_h + \left(\alpha\hat{L}\right)_{ij} - \left(\alpha\hat{L}\right)_{hj}\right] = 2\left(\frac{\sigma^2_{\alpha Y}}{K} + \frac{\sigma^2_{\alpha LY}}{K}\right) \quad (6)$$

K = number of years

Option 2: Means in the whole TPE

$$\mathrm{var}\left(\hat{\alpha}_i - \hat{\alpha}_h\right) = 2\left(\frac{\sigma^2_{\alpha L}}{J} + \frac{\sigma^2_{\alpha Y}}{K} + \frac{\sigma^2_{\alpha LY}}{JK}\right) \quad (7)$$

J = number of locations

Option 2 cont'd: For farmer at *j*-th location, there is also a bias

Expected deviation between the *ih*-th estimated variety difference in the TPE and the *ih*-th expected variety difference at the *j*-th location:

$$edev_{ihj} = (\alpha L)_{ij} - (\alpha L)_{hj} \tag{8}$$

Assuming that location *j* is a random draw from the TPE, we have

$$E\left(edev_{ihj}^2\right) = 2\sigma_{\alpha L}^2 \tag{9}$$

Hence, the mean squared error of a difference (MSED) is

$$MSED_{ihj}^{TPE} = \text{var}(\hat{\alpha}_i - \hat{\alpha}_h) + E\left(edev_{ihj}^2\right) = 2\left(\frac{\sigma_{\alpha L}^2}{J} + \frac{\sigma_{\alpha Y}^2}{K} + \frac{\sigma_{\alpha LY}^2}{JK}\right) + 2\sigma_{\alpha L}^2 \tag{10}$$

Comparison between Option 1 and Option 2

Option 2 is preferable when

$$2\left(\frac{\sigma^2_{\alpha Y}}{K}+\frac{\sigma^2_{\alpha LY}}{K}\right)>2\left(\frac{\sigma^2_{\alpha L}}{J}+\frac{\sigma^2_{\alpha Y}}{K}+\frac{\sigma^2_{\alpha LY}}{JK}\right)+2\sigma^2_{\alpha L}\Leftrightarrow\sigma^2_{\alpha LY}>K\left(J+1\right)\left(J-1\right)^{-1}\sigma^2_{\alpha L}$$

(11)

If the number of locations J is large, Option 2 is better if:

$$\sigma^2_{\alpha LY}>K\sigma^2_{\alpha L}$$

Single-year data (K = 1):

$$\sigma^2_{\alpha LY}>\sigma^2_{\alpha L}$$

Heroic assumption so far:

farmer's location = *j*-th location

More realistic assumption:

farmer's location $\neq$ *j*-th location

Worst-case scenario: *j*-th location and farm are random draw from the TPE:

$$MSED_{ihj}^{farm} = \text{var}\left[\hat{\alpha}_i - \hat{\alpha}_h + \left(\alpha\hat{L}\right)_{ij} - \left(\alpha\hat{L}\right)_{hj}\right] + 4\sigma_{\alpha L}^2 = 2\left(\frac{\sigma_{\alpha Y}^2}{K} + \frac{\sigma_{\alpha LY}^2}{K}\right) + 4\sigma_{\alpha L}^2 \quad (12)$$

$\Rightarrow MSED_{ihj}^{farm} > MSED_{ihj}^{TPE}$ always

Best-case scenario: Farmer can identify most similar location in trial network

Best-case scenario: Farmer can identify most similar location in trial network

$$MSED_{ihj}^{farm} = \text{var}\left[\hat{\alpha}_i - \hat{\alpha}_h + \left(\alpha\hat{L}\right)_{ij} - \left(\alpha\hat{L}\right)_{hj}\right] + C(J)\sigma_{\alpha L}^2 = 2\left(\frac{\sigma_{\alpha Y}^2}{K} + \frac{\sigma_{\alpha LY}^2}{K}\right) + C(J)\sigma_{\alpha L}^2 \quad (13)$$

where $4 \geq C(J) > 0$ with $C(1) = 4$

Simulation:

- Generate i.i.d. interactions (αL) from a standard normal distribution for *J* locations and a farm $j = 0$ for two varieties *i* and *h*
- Identify the smallest value of $\left[(\alpha L)_{ij} - (\alpha L)_{hj} - (\alpha L)_{i0} + (\alpha L)_{h0}\right]^2$ among all locations $j \in (1,2,...,J)$ in each simulation run
- Average over runs to estimate $C(J)$

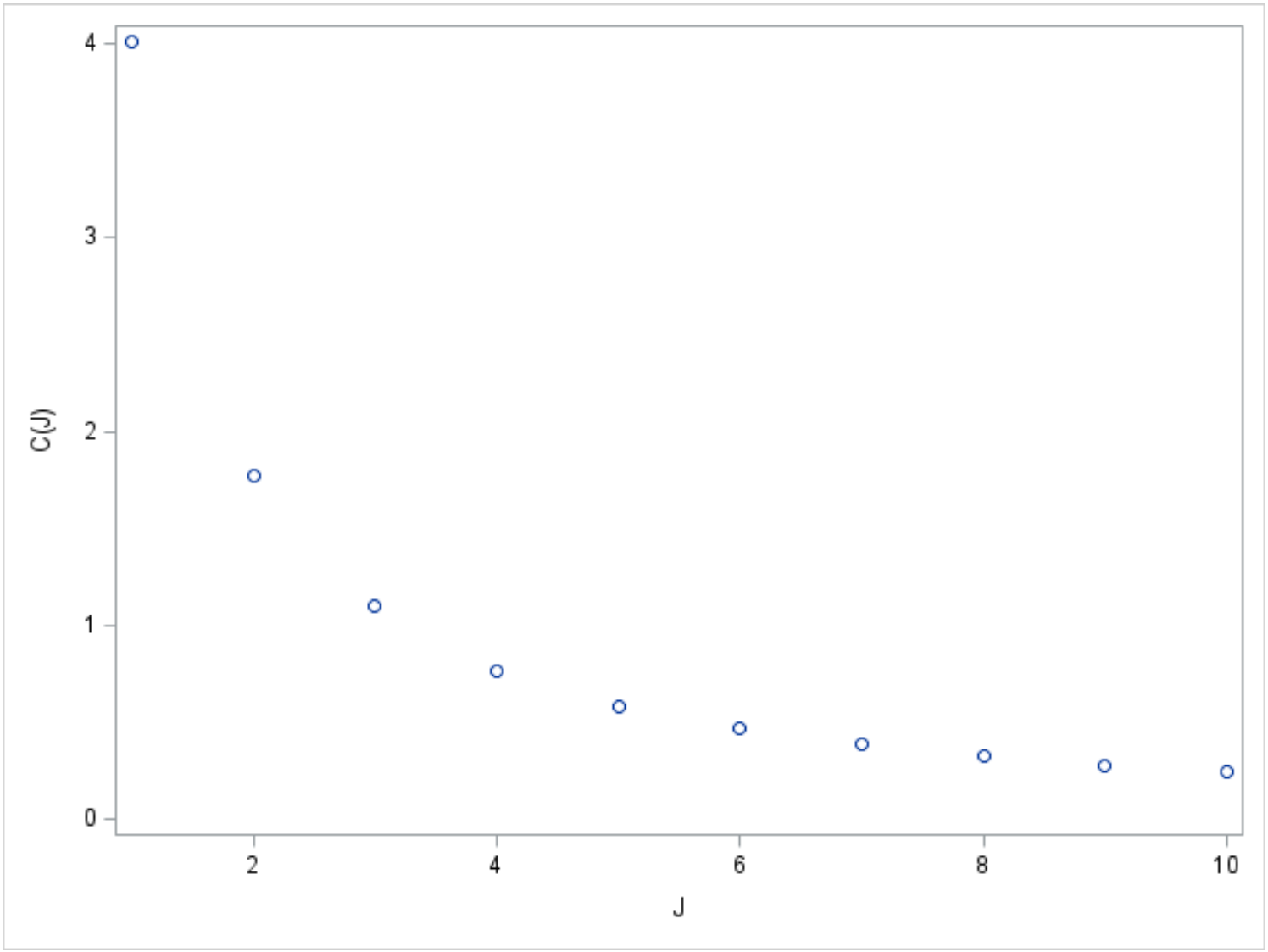
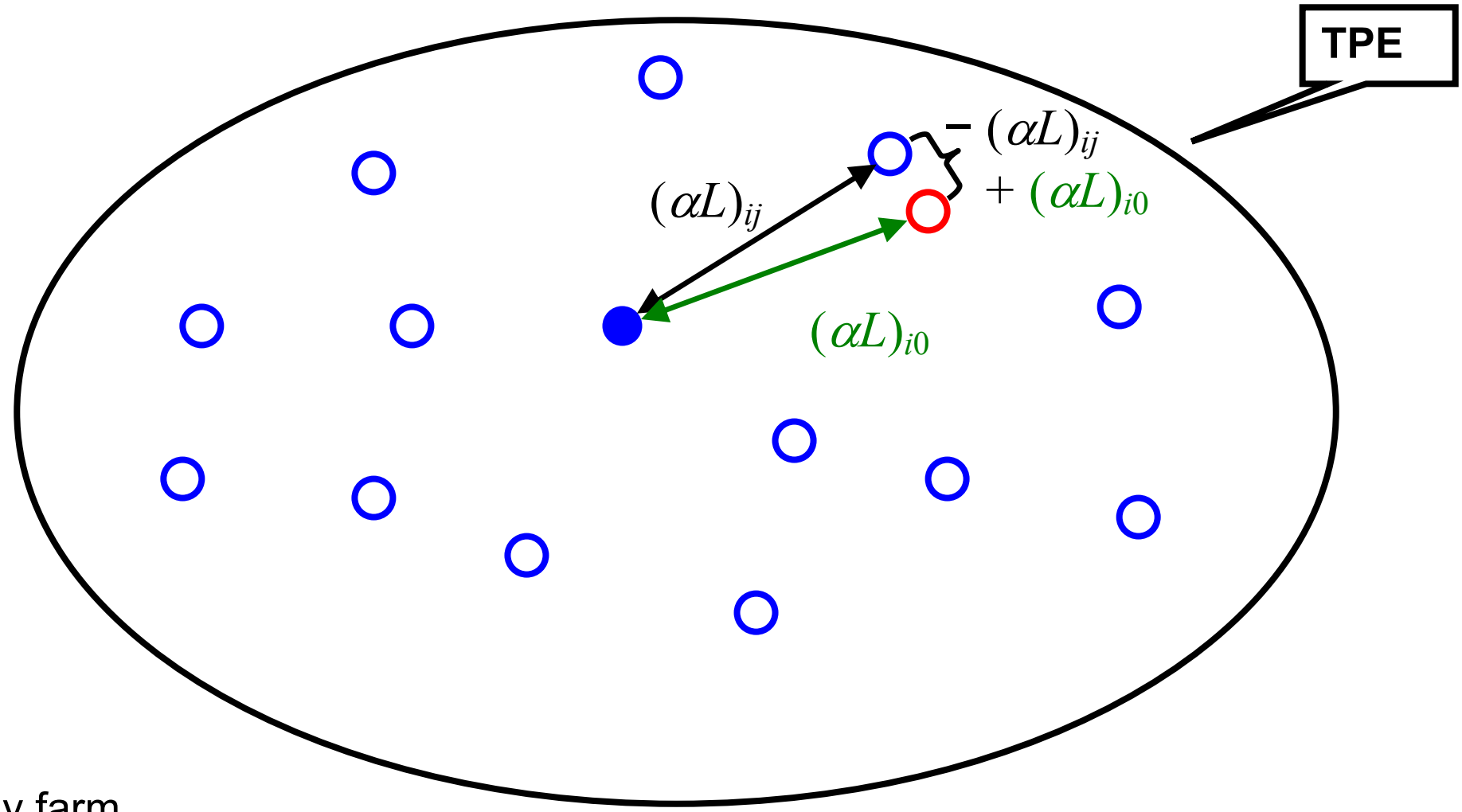


Figure 1: Simulated value of $C(J)$ for $J = 1,\ldots, 10$ based on 100,000 simulations

Observations:

- $C(1) = 4$
- potentially the bias can drop quite quickly with increasing J
- But we need to acknowledge that identification of the most similar trial location may not be easy for a farmer, especially when J is large
- Assessment assumes that only a single trait is under consideration, whereas variety choice is based on a wider array of traits in practice
- Assessment looks at just a single pair of varieties, whereas many pairs need to be looked at the select the best

$\Rightarrow$ eq. (13) just a lower bound on $MSED_{ihj}^{farm}$ that can be realized in practice



my farm

locations of trial network

"mean" of target population of environments (TPE)

Prediction for a new year

Prediction of a difference between *i*-th and *h*-th variety at the farm in a new year

$\Rightarrow$ add the variance $2\left(\sigma^2_{\alpha Y} + \sigma^2_{\alpha LY}\right)$ to the *MSED*

This amount is the same for both types of mean (TPE vs. *j*-th location)

$\Rightarrow$ no bearing on their relative merits

Two important general conclusions

(1) For a farmer the estimated variety mean in the TPE is likely to provide a more reliable basis for variety recommendations than the variety mean at the trial location considered to be most similar to the farmer's location

(2) Can judge the relative merits of TPE means versus location means via the MSED because environments (locations, years) are modelled as random

By comparison, when environment is a fixed factor:

- Can only estimate $\overline{\eta}_{i\bullet} = \alpha_i + \overline{\varepsilon}_{\bullet} + \overline{(\alpha\varepsilon)}_{i\bullet}$
- Cannot extrapolate to whole TPE

3. Random genotypic effects as an option

Two good reasons to model genotypes as random:

(1) Want to do genomic prediction (GBLUP)
(Bernardo 1994; Meuwissen et al. 2001)

(2) Want to borrow strength across zones
(Atlin et al. 2000)

Bottom line:

Random genotypes is an option

Random environments is a must

Random-effects model

$$\eta_{ijk} = \mu + a_i + L_j + Y_k + (LY)_{jk} + (aL)_{ij} + (aY)_{ik} + (aLY)_{ijk} \qquad \text{(i)}$$

a = random genotype

L = random location

Y = random year

Long-term mean of the *i*-th genotype at the *j*-th location:

$$\eta_{ij} = \mu + a_i + L_j + (aL)_{ij}$$

General form of estimator

$$\hat{\eta}_{ij} - \hat{\eta}_{hj} = S_a\left(\bar{y}_{i\bullet\bullet} - \bar{y}_{h\bullet\bullet}\right) + S_{aL}\left(\bar{y}_{ij\bullet} - \bar{y}_{hj\bullet}\right) \qquad \text{(ii)}$$

y_{ijk} = observed mean of the *i*-th genotype at the *j*-th location in the *k*-th year

Special cases:

(1) $S_a = 0$ & $S_{aL} = 1 \Rightarrow \bar{y}_{ij\bullet}$ = BLUE of η_{ij} when genotype and location are fixed

(2) $S_a = 1$ & $S_{aL} = 0 \Rightarrow \bar{y}_{i\bullet\bullet}$ = BLUE of TPE mean when genotype is fixed

Best linear unbiased prediction

$$\hat{\eta}_{ij} - \hat{\eta}_{hj} = S_a\left(\bar{y}_{i\bullet\bullet} - \bar{y}_{h\bullet\bullet}\right) + S_{aL}\left(\bar{y}_{ij\bullet} - \bar{y}_{hj\bullet}\right) \qquad \text{(ii)}$$

