Projecting results of zoned multi-environment trials to new locations using environmental covariates with random coefficient models: accuracy and precision

Harimurti Buntaran[1], Johannes Forkman[2], Hans-Peter Piepho[1]

[1]Biostatistics unit, Institute of Crop Science, University of Hohenheim
[2]Department of Crop Production Ecology, Swedish University of Agricultural Sciences

XII[th] Working Seminar on Statistical Methods in Variety Testing
COBORU, Słupia Wielka, 7 July 2022

Overview

Multi-environment trials (MET)

- To assess the performance of a set of genotypes in a target population of environments.
- To understand and exploit the pattern of genotype×environment interaction (GEI) in the target population of environments (TPE).
  - More targeted predictions and recommendations on cultivars.

Environmental covariates

- Useful to enhance the predictive capability of MET analyses (Heslot et al. 2014).
- Evaluate the adaptability of the genotypes to the new target environment.
- Incorporating environmental covariates in the GEI analysis: factorial regression (Denis 1988; Piepho et al. 1998; van Eeuwijk and Elgersma 1993).

Random coefficient (RC) models

- Regression on environmental covariates: usually modelled by fixed effects.
  - Only studying the pattern of GEI at the tested locations and making predictions in an unstructured TPE
- TPE is sub-divided into zones $\rightarrow$necessary to model genotypic effect as random $\rightarrow$to exploit genetic correlations between zones (Kleinknecht et al. 2013; Buntaran et al. 2019; Buntaran et al. 2020).
- Random genotype effect for genotype-specific regression coefficients, known as random coefficient (RC) models (Longford 1993; Milliken and Johnson 2002).

Predictions are not enough

- Mostly, trials in an MET do not coincide with the growers' fields, which can thus be considered as untested locations.
- Grower's fields, the real target of breeding, must be seen as new locations in the TPE.
  - More targeted predictions and recommendations on cultivars.
- Standard errors of prediction from MET analyses are not valid for untested locations.

Motivations and goals

- Compute valid standard errors of prediction for untested locations.
- Improve precision of prediction in untested locations by incorporating environmental covariates using random genotype effect.
  - Random genotype-specific regression coefficients, known as random coefficient (RC) models.
- Compare the predictive accuracy and precision of seven fixed-effect and seven random-effect genotype models.
- Compare precision:
  - Individual genotype: standard error of predicted values (SEPV)
  - Pairwise genotype differences: standard errors of predicted pairwise differences of genotypic values (SEPD).

Dataset

- Winter wheat, Swedish official cultivar testing in 2016.
- 25 genotypes tested in 18 locations. Laid out in an α-design per location.
- Locations are stratified into three zones: South, Middle, and North.
- 4 new locations (two in the North, two in the South).
- Covariates: pH, clay content, and humus content. Covariates are locations-specific. The pH & humus covariates are standardised. The clay covariate is scaled to $(\text{clay} - 40)/10$.

Figure: Swedish cultivar testing zones.

Statistical models: Stage 1 model

- Two-stage fully-efficient method (Damesa et al. 2017; Piepho et al. 2012)
- Stage I: per location analysis using mixed models

Stage I model

$$y_{ijk} = \underbrace{\mu + g_i}_{\text{fixed effects}} + \underbrace{r_j + b_{jk} + e_{ijk}}_{\text{random effects}}$$

y_{ijk} : plot-level observations, dry matter yield (DMY)

μ : grand mean

r_j : replicate effect

b_{jk} : block effect nested within a replicate

g_i : genotype main effect, empirical best linear unbiased estimator (EBLUE)

e_{ijk} : residual associated with y_{ijk}

Statistical models: Stage 2 model

- Stage II: using model comprises zone effects
- Forward full variance-covariance matrix of the adjusted means to stage II

Stage II model

$$\hat{\eta}_{ijk} = \underbrace{\mu + z_m}_{\text{fixed effects}} + \underbrace{l_{mp} + g_i + (gz)_{im} + (gl)_{imp} + e_{ijk}}_{\text{random effects}}$$

$\hat{\eta}_{ijk}$: genotype-location-treatment DMY means from Stage 1

z_m : zone effect

l_{mp} : location effect nested within a zone

g_i : genotype main effect, empirical best linear unbiased predictor (EBLUP)