BLUP:

$$S_a = \frac{JK\sigma_a^2}{JK\sigma_a^2 + K\sigma_{aL}^2 + J\sigma_{aY}^2 + \sigma_{aLY}^2}\left[1 - S_{aL}\,\frac{K\sigma_a^2 + \sigma_{aY}^2}{K\sigma_a^2}\right] \qquad \text{(iii)}$$

$$S_{aL} = \frac{K\sigma_{aL}^2}{K\sigma_{aL}^2 + \sigma_{aLY}^a} \qquad \text{(iv)}$$

Re-write estimator

$$\hat{\eta}_{ij} - \hat{\eta}_{hj} = S_a\left(\bar{y}_{i\bullet\bullet} - \bar{y}_{h\bullet\bullet}\right) + S_{aL}\left(\bar{y}_{ij\bullet} - \bar{y}_{hj\bullet}\right) \qquad \text{(ii)}$$

$\Rightarrow$ *index* computed from the genotype means in the TPE and at the *j*-th location:

$$S_a \bar{y}_{i\bullet\bullet} + S_{aL} \bar{y}_{ij\bullet}$$

$\Rightarrow$ analogous to the selection index combining family means and line-within-family means in plant breeding (Lush, 1947a,b; Piepho and Williams, 2006)

Re-write estimator (2)

$$\hat{\eta}_{ij} - \hat{\eta}_{hj} = \frac{S_a + JS_{aL}}{J}\left(\bar{y}_{ij\bullet} - \bar{y}_{hj\bullet}\right) + \frac{S_a}{J}\sum_{j' \neq j}\left(\bar{y}_{ij'\bullet} - \bar{y}_{hj'\bullet}\right) \qquad \text{(v)}$$

$\Rightarrow$ estimator for *j*-th location *borrows information* from other locations *j'*$\neq$ *j*

Mean squared error of a difference

$$MSED_{ihj}^{farm} = 2\left(S_a + S_{aL} - 1\right)^2 \sigma_a^2 + \frac{2\left(S_a + S_{aL}\right)^2}{K}\sigma_{aY}^2 + 2\left[\frac{S_a + JS_{aL}}{J} - 1\right]^2 \sigma_{aL}^2$$

$$+ 4\left(J-1\right)\left(\frac{S_a}{J}\right)^2 \sigma_{aL}^2 + 2\left(\frac{S_a + JS_{aL}}{J}\right)^2 K^{-1}\sigma_{aLY}^2 + 4\left(J-1\right)\left(\frac{S_a}{J}\right)^2 K^{-1}\sigma_{aLY}^2 + C\left(J\right)\sigma_{aL}^2$$

$\Rightarrow$ BLUP outperforms BLUE for *j*-th location mean

But: $C(J)$ does not go away

$\Rightarrow$ Unless farmer can make $C(J)$ small: better use TPE mean

Key issue:

4 interaction effects (aL) from farm to *j*-th location
2 interaction effects (aL) from farm to TPE mean

Idealized scenario

$\mu + \alpha_i$ and η_{ij} known without error ($\Rightarrow$ no estimator needed!)

$$MSED_{ihj}^{TPE} = 2\sigma_{\alpha L}^2$$

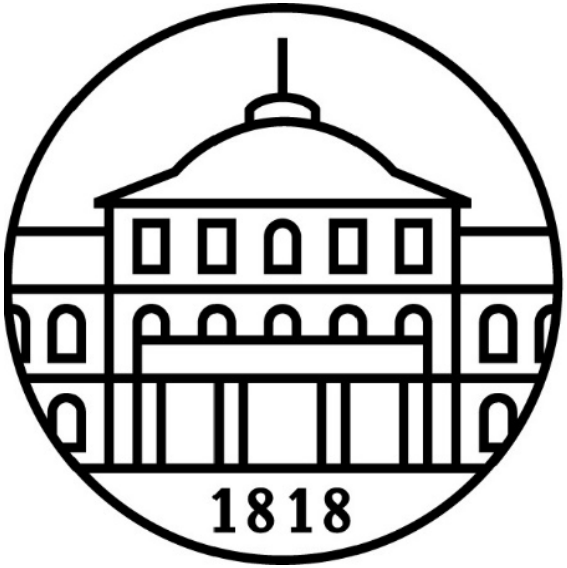
$$MSED_{ihj}^{farm} = C(J)\sigma_{\alpha L}^2 \text{, where } 0 < C(J) \leq 4$$

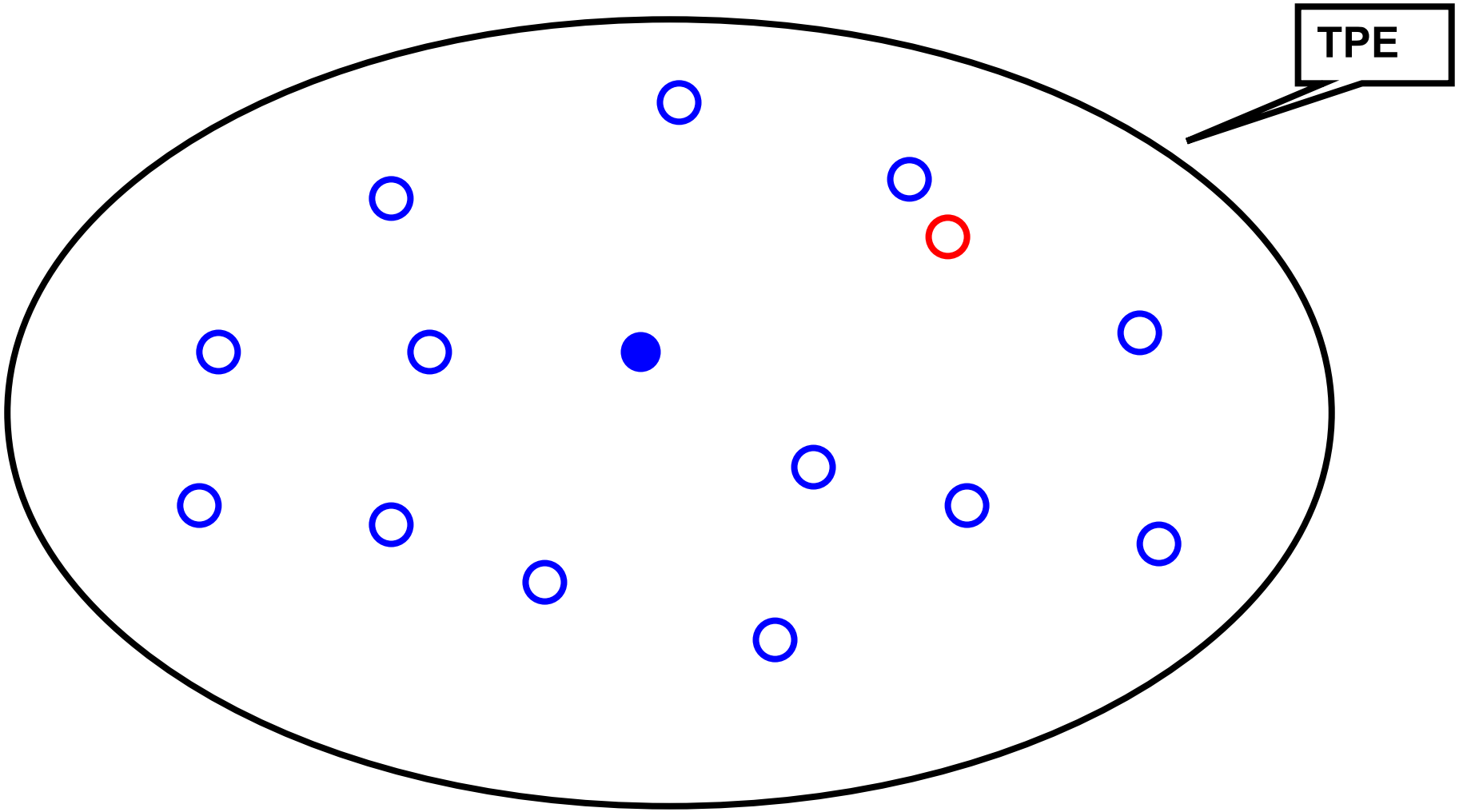
$\Rightarrow$ farmer better off using the TPE mean $(\bar{y}_{i\bullet\bullet})$ unless manages $C(J) < 2$

Factor-analytic variance-covariance structures for modelling genotype-environment interaction

Hans-Peter Piepho

Biostatistics Unit
University of Hohenheim
Germany

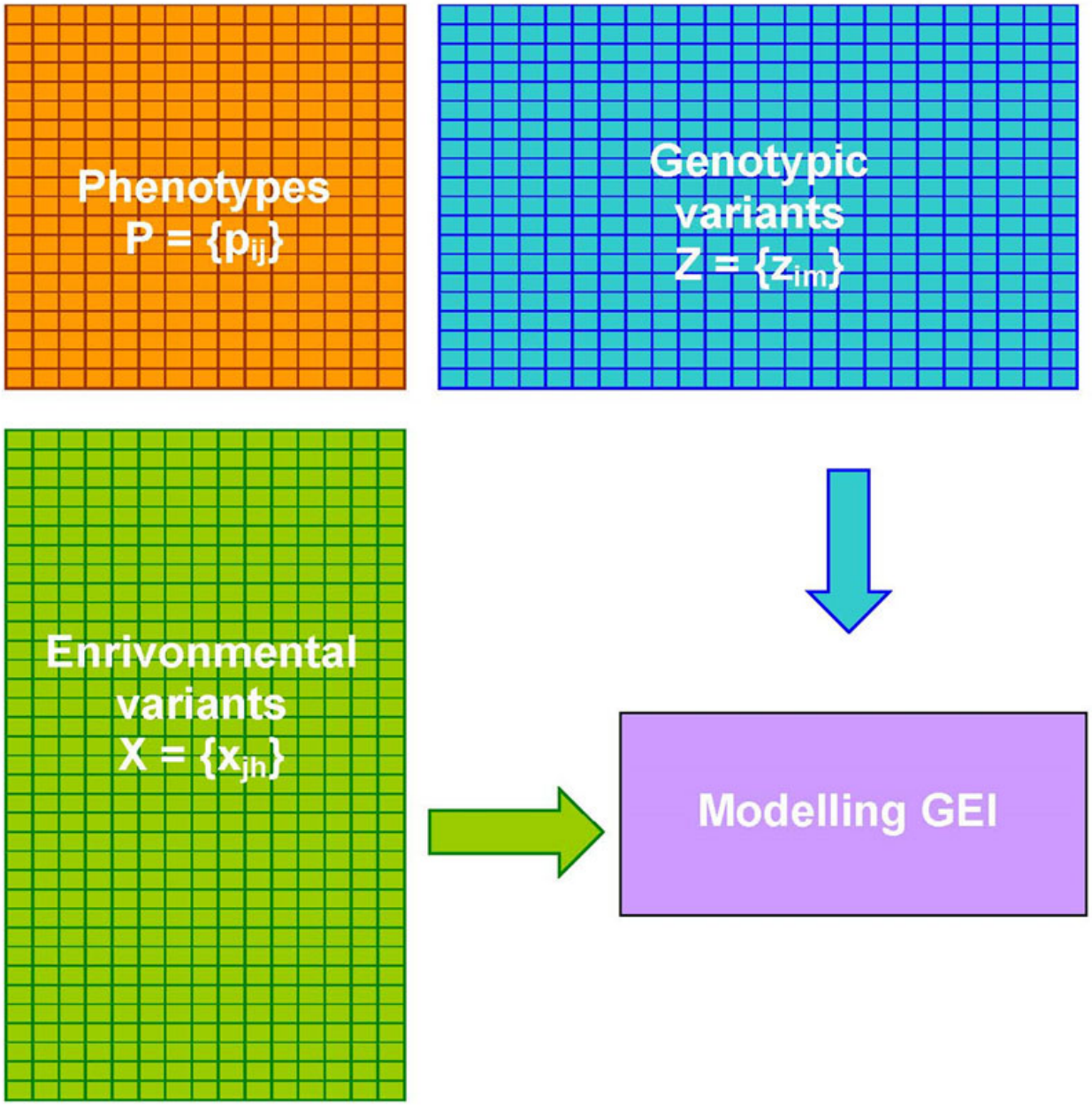




my farm

locations of trial network

“mean” of target population of environments (TPE)


Phenotypes
P = {pij}
Genotypic
variants
Z = {zim}
Enrivonmental
variants
X = {xjh}
Modelling GEI

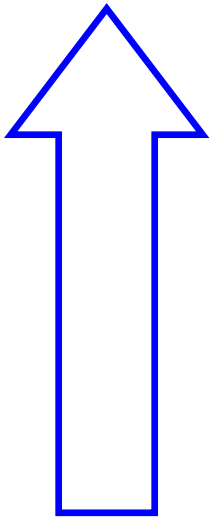
(Piepho, 2022)

Two types of models using environmental covariates (EC)

(1) Stratification of environments into environmental types (ET)
(Bustos-Korts et al. 2022)

(2) Regression models with EC:

- Factorial regression (Denis 1980)
- Extended Finlay-Wilkinson regression
(Piepho & Blancon 2023)
- Environmental kinship (Jarcquin et al. 2014)



Increasing
Complexity

Regression on latent environmental covariates

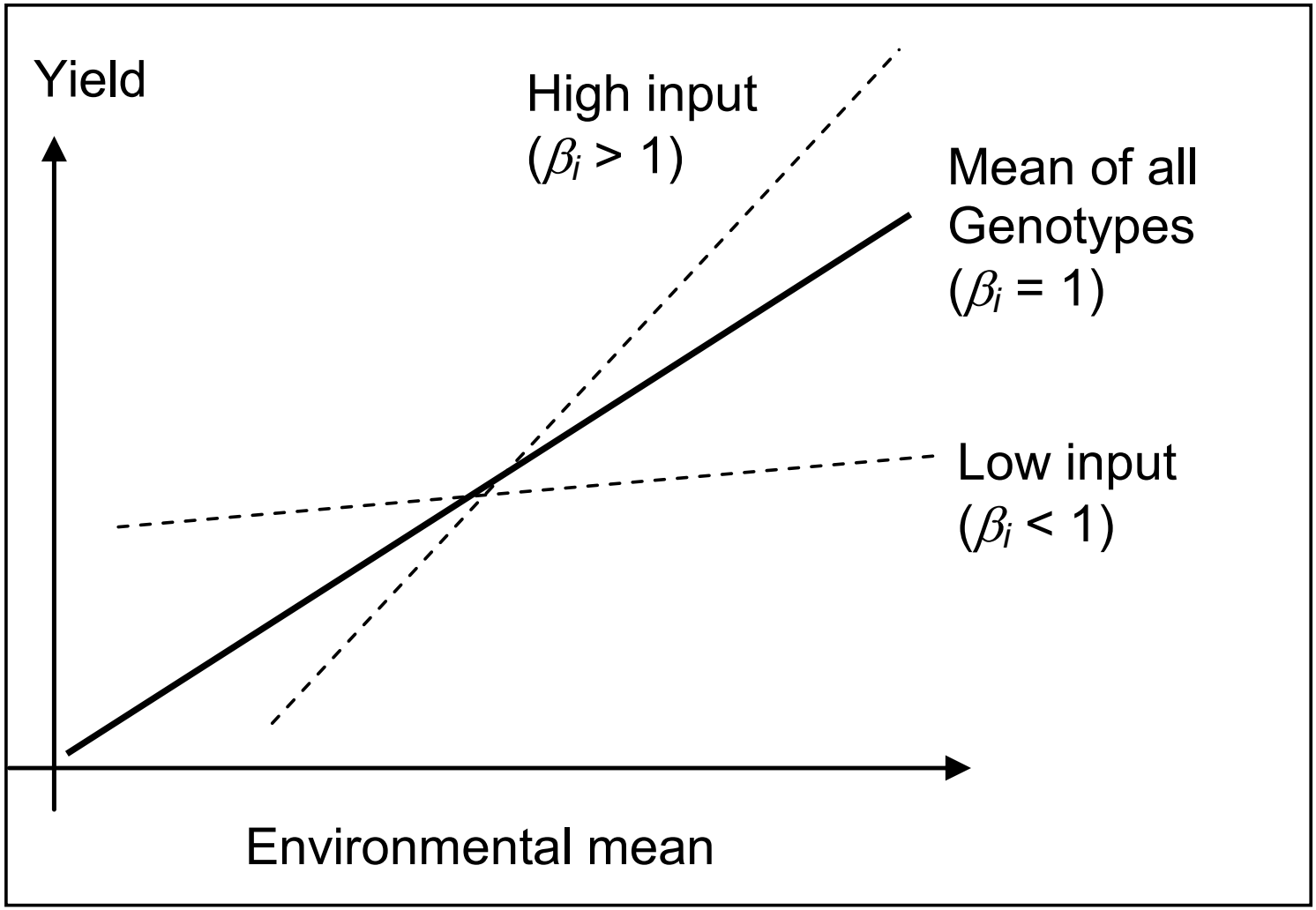
Mostly, I will not factor environments by years and locations for simplicity