(gz) : genotype-zone interaction effect

(gl) : genotype-location interaction effect

e_{ijk} : weighted residual associated with$\hat{\eta}_{ijk}$stage 1

Statistical models: Random coefficients model

- Introduce environmental covariates
- A model with environmental covariates, x_{mp} & x_{mp}^2

Random coefficients model: random coefficients in the genotype main effect

$$\hat{\eta}_{ijk} = \mu + z_m + \beta_1 x_{mp} + \beta_2 x_{mp}^2 + l_{mp} + \underbrace{(a_i + b_i x_{mp} + c_i x_{mp}^2)}_{\text{random coefficients}} + (gz)_{im} + (gl)_{imp} + e_{ijk}$$

x_{mp}, x_{mp}^2 : linear & quadratic trends for the covariate

β_1, β_2 : linear & quadratic regression slope coefficients of the covariate

a_i, b_i, c_i : random genotype-specific intercept, linear slope, & quadratic slope

Statistical models: Covariance structure for the random coefficients

- Unstructured covariance structure: Ensure invariance with respect to translation and scale transformation of the covariates (Longford, 1993; Wolfinger, 1996).

Unstructured covariance structure

$$\begin{bmatrix} a_i \\ b_i \\ c_i \end{bmatrix} \sim iid\ N(\mathbf{0}, \mathbf{G}_{g_i})$$

$$\mathbf{G}_{g_i} = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{ac} \\ \sigma_{ab} & \sigma_b^2 & \sigma_{bc} \\ \sigma_{ac} & \sigma_{bc} & \sigma_c^2 \end{bmatrix}$$

Predictions of genotypes in zones

- The intermediate inference space (McLean et al 1991): to determine which genotype is the best in the particular environment.
- The predictable functions for the EBLUPs in the zones can be expressed as:

Predictable functions

$$w = \mathbf{K}'\boldsymbol{\beta} + \mathbf{M}'\mathbf{u}$$

$\mathbf{K}'\boldsymbol{\beta}$: estimable function – fixed effects for zone

$\mathbf{M}'\mathbf{u}$: predictable function – genotype main effect and the genotype$\times$zone interaction effects

$\mathbf{K}', \mathbf{M}'$: consists of coefficients 1 and 0

$\boldsymbol{\beta}$: estimates of fixed effects

$\mathbf{u}$: estimates of random effects

Predictions of genotypes in new locations

- The predictable functions for the EBLUPs in the new location can be expressed as:

Predictable functions

$$w = \mathbf{K}'\boldsymbol{\beta} + \mathbf{M}'\mathbf{u} + \mathbf{M}_0'\mathbf{u_0}$$

$\mathbf{K}'\boldsymbol{\beta}$: estimable function – fixed effects for zone

$\mathbf{M}'\mathbf{u}$: predictable function – genotype main effect and the genotype$\times$zone interaction effects of the zone where the new locations located

$\mathbf{M}_0'\mathbf{u_0}$: predictable function – the location main effect and genotype$\times$location interaction effects for the new location, with $var(\mathbf{u_0}) = \mathbf{G_0}$

Predictions of genotypes in new locations with random coefficients

- The predictable functions for the EBLUPs with random coefficients in the new location can be expressed as:

Predictable functions

$$w = \mathbf{K}'\boldsymbol{\beta} + \mathbf{M}'\mathbf{u} + \mathbf{M}_0'\mathbf{u}_0$$

$\mathbf{K}'\boldsymbol{\beta}$: estimable function – fixed effects for zone and covariates

$\mathbf{M}'\mathbf{u}$: predictable function – genotype main effect and the genotype$\times$zone interaction effects of the zone where the new locations located, as well as any random covariate terms

$\mathbf{M}_0'\mathbf{u}_0$: predictable function – the location main effect and genotype$\times$location interaction effects for the new location, with $var(\mathbf{u}_0) = \mathbf{G}_0$

SEPV for untested locations

- Standard error of predicted values (SEPV): individual genotype

SEPV

$$\text{SEPV} = \sqrt{\text{var}(\hat{\eta}) + \text{var}(w|\boldsymbol{\beta}, \mathbf{u})}$$

$\hat{\eta}$: genotype-zone average (EBLUEs or EBLUPs)

w : deviation from $\hat{\eta}$(genotype in the untested locations)

$\text{var}(\hat{\eta})$: variance of genotype-zone average

$\text{var}(w|\boldsymbol{\beta}, \mathbf{u})$: variance of untested locations, $\sigma^2_l + \sigma^2_{gl}$. For RC models, includes variances and covariances for intercepts and slopes of random coefficients.