But it is understood that such factorization is usually needed for a full analysis

Part I

Origin in Finlay-Wilkinson regression

1. Finlay-Wilkinson regression



Simple regression

$$\eta_{ij} = \alpha_i + \beta_i x_j + d_{ij}$$

η_{ij} = expected performance of the *i*-th genotype in the *j*-th environment

α_i and β_i = intercept and slope for the *i*-th genotype

x_j = observable covariate for the *j*-th environment

d_{ij} = random deviation from regression line

Finlay-Wilkinson regression

$$\eta_{ij} = \alpha_i + \beta_i \bar{y}_{j\bullet} + d_{ij}$$

η_{ij} = expected performance of the *i*-th genotype in the *j*-th environment

α_i and β_i = intercept and slope for the *i*-th genotype; **constraint**: $\bar{\beta}_\bullet = 1$

$\bar{y}_{\bullet j}$ = sample mean of observed responses (!) y_{ij} in *j*-th environment

d_{ij} = random deviation from regression line

Finlay-Wilkinson regression

$$\eta_{ij} = \alpha_i + \beta_i w_j + d_{ij}$$

η_{ij} = expected performance of the *i*-th genotype in the *j*-th environment

α_i and β_i = intercept and slope for the *i*-th genotype; **constraint**: $\overline{\beta}_{\bullet} = 1$

w_j = latent effect of the *j*-th environment; $w_j \sim N\left(0, \sigma_w^2\right)$

d_{ij} = random deviation from regression line

Finlay-Wilkinson regression

$$\eta_{ij} = \alpha_i + \lambda_i w_j + d_{ij} \qquad (16)$$

η_{ij} = expected performance of the *i*-th genotype in the *j*-th environment

α_i and λ_i = intercept and slope for the *i*-th genotype

w_j = latent effect of the *j*-th environment; **constraint**: $w_j \sim N(0,1)$

d_{ij} = random deviation from regression line

(Patterson & Nabugoomu 1992; Gogel 1995)

Finlay-Wilkinson regression: variance-covariance structure

Assuming random environments:

$$\text{var}(\eta_j) = \lambda\lambda^T + \text{var}(d_j) \qquad (17)$$

where

$$\eta_j = (\eta_{1j}, \eta_{2j}, ..., \eta_{Ij})^T$$

$$\lambda = (\lambda_1, \lambda_2, ..., \lambda_I)^T$$

$$d_j = (d_{1j}, d_{2j}, ..., d_{Ij})^T$$

$\Rightarrow$ This is recognized as a factor-analytic (FA) model of order one
$\Rightarrow$ This variance-covariance structure can be fit using REML
$\Rightarrow$ Reduced-rank approximation of unstructured variance-covariance structure
$\Rightarrow$ More realistic variances and covariances than ANOVA-type models

(Jennrich & Schluchter 1986; Piepho 1997)

Eberhart-Russell regression

$$\mathrm{var}(d_j) = \begin{pmatrix} \sigma^2_{d1} & & & \\ & \sigma^2_{d2} & & \\ & & \ddots & \\ & & & \sigma^2_{dI} \end{pmatrix}$$

σ^2_{di} = stability variance of *i*-th genotype

(Eberhart & Russell 1966; Piepho 1997)

Prediction

w_j unknown for new environment $\Rightarrow$ model does not really help

Still can only estimate mean and difference in TPE:

$$E(\eta_{ij}) = \alpha_i \quad \text{and} \quad E(\eta_{ij} - \eta_{hj}) = \alpha_i - \alpha_h$$

Can quantify uncertainty based on

$$\text{var}(\eta_j) = \lambda\lambda^T + \text{var}(d_j) \qquad (17)$$

Hitchhiker's Guide to the Galaxy

Ford Prefect & Arthur Dent, stranded on prehistoric Earth and trying to get out:

Arthur Dent and Ford Prefect know exactly where they don't want to be. They don't want to be stranded on prehistoric Earth with a load of unwanted telephone sanitizers and advertising executives who have been thrown off their home planet of Golgafrincham, a world which has subsequently been wiped out by a particularly virulent disease contracted from an unexpectedly dirty telephone. Unfortunately that is precisely where they are. But fortunately they have found a way of coping with their predicament: they are drunk.

. . . . .

ARTHUR: Have you got an answer?

FORD: *No*, but I've got a different name for the problem!

2. Finlay-Wilkinson regression with intercepts for environments

$$\eta_{ij} = E_j + \lambda_i w_j + d_{ij} \qquad (18)$$

E_j = random intercept for *j*-th environment

Constraint: $\mathrm{E}(w_j) = 1$

Example: Heading dates in perennial ryegrass (*Lolium perenne*)

- DUS (distinctness, uniformity and stability) trials for cultivar registration
- I = 113 genotypes
- J = 7 years
- Balanced data
- Heading dates (number of days after April 1)

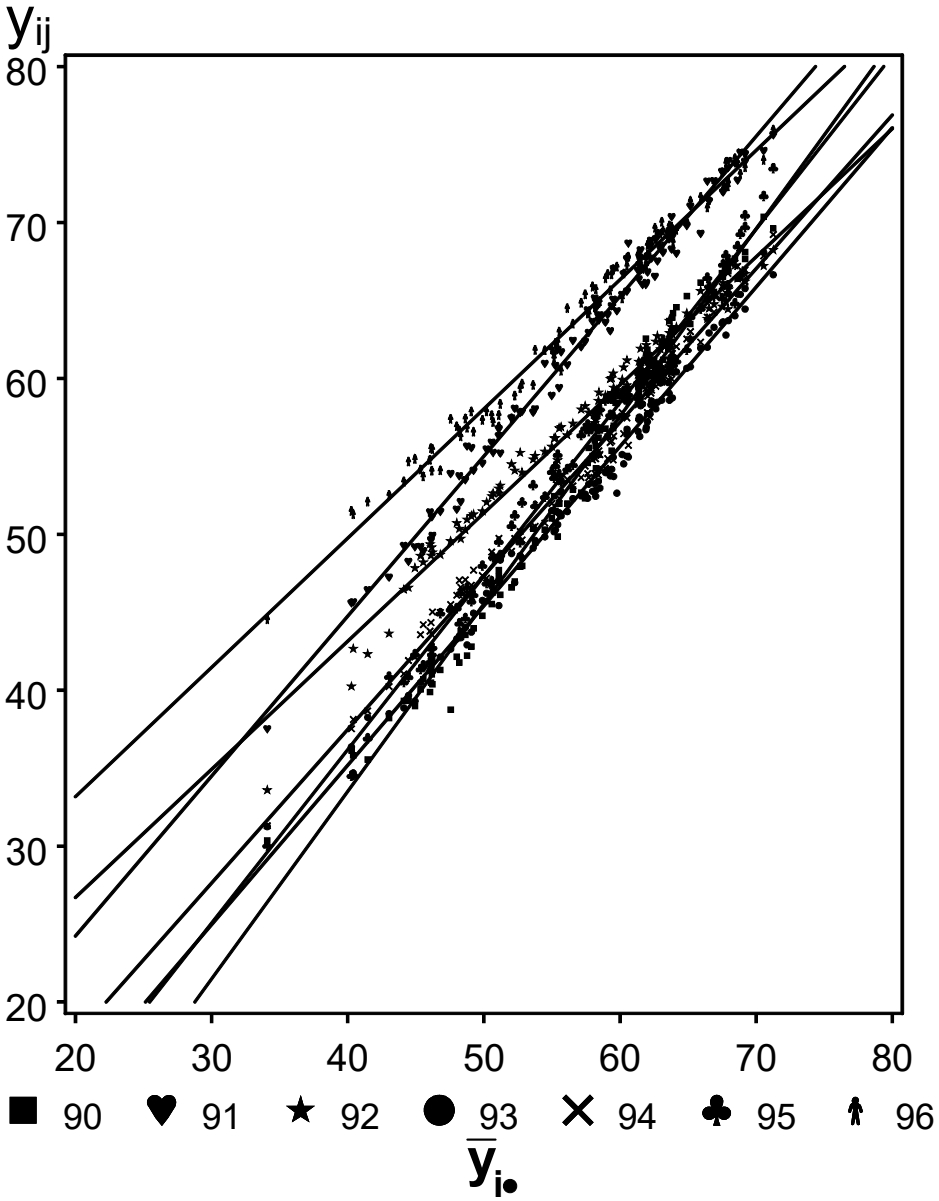


Fig.: Plot of observations y_{ij} vs. genotype means for years for heading dates in *Lolium*. (Piepho et al. 1998)

Effects (BLUPs):

Year	E_j	w_j
1990	***–14.50***	***1.201***
1991	3.78	1.025
1992	10.15	0.824
1993	–5.74	1.023
1994	–1.97	0.986
1995	–8.28	1.112
1996	***16.55***	***0.830***
(s.e.)	(4.168)	(0.0537)

Adding an overall intercept

$$\eta_{ij} = \mu + E_j + \lambda_i w_j + d_{ij} \qquad (19)$$

$$\mathrm{E}\left(E_j\right) = 0$$

$$\mathrm{E}(w_j) = 1$$

$$E\left(\eta_{ij}\right) = \mu + \lambda_i = \eta_i$$

The variance-covariance matrix for the *j*-th environment:

$$\mathrm{var}\left(\eta_j\right) = J\sigma_E^2 + \lambda\lambda^T\sigma_w^2 + \mathrm{var}\left(d_j\right) \qquad (20)$$

$\Rightarrow$ quadratic function of mean η_i !

$$\mathrm{var}\left(\eta_{ij}\right) = \sigma_E^2 + \left(\eta_i^2 - 2\eta_i\mu + \mu^2\right)\sigma_w^2 + \sigma_{di}^2$$

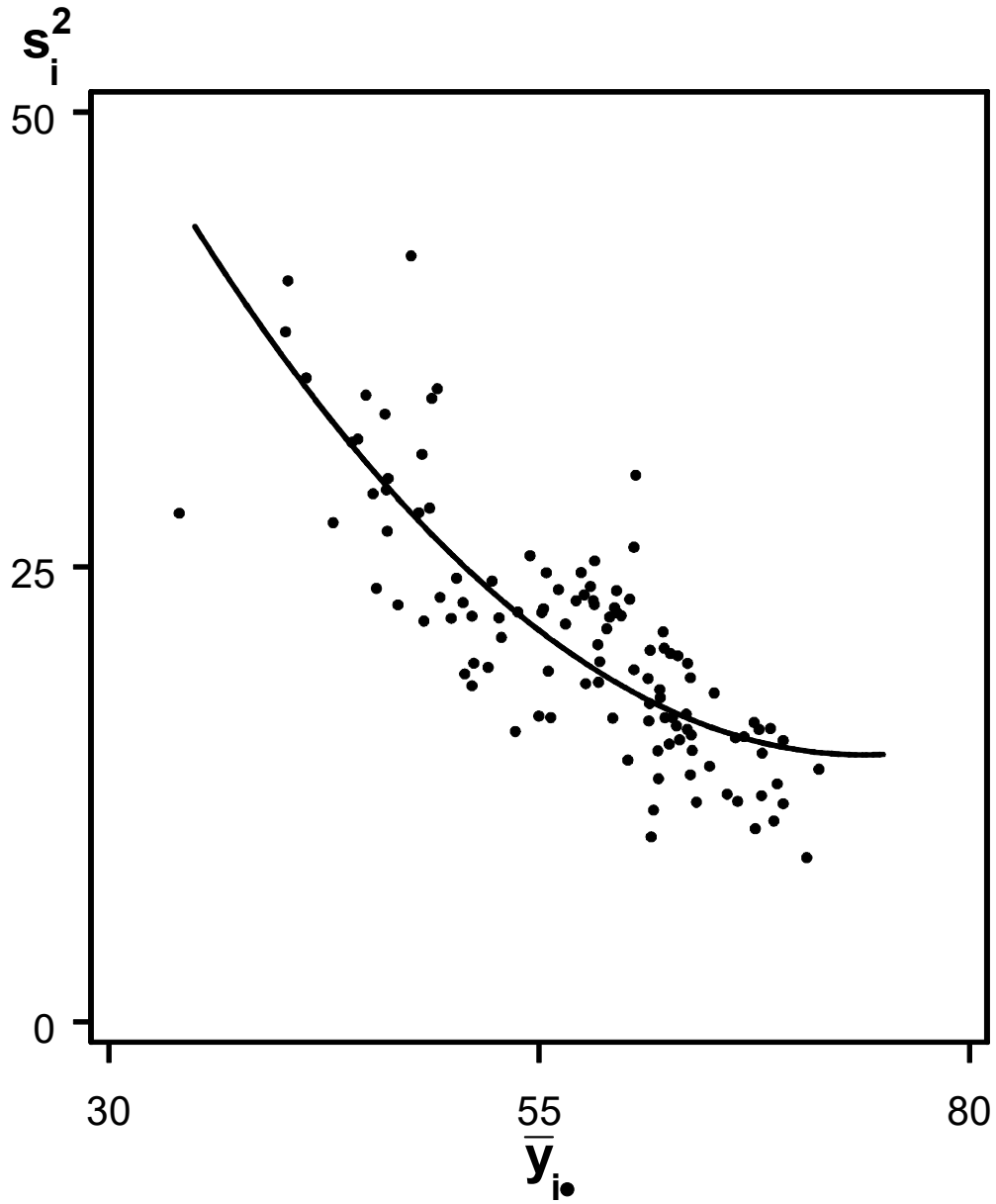


Figure: Plot of sample variance vs. sample mean across years for heading dates in *Lolium*. (Piepho et al. 1998)

Fitting this model

- Can not fit by REML because λ_i appears both in the mean and the variance
- Use full ML or approximate methods for nonlinear mixed models
- I do not think anybody has used this model after my publications

(Piepho et al. 1998; Piepho 1999)

3. Latent regression with intercepts for genotypes and environments

$$\eta_{ij} = \mu + \alpha_i + E_j + \lambda_i w_j + d_{ij} \qquad (21)$$

$$E(w_j) = 0 \quad (!) \quad \text{var}(w_j) = 1$$
$$E(\eta_{ij}) = \mu + \alpha_i$$

$\Rightarrow$ mixed model version of Additive Main effects Multiplicative Interaction (AMMI)

$\Rightarrow$ slope λ_i not involved in expectation

$\Rightarrow$ this is a linear mixed model

$\Rightarrow$ can use REML

(Piepho 1997)

4. Changes when genotypes are random and environments are fixed

$\Rightarrow$ Replace E_j with ε_j

$\Rightarrow$ Swap subscripts of letters used for effects in the multiplicative term

$$\eta_{ij} = \varepsilon_j + w_i\lambda_j + d_{ij} \qquad (22)$$

Constraint: $\mathrm{var}(w_i) = 1$

ε_j = fixed intercept for the *j*-th environment

λ_j = fixed slope for the *j*-th environment

(Piepho 1998a; Smith et al. 2001)

Important:

$$\neq \eta_{ij} = E_j + \lambda_i w_j + d_{ij} \qquad (18)$$

Genotype mean

$$\bar{\eta}_{i\bullet} = \bar{\varepsilon}_{\bullet} + w_i \bar{\lambda}_{\bullet} + \bar{d}_{i\bullet} \qquad (23)$$

$w_i \bar{\lambda}_{\bullet}$ = *generalized* main effect

$\Rightarrow$ measure of *overall performance* (OP)

(Smith & Cullis 2018; Tolhurst et al. 2022)

- The genotype mean in (23) is conditional on set of environments in MET $\Rightarrow$ only applies to the specific set of environments included in the trials
- For prediction of a new environment, might consider estimate of $\bar{\eta}_{i\bullet}$ in (23)
- But can't assess the uncertainty of this prediction because the loading λ_j for the new environment is unknown and it is a fixed effect

Adding a random genotypic main effect

$$\eta_{ij} = \mu + a_i + \varepsilon_j + w_i\lambda_j + d_{ij} \qquad (24)$$

a_i = random intercept of *i*-th genotype

Genotype mean:

$$\overline{\eta}_{i\bullet} = \mu + a_i + \overline{\varepsilon}_{\bullet} + w_i\overline{\lambda}_{\bullet} + \overline{d}_{i\bullet}$$

Genotype difference:

$$\overline{\eta}_{i\bullet} - \overline{\eta}_{h\bullet} = a_i - a_h + \left(w_i - w_h\right)\overline{\lambda}_{\bullet}$$

$\Rightarrow$ same inferential limitation as before

$\Rightarrow$ genotypic main effect a_i cannot be used for selection

5. Changes when both genotypes and environments are random

$$\eta_{ij} = \alpha_i + \lambda_i w_j + d_{ij} \qquad (16)$$

$$\lambda_i \sim N\left(\mu_\lambda, \sigma_\lambda^2\right) \quad , \quad w_j \sim N\left(\mu_w, \sigma_w^2\right)$$

$\Rightarrow$ η_{ij} is not normal!

$\Rightarrow$ Bayesian methods

$\Rightarrow$ iterative method of Nabugoomu et al. (1999)

$\Rightarrow$ A two-stage approach: (i) take genotypes as fixed to estimate α_i and λ_i

(ii) fit bivariate model $$\begin{pmatrix} \alpha_i \\ \lambda_i \end{pmatrix} \sim BVN\left[\begin{pmatrix} \mu_\alpha \\ \mu_\lambda \end{pmatrix}, \Sigma_{\alpha\lambda} = \begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha\lambda} \\ \sigma_{\alpha\lambda} & \sigma_\lambda^2 \end{pmatrix}\right] \qquad (25)$$

Including marker information

Fit variance-covariance structure

$$\text{var}\begin{pmatrix} \alpha_1 \\ \lambda_1 \\ \vdots \\ \alpha_I \\ \lambda_I \end{pmatrix} = K \otimes \Sigma_{\alpha\lambda}$$

6. Extensions to allow for more than one latent environmental factor

- The factor-analytic model can be extended >1 latent environmental factor
- Need to impose additional identifiability constraints on the loadings λ_1, λ_2, …
- For details see

  Jennrich & Schluchter (1986), Piepho (1998a), Smith et al. (2001), Meyer (2009)

7. Extensions factoring environments by locations and years

Nabugoomu et al. (1999):

$$\eta_{ijk} = \alpha_i + Y_k + (\alpha Y)_{ik} + \lambda_i v_{jk} + d_{ijk} \qquad (26)$$

Piepho & van Eeuwijk (2002)

$\Rightarrow$ fitted terms $\lambda_i w_j + \lambda_i' v_{jk}$

8. Including observable environmental covariates

Hardwick & Wood (1972), Guo et al. (2021), Piepho & Blancon (2023)

$$w_j = \theta_1 x_{1j} + \theta_2 x_{2j} + \ldots + \theta_p x_{pj} \qquad (27)$$

$\theta_m \; (m = 1, 2, \ldots, p)$ = regression coefficients

x_{mj} = value of *m*-th observable covariate at *j*-th location

$\Rightarrow$ can make specific predictions for unseen environments

Two types of covariates: (i) location-specific, constant over years

(ii) location-specific, varying over years

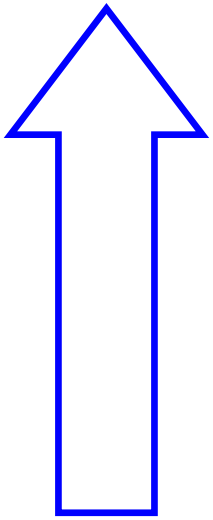
$\Rightarrow$ Need to model distribution of year-varying covariates to assess uncertainty

Two types of models using environmental covariates (EC)

(1) Stratification of environments into environmental types (ET)
(Bustos-Korts et al. 2022)

(2) Regression models with EC:

- Factorial regression (Denis 1980)
- Extended Finlay-Wilkinson regression
(Piepho & Blancon 2023)
- Environmental kinship (Jarcquin et al. 2014)



Increasing
Complexity

Summary

(1) Environments should be modelled as random $\Rightarrow$ inference for TPE

(2) For genotypes, fixed or random both are feasible options

(3) The most basic mixed model is a three-way ANOVA model

(4) Finlay-Wilkinson (FW) regression can be regarded as a regression on latent environmental variable w_j

(5) If w_j is modelled as random, FW regression induces a factor-analytic variance-covariance structure (FA)

(6) There are several variations of FA models

(7) When environments are modelled as fixed, FA models suffer from the same inferential limitations as simple ANOVA models

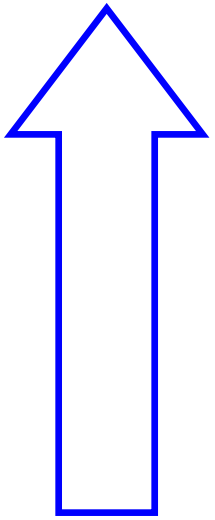
(8) The way forward is to model w_j using observable environmental covariates

Part II

Extending Finlay-Wilkinson regression using observable environmental covariates

Regression models with EC:

- Factorial regression (Denis 1980)
- Extended Finlay-Wilkinson regression (Piepho & Blancon 2023)
- Environmental kinship (Jarcquin et al. 2014)



Increasing Complexity

Three issues

(1) How are these models related?

(2) We need to factor environments by locations and years

(3) Uncertainty in covariates is the main problem in predictions

(Piepho et al. 2024)

Factorial regression

$$\eta_{ij} = \alpha_i + \gamma_{i1} x_{j1} + \gamma_{i2} x_{j2} + \ldots + \gamma_{ip} x_{jp}$$

where

η_{ij} = expected response for the *i*-th genotype (*i* = 1, …, *I*)
in the *j*-th environment (*j* = 1, …, *J*)

α_i = intercept for the *i*-th genotype

γ_{ik} = slope for the *k*-th environmental covariate (EC) for the *i*-th genotype

x_{jk} = value of the *k*-th covariate (*k* = 1, …, *p*) for the *j*-th environment

(Denis, 1988; Piepho & Blancon, 2023)

Full model for observed data

$$y_{ij} = \eta_{ij} + u_j + e_{ij}$$

where

y_{ij} = observed mean for the *i*-th genotype in the *j*-th environment

u_j = random main effect for *j*-th environment

e_{ij} = random residual

$\Rightarrow u$ and e will later be factored by years and locations

Now take genotypes as random

$$\alpha_i = \mu_\alpha + a_i$$

$$\gamma_{ik} = \mu_{\gamma k} + c_{ik}$$

$$\eta_{ij} = (\mu_\alpha + a_i) + (\mu_{\gamma 1} + c_{i1})x_{j1} + (\mu_{\gamma 2} + c_{i2})x_{j2} + \ldots + (\mu_{\gamma p} + c_{ip})x_{jp} \quad ,$$

where

$$\begin{pmatrix} a_i \\ c_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0_p \end{pmatrix}, \Sigma\right]$$

$$c_i = (c_{i1}, \ldots, c_{ip})^T$$

Rearranging terms:

$$\eta_{ij} = f\left(x_j\right) + g_i\left(x_j\right)$$

The fixed-effects part:

$$f\left(x_j\right) = \mu_\alpha + \mu_{\gamma 1} x_{j1} + \mu_{\gamma 2} x_{j2} + \ldots + \mu_{\gamma p} x_{jp}$$

$\Rightarrow$ mean regression across genotypes

The random part:

$$g_i\left(x_j\right) = a_i + c_{i1} x_{j1} + c_{i2} x_{j2} + \ldots + c_{ip} x_{jp}$$

$\Rightarrow$ random deviation of the *i*-th genotype from the mean regression $f(x_j)$

Including the random environmental main effect u_j

The full model for y_{ij} as a mixed main effect for environments:

$$\varepsilon_j = \mu_{\gamma 1} x_{j1} + \mu_{\gamma 2} x_{j2} + \ldots + \mu_{\gamma p} x_{jp} + u_j$$

$\Rightarrow$ the random effect u_j acts as a random deviation from the mean regression $f(x_j)$

Specifications for the variance-covariance structure for η_{ij}

Collect responses η_{ij} for the *i*-th genotype into a vector $\eta_i = (\eta_{i1}, \ldots, \eta_{iJ})^T$:

$$\eta_i = 1_J(\mu_\alpha + a_i) + X(\mu_\gamma + c_i)$$

where *J* = number of environments, 1_J = a *J*-vector of ones,

$$X = \{x_{jk}\} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & & x_{2p} \\ \vdots & & \ddots & \vdots \\ x_{J1} & x_{J2} & \cdots & x_{Jp} \end{pmatrix},$$

$$\mu_\gamma = (\mu_{\gamma 1}, \ldots, \mu_{\gamma p})^T$$

Distribution of η_i

$$\eta_i \sim N\left[1_J \mu_\alpha + X\mu_\gamma, \Omega\right]$$

where

$$\Omega = \left(1_J \; X\right)\Sigma\left(1_J \; X\right)^T$$

$\Rightarrow$ 3 different choices for Σ in

$$\begin{pmatrix} a_i \\ c_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0_p \end{pmatrix}, \Sigma\right]$$

(i) Unstructured model for Σ

$\Rightarrow$ random coefficients regression
$\Rightarrow$ translational invariance (Longford, 1993; Buntaran et al., 2021)
$\Rightarrow$ We may denote this approach in our context as *random FR*

(ii) The *kinship approach*

$$\Sigma = \begin{pmatrix} \sigma_\alpha^2 & 0 \\ 0 & I_p \sigma_\gamma^2 \end{pmatrix}$$

(Jarquin et al., 2014)

$\Rightarrow$ not translationally invariant
$\Rightarrow$ more parsimonious than *random FR*
$\Rightarrow$ regularization argument, implies ridge regression (Ruppert et al., 2003, p.66)

$$\Omega = 1 1_J^T \sigma_\alpha^2 + XX^T \sigma_\gamma^2$$

The variance-covariance structure for η_i under the kinship approach:

$$\Omega = 1 1_J^T \sigma_\alpha^2 + XX^T \sigma_\gamma^2$$

$K_E = XX^T$ = kinship matrix for environments

$\Rightarrow$ Standardizing the columns of X to zero mean and unit variance makes the kinship approach unique and ensures that each EC has equal influence on the regression, but does not resolve the lack of translational invariance issue

The model can be re-written in scalar form as

$$\eta_{ij} = \mu_\alpha + \mu_{\gamma 1} x_{j1} + \mu_{\gamma 2} x_{j2} + \ldots + \mu_{\gamma p} x_{jp} + a_i + w_{ij}$$

where

$$a_i \sim N\left(0, \sigma_\alpha^2\right)$$

is a genotype main effect and

$$w_i = \left(w_{i1}, w_{i2}, \ldots, w_{iJ}\right)^T \sim N\left(0, K_E \sigma_\gamma^2\right)$$

is the vector of interactions for the *i*-th genotype.

Note: The mean regression involving μ_γ is often omitted ($p > J$!)

(iii) *Reduced rank regression* (RRR)

$$\Sigma = \Lambda\Lambda^T$$

where Λ is a $J \times q$ matrix of factor loadings for q factors and J environments

$\Rightarrow$ not translationally invariant
$\Rightarrow$ but can be good approximation to unstructured model

(Buntaran et al., 2021; Tolhurst et al., 2022)

Can use the partition

$$\Lambda = \begin{pmatrix} \lambda_\alpha^T \\ \Lambda_\gamma \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \Lambda_\gamma \end{pmatrix} \begin{pmatrix} \lambda_\alpha^T \\ I_q \end{pmatrix}$$

where

$$\lambda_\alpha = \left(\lambda_{\alpha(1)}, \ldots, \lambda_{\alpha(q)}\right)^T$$

is a *q*-vector of loadings pertaining to the intercept and

$$\Lambda_\gamma = \left\{\lambda_{\gamma(kh)}\right\} \ \left(h = 1, \ldots, q\right)$$

is the $p \times q$ sub-matrix of Λ pertaining to the slopes

Key step

$$\begin{pmatrix} a_i \\ c_i \end{pmatrix} = \Lambda v_i \qquad \text{where } v_i \sim N\left(0_q, I_q\right).$$

After some re-arrangement we find

$$a_i = \lambda_\alpha^T v_i$$
$$c_i = \Lambda_\gamma v_i$$
$$Xc_i = X\Lambda_\gamma v_i = Zv_i$$
$$\eta_i = 1_J\left(\mu_\alpha + a_i\right) + X\mu_\gamma + Zv_i$$

where $Z = \left\{z_{jh}\right\} = X\Lambda_\gamma$ with

$$z_{jh} = \lambda_{\gamma(1h)}x_{j1} + \lambda_{\gamma(2h)}x_{j2} + \ldots + \lambda_{\gamma(ph)}x_{jp}$$

Synthetic covariate

$$z_{jh} = \lambda_{\gamma(1h)} x_{j1} + \lambda_{\gamma(2h)} x_{j2} + \ldots + \lambda_{\gamma(ph)} x_{jp}$$

is the value of the *h*-th synthetic environmental covariate (SC) for the *j*-th environment (Piepho & Blancon, 2023).