SEPD for untested locations

- Standard errors of predicted pairwise differences of genotypic values (SEPD): pairwise genotype differences

SEPD

$$\text{SEPD} = \sqrt{\text{VDIFF}(\hat{\eta}) + \text{VDIFF}(w|\boldsymbol{\beta}, \mathbf{u})}$$

$\text{VDIFF}(\hat{\eta})$: variance of genotype differences in the zone level

$\text{VDIFF}(w|\boldsymbol{\beta}, \mathbf{u})$: variance of genotype difference in the location level, $2 \times \sigma^2_{gl}$

Prediction intervals for the genotypes for untested locations

- The prediction interval for w is centred at η and the approximate $(1-\alpha)\times 100\%$ prediction interval is given by

$$\hat{\eta} \pm z_{1-\alpha/2}\text{SEPV}$$

- where $z_{1-\alpha/2}$ is the $(1-\alpha/2)\times 100\%$ quantile of the standard normal distribution.
- The prediction interval for pairwise differences between the i-th and i'-th genotype is given by:

$$\text{DIFF}(\hat{\eta}) \pm z_{1-\alpha/2}\text{SEPD}$$

- where $\text{DIFF}(\hat{\eta})$ is the difference in predictions between two genotypes.

Cross-validation (CV) & MSEP

- A leave-one-out CV for model comparison and selection.
- Left one location out at a time to mimic the prediction for new locations as the validation set.
- For the models with covariate: the covariate in the validation set was used for predictions.
- The assessment (mean squared error of prediction of difference, MSEP) was measured based on the discrepancies between observed $(y_{ij} - y_{i'j})$ and predicted $(z_{ij} - z_{i'j})$.
- The main interest in cultivar trials is in prediction of differences between genotypes rather than performance of individual genotype (Piepho, 1998).

Covariate selection

- Using fixed genotype effect model.

Effect	Numerator DF	Denominator DF	F Value	Pr >F
G	24	299.00	8.91	<.0001
Z	2	5.29	0.18	0.8394
G×Z	48	268.00	1.27	0.1212
Clay	1	7.69	3.26	0.1100
Clay squared	1	8.200	6.84	0.0303
pH	1	5.35	2.57	0.1656
pH squared	1	3.39	0.07	0.8094
Humus	1	6.87	0.01	0.9312
Humus squared	1	7.52	0.00	0.9993