Also note that

$$\operatorname{var}\begin{pmatrix} a_i \\ v_i \end{pmatrix} = \begin{pmatrix} \lambda_\alpha^T \lambda_\alpha & \lambda_\alpha^T \\ \lambda_\alpha & I_q \end{pmatrix} = \widetilde{\Lambda}\widetilde{\Lambda}^T$$

with

$$\widetilde{\Lambda}^T = \begin{pmatrix} \lambda_\alpha & I_q \end{pmatrix}$$

and

$$\Omega = (1_J Z)\widetilde{\Lambda}\widetilde{\Lambda}^T (1_J Z)^T$$

$\Rightarrow$ reduced rank regression using the synthetic covariates Z

$\Rightarrow$ two-stage approach to fit RRR for observable covariates X

Extended Finlay-Wilkinson regression: Piepho and Blancon (2023) suggested to obtain the synthetic covariates with a_i taken as fixed and assuming

$$\text{var}(c_i) = \Lambda_\gamma \Lambda_\gamma^T$$

From the fitted matrix Λ_γ one can then compute the SC using

$$Z = X\Lambda_\gamma$$

Subsequently, the model

$$\eta_i = 1_J \alpha_i + Z\beta_i$$

can be fitted, where

α_i = fixed intercept and $\beta_i = \left(\beta_{i(1)},\ldots,\beta_{i(q)}\right)^T$ = fixed regression coefficients

Adding in kinship for genotypes

The model for the environmental kinship

$$\eta_{ij} = \mu_\alpha + \mu_{\gamma 1} x_{j1} + \mu_{\gamma 2} x_{j2} + ... + \mu_{\gamma p} x_{jp} + a_i + w_{ij}$$

where

$$a_i \sim N\left(0, \sigma_\alpha^2\right)$$

is a genotype main effect and

$$w_i = \left(w_{i1}, w_{i2}, ..., w_{iJ}\right)^T \sim N\left(0, K_E \sigma_\gamma^2\right)$$

Vector of all interaction effects:

$$w = \left(w_1^T, w_2^T, ..., w_I^T \right)^T$$

Assumption so far:

$$w \sim N\left(0_{IJ}, I_I \otimes K_E \sigma_\gamma^2\right)$$

Alternative assumption:

$$w \sim N\left(0_{IJ}, K_G \otimes K_E \sigma_\gamma^2\right)$$

where K_G = genomic kinship matrix

Similarly, for the genotype main effect

$$a = \left(a_1, a_2, ..., a_I\right)^T$$

we may assume

$$a \sim N\left(0_I, K_G \sigma_\alpha^2\right)$$

The random FR and RRR approaches can be similarly modified.

Using regression models for predictions into new environments

The values of EC may not be known for a new environment

$\Rightarrow$ need a distributional model for the EC:

$$x_{lm} = \mu_x + L_{x(l)} + Y_{x(m)} + (LY)_{x(lm)}$$

where

x_{lm}	= value of the EC in the *l*-th location in the *m*-th year
μ_x	= intercept
$L_{x(l)} \sim N(0, \sigma^2_{x(L)})$	= random main effect for the *l*-th location
$Y_{x(m)} \sim N(0, \sigma^2_{x(Y)})$	= random main effect for the *m*-th year
$(LY)_{x(lm)} \sim N(0, \sigma^2_{x(LY)})$	= random location-year interaction for the *l*-th location and *m*-th year

A single regression term

$\gamma_i x$

γ_i = slope for *i*-th genotype
x = value of EC

Need plug-in value = expected value ξ for x:

$\gamma_i \xi$

Uncertainty associated with the value of $x \Rightarrow$ variance υ_x

Here: initially assume that both γ_i and ξ are known

$\Rightarrow$ contribution to the prediction variance for the response is $\gamma_i^2 \upsilon_x$

Four cases

Case	Target of prediction
1	Long-term mean in the TPE[§]
2	A new year at the mean of the TPE[§]
3	Long-term mean at new location (farm)
4	A new year at a new location (farm)

§ TPE = target population of environments

Case 1: *Long-term mean in the TPE*

$\Rightarrow$ evaluate regression at unconditional expectation of the EC in the TPE: by

$$\xi = E(x) = \mu_x$$

The regression term is

$$\gamma_i \mu_x$$

No uncertainty associated with the value we use for the EC:

$$\upsilon_x = 0$$

Case 2: A *new year at the mean of the TPE*

Ideally, we would want to replace x by the mean in the TPE in the new year m_0:

$$\xi_{m_0} = E\left(x \mid m_0\right) = \mu_x + Y_{x(m_0)}$$

Problem: EC may only be observed next year

$\Rightarrow$ need to use long-term mean ξ of x in the TPE:

$$\xi = \mu_x$$

$\Rightarrow$ deviates from ξ_{m_0} by amount $Y_{x(m_0)}$

$\Rightarrow$ From historical data on EC, can estimate $\mathrm{var}\left(Y_{x(m_0)}\right) = \sigma^2_{x(Y)}$

$\Rightarrow \upsilon_x = \gamma_i^2 \sigma^2_{x(Y)}$

Case 3: L*ong-term mean at a new location*

Long-term mean of the EC at location l_0:

$$\xi_{l_0} = E\left(x \mid l_0\right) = \mu_x + L_{x(l_0)}$$

As we are estimating a long-term mean, we have

$$\upsilon_x = 0$$

Case 4: *A new location in a new year*

As in previous case, need to use long-term mean of EC at location l_0

$$\xi_{l_0} = E\left(x \mid l_0\right) = \mu_x + L_{x(l_0)}$$

$\Rightarrow$ value of the EC for future year unavailable when prediction needed

However, this time

$$\upsilon_x = \gamma_i^2\left(\sigma_{x(Y)}^2 + \sigma_{x(LY)}^2\right)$$

Leave-one-environment-out cross-validation (CV)

Common practice:

- Leave out the lm-th environment for validation
- Fit model to the remaining environments
- The observed EC values for the left-out environment are plugged into the fitted model to obtain a prediction

This implies:

$$\xi = x_{lm}$$

But this is not a realistic scenario!

The closest scenario is Case 4 $\Rightarrow$ use

$$\xi = \xi_{l_0}$$

Example: Rice data from Bangladesh (BRRI)

- Stability trials
- Long-term data (1970 to 2020)
- 10 Locations
- 41 rice varieties in the winter season, 45 rice varieties in the monsoon season
- RCBD with three replications at each trial
- Leave-one-environment-out cross-validation &
  Leave-one-year-out cross-validation
- 47 weather and soil EC extracted from public databases (agERA5 and FAO)

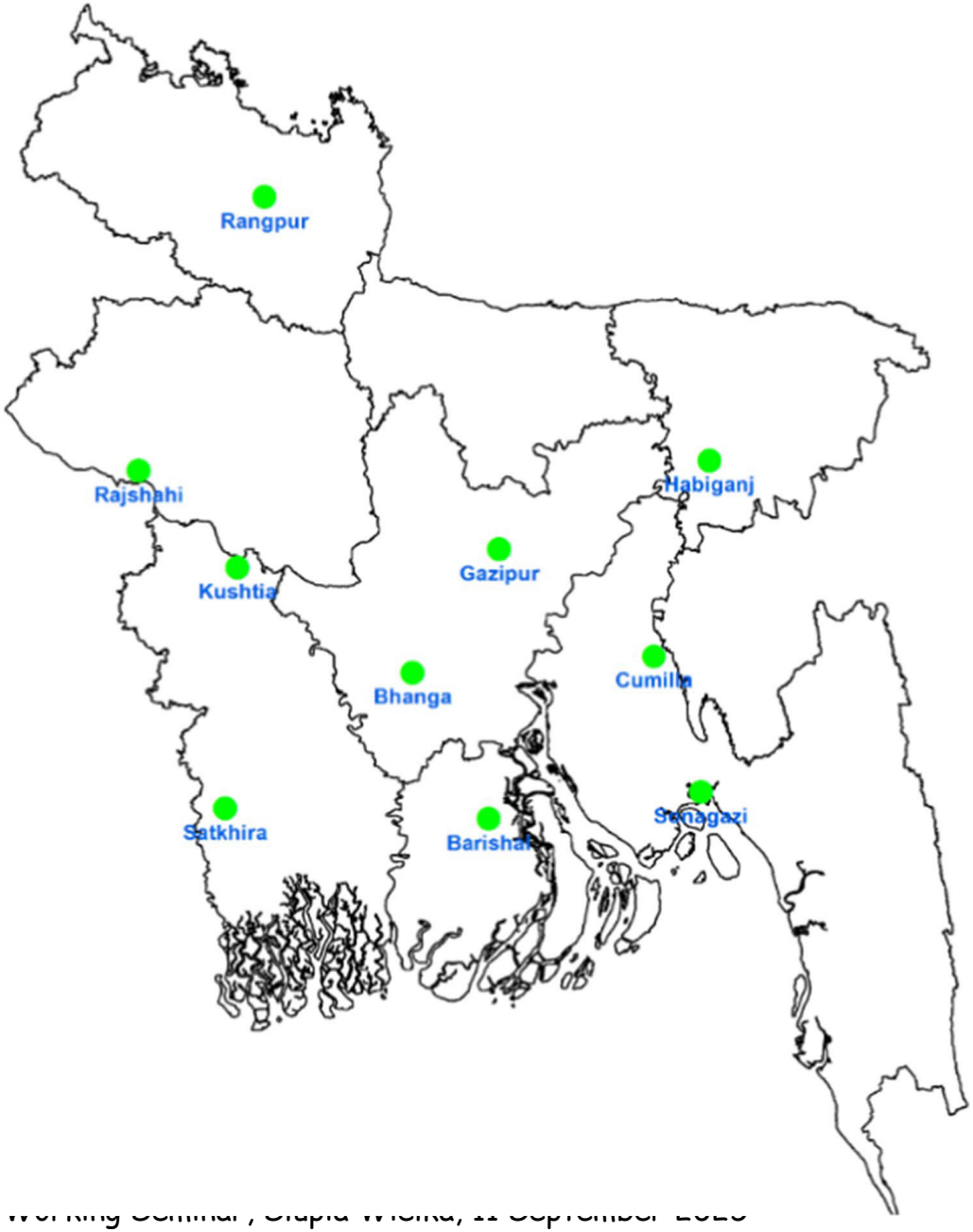

Rangpur
Rajshahi
Habiganj
Gazipur
Kushtia
Bhanga
Cumilla
Satkhira
Barishal
Sonagazi

Table 4. Model fit (summer rice) featuring the total number of parameters in the model (including those involved in synthetic covariates), full log-likelihood, and Akaike Information Criterion (AIC). Displayed values were rounded. Baseline – model without genotype-covariate interactions, Kinship – model with an environmental relationship matrix, RRR1 and RRR2 – reduced rank regression of rank one and two with observed covariates, RFR - random factorial regression with observed covariates, FW1-US and FW2-US – random factorial regression with one and two synthetic covariates respectively.

With the main EC effect				**Without the main EC effect**			
Model	**Parameters**	**LogLik**	**AIC**	**Model**	**Parameters**	**LogLik**	**AIC**
Baseline	48	-442	980	Baseline	8	-465	947
Kinship	49	-432	963	Kinship	9	-456	930
RRR1	88	-382	939	RRR1	48	-406	908
RRR2	128	-285	825	RRR2	88	-310	795
RFR	908	-144	2105	RFR	868	-168	2072
FW1-US	51	-369	840	FW1-US	50	-371	843
FW2-US	94	-306	800	FW2-US	92	-310	805

Table 6. Leave-one-environment-out (LOEO) and leave-one-year-and-location-out (LYLO) cross-validation means (summer rice). PCC – Pearson's correlation coefficient, MSEPD – mean squared error of predicted difference, MSPE – mean squared prediction error, Baseline – model without genotype-covariate interactions, Kinship – model with an environmental relationship matrix, RRR1 and RRR2 – reduced rank regression of rank one and two with observed covariates, RFR – random factorial regression with observed covariates, FW1-US and FW2-US – random factorial regression with one and two synthetic covariates.

Type	Model	Mean PCC		Mean MSEPD		Mean MSPE	
		LOEO	LYLO	LOEO	LYLO	LOEO	LYLO
With the main EC effect	Baseline	0.618	0.589	0.753	0.807	1.03	1.07
	Kinship	0.608	0.591	0.769	0.800	1.03	1.06
	RRR1	0.614	0.582	0.774	0.834	1.04	1.09
	RRR2	0.603	0.580	0.789	0.837	1.05	1.09
	RFR	0.601	0.577	0.791	0.831	1.05	1.07
	FW1-US	0.634	0.592	0.729	0.800	0.864	1.01
	FW2-US	0.638	0.593	0.708	0.800	0.857	0.988
Without the main EC effect	Baseline	0.618	0.589	0.753	0.807	0.868	1.02
	Kinship	0.609	0.585	0.768	0.811	0.875	1.02
	RRR1	0.614	0.583	0.774	0.833	0.882	1.04
	RRR2	0.603	0.580	0.789	0.836	0.895	1.05
	RFR	0.601	0.578	0.790	0.831	0.903	1.04
	FW1-US	0.634	0.592	0.729	0.800	0.864	1.03
	FW2-US	0.638	0.593	0.708	0.800	0.857	1.03

Table 7. Variance of the prediction (MVP) and mean squared prediction error (MSPE) from leave-one-year-and-location-out cross-validation (summer rice). Baseline – model without genotype-covariate interactions, Kinship – model with an environmental relationship matrix, RRR1 and RRR2 – reduced rank regression of rank one and two with observed covariates, RFR – random factorial regression with observed covariates, FW1-US and FW2-US – random factorial regression with one and two synthetic covariates.

Type	Model	MSPE		MVP	
		Mean	Median	Mean	Median
With the main EC effect	Baseline	1.07	0.776	0.994	0.948
	RRR1	1.09	0.792	0.979	0.906
	RRR2	1.10	0.824	1.05	0.988
	RFR	1.08	0.792	1.08	1.01
	FW1-US	1.01	0.679	0.923	0.925
	FW2-US	0.987	0.668	0.902	0.905
Without the main EC effect	Baseline	1.02	0.644	0.941	0.945
	RRR1	1.04	0.656	0.975	0.976
	RRR2	1.05	0.687	1.04	1.05
	RFR	1.04	0.688	1.09	1.08
	FW1-US	1.03	0.681	0.948	0.953
	FW2-US	1.03	0.679	0.948	0.952

Table 8. Variance of the predicted difference (VPD) and mean squared error of predicted difference (MSEPD) from the leave-one-year-and-location-out cross-validation (summer rice). Baseline – model without genotype-covariate interactions, Kinship – model with an environmental relationship matrix, RRR1 and RRR2 – reduced rank regression of rank one and two with observed covariates, RFR – random factorial regression with observed covariates, FW1-US and FW2-US – random factorial regression with one and two synthetic covariates. VPD for the baseline model was taken from the standard ASReml-R output.