Table: Fixed effect tests for covariate selection

Random coefficient models

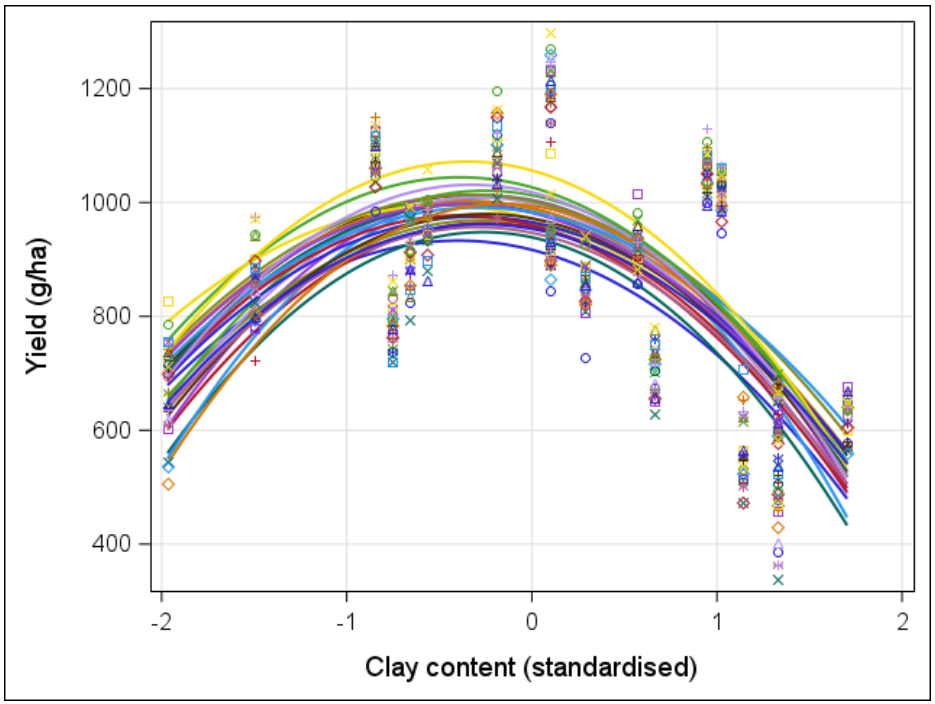

Figure: Genotype specific

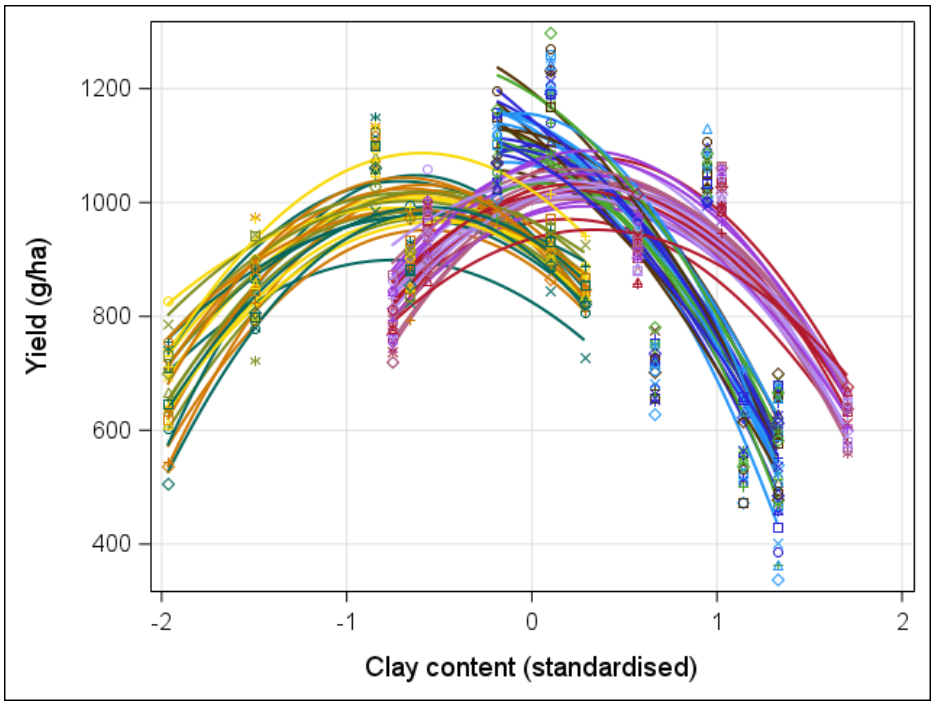

Figure: Genotype-Zone specific

SEPV & MSEP

- RC models reduced the average SEPV for all new locations by 30–38%.

Model	RC term	SEPV ($g.m^{-2}$)				$\sqrt{MSEP}$ ($g.m^{-2}$)
		Untested locations				
		N01	N02	S01	S02	
RC						
	gz	113	121	88	89	62
	g and gz	113	122	88	90	62
	g	114	122	88	89	62
EBLUP	×	174	174	142	142	61
EBLUE	×	236	236	232	232	85

Table: Average SEPV for 25 genotypes and MSEP (summary)

Prediction intervals

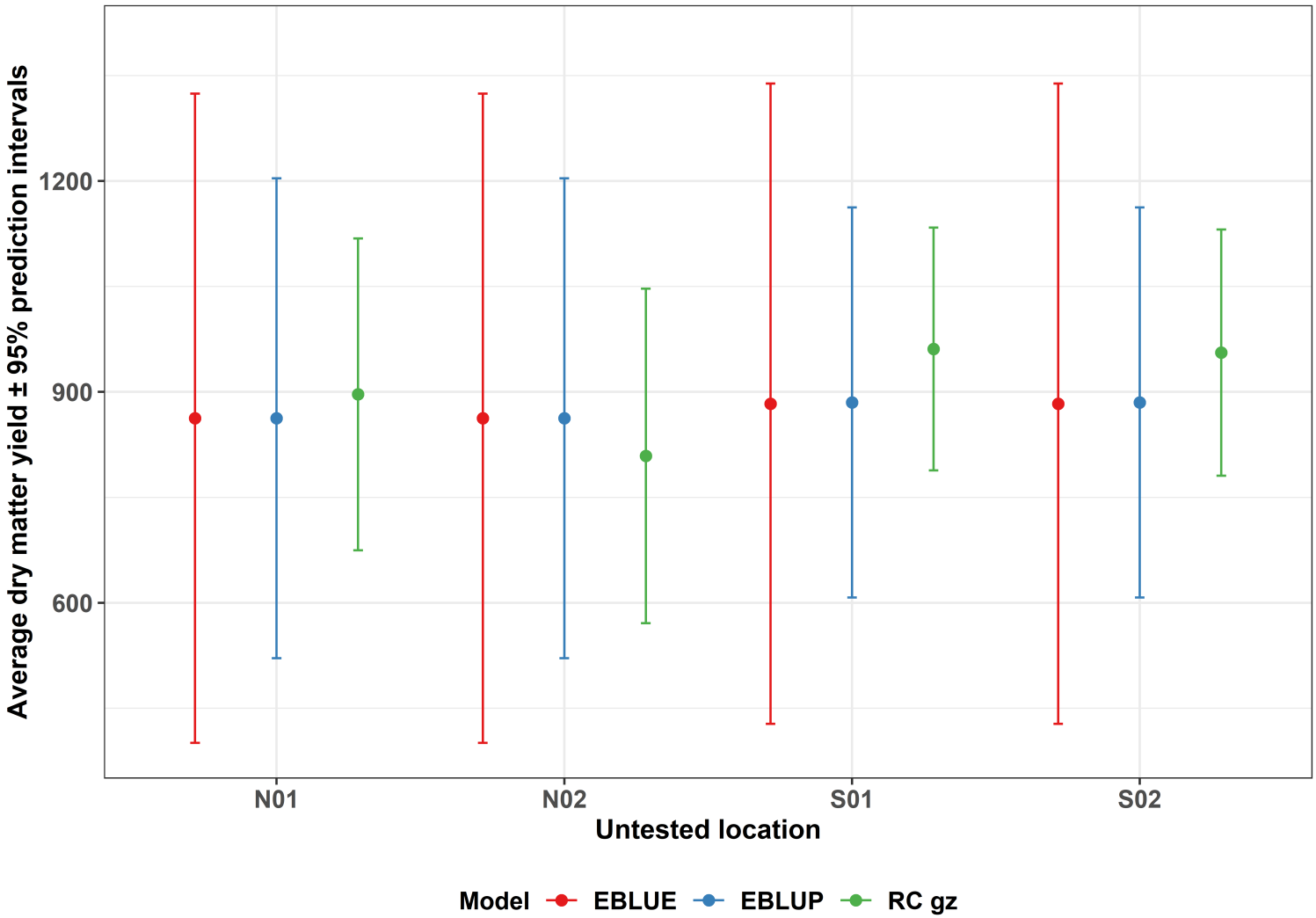

Figure: Average dry matter yield $\pm$ 95% prediction intervals

SEPD & MSEP

- RC models reduced the average SEPD for all new locations by 12–40%.

Model	RC term	SEPD ($g.m^{-2}$)				$\sqrt{MSEP}$ ($g.m^{-2}$)
		Untested locations				
		N01	N02	S01	S02	
RC						
	gz	30	33	29	31	62
	g and gz	30	33	29	31	62
	g	37	37	30	30	62
EBLUP	×	37	37	48	48	61
EBLUE	×	50	50	50	50	85

Table: Average SEPD for 25 genotypes and MSEP (summary)

Discussion

- Random coefficient models improved the precision of the predicted genotypic values and the pairwise differences between genotypes.
- Select the covariates (e.g. clay) and their functional forms (e.g. quadratic term) that improve precision.
- Humus content and pH were dropped due to non-significance.
- Covariate scaling is crucial to ensure positive definite covariance matrix and enhance model convergence.
  - E.g. clay was scaled to (clay $-$ 40/10) because this scaling resulted in non-negative variance estimates in the RC models.
- Different strategies are needed in the case of a large number of covariates, e.g., make synthetic covariates from some statistical methods.
- Covariates can initially be selected based on the biological considerations.
- R-square (R^2) for mixed models (Piepho, 2019) is an option for covariate selection, demonstrated in Hadasch et al. (2020).