Type	Model	MSEPD		VPD	
		Mean	Median	Mean	Median
With the main EC effect	Baseline	0.807	0.700	0.652	0.651
	RRR1	0.835	0.707	0.710	0.705
	RRR2	0.837	0.721	0.858	0.837
	RFR	0.837	0.724	0.975	0.974
	FW1-US	0.800	0.705	0.640	0.630
	FW2-US	0.800	0.705	0.626	0.617
Without the main EC effect	Baseline	0.807	0.699	0.652	0.650
	RRR1	0.835	0.706	0.713	0.707
	RRR2	0.836	0.717	0.853	0.834
	RFR	0.837	0.717	0.974	0.974
	FW1-US	0.800	0.705	0.639	0.629
	FW2-US	0.800	0.705	0.625	0.616

Estimating the prediction variance

The overall prediction variance associated with $\gamma_i x$ has two components:

(i) Both γ_i and ξ in the product $\gamma_i \xi$ need to be estimated

(ii) $\upsilon_x = \gamma_i^2 \sigma_x^2$

(i) Variance due to the estimation of $\gamma_i \xi$

$$\mathrm{var}\left(\hat{\gamma}_i \hat{\xi}\right) = \gamma_i^2 \,\mathrm{var}\left(\hat{\xi}\right) + \xi^2 \,\mathrm{var}\left(\hat{\gamma}_i\right) + \mathrm{var}\left(\hat{\gamma}_i\right)\mathrm{var}\left(\hat{\xi}\right) \qquad \text{(Goodman, 1960)}$$

The naïve plug-in estimator of γ_i^2 is biased

$$\mathrm{var}(\hat{\gamma}_i) = E(\hat{\gamma}_i^2) - [E(\hat{\gamma}_i)]^2 = E(\hat{\gamma}_i^2) - \gamma_i^2$$

$\Rightarrow$

$$E(\hat{\gamma}_i^2) = \gamma_i^2 + \mathrm{var}(\hat{\gamma}_i)$$

$\Rightarrow$

Estimate γ_i^2 by $\widetilde{\gamma}_i^2 = \hat{\gamma}_i^2 - \mathrm{var}(\hat{\gamma}_i)$ and ξ^2 by $\widetilde{\xi}^2 = \hat{\xi}^2 - \mathrm{var}(\hat{\xi})$

$\Rightarrow$

$$est.\mathrm{var}(\hat{\gamma}_i\hat{\xi}) = \hat{\gamma}_i^2\,\mathrm{var}(\hat{\xi}) + \hat{\xi}^2\,\mathrm{var}(\hat{\gamma}_i) - \mathrm{var}(\hat{\gamma}_i)\mathrm{var}(\hat{\xi})$$

(ii) Estimation of $\upsilon_x = \gamma_i^2 \sigma_x^2$

The naïve plug-in estimator $\hat{\upsilon}_x = \hat{\gamma}_i^2 \hat{\sigma}_x^2$ is biased

Bias may be reduced by using the estimator

$$\tilde{\upsilon}_x = \left[\hat{\gamma}_i^2 - \mathrm{var}\left(\hat{\gamma}_i\right)\right]\hat{\sigma}_x^2$$

The **overall prediction variance** is obtained by adding the variances in **(i)** and **(ii)**

Conjecture: the main problem is υ_x

Extension to multiple EC and inclusion of intercept

This is straightforward but will be omitted here

Jointly consider intercept and all slopes:

$$\gamma_i \rightarrow \gamma_i'$$
$$\xi \rightarrow \xi'$$
$$x \rightarrow x'$$

Overall prediction variance

$$\upsilon = \mathrm{var}\left(\hat{\gamma}_i'^T \hat{\xi}'\right) + \upsilon_{x'} + \upsilon_R$$

υ_R = variance of deviations from regression

Table 2: Explicit expressions for ξ', $\Sigma_{x'}$, $\upsilon_{x'}$ and υ_R for the four scenarios (Case 1 to 4)

Case	ξ'	$\Sigma_{x'}$	$\upsilon_{x'}$	υ_R
1	$\mu_{x'}$	0	0	0
2	$\mu_{x'}$	$\Sigma_{x'(Y)}$	$\gamma_i'^T \Sigma_{x'(Y)} \gamma_i'$	$\sigma_Y^2 + \sigma_{\alpha Y}^2$
3	$\mu_{x'} + L_{x'(l_0)}$	0	0	$\sigma_L^2 + \sigma_{\alpha L}^2$
4	$\mu_{x'} + L_{x'(l_0)}$	$\Sigma_{x'(Y)} + \Sigma_{x'(LY)}$	$\gamma_i'^T \left(\Sigma_{x'(Y)} + \Sigma_{x'(LY)}\right) \gamma_i'$	$\sigma_L^2 + \sigma_{\alpha L}^2 + \sigma_Y^2 + \sigma_{\alpha Y}^2 + \sigma_{LY}^2 + \sigma_{\alpha LY}^2$

Contribution of the deviations from regression

Residual terms u_j and e_{ij} partitioned by year and location:

$$u_{lm} = L_l + Y_m + (LY)_{lm} \qquad e_{ilm} = (\alpha L)_{il} + (\alpha Y)_{im} + (\alpha LY)_{ilm}$$

where *L*, *Y* and α denote the factors location, year and genotype

$$L_l \sim N(0, \sigma_L^2) \qquad (\alpha L)_{il} \sim N(0, \sigma_{\alpha L}^2)$$
$$Y_m \sim N(0, \sigma_Y^2) \qquad (\alpha Y)_{im} \sim N(0, \sigma_{\alpha Y}^2)$$
$$(LY)_{lm} \sim N(0, \sigma_{LY}^2) \qquad (\alpha LY)_{ilm} \sim N(0, \sigma_{\alpha LY}^2)$$

A subset of these variances will contribute to the overall uncertainty
The variance of this contribution will be denoted as υ_R

Residual prediction variance υ_R in the four cases

Table 2: Explicit expressions υ_R for the four scenarios (Case 1 to 4)

Case	Target of prediction	υ_R
1	Long-term mean in TPE	0
2	A new year in TPE	$\sigma_Y^2 + \sigma_{\alpha Y}^2$
3	Long-term mean at farm	$\sigma_L^2 + \sigma_{\alpha L}^2$
4	A new year at farm	$\sigma_L^2 + \sigma_{\alpha L}^2 + \sigma_Y^2 + \sigma_{\alpha Y}^2 + \sigma_{LY}^2 + \sigma_{\alpha LY}^2$

Pairwise differences

To compute υ_x for the pairwise difference of two genotypes i and i', we replace γ_i with

$$\delta_{ii'} = \gamma_i - \gamma_{i'}$$

In the variance υ_R, we drop σ^2_L, σ^2_Y and σ^2_{LY}, because the corresponding random effects drop out in the pairwise difference.

Hitchhiker's Guide to the Galaxy

. . . . .

ARTHUR: **Have you got an answer?**

FORD: ***No*, but I've got a different name for the problem!**

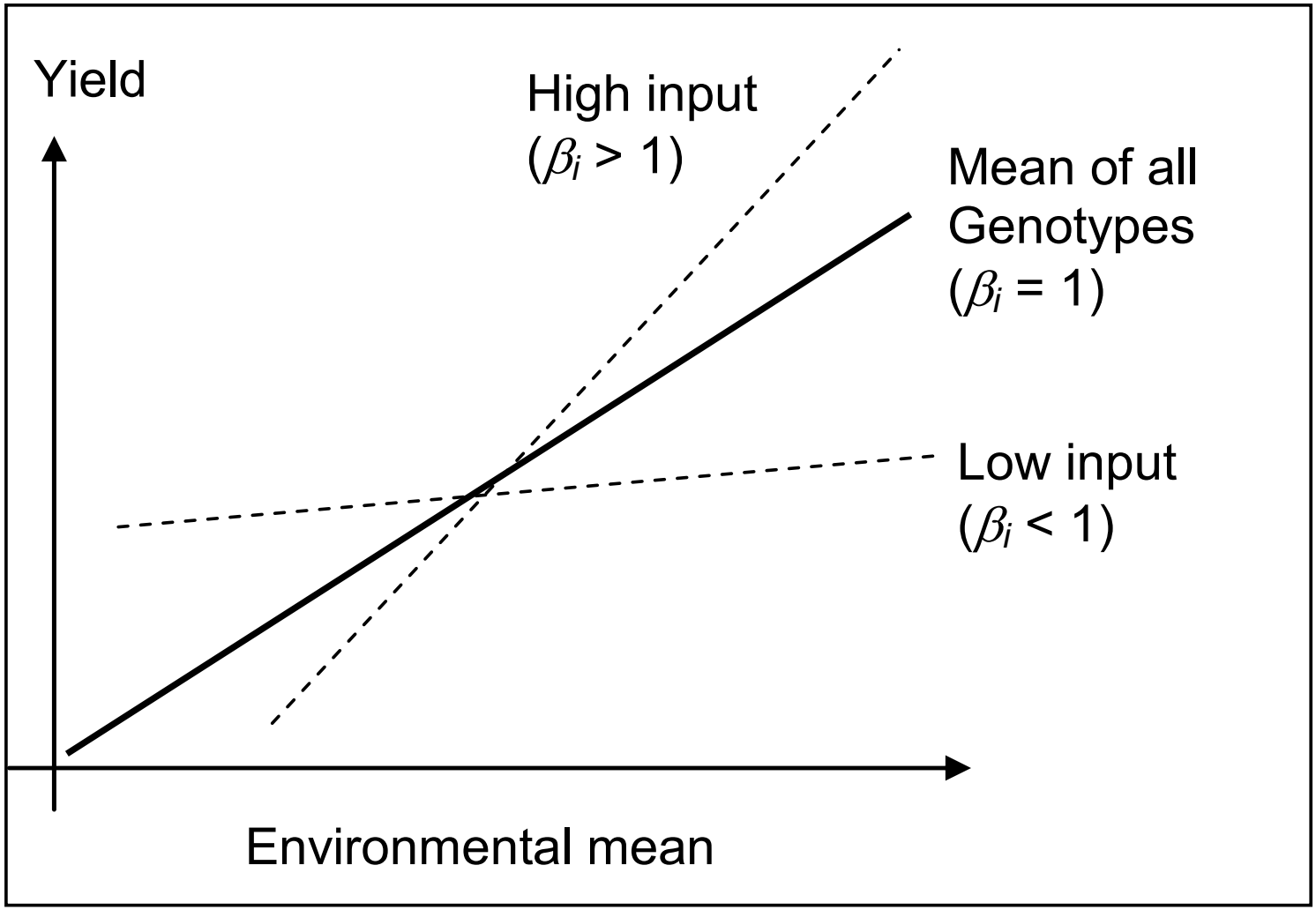
References

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

Piepho, H.P., Williams, E.R. (2024): Factor-analytic variance-covariance structures for prediction into a target population of environments. *Biometrical Journal* **66**, e202400008. https://doi.org/10.1002/bimj.202400008

Hrachov, M., Piepho, H.P., Rahman, N.F., Malik, W. (2025): Regression approaches for modelling genotype-environment interaction and making predictions into a target population of environments. https://doi.org/10.48550/arXiv.2507.18125

Finlay-Wilkinson regression



Finlay-Wilkinson regression

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

η_{ij} = expected performance of the *i*-th genotype in the *j*-th environment

α_i and β_i = intercept and slope for the *i*-th genotype

w_j = latent effect of the *j*-th environment

Estimation

Balanced data:

(1) $\hat{w}_j = \bar{y}_{\bullet j}$ (Finlay & Wilkinson 1963)

(2) Singular value decomposition (SVD) of the matrix $\{y_{ij} - \bar{y}_{i\bullet}\}$
(Williams, 1952; Hardwick & Wood 1972; Yan & Kang 2003)

Unbalanced data:

(3) Criss-cross regression (CCR) (Digby 1979; Gabriel & Zamir 1979)

$\eta_{ij} = \alpha_i + \beta_i w_j$ (criss: fix w_j) $\Leftrightarrow$ $\eta_{ij} = \alpha_i + \beta_i w_j$ (cross: fix α_i & β_i)

(4) Non-linear least squares (Ng & Grunwald 1997; Ng & Williams 2001)

(5) Expectation-maximization (EM) algorithm (Gauch & Zobel 1990)

Extending F-W regression with covariates

Regressing w_j on p observable covariates x_{jk} $(k = 1, \ldots, p)$, i.e.,

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

where

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

(Li et al. 2018; Guo et al. 2021)

Reparameterized (and equivalent) model

$$\eta_{ij} = \alpha_i + \beta_i z_j$$

where

$$z_j = \theta_1 x_{j1} + \theta_2 x_{j2} + ... + \theta_p x_{jp}$$

is a synthetic covariate

Estimation with balanced data

Assume $y_{ij} = \eta_{ij} + e_{ij}$ with $e_{ij} \sim \mathrm{N}(0, \sigma^2)$

(i) Consider the environmental averages:

$$\overline{\eta}_{\bullet j} = \overline{\alpha}_{\bullet} + \overline{\beta}_{\bullet}\left(\theta_0 + \theta_1 x_{j1} + \theta_2 x_{j2} + ... + \theta_p x_{jp}\right)$$

$\Rightarrow$ multiple regression of observed environmental means $\overline{y}_{\bullet j}$ on the covariates provides estimates of slopes $\widetilde{\theta}_k = \overline{\beta}_{\bullet}\theta_k$ for covariates $x_{jk}\left(k = 1,..., p\right)$ and the intercept $\overline{\alpha}_{\bullet} + \overline{\beta}_{\bullet}\theta_0$. Without loss of generality, we may then use

$$z_j = \widetilde{\theta}_1 x_{j1} + \widetilde{\theta}_2 x_{j2} + ... + \widetilde{\theta}_p x_{jp}$$

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

Estimation with balanced data (cont'd)

(ii) Criss-cross regression (CCR) (Digby 1979; Gabriel & Zamir 1979)

$\eta_{ij} = \alpha_i + \beta_i w_j$ (criss: fix w_j) $\Leftrightarrow$ $\eta_{ij} = \alpha_i + \beta_i w_j$ (cross: fix α_i & β_i)

with $w_j = \theta_0 + \theta_1 x_{j1} + \theta_2 x_{j2} + ... + \theta_p x_{jp}$

This method provides a least squares fit.

(iii) Redundancy analysis (RA): Fit factorial regression (FR)

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

and subsequently subject the matrix of fitted terms $\left\{\hat{\gamma}_{i1} x_{j1} + \hat{\gamma}_{i2} x_{j2} + ... + \hat{\gamma}_{ip} x_{jp}\right\}$ to an SVD. The first term of this decomposition provides the least squares fit for $\beta_i z_j$. (Hardwick & Wood 1972; Wood 1976; Davies & Tso 1982; van Eeuwijk et al. 1992)

Estimation with balanced data (cont'd)

(iv) Estimate

$$z_j = \theta_1 x_{j1} + \theta_2 x_{j2} + ... + \theta_p x_{jp}$$

using the first factor extracted by partial least squares (PLS; Aastveit & Martens 1986), regarding the genotypic responses in an environment as a single *multivariate* response.

(v) Fit

$$\eta_{ij} = \alpha_i + \beta_i z_j$$

by nonlinear least squares (Ng & Grunwald 1997; Ng & Williams 2001).

Estimation with unbalanced data

(ii) Criss-cross regression (CCR) (Digby 1979; Gabriel & Zamir 1979)

$\eta_{ij} = \alpha_i + \beta_i w_j$ (criss: fix w_j) $\Leftrightarrow$ $\eta_{ij} = \alpha_i + \beta_i w_j$ (cross: fix α_i & β_i)

with $w_j = \theta_0 + \theta_1 x_{j1} + \theta_2 x_{j2} + ... + \theta_p x_{jp}$

(iii) Redundancy analysis (RA): Fit factorial regression (FR)

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

and subsequently subject the matrix of fitted terms $\left\{\hat{\gamma}_{i1} x_{j1} + \hat{\gamma}_{i2} x_{j2} + ... + \hat{\gamma}_{ip} x_{jp}\right\}$ to an SVD, including those for empty cells. The first term of this decomposition provides the least squares fit for $\beta_i z_j$.