Discussion

- Location effect has to be random
  - To compute valid standard errors for the untested locations.
  - To account for the uncertainty of effects for the untested locations (variance component estimate of location).
  - Random location main effect, as well as the random cultivar-location interaction effects, are part of the deviations from the regression curves.
  - Fixed effect of location will give no variance component estimate of location. Only provide valid standard error for trials' locations.
- Breeders can use RC models to determine the adaptability of tested genotypes in new environments.
- Agronomists and growers, can use RC models to identify the best locally adapted genotypes.

Conclusion

- The RC model was competitive, with regards to MSEP, compared to the EBLUP models.
- The RC model with random coefficients of linear and quadratic terms in the G×Z effect, can be recommended based on joint consideration of precision in predictions and accuracy.
- The RC models improved the precision of the predictions for a new location by **utilising covariate information in the new location** in the random effects part, and **by borrowing information from other zones via genetic correlation between zones**.
- The scale of the covariate is essential to obtain reliable variance component estimates and avoid convergence issues.

More details

- Buntaran, H., Forkman, J. Piepho, HP. Theor Appl Genet 134, 1513–1530 (2021). https://doi.org/10.1007/s00122-021-03786-2

References

- Buntaran H, Piepho H-P, Hagman J, Forkman J (2019) A cross-validation of statistical models for zoned-based prediction in cultivar testing. Crop Science 59:1544-1553. https://doi:10.2135/cropsci2018.10.0642
- Buntaran H, Piepho H-P, Schmidt P, Rydén J, Halling M, Forkman J (2020) Cross-validation of stage-wise mixed-model analysis of Swedish variety trials with winter wheat and spring barley. Crop Science 60:2221-2240. https://doi:10.1002/csc2.20177
- Damesa TM, Möhring J, Worku M, Piepho H-P (2017) One step at a time: stage-wise analysis of a series of experiments. Agron J 109:845–857. https://doi.org/10.2134/agron j2016.07.0395
- Denis JB (1988) Two-way analysis using covariates. Statistics 19:123-132. doi:10.1080/02331888808802080
- Hadasch S, Laidig F, Macholdt J, Bönecke E, Piepho HP (2020) Trends in mean performance and stability of winter wheat and winter rye yields in a long-term series of variety trials Field Crops Research 252:107792. https://doi.org/10.1016/j.fcr.2020.107792

References

- Heslot N, Akdemir D, Sorrells ME, Jannink J-L (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theoret Appl Genetics 127:463-480. https://doi:10.1007/s00122-013-2231-5
- Kleinknecht K, Möhring J, Singh KP, Zaidi PH, Atlin GN, Piepho HP (2013) Comparison of the performance of best linear unbiased estimation and best linear unbiased prediction of genotype effects from zoned Indian maize data. Crop Science 53:1384-1391. https://doi:10.2135/cropsci2013.02.0073
- Longford NT (1993) Random coefficient models. Oxford University Press, New York
- McLean RA, Sanders WL, Stroup WW (1991) A unified approach to mixed linear models. Am Stat 45:54–64. https://doi.org/10.1080/00031305.1991.10475767
- Milliken GA, Johnson DE (2002) Analysis of messy data, volume III: analysis of covariance. CRC Press, Boca Raton
- Piepho, H-P (1998) Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. Theoret Appl Genetics 97:195-201. https://doi:10.1007/s001220050885

References

- Piepho H-P, Denis J-B, van Eeuwijk FA (1998) Predicting cultivar differences using covariates. Journal of Agricultural, Biological, and Environmental Statistics 3:151-162. https://doi:10.2307/1400648
- Piepho H-P (2019) A coefficient of determination (R2) for generalized linear mixed models. Biometrical Journal 61:860-872. https://doi.org/10.1002/bimj.201800270
- Piepho H-P, Möhring J, Schulz-Streeck T, Ogutu JO (2012) A stagewise approach for the analysis of multi-environment trials. Biom J 54:844–860. https://doi.org/10.1002/bimj.20110 0219
- van Eeuwijk FA, Elgersma A (1993) Incorporating environmental information in an analysis of genotype by environment interaction for seed yield in perennial ryegrass. Heredity 70:447-457. https://doi:10.1038/hdy.1993.66
- Wolfinger RD (1996) Heterogeneous variance: covariance structures for repeated measures. Journal of Agricultural, Biological, and Environmental Statistics 1:205-230. https://doi:10.2307/1400366