Estimation with unbalanced data (cont'd)

(iv) EM-method to obtain PLS-estimate of

$$z_j = \theta_1 x_{j1} + \theta_2 x_{j2} + ... + \theta_p x_{jp}$$

Nelson et al. (1996)

More than one synthetic environmental covariate

The model may be extended as

$$\eta_{ij} = \alpha_i + \beta_{i1} z_{j1} + ... + \beta_{iq} z_{jq}$$

where

$z_{j1},....,z_{jq}$ = q synthetic environmental covariates

$\beta_{i1},....,\beta_{iq}$ = q genotype-specific slopes

and

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

$\Rightarrow$ Estimation by the same methods as for $q = 1$

$\Rightarrow$ Estimability constraints need to be imposed

Random-effects extensions of the model

Environments are a random factor throughout (see above)

Genotypes are fixed or random, depending on objectives:

- Using markers/kinship in GBLUP require random effects
- many genotypes make random effects convenient
- subdivided TPE require random effects to borrow strength across zones

All parameters in η_{ij} are fixed

$\Rightarrow$ genotypes fixed

$$y_{ij} = \eta_{ij} + u_j + e_{ij}$$

where

u_j = random environment main effect with variance σ_u^2

e_{ij} = random with stability variance $\sigma_{e(i)}^2$ for the *i*-th genotype

(Eberhart & Russell 1966; Shukla 1972)

Estimation with random effects

(ii) CCR is easily extended in a mixed model framework. In the criss-step, estimating α_i and β_{ih} $(i = 1,...,n; h = 1,...,q)$, we fix the variance parameters at their current estimates because this usually has more parameters to be estimated as fixed effects than the cross-step. These variance parameters are re-estimated in the cross-step, estimating θ_{hk} $(k = 1,..., p; h = 1,...,q)$, using REML.

(Nabugoomu et al. 1999; Macholdt et al. 2022)

(iii) Redundancy analysis (RA): Fit FR model under our assumed random-effects specification using REML. From the fitted model we 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 ones with missing data. This matrix is subjected to an SVD.

Estimation with random effects (cont'd)

(v) Fit model directly using full maximum likelihood (ML) [e.g., using NLMIXED in SAS] $\Leftrightarrow$ nonlinear least squares in fixed effects case

Note that we can not use REML because the model is non-linear in the parameters and hence the fixed effects cannot be removed by linear contrasts as 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 our nonlinear mixed model. 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.

(Piepho 1999)

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

Not as straightforward as it may seem!

$\Rightarrow$ Assume one synthetic covariate

$\Rightarrow$ Take both the intercept α_i and the slope β_i to be random

$\alpha_i = \mu_\alpha + a_i$ with $E(a_i) = 0$ and $\text{var}(a_i) = \sigma_a^2 = \text{var}(\alpha_i)$.

$\beta_i = \mu_\beta + b_i$ with $E(b_i) = 0$ and $\text{var}(b_i) = \sigma_b^2 = \text{var}(\beta_i)$.

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

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

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

The main challenge: z_j (involves parameters!) now appears both in the fixed-effects term $\mu_\beta z_j$ and the random effects term $b_i z_j$.

We cannot assume $\mu_\beta = 0$ without loss of generality, because there is no fixed environmental main effect in that would absorb this term.

Also, $\mu_\beta z_j$ is a multiplicative regression term that may be worth fitting explicitly.

Furthermore, to ensure invariance to shift-scale transformations of the observable covariates, we need to allow for a covariance between intercept and slope, i.e., $\mathrm{cov}(a_i, b_i) = \sigma_{ab}$. (Piepho 1999)

Estimation with random genotypic effects

(1) A stage-wise approach by which we first model all parameters as fixed, using either of the methods (ii) to (iv) to estimate z_j. This is then held fixed, using REML to fit μ_α, a_i, σ_a^2, μ_β, b_i, and σ_b^2.

(2) CCR (Nabugoomu et al. 1999).

(3) A new factor-analytic (FA) model (next slides) (Piepho & Blancon 2023)

A new factor-analytic (FA) approach

Model a_i as fixed, while slopes b_i are random (Piepho & Ogutu 2002).

Add a fixed environmental main effect ε_j that absorbs $\mu_\beta z_j$.

Hence, the only random effect in η_{ij} is the slope b_i, and the model can be written as

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

with $E(b_i) = 0$ and $\text{var}(b_i) = \sigma_b^2 = \text{var}(\beta_i)$.

A new factor-analytic (FA) approach

Can easily extend to several synthetic environmental covariates:

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

Re-write the random terms

$$b_{i1} z_{j1} + \ldots + b_{iq} z_{jq} = \boldsymbol{z}_j^{\boldsymbol{T}} \boldsymbol{b}_{\boldsymbol{i}}$$

where $\boldsymbol{z}_{\boldsymbol{j}} = \left(z_{j1}, z_{j2}, \ldots, z_{jq}\right)^T$ and $\boldsymbol{b}_{\boldsymbol{i}} = \left(b_{i1}, b_{i2}, \ldots, b_{iq}\right)^T$.

$\Rightarrow$ This model 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 $\boldsymbol{x_j} = \left(x_{j1}, x_{j2}, ..., x_{jp}\right)^T$, so that $z_{jh} = \boldsymbol{x_j^T}\boldsymbol{\theta_h}$ with $\boldsymbol{\theta_h} = \left(\theta_{1h}, \theta_{2h}, ..., \theta_{ph}\right)^T$.

Consider a random coefficient regression of the form

$$\eta_{ij} = \alpha_i + \varepsilon_j + \boldsymbol{x_j^T}\boldsymbol{c_i}$$

where $\boldsymbol{c_i} = \left(c_{i1}, c_{i2}, ..., c_{ip}\right)^T$ with $E\left(\boldsymbol{c_i}\right) = \boldsymbol{0}$ and $\mathrm{var}\left(\boldsymbol{c_i}\right) = \boldsymbol{\Sigma_c}$ (Longford 1995).

Next, approximate $\boldsymbol{\Sigma_c}$ by

$$\boldsymbol{\Sigma_c} = \boldsymbol{\Lambda\Lambda^T}$$ [FA0(*q*) in SAS]

where $\boldsymbol{\Lambda} = \left\{\lambda_{kh}\right\}$ is a $p \times q$ matrix of factor loadings.

(Buntaran et al. 2021; Tolhurst et al. 2022)

Now the key step

$$\boldsymbol{c}_i = \Lambda \boldsymbol{v}_i$$

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

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

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

Practical upshot:

Simply fit random regression on $\boldsymbol{x}_j$ with a FA0(q) structure $\Sigma_c = \Lambda\Lambda^T$ for $\boldsymbol{c}_i$, extract the estimate of Λ and use this to compute the q synthetic covariates $\boldsymbol{z}_j$.

Example for random effects modelling

- 8 lettuce genotypes
- Randomized complete block design
- Nitrate concentrations (g/l) measured at 18 time points = environments
- 8 environmental covariates assessed at each time point

(van Eeuwijk 1992)

Example for random effects modelling

Table 3. Mean nitrate concentrations (g/l) over the eight replicates of a randomized blocks design for the genotypes from Table 1 in the environments of Table 2

Environment	Genotype							
	Pa	DM	Pi	GT	RW	Wi	Tr	Ls
1	3.113	2.835	2.629	1.988	2.199	2.414	1.248	2.380
2	3.379	3.222	2.848	2.823	3.002	2.950	2.176	3.196
3	3.067	2.326	2.511	2.120	2.692	2.598	1.032	2.355
4	3.202	2.663	2.230	1.638	2.187	2.171	1.062	1.599
5	3.921	3.365	3.028	2.653	2.935	2.931	2.007	2.942
6	4.153	3.970	3.444	2.813	2.865	3.232	2.341	3.289
7	4.851	4.512	4.010	3.504	3.135	3.624	3.080	3.612
8	4.547	4.203	3.429	2.944	2.616	3.052	2.817	3.070
9	3.721	3.505	3.337	2.425	2.177	2.525	1.917	2.830
10	3.581	3.298	3.287	2.389	2.159	2.681	1.744	2.726
11	3.312	3.130	2.959	2.280	1.797	2.152	1.365	2.178
12	3.439	3.329	3.254	2.561	2.843	3.035	1.927	3.058
13	3.195	3.047	2.948	2.696	2.610	2.902	1.914	3.138
14	2.890	2.297	2.295	2.237	1.930	2.414	1.462	2.274
15	2.700	2.430	2.172	2.004	2.194	2.392	1.374	2.144
16	3.143	2.710	2.429	2.260	2.406	2.438	1.536	2.464
17	2.746	2.470	2.226	2.126	2.332	2.185	1.287	2.621
18	3.273	2.384	2.555	2.167	2.545	2.386	1.616	2.813

(van Eeuwijk 1992)

Example for random effects modelling

Table 5. Values of the environmental variables of Table 4 in the environments of Table 2

Environment	Environmental variable							
	1	2	3	4	5	6	7	8
1	2.1	1,136	993	911	881	14.75	11.23	13.15
2	2.1	1,345	1,277	1,250	1,815	11.78	13.28	14.93
3	2.1	1,700	2,191	2,586	2,556	16.14	16.48	16.37
4	2.1	1,076	1,090	1,323	1,065	14.81	13.61	12.74
5	2.2	960	779	539	457	13.61	12.46	10.89
6	2.0	316	482	421	556	12.81	11.23	9.04
7	2.0	145	117	102	42	11.91	10.29	8.05
8	1.9	109	93	127	42	10.96	8.71	7.73
9	2.2	555	504	415	383	9.64	8.50	9.90
10	2.0	641	663	596	780	8.45	7.98	11.78
11	2.0	676	666	541	546	7.70	9.04	12.60
12	1.6	1,951	2,427	2,413	2,286	9.22	12.05	14.39
13	1.5	1,651	1,789	1,276	1,518	11.78	13.01	15.21
14	1.5	2,281	2,359	2,376	2,514	13.15	14.45	15.62
15	1.5	1,244	1,456	1,604	1,398	14.07	15.37	16.23
16	2.3	1,398	1,852	2,719	2,975	14.51	15.96	16.45
17	1.5	2,041	1,515	1,350	988	14.93	16.23	16.48
18	1.5	1,326	1,416	1,779	1,580	15.26	16.39	16.41

(van Eeuwijk 1992)

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

Model	Random effects§	Structure for random effects	Residual error structure	Residual log-likelihood	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

§ The covariates were standardized to zero mean and unit variance.

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.	F-value	p-value
Genotype	7	15.5	166.45	<.0001
z	1	16.7	8.10	0.0113
Genotype × z	7	42.3	12.58	<.0001

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.

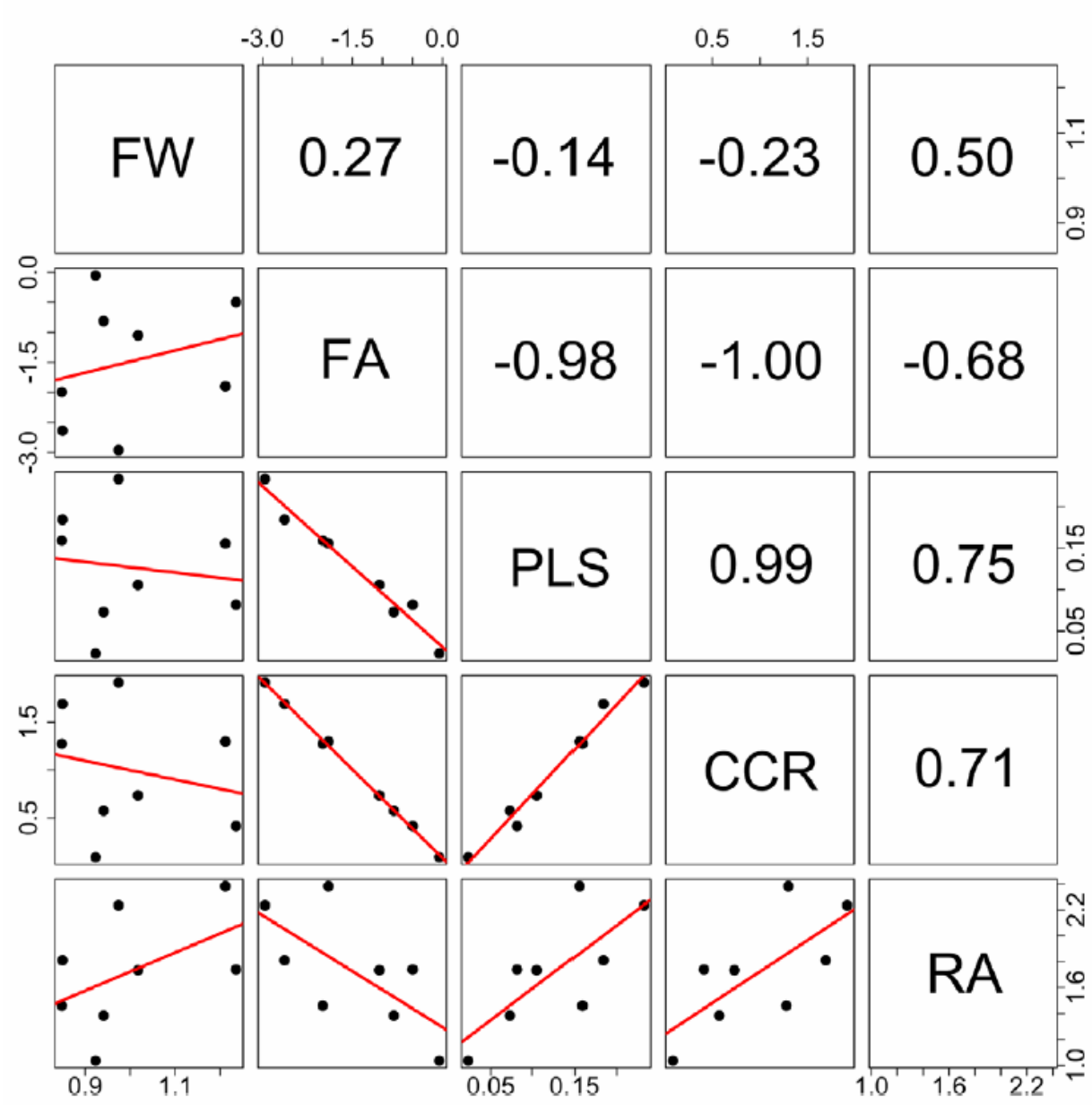
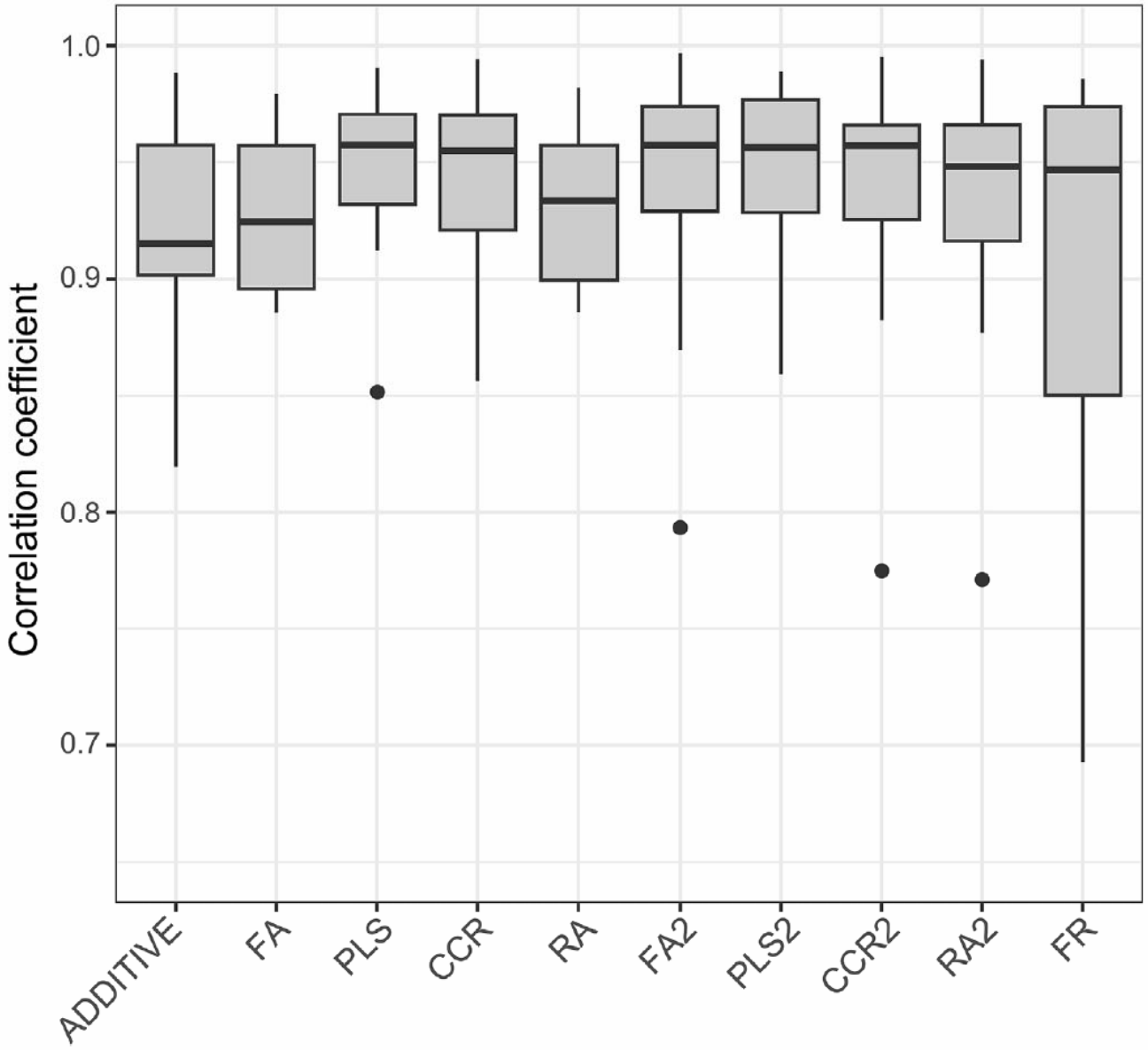


FIGURE 1 Correlation between genotypic slopes (β_i) estimated with different models. The lower triangle shows the link between the slopes estimated with the different approaches, with the regression line in red, while the upper triangle shows the corresponding coefficient of correlation. [Color figure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

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.



Summary

- Finlay-Wilkinson regression still very popular
- Can take this as a starting point for incorporating environmental covariates
- Modelling risk / stability / resilience requires random environments
- Accounting for experimental design and structure of environments requires random effects

  $\Rightarrow$ need mixed model
- Several approximate method of estimation available
- CCR and FA approach look promising
- Large number of covariates $\Rightarrow$ PLS
- Empirical comparison needed in future, also with contenders such as ML

Another example with synthetic covariates

- Sorghum, Ethiopia (Tadese et al. 2024)
- 6 locations
- One year
- Use PLS to derive synthetic covariates
- Genotypes modelled as random
- Pedigree information available
- Leave-one-location-out cross-validation

$$MSEPD = \frac{\sum_{j=1}^{J}\sum_{i=1}^{I}\sum_{h\neq i}^{I}\left[\left(\bar{y}_{ij} - \bar{y}_{hj}\right) - \left(z_{ij} - z_{hj}\right)\right]^2}{JI\left(I-1\right)}$$

$\bar{y}_{ij}$ = observed yield of *i*-th variety in *j*-th location

z_{ij} = predicted yield of *i*-th variety in *j*-th location

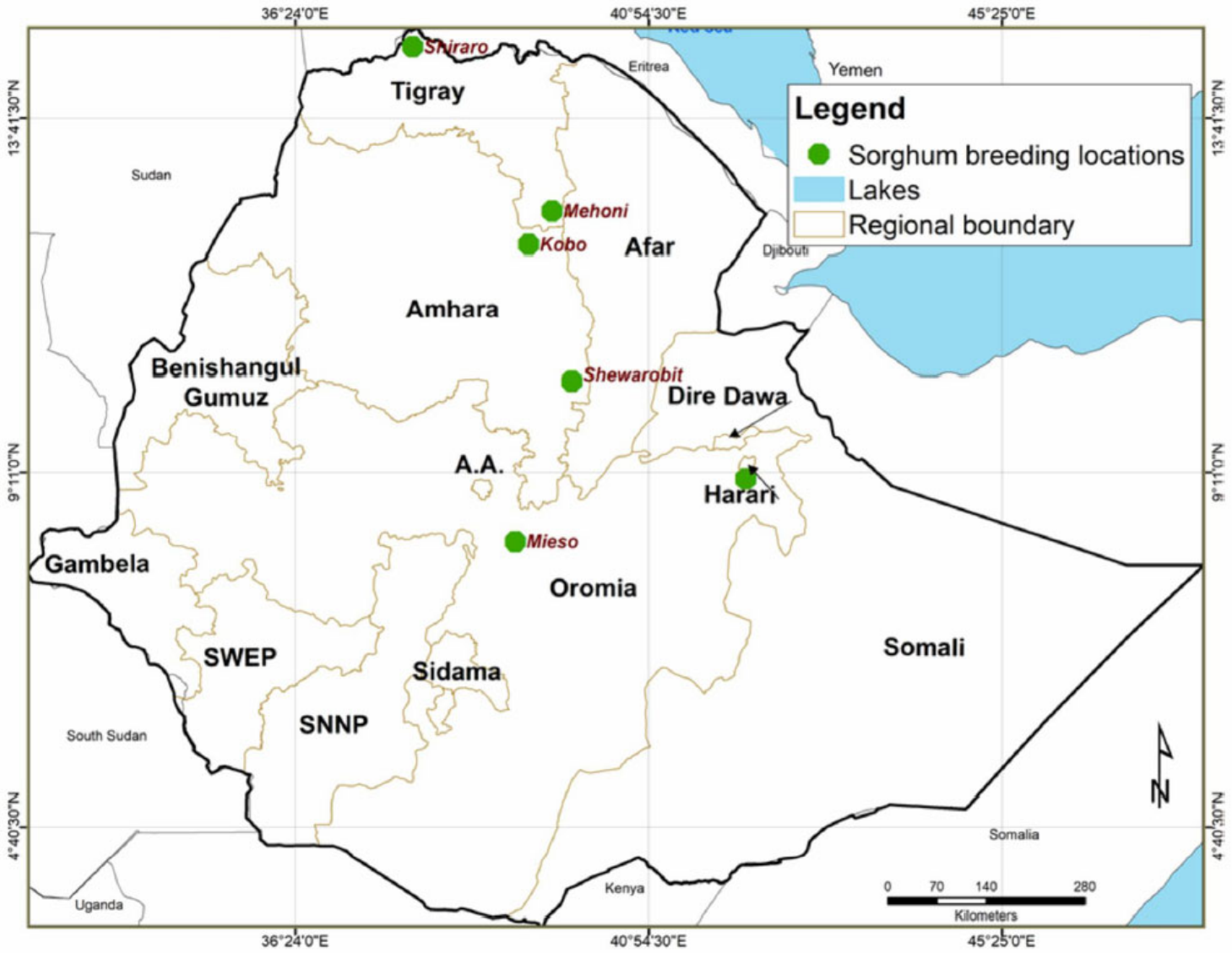


Fig. 1 Map of Ethiopia including the six sorghum breeding locations used in this study

Table 2 Description of the soil information at different soil layers and weather data taken from each location

Description of the covariates	Layers	Acronyms
Soil organic carbon concentration	6	orc1, orc2, …, orc6
Soil pH	6	ph1, ph2, …, ph6
Coarse fragments volumetric	6	CECrf, Corf, Curf, Ompctrf, CaCO3rf, Mnrt
Soil texture fraction sand	6	sand1, sand2, …, sand6
Soil texture fraction silt	6	silt1, silt2, …, silt6
Soil texture fraction clay	6	clay1, clay2, …, clay6
Cation Exchange Capacity	6	CEC1, CEC2, …, CEC6
Total nitrogen	1	cton
Aluminum concentration	2	MgAIrt, Morf
Exchangeable acidity	3	Prt, pHrt, Npctrt
Exchangeable calcium	2	catomg, Casatrf
Exchangeable magnesium	2	ktomg, Mgsatrf
Exchangeable sodium	2	Srt, Sirt
Sum of exchangeable bases	3	Znrf, ECrf, Krf
Electrical conductivity	6	EC1, EC2, …, EC6
Temperature	–	Temp
Rainfall	–	RF

Models

a = intercepts of *I* genotypes
b = slopes of *I* genotypes

I = identity matrix
A = pedigree matrix
G = 2 × 2 unstructured matrix

$$a \sim N\left(0, I\sigma_a^2\right) \quad \Rightarrow \text{M1, M5, M9}$$

$$a \sim N\left(0, A\sigma_a^2\right) \quad \Rightarrow \text{M2, M6, M10}$$

$$\begin{pmatrix} a \\ b \end{pmatrix} \sim N\left(0, G \otimes I\right) \quad \Rightarrow \text{M3, M6, M11}$$

$$\begin{pmatrix} a \\ b \end{pmatrix} \sim N\left(0, G \otimes A\right) \quad \Rightarrow \text{M4, M8, M2}$$

Table 3 Summary of the 12 models used to predict genotypes performance in the new locations using the first SC and pedigree information

Models	Fixed effects	Random effects	Variance–covariance matrix of		
			$\boldsymbol{a}$	$\boldsymbol{l}$	$\boldsymbol{s}$
M1	μ	$\boldsymbol{a},\boldsymbol{l},\boldsymbol{s}$	$\mathbf{I}_I\sigma_a^2$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{I}_I \otimes \mathbf{I}_M\sigma_s^2$
M2	μ	$\boldsymbol{a},\boldsymbol{l},\boldsymbol{s}$	$\mathbf{A}\sigma_a^2$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{A}\otimes\mathbf{I}_M\sigma_s^2$
M3	μ,β	$\boldsymbol{a},\boldsymbol{b},\boldsymbol{l},\boldsymbol{s}$	$\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} \sim N(0,\mathbf{I}_I \otimes \boldsymbol{G})\boldsymbol{G} = \begin{bmatrix}\sigma_a^2 & \sigma_{ab}\\ \sigma_{ab} & \sigma_b^2\end{bmatrix}$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{I}_I \otimes \mathbf{I}_M\sigma_s^2$
M4	μ,β	$\boldsymbol{a},\boldsymbol{b},\boldsymbol{l},\boldsymbol{s}$	$\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} \sim N(0,\mathbf{A} \otimes \boldsymbol{G})\boldsymbol{G} = \begin{bmatrix}\sigma_a^2 & \sigma_{ab}\\ \sigma_{ab} & \sigma_b^2\end{bmatrix}$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{A}\otimes\mathbf{I}_M\sigma_s^2$
M5	μ	$\boldsymbol{a},\boldsymbol{l},\boldsymbol{s}$	$\mathbf{I}_I\sigma_a^2$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{I}_I \otimes \boldsymbol{\Phi}$
M6	μ	$\boldsymbol{a},\boldsymbol{l},\boldsymbol{s}$	$\mathbf{A}\sigma_a^2$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{A} \otimes \boldsymbol{\Phi}$
M7	μ,β	$\boldsymbol{a},\boldsymbol{b},\boldsymbol{l},\boldsymbol{s}$	$\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} \sim N(0,\mathbf{I}_I \otimes \boldsymbol{G})\boldsymbol{G} = \begin{bmatrix}\sigma_a^2 & \sigma_{ab}\\ \sigma_{ab} & \sigma_b^2\end{bmatrix}$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{I}_I \otimes \boldsymbol{\Phi}$
M8	μ,β	$\boldsymbol{a},\boldsymbol{b},\boldsymbol{l},\boldsymbol{s}$	$\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} \sim N(0,\mathbf{A} \otimes \boldsymbol{G})\boldsymbol{G} = \begin{bmatrix}\sigma_a^2 & \sigma_{ab}\\ \sigma_{ab} & \sigma_b^2\end{bmatrix}$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{A} \otimes \boldsymbol{\Phi}$
M9	μ	$\boldsymbol{a},\boldsymbol{l},\boldsymbol{s}$	$\mathbf{I}_I\sigma_a^2$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{I}_I \otimes \boldsymbol{\Sigma}$
M10	μ	$\boldsymbol{a},\boldsymbol{l},\boldsymbol{s}$	$\mathbf{A}\sigma_a^2$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{A} \otimes \boldsymbol{\Sigma}$
M11	μ,β	$\boldsymbol{a},\boldsymbol{b},\boldsymbol{l},\boldsymbol{s}$	$\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} \sim N(0,\mathbf{I}_I \otimes \boldsymbol{G})\boldsymbol{G} = \begin{bmatrix}\sigma_a^2 & \sigma_{ab}\\ \sigma_{ab} & \sigma_b^2\end{bmatrix}$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{I}_I \otimes \boldsymbol{\Sigma}$
M12	μ,β	$\boldsymbol{a},\boldsymbol{b},\boldsymbol{l},\boldsymbol{s}$	$\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} \sim N(0,\mathbf{A} \otimes \boldsymbol{G})\boldsymbol{G} = \begin{bmatrix}\sigma_a^2 & \sigma_{ab}\\ \sigma_{ab} & \sigma_b^2\end{bmatrix}$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{A} \otimes \boldsymbol{\Sigma}$

The matrix $\mathbf{I}$ represent identity matrix, with the subscript denoting the dimension,$\boldsymbol{\Phi}$ represents diagonal matrix for the genotype-by-location interactions, $\otimes$ represents the Kronecker product, $\boldsymbol{\Sigma}$ is the factor-analytic variance–covariance structure, A is the kinship matrix, $\boldsymbol{a}$ and $\boldsymbol{b}$ are vectors of random coefficients for genotypes, $\boldsymbol{l}$ is a vector for location main effects, $\boldsymbol{s}$ is a vector for genotype-by-location interactions,β is the slope for the regression on $\boldsymbol{t}$, where $\boldsymbol{t}$ is the vector of the synthetic covariate, I is the number of genotypes, M is the number of locations

Table 5 The mean squared error of predicted differences (MSEPD) and average rank correlation between adjusted means and predicted values across locations for each model without SC, with one and two SCs of the cross-validation. The values in brackets in the correlation columns are the standard errors of the correlation for each model across locations

Models	Without SC		With one SC		With two SC	
	MSEPD (t^2/ha^2)	Correlation	MSEPD (t^2/ha^2)	Correlation	MSEPD (t^2/ha^2)	Correlation
M1	1.02248	0.42829 (0.187)	–	–	–	–
M2	1.02311	0.4204 (0.096)	–	–	–	–
M3	–	–	0.95536	0.46375 (0.167)	1.032472	0.425681 (0.098)
M4	–	–	1.02599	0.42105 (0.082)	1.028033	0.420002 (0.083)
M5	1.0464	0.42827 (0.101)	–	–	–	–
M6	1.04629	0.42381 (0.092)	–	–	–	–
M7	–	–	0.71342	0.50913 (0.259)	0.997379	0.47778 (0.168)
M8	–	–	1.0321	0.42548 (0.097)	1.035721	0.423737 (0.089)
M9	1.09375	0.43914 (0.132)	–	–	–	–
M10	1.03831	0.4308 (0.083)	–	–	–	–
M11	–	–	0.90766	0.58367 (0.187)	0.879093	0.588384 (0.196)
M12	–	–	1.00929	0.43857 (0.096)	1.015214	0.437673 (0.097)

M1–M12 are as defined in the Table 3